# Learning to Recover 3D Scene Shape from a Single Image

Wei Yin[†], Jianming Zhang[‡], Oliver Wang[‡], Simon Niklaus[‡], Long Mai[‡], Simon Chen[‡], Chunhua Shen[†*]

[†] The University of Adelaide, Australia [‡] Adobe Research

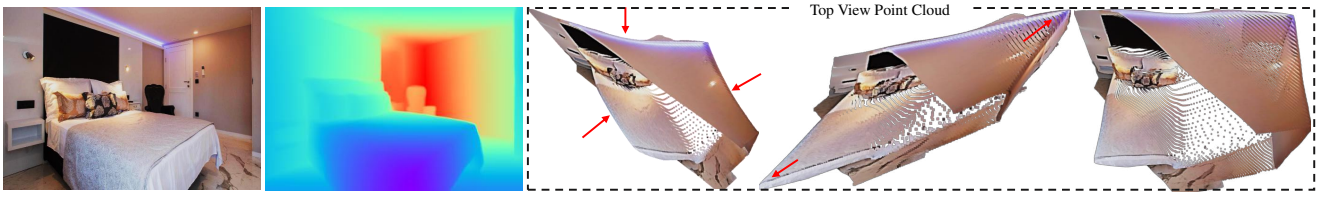


**Figure 1:** 3D scene structure distortion of projected point clouds. While the predicted depth map is correct, the 3D scene shape of the point cloud suffers from noticeable distortions due to an unknown depth shift and focal length (third column). Our method recovers these parameters using 3D point cloud networks. With recovered depth shift, the walls and bed edges become straight, but the overall scene is stretched (fourth column). Finally, with recovered focal length, an accurate 3D scene can be reconstructed (fifth column).

## Abstract

*Despite significant progress in monocular depth estimation in the wild, recent state-of-the-art methods cannot be used to recover accurate 3D scene shape due to an unknown depth shift induced by shift-invariant reconstruction losses used in mixed-data depth prediction training, and possible unknown camera focal length. We investigate this problem in detail, and propose a two-stage framework that first predicts depth up to an unknown scale and shift from a single monocular image, and then use 3D point cloud encoders to predict the missing depth shift and focal length that allow us to recover a realistic 3D scene shape. In addition, we propose an image-level normalized regression loss and a normal-based geometry loss to enhance depth prediction models trained on mixed datasets. We test our depth model on nine unseen datasets and achieve state-of-the-art performance on zero-shot dataset generalization. Code is available at:* https://git.io/Depth

## 1. Introduction

3D scene reconstruction is a fundamental task in computer vision. The established approach to address this task is SLAM or SfM [16], which reconstructs 3D scenes based on feature-point correspondence with consecutive frames or multiple views. In contrast, this work aims to achieve *dense 3D scene shape reconstruction from a single in-the-wild image*. Without multiple views available, we rely on monocular depth estimation. However, as shown in Fig. 1, existing monocular depth estimation methods [10, 38, 48, 40] alone are unable to faithfully recover an accurate 3D point cloud.

Unlike multi-view reconstruction methods, monocular depth estimation requires leveraging high level scene priors, so data-driven approaches have become the *de facto* solution to this problem [24, 29, 37, 49, 47]. Recent works have shown promising results by training deep neural networks on diverse in-the-wild data, *e.g.* web stereo images and stereo videos [5, 7, 29, 37, 43, 44, 49]. However, the diversity of the training data also poses challenges for the model training, as training data captured by different cameras can exhibit significantly different image priors for depth estimation [11]. Moreover, web stereo images and videos can only provide depth supervision up to a scale and shift due to the unknown camera baselines and stereoscopic post processing [23]. As a result, state-of-the-art in-the-wild monocular depth models use various types of losses invariant to scale and shift in training. While an unknown scale in depth will not cause any shape distortion, as it scales the 3D scene uniformly, an unknown depth shift will (see Sec. 3.1 and Fig. 1). In addition, the camera focal length of a given image may not be accessible at test time, leading to more distortion of the 3D scene shape. This scene shape distortion is a critical problem for downstream tasks such as 3D view synthesis and 3D photography.

To address these challenges, we propose a novel monocular scene shape estimation framework that consists of a

*Correspondence should be addressed to C. Shen.

depth prediction module and a point cloud reconstruction module. The depth prediction module is a convolutional neural network trained on a mixture of existing datasets that predicts depth maps up to a scale and shift. The point cloud reconstruction module leverages point cloud encoder networks that predict shift and focal length adjustment factors from an initial guess of the scene point cloud reconstruction. A key observation that we make here is that, *when operating on point clouds derived from depth maps, and not on images themselves, we can train models to learn 3D scene shape priors using synthetic 3D data or data acquired by 3D laser scanning devices. The domain gap is significantly less of an issue for point clouds than that for images,* although these data sources are significantly less diverse than internet images.

We empirically show that these point cloud encoders generalize well to unseen datasets.

Furthermore, to train a robust monocular depth prediction model on mixed data from multiple sources, we propose a simple but effective image-level normalized regression loss, and a pair-wise surface normal regression loss. The former loss transforms the depth data to a canonical scale-shift-invariant space for more robust training, while the latter improves the geometry of our predicted depth maps. To summarize, our main contributions are:

- A novel framework for in-the-wild monocular 3D scene shape estimation. To the best of our knowledge, this is the first fully data-driven method for this task, and the first method to leverage 3D point cloud neural networks for improving the structure of point clouds derived from depth maps.
- An image-level normalized regression loss and a pair-wise surface normal regression loss for improving monocular depth estimation models trained on mixed multi-source datasets.

Experiments show that our point cloud reconstruction module can recover accurate 3D shape from a single image, and that our depth prediction module achieves state-of-the-art results on zero-shot dataset transfer to 9 unseen datasets.

## 2. Related Work

**Monocular depth estimation in the wild.** This task has recently seen impressive progress [5, 6, 7, 24, 37, 40, 43, 44, 49, 47]. The key properties of such approaches are what data can be used for training, and what objective function makes sense for that data. When metric depth supervision is available, networks can be trained to directly regress these depths [10, 25, 48]. However, obtaining metric ground truth depth for diverse datasets is challenging. As an alternative, Chen *et al*. [5] collect diverse *relative* depth annotations for internet images, while other approaches propose to scrape stereo images or videos from the internet [29, 37, 43, 44, 49]. Such diverse data is important for generalizability, but as the metric depth is not available, direct depth regression losses cannot be used. Instead, these methods rely either on ranking losses which evaluate relative depth [5, 43, 44] or scale and shift invariant losses [29, 37] for supervision. The later methods produce especially robust depth predictions, but as the camera model is unknown and an unknown shift resides in the depth, the 3D shape cannot be reconstructed from the predicted depth maps. In this paper, we aim to reconstruct the 3D shape from a single image in the wild.

**3D reconstruction from a single image.** A number of works have addressed reconstructing different types of objects from a single image [2, 39, 42], such as humans [30, 31], cars, planes, tables, etc. The main challenge is how to best recover objects details, and how to represent them with limited memory. Pixel2Mesh [39] proposes to reconstruct the 3D shape from a single image and express it in a triangular mesh. PIFu [30, 31] proposes an memory-efficient implicit function to recover high-resolution surfaces, including unseen/occluded regions, of humans. However, all these methods rely on learning priors specific to a certain object class or instance, typically from 3D supervision, and can therefore not work for full scene reconstruction.

On the other hand, several works have proposed reconstructing 3D scene structure from a single image. Saxena *et al*. [32] assume that the whole scene can be segmented into several pieces, of which each one can be regarded as a small plane. They predict the orientation and the location of the planes and stitch them together to represent the scene. Other works propose to use image cues, such as shading [28] and contour edges [20] for scene reconstruction. However, these approaches use hand-designed priors and restrictive assumptions about the scene geometry. Our method is fully data driven, and can be applied to a wide range of scenes.

**Camera intrinsic parameter estimation.** Recovering a camera's focal length is an important part of 3D scene understanding. Traditional methods utilize reference objects such as a planar calibration grids [51] or vanishing points [9], which can then be used to estimate a focal length. Other methods [18, 41] propose a data driven approach where a CNN recovers the focal length on in-the-wild data directly from an image. In contrast, our point cloud module estimates the focal length directly in 3D, which we argue is an easier task than operating on natural images directly.

## 3. Our Method

Our two-stage single image 3D shape estimation pipeline is illustrated in Fig. 2. It is composed of a depth prediction module (DPM) and a point cloud module (PCM). The two modules are trained separately on different data sources, and are then combined together at inference time. The DPM takes an RGB image and outputs a depth map [49] with un-

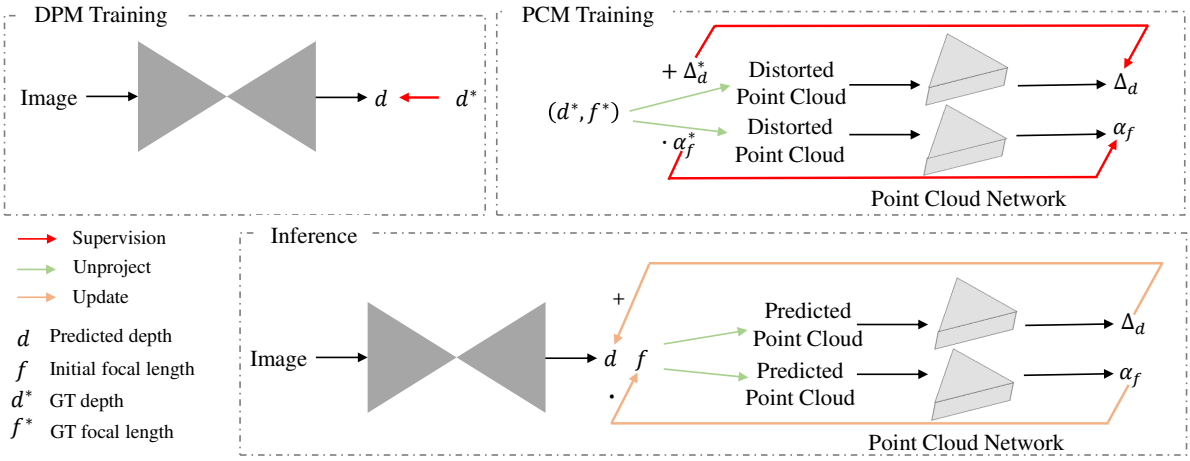


**Figure 2: Method Pipeline.** During training, the depth prediction model (top left) and point cloud module (top right) are trained separately on different sources of data. During inference (bottom), the two networks are combined together to predict depth $d$ and from that, the depth shift $\Delta_d$ and focal length $f \cdot \alpha_f$ that together allow for an accurate scene shape reconstruction. Note that we employ point cloud networks to predict shift and focal length scaling factor separately. Please see the text for more details.

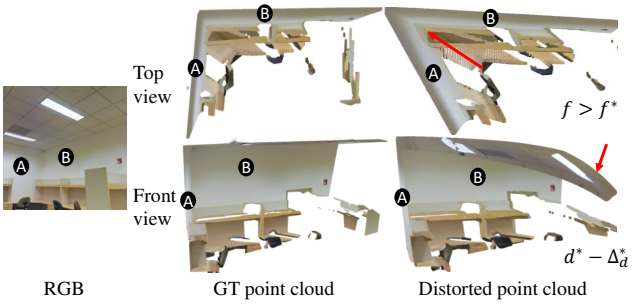


**Figure 3:** Illustration of the distorted 3D shape caused by incorrect shift and focal length. A ground truth depth map is projected in 3D and visualized. When the focal length is incorrectly estimated ($f > f^*$), we observe significant structural distortion, e.g., see the angle between two walls A and B. Second row (front view): a shift ($d^* + \Delta_d$) also causes the shape distortion, see the roof.

known scale and shift in relation to the true metric depth map. The PCM takes as input a distorted 3D point cloud, computed using a predicted depth map $d$ and an initial estimation of the focal length $f$, and outputs shift adjustments to the depth map and focal length to improve the geometry of the reconstructed 3D scene shape.

### 3.1. Point Cloud Module

We assume a pinhole camera model for the 3D point cloud reconstruction, which means that the unprojection from 2D coordinates and depth to 3D points is:

$$\begin{cases} x = \frac{u-u_0}{f} d \\ y = \frac{v-v_0}{f} d \\ z = \quad d \end{cases} \tag{1}$$

where $(u_0, v_0)$ are the camera optical center, $f$ is the focal length, and $d$ is the depth. The focal length affects the point cloud shape as it scales $x$ and $y$ coordinates, but not $z$. Similarly, a shift of $d$ will affect the $x$, $y$, and $z$ coordinates non-uniformly, which will result in shape distortions.

For a human observer, these distortions are immediately recognizable when viewing the point cloud at an oblique angle (Fig. 3), although they cannot be observed looking at a depth map alone. As a result, we propose to directly analyze the point cloud to determine the unknown shift and focal length parameters. We tried a number of network architectures that take unstructured 3D point clouds as input, and found that the recent PVCNN [26] performed well for this task, so we use it in all experiments here.

During training, a perturbed input point cloud with incorrect shift and focal length is synthesized by perturbing the known ground truth depth shift and focal length. The ground truth depth $d^*$ is transformed by a shift $\Delta_d^*$ drawn from $\mathcal{U}(-0.25, 0.8)$, and the ground truth focal length $f^*$ is transformed by a scale $\alpha_f^*$ drawn from $\mathcal{U}(0.6, 1.25)$ to keep the focal length positive and non-zero.

When recovering the depth shift, the perturbed 3D point cloud is $\mathcal{F}(u_0, v_0, f^*, d^* + \Delta_d^*)$ is given as input to the shift point cloud network $\mathcal{N}_d(\cdot)$, trained with the objective:

$$L = \min_\theta |\mathcal{N}_d(\mathcal{F}(u_0, v_0, f^*, d^* + \Delta_d^*), \theta) - \Delta_d^*| \tag{2}$$

where $\theta$ are network weights and $f^*$ is the true focal length.

Similarly, when recovering the focal length, the point cloud $\mathcal{F}(u_0, v_0, \alpha_f^* f^*, d^*)$ is fed to the focal length point cloud network $\mathcal{N}_f(\cdot)$, trained with the objective:

$$L = \min_\theta \left|\mathcal{N}_f(\mathcal{F}(u_0, v_0, \alpha_f^* f^*, d^*), \theta) - \alpha_f^*\right| \tag{3}$$

During inference, the ground truth depth is replaced with the predicted affine-invariant depth $d$, which is normalized to $[0, 1]$ prior to the 3D reconstruction. We use an initial guess of focal length $f$, giving us the reconstructed point cloud $\mathcal{F}(u_0, v_0, f, d)$, which is fed to $\mathcal{N}_d(\cdot)$ and $\mathcal{N}_f(\cdot)$ to predict the shift $\Delta_d$ and focal length scaling factor $\alpha_f$ respectively. In our experiments we simply use an initial focal length with a field of view (FOV) of $60°$. We have also tried to employ a single network to predict both the shift and the scaling factor, but have empirically found that two separate networks can achieve a better performance.

### 3.2. Monocular Depth Prediction Module

We train our depth prediction on multiple data sources including high-quality LiDAR sensor data [50], and low-quality web stereo data [29, 37, 44] (see Sec. 4). As these datasets have varied depth ranges and web stereo datasets contain unknown depth scale and shift, we propose an image-level normalized regression (ILNR) loss to address this issue. Moreover, we propose a pair-wise normal regression (PWN) loss to improve local geometric features.

**Image-level normalized regression loss.** Depth maps of different data sources can have varied depth ranges. Therefore, they need to be normalized to make the model training easier. Simple Min-Max normalization [13, 35] is sensitive to depth value outliers. For example, a large value at a single pixel will affect the rest of the depth map after the Min-Max normalization. We investigate more robust normalization methods and propose a simple but effective image-level normalized regression loss for mixed-data training.

Our image-level normalized regression loss transforms each ground truth depth map to a similar numerical range based on its individual statistics. To reduce the effect of outliers and long-tail residuals, we combine tanh normalization [35] with a trimmed Z-score, after which we can simply apply a pixel-wise mean average error (MAE) between the prediction and the normalized ground truth depth maps. The ILNR loss is formally defined as follows.

$$L_{\text{ILNR}} = \frac{1}{N}\sum_{i}^{N}\left|d_i - \overline{d}_i^*\right| + \left|\tanh(d_i/100) - \tanh(\overline{d}_i^*/100)\right|$$

where $\overline{d}_i^* = (d_i^* - \mu_{\text{trim}})/\sigma_{\text{trim}}$ and $\mu_{\text{trim}}$ and $\sigma_{\text{trim}}$ are the mean and the standard deviation of a trimmed depth map which has the nearest and farthest $10\%$ of pixels removed, $d$ is the predicted depth, and $d^*$ is the ground truth depth map. We have tested a number of other normalization methods such as Min-Max normalization [35], Z-score normalization [12], and median absolute deviation normalization (MAD) [35]. In our experiments, we found that our proposed ILNR loss achieves the best performance.

**Pair-wise normal loss.** Normals are an important geometric property, which have been shown to be a complementary modality to depth [34]. Many methods have been proposed to use normal constraints to improve the depth quality, such as the virtual normal loss [48]. However, as the virtual normal only leverages global structure, it cannot help improve the local geometric quality, such as depth edges and planes. Recently, Xian *et al.* [44] proposed a structure-guided ranking loss, which can improve edge sharpness. Inspired by these methods, we follow their sampling method but enforce the supervision in surface normal space. Moreover, our samples include not only edges but also planes. Our proposed pair-wise normal (PWN) loss can better constrain both the global and local geometric relations.

The surface normal is obtained from the reconstructed 3D point cloud by local least squares fitting [48]. Before calculating the predicted surface normal, we align the predicted depth and the ground truth depth with a scale and shift factor, which are retrieved by least squares fitting [29]. From the surface normal map, the planar regions where normals are almost the same and edges where normals change significantly can be easily located. Then, we follow [44] and sample paired points on both sides of these edges. If planar regions can be found, paired points will also be sampled on the same plane. In doing so, we sample $100$K paired points per training sample on average. In addition, to improve the global geometric quality, we also randomly sample paired points globally. The sampled points are $\{(A_i, B_i), i = 0, ..., N\}$, while their corresponding normals are $\{(n_{Ai}, n_{Bi}), i = 0, ..., N\}$. The PWN loss is:

$$L_{\text{PWN}} = \frac{1}{N}\sum_{i}^{N}|n_{Ai}\cdot n_{B_i} - n^*_{Ai}\cdot n^*_{Bi}| \tag{4}$$

where $n^*$ denotes ground truth surface normals. As this loss accounts for both local and global geometry, we find that it improves the overall reconstructed shape.

Finally, we also use a multi-scale gradient loss [24]:

$$L_{\text{MSG}} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N}\left|\nabla_x^k d_i - \nabla_x^k \overline{d}_i^*\right| + \left|\nabla_y^k d_i - \nabla_y^k \overline{d}_i^*\right| \tag{5}$$

| Dataset | Structure guided ranking loss | ILNR | PWN (plane) | PWN (local) | Multi-scale gradient loss |
|---|---|---|---|---|---|
| Taskonomy | ✓ | ✓ | ✓ | ✓ | ✓ |
| 3D Ken Burns | ✓ | ✓ | ✓ | ✓ | ✓ |
| DIML | ✓ | ✓ | ✓ | | ✓ |
| HRWSI+Holopix | ✓ | | | | |
| Weight | 1 | 1 | 1 | 1 | 0.5 |

**Table 1:** Losses on different datasets.

Different losses are enforced on different dataset (see Table 1).

## 4. Experiments

**Datasets and implementation details.** To train the PCM, we sampled 100K Kinect-captured depth maps from ScanNet, 114K LiDAR-captured depth maps from Taskonomy,

| Dataset | # Img | Scene Type | Evaluation Metric | Supervision Type |
|---|---|---|---|---|
| NYU | 654 | Indoor | AbsRel & $\delta_1$ | Kinect |
| ScanNet | 700 | Indoor | AbsRel & $\delta_1$ | Kinect |
| 2D-3D-S | 12256 | Indoor | LSIV | LiDAR |
| iBims-1 | 100 | Indoor | AbsRel & $\varepsilon_{\text{PE}}$ &$\varepsilon_{\text{DBE}}$ | LiDAR |
| KITTI | 652 | Outdoor | AbsRel & $\delta_1$ | LiDAR |
| Sintel | 641 | Outdoor | AbsRel & $\delta_1$ | Synthetic |
| ETH3D | 431 | Outdoor | AbsRel & $\delta_1$ | LiDAR |
| YouTube3D | 58054 | In the Wild | WHDR | SfM, Ordinal pairs |
| OASIS | 10000 | In the Wild | WHDR & LSIV | User clicks, Small patches with GT |
| DIODE | 771 | Indoor & Outdoor | AbsRel & $\delta_1$ | LiDAR |

**Table 2:** Overview of the test sets in our experiments.

and 51K synthetic depth maps from the 3D Ken Burns paper [27]. We train the network using SGD with a batch size of 40, an initial learning rate of 0.24, and a learning rate decay of 0.1. For parameters specific to PVCNN, such as the voxel size, we follow the original work [26].

To train the DPM, we sampled 114K RGBD pairs from LiDAR-captured Taskonomy [50], 51K synthetic RGBD pairs from the 3D Ken Burns paper [27], 121K RGBD pairs from calibrated stereo DIML [21], 48K RGBD pairs from web-stereo Holopix50K [19], and 20K web-stereo HRWSI [44] RGBD pairs. Note that when doing the ablation study about the effectiveness of PWN and ILNR, we sampled a smaller dataset which is composed of 12K images from Taskonomy, 12K images from DIML, and 12K images from HRWSI. During training, 1000 images are withheld from all datasets as a validation set. We use the depth prediction architecture proposed in Xian *et al.* [44], which consists of a standard backbone for feature extraction (*e.g.*, ResNet50 [17] or ResNeXt101 [46]), followed by a decoder, and train it using SGD with a batch size of 40, an initial learning rate 0.02 for all layer, and a learning rate decay of 0.1. Images are resized to 448×448, and flipped horizontally with a 50% chance. Following [49], we load data from different datasets evenly for each batch.

**Evaluation details.** The focal length prediction accuracy is evaluated on 2D-3D-S [1] following [18]. Furthermore, to evaluate the accuracy of the reconstructed 3D shape, we use the Locally Scale Invariant RMSE (LSIV) [7] metric on both OASIS [7] and 2D-3D-S [1]. It is consistent with the previous work [7]. The OASIS [7] dataset only has the ground truth depth on some small regions, while 2D-3D-S has the ground truth for the whole scene.

To evaluate the generalizability of our proposed depth prediction method, we take 9 datasets which are unseen during training, including YouTube3D [6], OASIS [7], NYU [34], KITTI [14], ScanNet [8], DIODE [36], ETH3D [33], Sintel [4], and iBims-1 [22]. On OASIS and YouTube3D, we use the Weighted Human Disagreement Rate (WHDR) [43] for evaluation. On other datasets, except for iBims-1, we evaluate the absolute mean relative error (AbsRel) and the percentage of pixels with $\delta_1 = \max(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}) < 1.25$. We follow Ranftl *et al.* [29] and align the scale and shift before evaluation. To evaluate the geometric quality of the depth, i.e. the quality of edges and planes, we follow [27, 44] and evaluate the depth boundary error [22] ($\varepsilon_{\text{DBE}}^{\text{acc}}, \varepsilon_{\text{DBE}}^{\text{comp}}$) as well as the planarity error [22] ($\varepsilon_{\text{PE}}^{\text{plan}}, \varepsilon_{\text{PE}}^{\text{orie}}$) on iBims-1. $\varepsilon_{\text{PE}}^{\text{plan}}$ and $\varepsilon_{\text{PE}}^{\text{orie}}$ evaluate the flatness and orientation of reconstructed 3D planes compared to the ground truth 3D planes respectively, while $\varepsilon_{\text{DBE}}^{\text{acc}}$ and $\varepsilon_{\text{DBE}}^{\text{comp}}$ demonstrate the localization accuracy and the sharpness of edges respectively. More details as well as a comparison of these test datasets are summarized in Table 2

| Method | ETH3D | NYU | KITTI | Sintel | DIODE |
|---|---|---|---|---|---|
| | AbsRel ↓ | | | | |
| Baseline | 23.7 | 25.8 | 23.3 | 47.4 | 46.8 |
| Recovered Shift | **15.9** | **15.1** | **17.5** | **40.3** | **36.9** |

**Table 3:** Effectiveness of recovering the shift from 3D point clouds with the PCM. Compared with the baseline, the AbsRel is much lower after recovering the depth shift over all test sets.

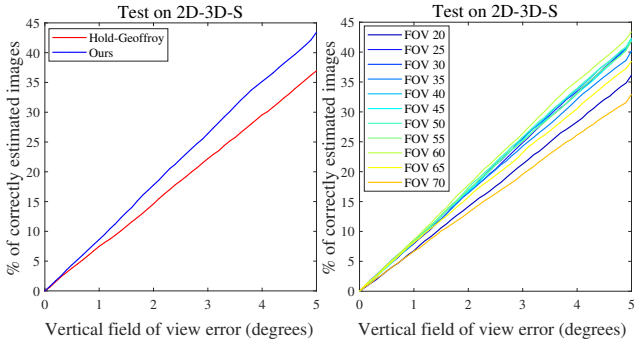


**Figure 4:** Comparison of recovered focal length on the 2D-3D-S dataset. Left, our method outperforms Hold-Geoffroy *et al.* [18]. Right, we conduct an experiment on the effect of the initialization of field of view (FOV). Our method remains robust across different initial FOVs, with a slight degradation in quality past 25° and 65°.

### 4.1. 3D Shape Reconstruction

**Shift recovery.** To evaluate the effectiveness of our depth shift recovery, we perform zero-shot evaluation on 5 datasets unseen during training. We recover a 3D point cloud by unprojecting the predicted depth map, and then compute the depth shift using our PCM. We then align the unknown scale [3, 15] of the original depth and our shifted depth to the ground truth, and evaluate both using the AbsRel error. The results are shown in Table 3, where we see that, on all test sets, the AbsRel error is lower after recovering the shift. We also trained a standard 2D convolutional neural network to predict the shift given an image composed

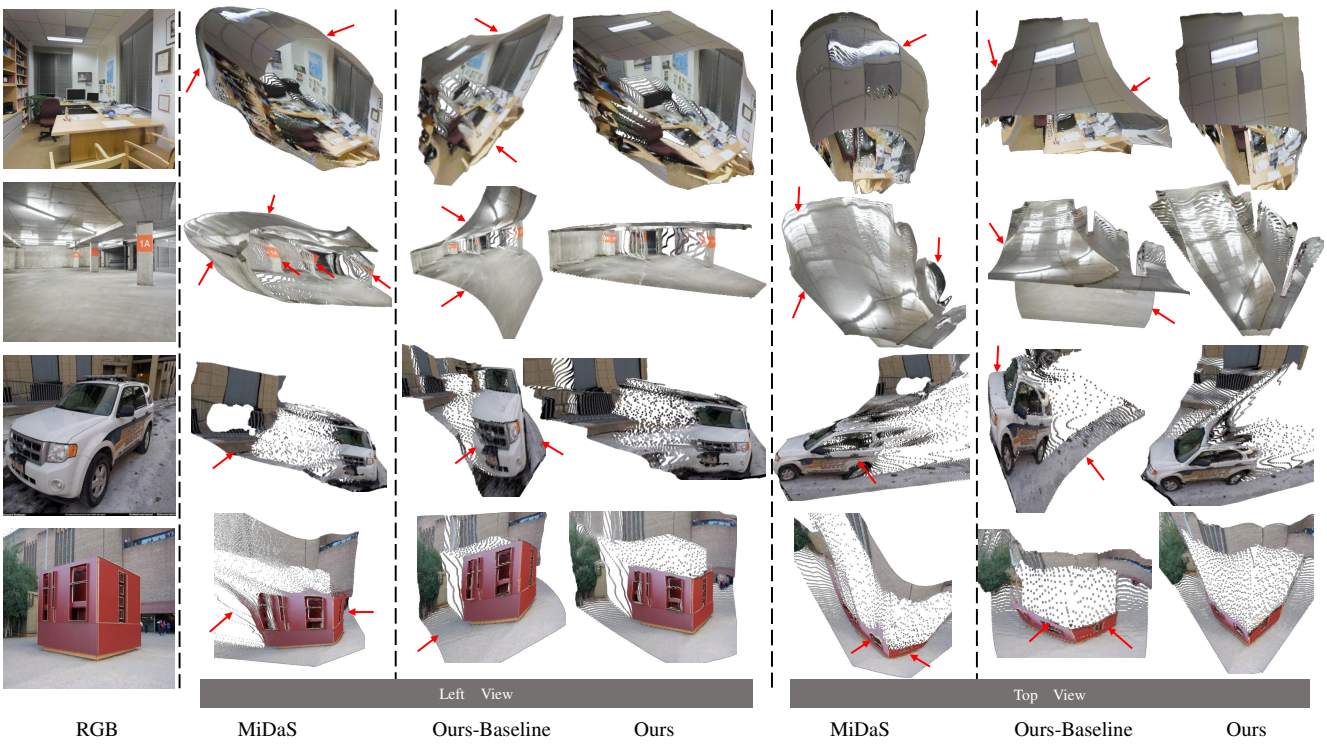


**Figure 5:** Qualitative comparison. We compare the reconstructed 3D shape of our method with several baselines. As MiDaS [29] does not estimate the focal length, we use the focal length recovered from [18] to convert the predicted depth to a point cloud. "Ours-Baseline" does not recover the depth shift or focal length and uses an orthographic camera, while "Ours" recovers the shift and focal length. We can see that our method better reconstructs the 3D shape, especially at edges and planar regions (see arrows).

of the unprojected point coordinates, but this approach did not generalize well to samples from unseen datasets.

**Focal length recovery.** To evaluate the accuracy of our recovered focal length, we follow Hold-Geoffroy *et al*. [18] and compare on the 2D-3D-S dataset, which is unseen during training for both methods. The model of [18] is trained on the in-the-wild SUN360 [45] dataset. Results are illustrated in Fig. 4, where we can see that our method demonstrates better generalization performance. Note that PVCNN is very lightweight and only has $5.5M$ parameters, but shows promising generalizability, which could indicate that there is a smaller domain gap between datasets in the 3D point cloud space than in the image space where appearance variation can be large.

Furthermore, we analyze the effect of different initial focal lengths during inference. We set the initial field of view (FOV) from $20°$ to $70°$ and evaluate the accuracy of the recovered focal length, Fig. 4 (right). The experimental results demonstrate that our method is not particularly sensitive to different initial focal lengths.

**Evaluation of 3D shape quality.** Following OASIS [7], we use LSIV for the quantitative comparison of recovered 3D shapes on the OASIS [7] dataset and the 2D-3D-S [1] dataset. OASIS only provides the ground truth point cloud on small regions, while 2D-3D-S covers the whole 3D scene. Following OASIS [7], we evaluate the reconstructed 3D shape with two different camera models, i.e. the orthographic projection camera model [7] (infinite focal length) and the (more realistic) pinhole camera model.

| Method | OASIS LSIV ↓ | 2D-3D-S LSIV↓ |
|---|---|---|
| Orthographic Camera Model | | |
| MegaDepth [24] | 0.64 | 2.68 |
| MiDaS [29] | 0.63 | 2.65 |
| Ours-DPM | 0.63 | 2.65 |
| Pinhole Camera Model | | |
| MegaDepth [24] + Hold-Geoffroy [18] | 1.69 | 1.81 |
| MiDaS [29] + Hold-Geoffroy [18] | 1.60 | 0.94 |
| MiDaS [29] + Ours-PCM | 1.32 | 0.94 |
| Ours + Hold-Geoffroy [18] | 2.66 | 0.99 |
| Ours | **0.52** | **0.80** |

**Table 4:** Quantitative evaluation of the reconstructed 3D shape quality on OASIS and 2D-3D-S. Our method can achieve better performance than previous methods. Compared with the orthographic projection, using the pinhole camera model can obtain better performance. DPM and PCM refers to our depth prediction module and point cloud module respectively.

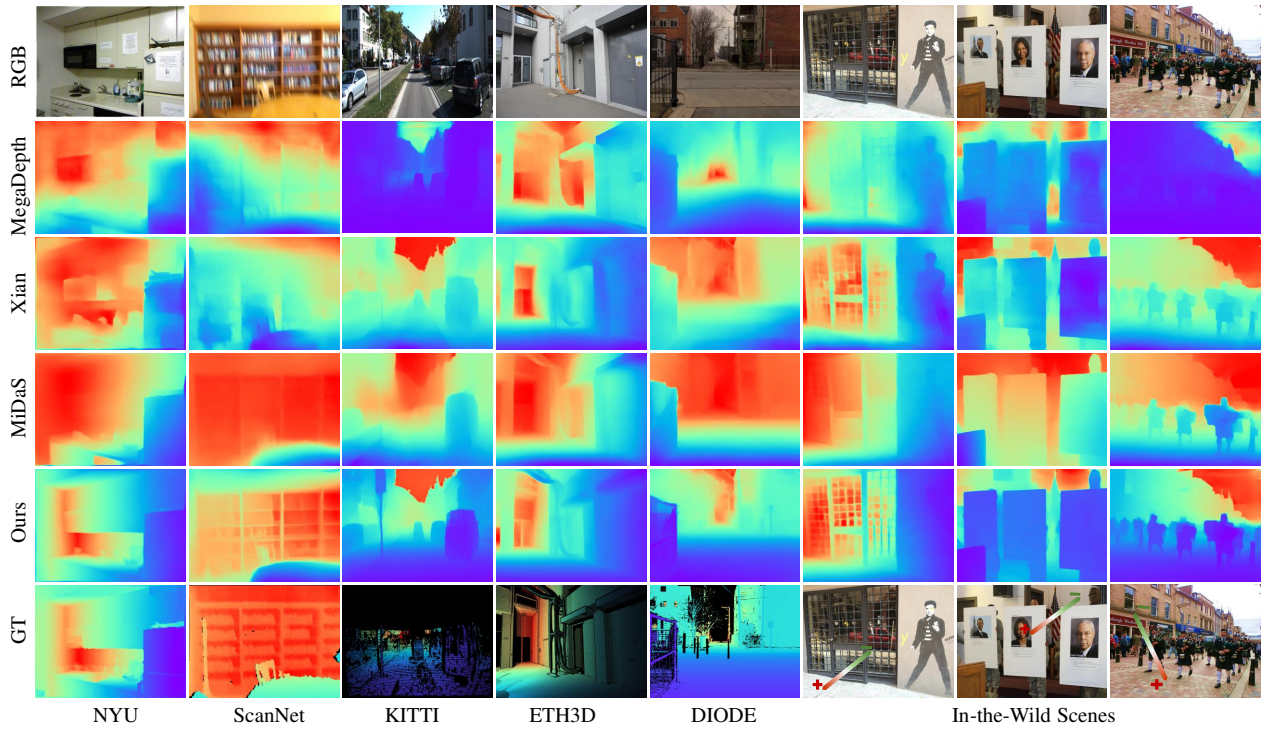


**Figure 6:** Qualitative comparisons with state-of-the-art methods, including MegaDepth [24], Xian *et al*. [44], and MiDaS [29]. It shows that our method can predict more accurate depths at far locations and regions with complex details. In addition, we see that our method generalizes better on in-the-wild scenes.

As MiDaS [29] and MegaDepth [24] do not estimate the focal length, we use the focal length recovered from Hold-Geoffroy [18] to convert the predicted depth to a point cloud. We also evaluate a baseline using MiDaS instead of our DPM with the focal length predicted by our PCM ("MiDaS + Ours-PCM"). From Table 4 we can see that with an orthographic projection, our method ("Ours-DPM") performs roughly as well as existing state-of-the-art methods. However, for the pinhole camera model our combined method significantly outperforms existing approaches. Furthermore, comparing "MiDaS + Ours-PCM" and "MiDaS + Hold-Geoffroy", we note that our PCM is able to generalize to different depth prediction methods.

A qualitative comparison of the reconstructed 3D shape on in-the-wild scenes is shown in Fig. 5. It demonstrates that our model can recover more accurate 3D scene shapes. For example, planar structures such as walls, floors, and roads are much flatter in our reconstructed scenes, and the angles between surfaces (*e.g.*, walls) are also more realistic. Also, the shape of the car has less distortions.

## 4.2. Depth prediction

In this section, we conduct several experiments to demonstrate the effectiveness of our depth prediction method, including a comparison with state-of-the-art methods, an ablation study of our image-level normalized regression loss, and an analysis of the effectiveness of our pair-wise normal regression loss.

**Comparison with state-of-the-art methods.** In this comparison, we test on datasets unseen during training. We compare with methods that have been shown to best generalize to in-the-wild scenes. Their results are obtained by running the publicly released code and weight. When comparing the AbsRel error, we follow Ranftl *et al*. [29] to align the scale and shift before the evaluation. From Table 6, we can see that our method outperforms prior works, and using ResNeXt101 backbone further improves the results. Fig. 6 shows the qualitative comparison.

**Pair-wise normal loss.** To evaluate our full method on edges and planes, we compare our depth model with pre-

| Method | iBims-1 | | | | |
|---|---|---|---|---|---|
| | $\varepsilon_{\text{DBE}}^{\text{acc}}$ ↓ | $\varepsilon_{\text{DBE}}^{\text{comp}}$ ↓ | $\varepsilon_{\text{PE}}^{\text{plan}}$ ↓ | $\varepsilon_{\text{PE}}^{\text{orie}}$ ↓ | AbsRel↓ |
| Xian *et al*. [44] | 7.72 | 9.68 | 5.00 | 44.77 | 0.301 |
| MegaDepth [24] | 4.09 | 8.28 | 7.04 | 33.03 | 0.20 |
| MiDaS [29] | 1.91 | 5.72 | 3.43 | 12.78 | 0.104 |
| 3D Ken Burns [27] | 2.02 | **5.44** | 2.19 | 10.24 | 0.097 |
| Ours† w/o PWN | 2.05 | 6.10 | 3.91 | 13.47 | 0.106 |
| Ours† | 1.91 | 5.70 | 2.95 | 11.59 | 0.101 |
| Ours Full | **1.90** | 5.73 | **2.0** | **7.41** | **0.079** |

**Table 5:** Quantitative comparison of the quality of depth boundaries (DBE) and planes (PE) on the iBims-1 dataset. † indicates when a method was trained on the small training subset.

| Method | Backbone | OASIS | YT3D | NYU | | KITTI | | DIODE | | ScanNet | | ETH3D | | Sintel | | Rank |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | WHDR↓ | | AbsRel↓ | $\delta_1$ ↑ | AbsRel↓ | $\delta_1$ ↑ | AbsRel↓ | $\delta_1$ ↑ | AbsRel↓ | $\delta_1$ ↑ | AbsRel↓ | $\delta_1$ ↑ | AbsRel↓ | $\delta_1$ ↑ | |
| OASIS [7] | ResNet50 | 32.7 | 27.0 | 21.9 | 66.8 | 31.7 | 43.7 | 48.4 | 53.4 | 19.8 | 69.7 | 29.2 | 59.5 | 60.2 | 42.9 | 6.7 |
| MegaDepth [24] | Hourglass | 33.5 | 26.7 | 19.4 | 71.4 | 20.1 | 66.3 | 39.1 | 61.5 | 19.0 | 71.2 | 26.0 | 64.3 | 39.8 | 52.7 | 6.7 |
| Xian *et al*. [44] | ResNet50 | 31.6 | 23.0 | 16.6 | 77.2 | 27.0 | 52.9 | 42.5 | 61.8 | 17.4 | 75.9 | 27.3 | 63.0 | 52.6 | 50.9 | 6.7 |
| WSVD [37] | ResNet50 | 34.8 | 24.8 | 22.6 | 65.0 | 24.4 | 60.2 | 35.8 | 63.8 | 18.9 | 71.4 | 26.1 | 61.9 | 35.9 | 54.5 | 6.6 |
| Chen *et al*. [6] | ResNet50 | 33.6 | 20.9 | 16.6 | 77.3 | 32.7 | 51.2 | 37.9 | 66.0 | 16.5 | 76.7 | 23.7 | 67.2 | 38.4 | 57.4 | 5.6 |
| DiverseDepth [49, 47] | ResNeXt50 | 30.9 | 21.2 | 11.7 | 87.5 | 19.0 | 70.4 | 37.6 | 63.1 | 10.8 | 88.2 | 22.8 | 69.4 | 38.6 | 58.7 | 4.4 |
| MiDaS [29] | ResNeXt101 | 29.5 | 19.9 | 11.1 | 88.5 | 23.6 | 63.0 | 33.2 | 71.5 | 11.1 | 88.6 | 18.4 | 75.2 | 40.5 | 60.6 | 3.5 |
| Ours | ResNet50 | 30.2 | 19.5 | 9.1 | 91.4 | **14.3** | **80.0** | 28.7 | 75.1 | 9.6 | 90.8 | 18.4 | 75.8 | 34.4 | 62.4 | 1.9 |
| Ours | ResNeXt101 | **28.3** | **19.2** | **9.0** | **91.6** | 14.9 | 78.4 | **27.1** | **76.6** | **9.5** | **91.2** | **17.1** | **77.7** | **31.9** | **65.9** | **1.1** |

**Table 6:** Quantitative comparison of our depth prediction with state-of-the-art methods on eight zero-shot (unseen during training) datasets. Our method achieves better performance than existing state-of-the-art methods across all test datasets.


**Figure 7:** Qualitative comparison of reconstructed point clouds. Using the pair-wise normal loss (PWN), we can see that edges and planes are better reconstructed (see highlighted regions).

vious state-of-the-art methods on the iBims-1 dataset. In addition, we evaluate the effect of our proposed pair-wise normal (PWN) loss through an ablation study. The ablation is conducted on the small training subset to to limit the computational overhead. The results are shown in Table 5. We can see that our full method outperforms prior work for this task. In addition, under the same settings, both edges and planes are improved by adding the PWN loss. We further show a qualitative comparison in Fig. 7.

**Image-level normalized regression loss.** To show the effectiveness of our proposed ILNR loss, we compare it with the scale-shift invariant loss (SSMAE) [29] and the scale-invariant multi-scale gradient loss [37]. All methods are trained on the small training subset to limit the computational overhead, and are compared on datasets that are unseen during training. All models have been trained for 50 epochs, and we have verified that all models fully converged by then. The quantitative comparison is shown in Table 7, where we can see an improvement of ILNR over other scale and shift invariant losses. Furthermore, we also analyze different options for normalization, including image-level Min-Max (MinMax) normalization and median absolute deviation (MAD) normalization, and find that our ILNR performs a bit better. Note that tanh term is not enforced for other normalization methods.

| Method | RedWeb | NYU | KITTI | ScanNet | DIODE |
|---|---|---|---|---|---|
| | WHDR↓ | AbsRel↓ | | | |
| SMSG [37] | 19.1 | 15.6 | 16.3 | 13.7 | 36.5 |
| SSMAE [29] | 19.2 | 14.4 | 18.2 | 13.3 | 34.4 |
| MinMax | 19.1 | 15.0 | 17.1 | 13.3 | 46.1 |
| MAD | 18.8 | 14.8 | 17.4 | 12.5 | 34.6 |
| ILNR | **18.7** | **13.9** | **16.1** | **12.3** | **34.2** |

**Table 7:** Quantitative comparison of different losses on zero shot generalization to 5 datasets unseen during training.

## 5. Limitations

We observed a few limitations of our method. For example, our PCM cannot recover accurate focal length or depth shift when the scene does not have enough geometric cues, *e.g*. when the whole image is mostly a sky region. The accuracy of our method will also decrease with images taken from uncommon view angles (e.g., top-down) or extreme focal lengths. More diverse training data may address these failure cases. Furthermore, we do not model the radial distortion and thus the reconstructed scene shape can be distorted in cases with severe radial distortion. Studying how to recover the radial distortion parameters using PCM can be an interesting future direction.

**Conclusion** In summary, we presented, to our knowledge, the first fully data driven method that reconstructs 3D scene shape from a monocular image. To recover the shift and focal length for 3D reconstruction, we proposed to use point cloud networks trained on datasets with known global depth shifts and focal lengths. This approach showed strong generalization capabilities and we are under the impression that it may be helpful for related depth-based tasks. Extensive experiments demonstrated the effectiveness of our scene shape reconstruction method and the superior generalizability to unseen data.

## References


[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv: Comp. Res. Repository*, page 1702.01105, 2017. 5, 6

[2] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(8):1670–1687, 2014. 2

[3] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 35–45, 2019. 5

[4] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *Proc. Eur. Conf. Comp. Vis.*, pages 611–625. Springer, 2012. 5

[5] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 730–738, 2016. 1, 2

[6] Weifeng Chen, Shengyi Qian, and Jia Deng. Learning single-image depth from videos using quality assessment networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5604–5613, 2019. 2, 5, 8

[7] Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. Oasis: A large-scale dataset for single image 3d in the wild. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 679–688, 2020. 1, 2, 5, 6, 8

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5828–5839, 2017. 5

[9] Jonathan Deutscher, Michael Isard, and John MacCormick. Automatic camera calibration from a single manhattan image. In *Proc. Eur. Conf. Comp. Vis.*, pages 175–188. Springer, 2002. 2

[10] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 2366–2374, 2014. 1, 2

[11] Jose M Facil, Benjamin Ummenhofer, Huizhong Zhou, Luis Montesano, Thomas Brox, and Javier Civera. Camconvs: camera-aware multi-scale convolutions for single-view depth. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 11826–11835, 2019. 1

[12] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Elsevier, 2013. 4

[13] Salvador García, Julián Luengo, and Francisco Herrera. *Data preprocessing in data mining*. Springer, 2015. 4

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3354–3361. IEEE, 2012. 5

[15] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2019. 5

[16] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 770–778, 2016. 5

[18] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Jonathan Eisenmann, Matthew Fisher, Emiliano Gambaretto, Sunil Hadap, and Jean-François Lalonde. A perceptual measure for deep single image camera calibration. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2354–2363, 2018. 2, 5, 6, 7

[19] Yiwen Hua, Puneet Kohli, Pritish Uplavikar, Anand Ravi, Saravana Gunaseelan, Jason Orozco, and Edward Li. Holopix50k: A large-scale in-the-wild stereo image dataset. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, June 2020. 5

[20] Olga A Karpenko and John Hughes. Smoothsketch: 3d free-form shapes from complex sketches. In *ACM. T. Graph. (SIGGRAPH)*, pages 589–598. 2006. 2

[21] Youngjung Kim, Hyungjoo Jung, Dongbo Min, and Kwanghoon Sohn. Deep monocular depth estimation via integration of global and local predictions. *IEEE Trans. Image Process.*, 27(8):4131–4144, 2018. 5

[22] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of CNN-based single-image depth estimation methods. In *Eur. Conf. Comput. Vis. Worksh.*, pages 331–348, 2018. 5

[23] Manuel Lang, Alexander Hornung, Oliver Wang, Steven Poulakos, Aljoscha Smolic, and Markus Gross. Nonlinear disparity mapping for stereoscopic 3d. *ACM Trans. Graph.*, 29(4):1–10, 2010. 1

[24] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2041–2050, 2018. 1, 2, 4, 6, 7, 8

[25] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(10):2024–2039, 2015. 2

[26] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. In *Proc. Advances in Neural Inf. Process. Syst.*, 2019. 3, 5

[27] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM Trans. Graph.*, 38(6):184:1–184:15, 2019. 5, 7

[28] Emmanuel Prados and Olivier Faugeras. Shape from shading: a well-posed problem? In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, volume 2, pages 870–877. IEEE, 2005. 2

[29] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020. 1, 2, 4, 5, 6, 7, 8

[30] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2304–2314, 2019. 2

[31] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 84–93, 2020. 2

[32] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(5):824–840, 2008. 2

[33] Thomas Schops, Johannes L Schonberger, 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3260–3269, 2017. 5

[34] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proc. Eur. Conf. Comp. Vis.*, pages 746–760. Springer, 2012. 4, 5

[35] Dalwinder Singh and Birmohan Singh. Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, page 105524, 2019. 4

[36] 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: Comp. Res. Repository*, page 1908.00463, 2019. 5

[37] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. In *Int. Conf. 3D. Vis.*, pages 348–357. IEEE, 2019. 1, 2, 4, 8

[38] Lijun Wang, Jianming Zhang, Oliver Wang, Zhe Lin, and Huchuan Lu. Sdc-depth: Semantic divide-and-conquer network for monocular depth estimation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 541–550, 2020. 1

[39] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single RGB images. In *Proc. Eur. Conf. Comp. Vis.*, pages 52–67, 2018. 2

[40] Xinlong Wang, Wei Yin, Tao Kong, Yuning Jiang, Lei Li, and Chunhua Shen. Task-aware monocular depth estimation for 3d object detection. In *Proc. AAAI Conf. Artificial Intell.*, 2020. 1, 2

[41] Scott Workman, Connor Greenwell, Menghua Zhai, Ryan Baltenberger, and Nathan Jacobs. Deepfocal: A method for direct focal length estimation. In *Proc. IEEE Int. Conf. Image Process.*, pages 1369–1373. IEEE, 2015. 2

[42] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William Freeman, and Joshua Tenenbaum. Learning shape priors for single-view 3d completion and reconstruction. In *Proc. Eur. Conf. Comp. Vis.*, pages 646–662, 2018. 2

[43] 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 311–320, 2018. 1, 2, 5

[44] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 611–620, 2020. 1, 2, 4, 5, 7, 8

[45] Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2695–2702. IEEE, 2012. 6

[46] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1492–1500, 2017. 5

[47] Wei Yin, Yifan Liu, and Chunhua Shen. Virtual normal: Enforcing geometric constraintsfor accurate and robust depth prediction. *arXiv: Comp. Res. Repository*, page 2103.04216, 2021. 1, 2, 8

[48] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2019. 1, 2, 4

[49] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv: Comp. Res. Repository*, page 2002.00569, 2020. 1, 2, 5, 8

[50] Amir Zamir, Alexander Sax, , William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* IEEE, 2018. 4, 5

[51] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(11):1330–1334, 2000. 2