

Recovering Detail in 3D Shapes Using Disparity Maps

Marissa Ramirez de Chanlatte¹, Matheus Gadelha², Thibault Groueix², and
Radomir Mech²

¹ University of California, Berkeley

² Adobe Research

Abstract. We present a fine-tuning method to improve the appearance of 3D geometries reconstructed from single images. We leverage advances in monocular depth estimation to obtain disparity maps and present a novel approach to transforming 2D normalized disparity maps into 3D point clouds by using shape priors to solve an optimization on the relevant camera parameters. After creating a 3D point cloud from disparity, we introduce a method to combine the new point cloud with existing information to form a more faithful and detailed final geometry. We demonstrate the efficacy of our approach with multiple experiments on both synthetic and real images.

1 Introduction

Reconstructing full 3D shapes from single RGB images is a longstanding computer vision problem. A system capable of performing such task needs to have some understanding of *pixels* – how the scene illumination interacts with the geometry and materials to create that image – and *shapes* – their structure, how they are usually decomposed and which symmetries arise from their function and style. To develop such a system, a learning-based procedure would require an abundance of *paired* data in those two forms. Unfortunately, a practical limitation comes from the fact that we have significantly more access to images (pixels) than 3D models (shapes).

There are many learning-based techniques that encode knowledge about *shapes* [1–3]. Most of them are trained using data from ShapeNet [4], a 3D mesh dataset with tens of thousands of shapes. Models trained using this data can perform various conditional shape generation tasks, like shape interpolation, completion and reconstruction from single views. Most of the approaches trained on ShapeNet have only a very rudimentary knowledge of pixels from real images. While we are able to generate images using 3D meshes (i.e. rendering), those images are very different from the ones we capture from our cameras in the real world. Despite the rich behaviors that can be simulated through computer graphics, the research community is far from recreating the abundant visual experience we encounter in our lives. On the other hand, monocular depth estimation models are capable of inferring the partial scene geometry quite well,

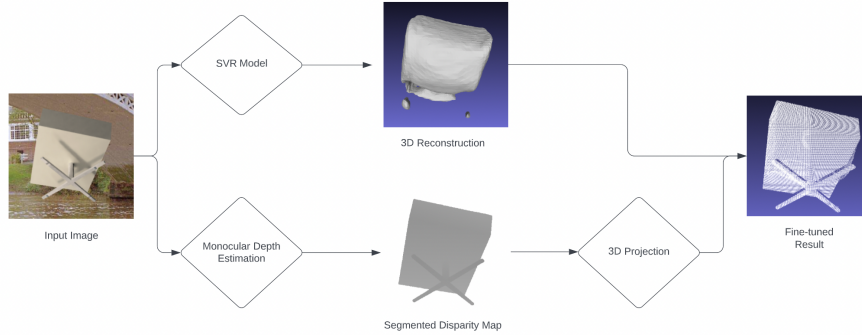


Fig. 1: Our method projects the output of a monocular depth estimator to 3D, using it to improve the appearance of a 3D geometry from a single view reconstruction model.

even on real images. Unlike models trained on ShapeNet, depth estimation techniques employ a variety of training data – synthetic scenes, scanners, stereoscopic images, structure from motion, and so on. The variety of training data leads to models capable of performing the task remarkably well on a variety of scenarios including real images.

The central proposition of this work is to use the knowledge of *pixels* from depth estimation approaches to improve models estimating *shapes* from single RGB images. A key challenge lies in the fact that, despite their name, monocular depth estimation techniques actually predict disparity, the distance between two points in a stereo image pair, rather than depth. *In order to recover true depth, one still needs to estimate intrinsic camera parameters and solve scale and distance ambiguities.* This small but important step has been understudied. Whereas other works have attempted to convert disparity information to depth [5], they are restricted to visible portions of full scenes and do not leverage any knowledge of particular shapes. Our solution, which to our knowledge is the first to address converting disparity to depth on single shapes, is to use shape estimates coming from vanilla single-view reconstruction models to recover image parameters that align the disparity estimations with the object coordinate system. The geometry extracted from the disparity estimation models can then be used to refine 3D shapes from single-view reconstruction techniques. Notably, our method can be applied to any view-centric single-view reconstruction technique without needing extra training or additional data, unlike data-driven approaches such as [5]. Similarly, as monocular depth estimation models get better, their improvements can be immediately translated into their single-view reconstruction counterparts. Finally, we demonstrate the efficacy of our technique in experiments with both synthetic and real images.

2 Related Work

Single-view 3D reconstruction (SVR). Computational approaches to SVR date back to at least Roberts’ PhD thesis [6] in 1963, and the problem was already known to be ill-posed by the scientist Alhazen in the 11th century. Most related to our work are recent deep learning approaches which currently present state-of-the-art results for this problem.

Deep Networks for 3D Shape Representation. The choice of 3D data representation is one of the main discriminating factors to analyze deep learning approaches for SVR. Choy et al. [1] propose a volumetric representation using 3D voxel-grids [2, 3], as a natural extension to 2D pixel grids on which to perform 3D convolutions. Several solutions have been explored to mitigate the prohibitive memory consumption of using 3D grids in deep learning. Previously, these solutions were focused on occupancy grids [7–9], but several more recent radiance fields alternatives using hash tables and tensor decomposition [10–13] were also proposed. Instead of using grids, three papers [14–16] concurrently pioneered the use coordinate-based Multi-Layered Perceptrons to model a 3D volume [17]. Several other approaches model a surface instead of a volume. Fan et al. [18] pioneered an approach to generate point clouds on a surface [19, 20]. Other techniques model 3D surfaces via parametric deformations from a reference surface [21–23]. Despite a lot of progress in representing shapes, all the aforementioned SVR methods suffer when applied to real images. Our approach can be thought of as a modular extension to some of these methods that leverages the knowledge of monocular depth estimation techniques, which can be trained on real images, thus generalizing better to more realistic scenarios.

Single-view 3D reconstruction from real images. There are two main strategies in SVR to target performance on real images. The first one is through the training data *i.e.* to have real images in the training set. Existing datasets with associated ground truth 3D model have clear limitations. Pix3D [24] does not have good diversity in terms of shapes (only about 700) and mostly contains furniture. ObjectNet [25] and Pascal3D [26] don’t have good image-shape alignment. Some approaches aim to learn SVR using image-only datasets using differentiable reprojection losses [27–33], though this is an extremely challenging task. Another strategy is to use domain adaption techniques to bridge the gap between synthetic renderings and real images, for instance by imposing depth and normal as an intermediate representation between RGB images and the full 3D geometry [34–37]. These techniques show some improvement on real images but within the scope of the categories spanned by the 3D dataset. They thus don’t fully leverage the generality of monocular depth estimation methods. Whereas those approaches perform better than vanilla models trained only on ShapeNet or Pix3D, they still fail to generalize to real scenarios. Our method is complementary to these and can be used along monocular depth estimation in addition to domain adaptation techniques.

Monocular depth estimation. In their seminal work make3D, Saxena et al. [38] cast depth estimation as a supervised learning problem trained on a dataset of

laser scans. Several subsequent papers have improved the architectures [39–43], the losses [44–46] and post-processing steps [47, 48]. We organize our discussion around the training data since the robustness and generality of monocular depth estimation approaches stems from the diversity of the training datasets [49]. Laser scanners [50–52] based on time-of-flight as well as sensors based on structured light [53, 54] provide ground truth depth but sparse annotations for dynamic scenes. Structure-from-motion can also be used to obtain sparse 3D ground-truth, up to scale, from multi-view images of a static scene [55]. Garg et al. [56] propose to use rectified stereo pairs as supervision [57–59] but the corresponding datasets are not all calibrated, providing disparity up to scale and shift. Chen et al. created a dataset where ordinal relationships between pixels are manually annotated [60]. MidasNet [49] pioneered leveraging those diverse sources of data by estimating normalized disparity and achieving breakthrough results in generalization capabilities of monocular depth estimation. Recent monocular depth estimation models [5, 48] show remarkable performance in various scenarios. Those models can capture the visible part of the scene geometry but have little knowledge about shapes. For example if three legs of a chair are visible, the depth estimation models will not be very helpful for completing the shape and generating the remaining leg. On the other hand, SVR models might miss fine details of visible parts of the chair that were captured by the depth estimation network, but will most certainly create a four-legged chair because it is capable of reasoning about the overall shape structure – it was supervised using complete chair meshes during training. Our goal is to try to get the best of both worlds: more details when reconstructing shapes from real images while maintaining the coarse estimation from SVR models.

3 Method

We start with an off-the-shelf and pre-trained single-image predictor that produces an unsatisfactory result and aim to fine-tune that result to better match the input image. Take for example an image of a chair. The network may produce a geometry that looks like a chair, but perhaps the shape of the legs or detail in the back is inaccurate. Our method takes that initial prediction and uses a disparity map to improve the appearance of the geometry.

The fine-tuning procedure is as follows:

1. Get a point cloud, Λ , representation of the initial prediction.
2. Split Λ into two sub-pointclouds of points that are visible, Λ_{vis} , and occluded, Λ_{occ} , from the camera viewpoint of the input image.
3. Given a 2D disparity map of image I called I_{disp} find the focal length, scale, displacement, and translation constants, $\omega = \{fov, s, t, z_t\}$, necessary to project I_{disp} into Ψ , the 3D point cloud most closely resembling Λ_{vis} .
4. Combine Λ_{occ} and Ψ and remove any points from Λ_{occ} that are now visible. Call this cleaned and combined final point cloud Ψ' .

In step 1, we start with a point cloud representation of the initial prediction of our object. This point cloud can come from any geometry including a mesh or implicit function. We are not limited to networks that produce any particular type of geometry.

After we have our initial prediction, we split the point cloud into visible and occluded points. The idea here is to leverage the information we have from the disparity map while keeping the points for which we have no disparity information (the occluded points) the same.

In step 3, we take the predicted 2D disparity map and convert it into a 3D point cloud that will replace the visible part of the object. This is the most technically challenging step and our key contribution. Because of its importance we will discuss the details of this step further in the next section

The final step is to combine the occluded point cloud, Λ_{occ} from the initial prediction with the projected disparity points, Ψ . Then a final cleaning is performed, to remove any occluded points that are no longer occluded by the new visible geometry. This is done by iteratively removing points from Λ_{occ} that are not occluded by the projected disparity points, Ψ . This is to account for errors in the initial prediction. Imagine looking at the back of a chair that is mostly open, connected only by spokes. If the initial prediction erroneously reconstructed the chair with a solid back, replacing the visible points with the projected disparity map should restore this detail, but only on the visible side of the back of the chair. The other side will still be solid. We must also remove extraneous points from the original set of occluded points, Λ_{occ} . The final result is the point cloud Ψ' . For simplicity, we leave the geometry as a point cloud in the experiments, but it can easily be transformed into a mesh [61] or an implicit function [62] as desired because Ψ' is dense. This meshing step will also get rid of any discontinuity between the previous occluded points and the new visible points added from the depth map.

3.1 Disparity Projection

The heart of this method is the disparity projection. In the literature, people often discuss the task of “depth” prediction of a single image, but in most cases what is being predicted is disparity, or the difference between two matching points in two stereo images. Disparity is related to inverse depth by an affine transformation and is used in monocular depth estimation [63, 64, ?] for several reasons. The first being the availability of disparity data and the difficulty of converting disparity to depth. In mixed datasets it is easier to convert depth maps to disparity maps than the other way around, so disparity maps are popular training sets. Second, because of its relationship to inverse depth, using a disparity map for training effectively weights the loss function by depth, biasing towards objects in the foreground.

Without knowledge of several camera parameters, it is impossible to get true depth from normalized disparity. We assume we are starting with a normalized disparity map d which is related to true depth, Z , through the following

relationship:

$$\frac{1}{Z} = \frac{d - (c_x^R - c_x^L)}{fb}, \quad (1)$$

where b is the camera baseline, f is the focal length, and c_x^R, c_x^L are the horizontal coordinates of the principal points from the stereo camera pair. For simplicity, we rewrite this equation as

$$\frac{1}{Z} = s * d + t, \quad (2)$$

where $s = \frac{1}{fb}$ and $t = -\frac{(c_x^R - c_x^L)}{fb}$.

These equations are used to convert disparity to depth, but then depth coordinates (u, v) must be converted to 3D coordinates (X, Y, Z) to produce the point cloud necessary for our method. To do this, we invert the perspective projection equation in the pinhole camera model:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \frac{1}{Z} \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \quad (3)$$

where (f_x, f_y) is the focal length and (u_0, v_0) is the optical center of the image plane.

Solving for (X, Y, Z) we get

$$u = \frac{f}{Z}X + u_0 \implies X = \frac{(u - u_0)}{f_x}Z \quad (4)$$

$$v = \frac{f}{Z}Y + v_0 \implies Y = \frac{(v - v_0)}{f_y}Z \quad (5)$$

where f is the focal length defined as $f = \frac{x}{2\tan(fov/2)}$ and fov is the field of view and x is the diagonal of the image. We assume (u_0, v_0) to be $(0, 0)$. When we are given only a normalized disparity map, s, t , and fov are all unknown, so we must find acceptable approximations of these parameters in order to get the desired 3D point cloud.

The pinhole camera model assumes the origin of the coordinate system to be the camera position, while most view-centered reconstruction models assume the reconstructed geometry to be placed at the origin. To relate the two coordinate systems, we need to predict an additional parameter, a translation constant z_t . We translate our projection to get $Z' = Z + z_t$ where Z' is aligned with the initial object reconstruction. In total this gives a set four parameters, $\omega = \{s, t, fov, z_t\}$ that we must predict.

Combined, the equations above form a function $f(s, t, fov, z_t, d) = \Psi$ which takes a normalized disparity map along with other parameters and transforms it into the corresponding 3D point cloud, Ψ . To approximate the parameters ω , we fix the disparity map, d , and use stochastic gradient descent to minimize the Chamfer distance between Ψ and the initially predicted visible points A_{vis} .

$$loss = \frac{1}{|\Psi|} \sum_{x \in \Psi} \min_{y \in \Lambda_{vis}} \|x - y\|_2^2 + \frac{1}{|\Lambda_{vis}|} \sum_{y \in \Lambda_{vis}} \min_{x \in \Psi} \|x - y\|_2^2 \quad (6)$$

The parameters that minimize Eq. 6 are used to create the final point cloud used throughout the rest of the fine-tuning procedure.

The intuition for the approach comes from the assumption that the initial prediction is roughly the right general shape and size as the desired object, and is only lacking in detail. By fitting the disparity map to the initial visible points, we should recover camera parameters that lead to a point cloud of roughly the right shape.

It is important to note that this is an overdetermined function. There are infinite choices of ω that will result in the same point cloud. There are also many local minima. This makes this approach very sensitive to initialization. To improve robustness, we perform the fitting several times, each time randomly initializing, and then choosing ω that resulted from initialization with best loss as the final parameters.

4 Experiments

To validate our method, we perform experiments on two different off-the-shelf reconstruction networks that were trained on two different data sets. The first dataset is a synthetic dataset formed by rendering ShapeNet objects in different poses and lighting conditions. The second is Pix3d, a real-world data set. In both cases we focus on the category of chairs.

We evaluated the final result using two metrics: f-score and Chamfer distance.

4.1 Sensitivity Studies

Before discussing full-scale experiments, we first present two studies to examine the sensitivity of our method to errors in the disparity map and the initialization of parameters.

To examine the effect of error in disparity prediction, we simulate error by uniformly adding noise to ground truth disparity maps. Using our disparity projection method, we estimate camera parameters. We then use those camera parameters to project the ground truth disparity and measure the Chamfer distance from the ground truth object. As can be seen in Fig. 2, the Chamfer distance stays pretty small for small amounts of noise, but as more noise is added, the Chamfer distance increases, indicating some level of sensitivity to errors in prediction. However, it is important to remember that errors in disparity prediction are often not uniform. The more common failure mode is that errors are

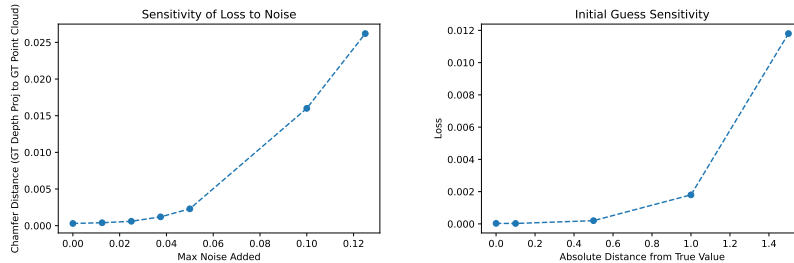


Fig. 2: **Left:** our 3D projection is robust to some amount of noise in the input disparity map, but deteriorates if the noise level goes beyond 0.05. **Right:** the error in the projection grows as a function of the distance between the initial guess and the true camera parameters, because of local minima in our optimization problem. We address this issue by randomly choosing multiple initializations.

concentrated in one area (i.e. the back leg predicted to be closer than it actually is, or fading into the background completely).

The second case of sensitivity we examine is sensitivity to initialization. To measure this we pick a case in which the initial parameters are known and perturb the initial guess steadily further away from the known value. As can be seen in Fig. 2, the final loss grows exponentially with distance from the initial value, indicating a significant sensitivity. This finding motivates the random initialization seen in our algorithm. We run the fitting multiple times and pick the final result with the lowest loss.

4.2 Synthetic Images

We present two oracle experiments on synthetic data to clarify the sources of errors in our approach. In this section, we use a dataset of renderings of ShapeNet [4] objects from 3DShapeGen authors [17]. We use their image-only network to perform single-image reconstruction.

First, we try merging the visible points of the ground truth to the 101 initial predictions provided by 3DShapeGen [17]. This achieves the maximum benefit possible with a perfect disparity map and a perfect projection of this map to 3D. On average, each object saw an average improvement of 180% in the f-score at 1 and a 40% improvement in Chamfer distance. This result demonstrates the potential for depth maps to improve the overall appearance and detail of a reconstructed object by an SVR method.

As was discussed in section 3.1, the challenge is that in most cases, monocular depth estimation does not predict absolute depth maps, but instead normalized disparity maps. We use ground truth maps to eliminate any confusion about the source of error. We further test our parameter fitting method on ground truth disparity maps, assuming an oracle monocular depth estimation. We took the same 101 object predictions from 3DShapeGen, applied our method on the

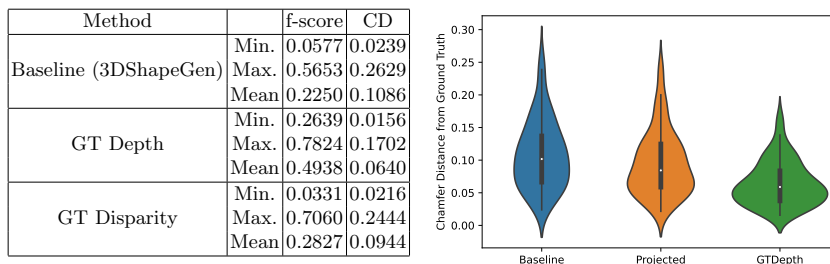


Fig. 3: We compare a baseline SVR model, 3DShapeGen [17], with two oracle versions of our approach using ground truth depth, and projected ground truth disparities, on synthetic renderings of ShapeNet chairs. Using ground truth depth yield a 180% improvement in the f-score@1 in average. Despite not knowing camera parameters perfectly, our method still provides a 32% improvement in the f-score@1 in average.

ground truth normalized disparity maps and merged the initial prediction and projected disparity maps. We found a 32% improvement in f-score at 1 and a 12% improvement in Chamfer distance. While the magnitude of the improvement is not as large as with the ground truth, there is still a significant improvement. Figure 3 compares the performance of the baseline 3DShapeGen prediction, the oracle with ground truth depth and the oracle with ground truth disparity. We show qualitative examples in Figure 7.

4.3 Real Images

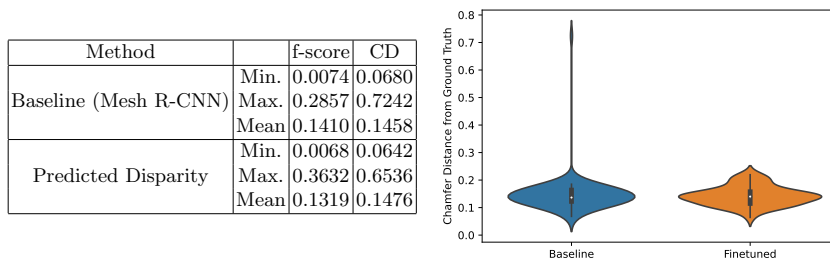


Fig. 4: We compare the performance of Mesh R-CNN [23] with and without our fine-tuning, on real images from the Pix3d dataset [24]. We use AdelaiDepth [5] to predict disparities. See section 4.4 for a detailed discussion on performance.

We proceed to test our method on real images. We use 88 images of chairs from Pix3D [24] as our dataset and Mesh R-CNN [23] as our initial method. In this experiment, we use AdelaiDepth [5] to predict disparity, but use ground truth segmentation masks, again to isolate error. In Figure 4, we show the distribution of Chamfer distances among the objects in the dataset. After application of our method you can see a slight overall improvement in Chamfer distance.

We found that 44% of objects improved in Chamfer distance from the ground truth with application of our method. In cases that did not see improvement, it was most often from clear errors in the depth estimation. We show qualitative examples in Figure 8.

4.4 Discussion

The experiments demonstrate the great potential of depth for fine-tuning 3D geometries and the success of our disparity projection approach. In particular, our method is good for recovering fine detail that SVR models often miss. An example is shown in Fig. 7. Take a look at the chairs in the first and fourth rows. The single-view reconstruction model, reconstructs two fairly similar looking chairs with rounded back and no arms, roughly accurate in shape, but missing most of the detail. Applying our fine-tuning method restores the fine detail that is necessary to distinguish the two chairs.

Another great use case is fine detail in the legs. Again looking at Fig. 7, rows two, five, six, and seven are all examples of chairs with very thin, short, or otherwise unusual legs that the single view reconstruction model failed to reconstruct. Our model was able to reconstruct these legs, creating a more complete and distinctive object.

The objects that fail to see significant improvement in appearance after our fine tuning method tend to fall into two categories: insurmountable errors in the initial prediction and insurmountable errors in depth estimation.

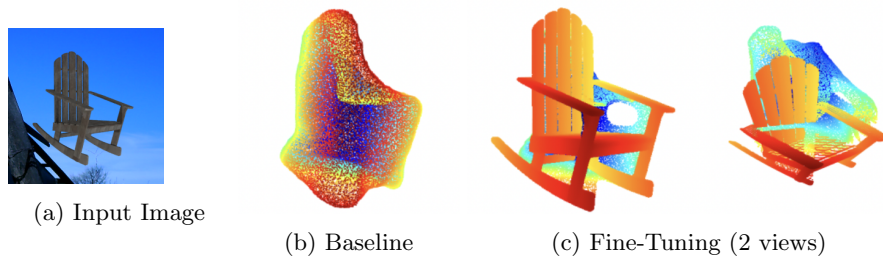


Fig. 5: An example of a flattened projection due to errors in the initial prediction.

Fig. 5 shows a typical example of an insurmountable error in the initial prediction, where the network creates a blobby initial prediction. Because of this blobby shape, the visible points of the initial prediction are mostly flat and

the estimated disparity is thus projected to a flat point cloud even though the actual visible pixels are not flat.

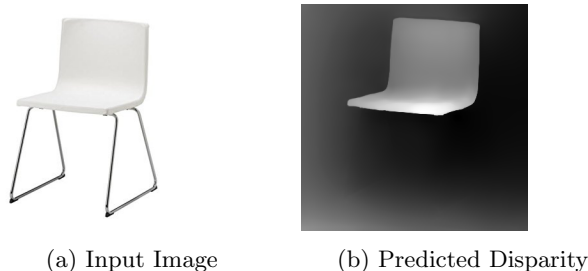


Fig. 6: An insurmountable mistake in disparity prediction: the legs are treated as background.

The second major source of errors lies in the disparity estimation. Our method is fairly robust to minor errors which will occur in any depth estimation method, but there are large fundamental errors that are insurmountable with our approach. An example is the chair in Fig. 6 where the monocular depth estimation failed to detect the legs of the chair, so they blend in with the background. Even with perfect segmentation and camera parameter estimation, the legs of the chair would be placed in the background away from the chair. These types of errors account for most of the objects in our dataset that failed to improve, as evidenced by the difference in performance between the ground truth normalized disparity maps and the predicted ones. As monocular depth estimation continues to improve, these errors will become less common.

5 Conclusion

In summary, we have demonstrated how to improve the predictions of single-view reconstruction models using monocular-depth estimation. We presented a method to approximate the relevant camera parameters to convert normalized disparity, from an off-the-shelf monocular depth model, to a 3D point cloud. We tested our method via multiple experiments, quantified and analysed the improvement that disparity maps can provide on real images.

Acknowledgement

This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231.

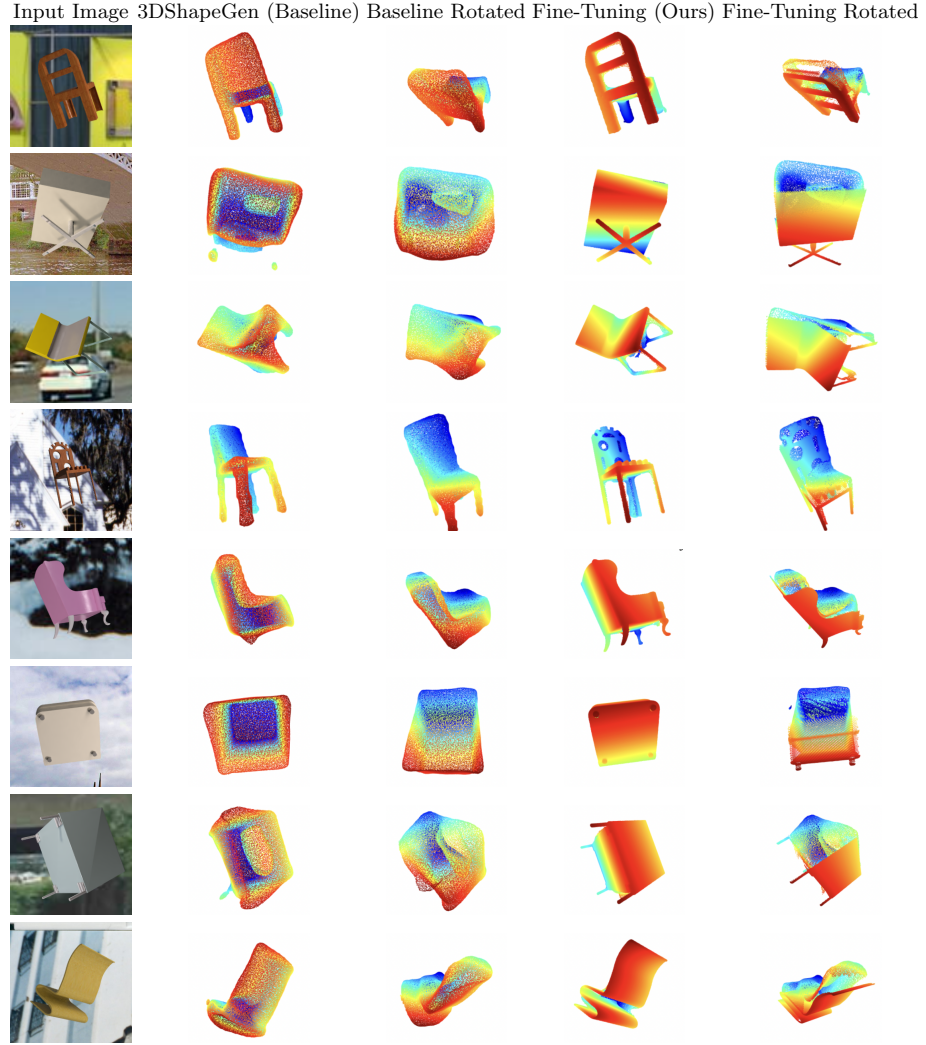


Fig. 7: Selected qualitative results from study on synthetic images with ground truth normalized disparity. Our method successfully adds much needed detail to the baseline reconstruction. That detail is preserved not just from the view of the image, but upon rotation as well.

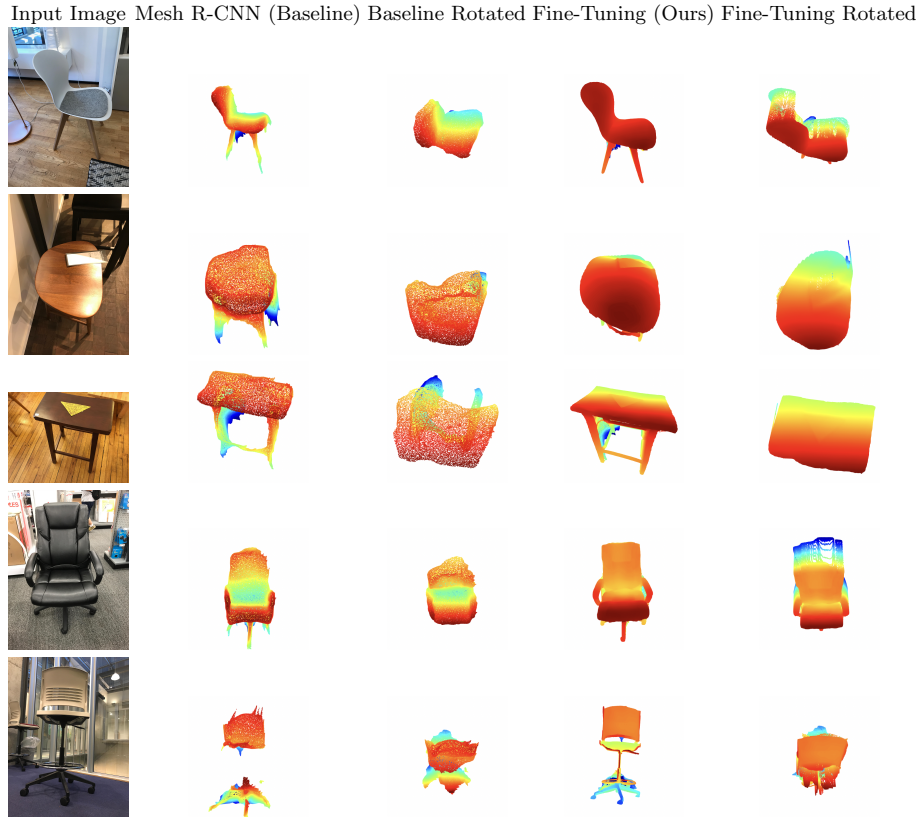


Fig. 8: Selected qualitative results from the study on real images from Pix3D with estimated normalized disparity.

References

1. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In: *eccv*. (2016)
2. Mittal, P., Cheng, Y.C., Singh, M., Tulsiani, S.: Autosdf: Shape priors for 3d completion, reconstruction and generation. In: *CVPR*. (2022)
3. Wu, J., Zhang, C., Xue, T., Freeman, W.T., Tenenbaum, J.B.: Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In: *Advances in Neural Information Processing Systems*. (2016) 82–90
4. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: An Information-Rich 3D Model Repository. In: *Arxiv*. (2015)
5. Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., Shen, C.: Learning to recover 3d scene shape from a single image. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. (June 2021) 204–213
6. Roberts, L.G.: Machine perception of three-dimensional solids. (1963)

7. Häne, C., Tulsiani, S., Malik, J.: Hierarchical surface prediction for 3D object reconstruction. In: 3DV. (2017)
8. Riegler, G., Ulusoy, A.O., Bischof, H., Geiger, A.: OctNetFusion: Learning depth fusion from data. In: 3DV. (2017)
9. Tatarchenko, M., Dosovitskiy, A., Brox, T.: Octree generating networks: Efficient convolutional architectures for high-resolution 3D outputs. In: iccv. (2017)
10. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: CVPR. (2020)
11. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. (2022)
12. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.* **41**(4) (July 2022) 102:1–102:15
13. Sara Fridovich-Keil and Alex Yu, Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: CVPR. (2022)
14. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: DeepSDF: Learning continuous signed distance functions for shape representation. In: CVPR. (2019)
15. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: CVPR. (2019)
16. Chen, Z., Zhang, H.: Learning implicit fields for generative shape modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019)
17. Thai, A., Stojanov, S., Upadhyay, V., Rehg, J.M.: 3d reconstruction of novel object shapes from single images. In: 2021 International Conference on 3D Vision (3DV), IEEE (2021) 85–95
18. Fan, H., Su, H., Guibas, L.: A point set generation network for 3D object reconstruction from a single image. In: CVPR. (2017)
19. Gadelha, M., Wang, R., Maji, S.: Multiresolution tree networks for 3d point cloud processing. In: ECCV. (2018)
20. Luo, S., Hu, W.: Diffusion probabilistic models for 3d point cloud generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (June 2021) 2837–2845
21. Groueix, T., Fisher, M., Kim, V.G., Russell, B., Aubry, M.: AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). (2018)
22. Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2mesh: Generating 3d mesh models from single rgb images. In: ECCV. (2018)
23. Gkioxari, G., Malik, J., Johnson, J.: Mesh r-cnn. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 9785–9795
24. Sun, X., Wu, J., Zhang, X., Zhang, Z., Zhang, C., Xue, T., Tenenbaum, J.B., Freeman, W.T.: Pix3d: Dataset and methods for single-image 3d shape modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2974–2983
25. Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J.B., Katz, B.: Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In: NeurIPS. (2019)
26. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond pascal: A benchmark for 3d object detection in the wild. In: IEEE Winter Conference on Applications of Computer Vision (WACV). (2014)

27. Vasudev, K.A., Gupta, A., Tulsiani, S.: Pre-train, self-train, distill: A simple recipe for supersizing 3d reconstruction. In: Computer Vision and Pattern Recognition (CVPR). (2022)
28. Li, X., Liu, S., Kim, K., De Mello, S., Jampani, V., Yang, M.H., Kautz, J.: Self-supervised single-view 3d reconstruction via semantic consistency. In: ECCV. (2020)
29. Kanazawa, A., Tulsiani, S., Efros, A.A., Malik, J.: Learning category-specific mesh reconstruction from image collections. In: ECCV. (2018)
30. Goel, S., Kanazawa, A., , Malik, J.: Shape and viewpoints without keypoints. In: ECCV. (2020)
31. Tulsiani, S., Kulkarni, N., Gupta, A.: Implicit mesh reconstruction from unannotated image collections. arXiv preprint arXiv:2007.08504 (2020)
32. Lin, C.H., Wang, C., Lucey, S.: Sdf-srn: Learning signed distance 3d object reconstruction from static images. In: Advances in Neural Information Processing Systems (NeurIPS). (2020)
33. Ye, Y., Tulsiani, S., Gupta, A.: Shelf-supervised mesh prediction in the wild. In: Computer Vision and Pattern Recognition (CVPR). (2021)
34. Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, W.T., Tenenbaum, J.B.: MarrNet: 3D Shape Reconstruction via 2.5D Sketches. In: Advances In Neural Information Processing Systems. (2017)
35. Wu, J., Zhang, C., Zhang, X., Zhang, Z., Freeman, W.T., Tenenbaum, J.B.: Learning 3D Shape Priors for Shape Completion and Reconstruction. In: European Conference on Computer Vision (ECCV). (2018)
36. Zhang, X., Zhang, Z., Zhang, C., Tenenbaum, J.B., Freeman, W.T., Wu, J.: Learning to Reconstruct Shapes From Unseen Classes. In: Advances in Neural Information Processing Systems (NeurIPS). (2018)
37. Shin, D., Fowlkes, C., Hoiem, D.: Pixels, voxels, and views: A study of shape representations for single view 3d object shape prediction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018)
38. Saxena, A., Sun, M., Ng, A.Y.: Make3d: Learning 3d scene structure from a single still image. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. (2009)
39. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. (2014)
40. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 3D Vision (3DV), 2016 Fourth International Conference on, IEEE (2016) 239–248
41. Roy, A., Todorovic, S.: Monocular depth estimation using neural regression forest. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 5506–5514
42. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. (2015)
43. Li, R., Xian, K., Shen, C., Cao, Z., Lu, H., Hang, L.: Deep attention-based classification network for robust depth prediction. In: Asian Conference on Computer Vision, Springer (2018) 663–678
44. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: 2015 IEEE International Conference on Computer Vision (ICCV). (2015)

45. Su, W., Zhang, H., Li, J., Yang, W., Wang, Z.: Monocular depth estimation as regression of classification using piled residual networks. (2019)
46. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018)
47. Bo Li, Chunhua Shen, Yuchao Dai, van den Hengel, A., Mingyi He: Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
48. Miangoleh, S.M.H., Dille, S., Mai, L., Paris, S., Aksoy, Y.: Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. (2021)
49. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2020)
50. Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
51. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. (2012) 3354–3361
52. Hansard, M., Lee, S., Choi, O., Horaud, R.: Time of Flight Cameras: Principles, Methods, and Applications. (10 2012)
53. Khoshelham, K., Elberink, S.O.: Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors (Basel, Switzerland)* **12** (2012) 1437 – 1454
54. Fankhauser, P., Bloesch, M., Rodriguez, D., Kaestner, R., Hutter, M., Siegwart, R.: Kinect v2 for mobile robot navigation: Evaluation and modeling. In: 2015 International Conference on Advanced Robotics (ICAR). (2015) 388–394
55. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Computer Vision and Pattern Recognition (CVPR). (2018)
56. Garg, R., Kumar, B.V., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: European Conference on Computer Vision, Springer (2016) 740–756
57. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR. (2017)
58. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth prediction. (October 2019)
59. Yue Luo, Jimmy Ren, M.L.J.P.W.S.H.L.L.L.: Single view stereo matching. In: CVPR. (2018)
60. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R., eds.: *Advances in Neural Information Processing Systems*. Volume 29., Curran Associates, Inc. (2016)
61. Kazhdan, M., Hoppe, H.: Screened poisson surface reconstruction. *ACM Trans. Graph.* **32**(3) (jul 2013)
62. Baorui, M., Zhizhong, H., Yu-Shen, L., Matthias, Z.: Neural-pull: Learning signed distance functions from point clouds by learning to pull space onto surfaces. In: International Conference on Machine Learning (ICML). (2021)

- 63. Wang, C., Lucey, S., Perazzi, F., Wang, O.: Web stereo video supervision for depth prediction from dynamic scenes. In: 2019 International Conference on 3D Vision (3DV), IEEE (2019) 348–357
- 64. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence (2020)