Floating No More: Object-Ground Reconstruction from a Single Image

Supplementary Material

A. Implementation Details

Here we provide more details regarding the implementation and training of our model.

Backbone and Decoder. We use PVTv2-b3 [48] pre-trained on the COCO dataset [23] as our encoder backbone. And we use a decoder of a similar design to SegFormer [56], which consists of four multi-layer perceptron layers (MLP) to extract feature maps of different scales. We predict two dense fields with five channels: two for the front and back surface pixel height map, one for latitude field, and two for gravity field.

Data Normalization. For pixel height estimation, we normalize the ground truth maps by dividing them with the height of the image, which roughly turns the range of the pixel heights into $[0, 1]$ such that our model is not affected by objects at different scale. For two perspective fields, we normalize the latitude field into $[0, 1]$ and we represent the gravity (up-vector) field with a (sine, cosine) tuple as described in Section 3.2 in the main paper. The estimation of all three representations are formulated as regression problems and trained by MSE loss. Similar to existing methods [34, 38, 39, 61], due to the estimation of a normalized pixel height representation, our reconstructed models (Section 3.3) preserve the 3D geometry of the original objects but are scale-ambiguous. We calibrate the objects reconstructed by our methods and prior method using a linear scaling following LeReS [61].

Object Mask. All the datasets we use for training and quantitative evaluation come with object masks, which are from human annotation or off-the-shelf segmentation models. When evaluating web images, we utilize the Rembg segmentation model with u2net backbone [13] to obtain the foreground mask.

Data Generation. We use the physically-based rendering engine Blender [3] to render realistic RGB channel results. The front and back surface pixel height is calculated by our ray tracer. In detail, we shoot one ray to each pixel, find the first and last intersection points of the ray-object, and calculate their relevant 3D foot points (z=0). Then we project the intersection points and their footpoints onto the camera. The pixel heights are calculated by measuring the distances of the projected intersection points and their projected foot points in pixel units. Our pixel height calculation is efficient and can be computed in real time.

Training and Scheduling. The model is trained with the AdamW [27] optimizer with initial learning rate 0.0005 and a weight decay $1e$-2 for 60K steps with batch size 8 on a 4-A100 machine. We schedule the multi-step training stages at steps 30K, 40K, and 50K, with a learning rate decreasing $10\times$ each time. We resize the images to $(512, 512)$ resolution. We use horizontal flipping, random cropping, and color jittering augmentation during training. And because horizontal flipping, random cropping, and resizing will affect the values of our representations, we update the ground truth maps accordingly. The whole model is implemented using the PyTorch framework [31].

B. More Qualitative Analysis

Here we demonstrate more visualization examples of ORG. We show more diverse categories of objects with different camera viewpoints on random web images, and also full object geometry reconstruction results.

Diverse Categories. In Figure A, we show our direct estimation of pixel height and prospective fields, and also visualize the reprojected depth maps and reconstructed object-ground point clouds of diverse categories of objects from web images. The categories include common objects like microphone, plant, car, and tripod, as well as cartoon figures. The results show a great generalizability and robustness of our method in the wild.

Object-Ground Reconstruction. In addition to our previous analyses, we present a detailed visualization of the complete 3D geometry of the reconstructed objects and the ground in Figure B. Here, the objects are represented using 3D point clouds. Despite employing a simplified geometric model in our approach, our results effectively showcase superior reconstruction quality, particularly for objects with relatively straightforward geometric structures. This aspect of ORG highlights the balance between model simplicity and the ability to achieve high-fidelity reconstructions, even with less complex geometries.

C. Limitations and Future Work

Primarily, our approach relies on a simplified object shape assumption, optimizing for efficient image editing (*e.g.*, reflection, shadow generation, and ground-aware object pose change). However, this simplification may yield less than satisfactory 3D reconstruction results for objects with intricate geometries, particularly in estimating their back surfaces. Additionally, our method focuses solely on the geometric aspects of objects, excluding considerations of color and texture. We propose that leveraging our estimated geometry as a conditioned prior could significantly enhance

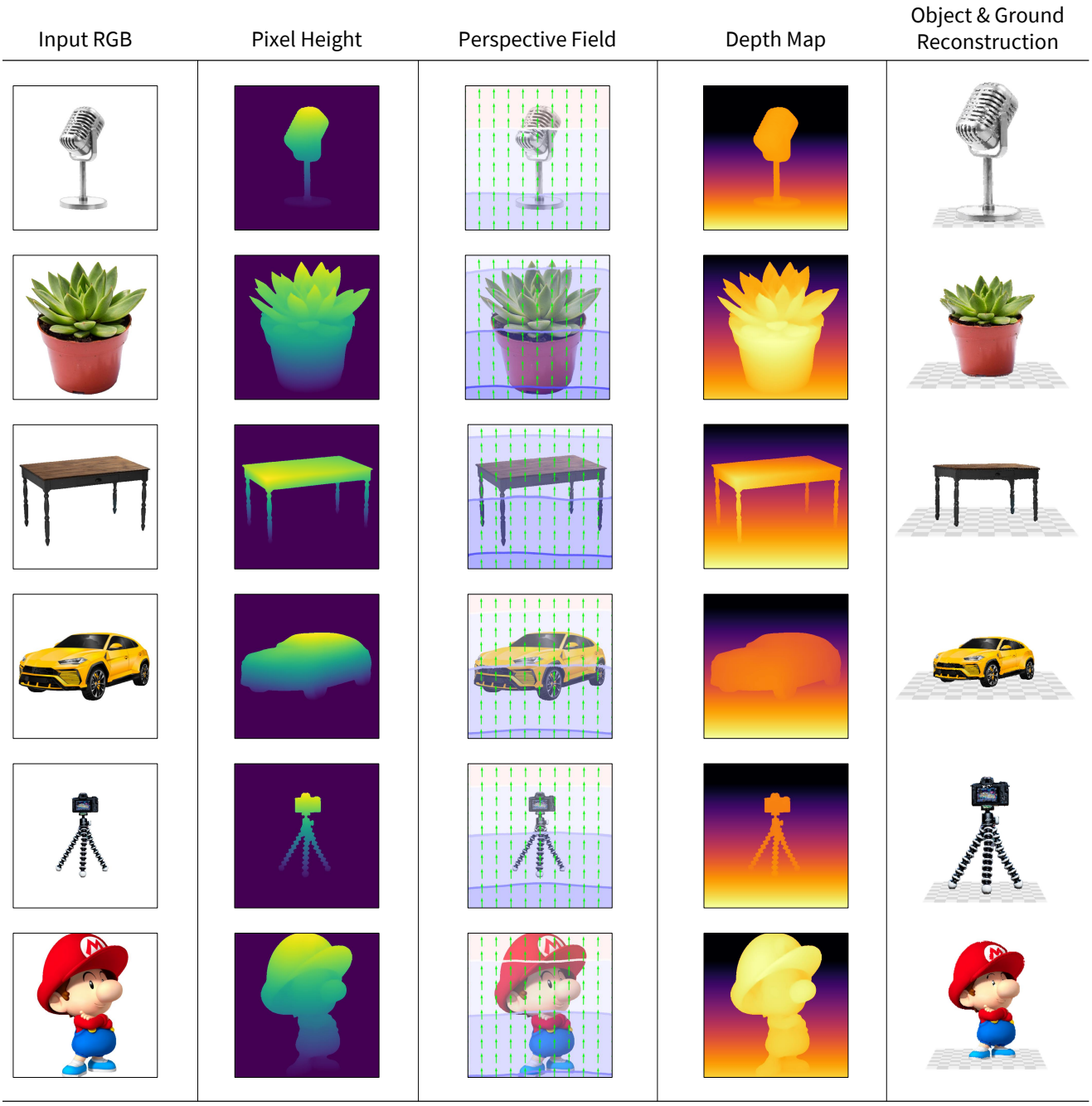

Figure A. Visualization of ORG on pixel height, (foreground) perspective fields, depth map, and object-ground reconstruction results. The results demonstrate that our work generalizes to various categories of objects.

image-inpainting processes, presenting a promising direction for future research.

References

[1] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *CVPR*, 2021. 5

[2] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 5, 6, 8

[3] Blender Online Community. Blender - a 3D modelling and rendering package, 1994. 5, 1

[4] Thomas J Cashman and Andrew W Fitzgibbon. What shape are dolphins? building 3D morphable models from 2D images. *TPAMI*, 2012. 3

[5] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3D: Disentangling geometry and appearance for high-quality text-to-3D content creation. In *ICCV*, 2023. 3

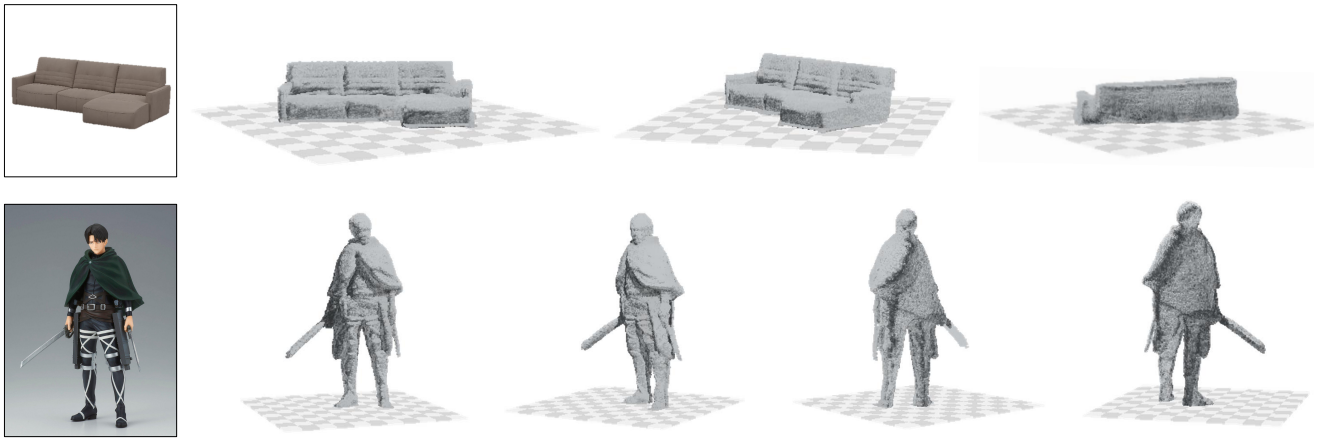

Figure B. Visualization of ORG on the full object geometry (front surface and back surface) with the ground plane.

[6] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *NeurIPS*, 2016. 1, 2

[7] Weifeng Chen, Shengyi Qian, and Jia Deng. Learning single-image depth from videos using quality assessment networks. In *CVPR*, 2019. 1, 2

[8] 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 *CVPR*, 2020. 2, 6

[9] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander Schwing, and Liangyan Gui. SDFusion: Multimodal 3D Shape Completion, Reconstruction, and Generation. In *CVPR*, 2023. 2

[10] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3D objects. In *CVPR*, 2023. 2, 5

[11] Jonathan Deutscher, Michael Isard, and John MacCormick. Automatic camera calibration from a single manhattan image. In *ECCV*, 2002. 3

[12] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014. 2

[13] Daniel Gatis. Rembg. `https://github.com/danielgatis/rembg`, 2023. 1

[14] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *ECCV*, 2020. 3

[15] 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 *CVPR*, 2018. 3

[16] Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F Fouhey. Perspective fields for single image camera calibration. In *CVPR*, 2023. 2, 3

[17] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 3

[18] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *CVPR*, 2015. 3

[19] Jinwoo Lee, Hyunsung Go, Hyunjoon Lee, Sunghyun Cho, Minhyuk Sung, and Junho Kim. Ctrl-C: Camera calibration transformer with line-classification. In *ICCV*, 2021. 3, 5, 6, 7, 8

[20] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3D reconstruction via semantic consistency. In *ECCV*, 2020. 3

[21] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 5, 8

[22] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3D: High-resolution text-to-3D content creation. In *CVPR*, 2023. 3

[23] 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 *ECCV*, 2014. 4, 1

[24] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *TPAMI*, 2015. 2

[25] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3D mesh in 45 seconds without per-shape optimization. In *NeurIPS*, 2023. 2, 3

[26] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In *ICCV*, 2023. 1, 2, 3, 5, 6

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2017. 4, 1

[28] Yunze Man, Xinshuo Weng, Xi Li, and Kris Kitani. GroundNet: Monocular ground plane normal estimation with geometric consistency. In *ACMMM*, 2019. 3

[29] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. RealFusion: 360 reconstruction of any object from a single image. In *CVPR*, 2023. 2, 3

[30] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks:

Learning 3D reconstruction in function space. In *CVPR*, 2019. 3

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 1

[32] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *ICLR*, 2022. 3

[33] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 2, 3

[34] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020. 1, 2, 4, 5, 6, 8

[35] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 5, 6, 8

[36] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common Objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction. In *ICCV*, 2021. 5

[37] Lawrence G Roberts. *Machine perception of three-dimensional solids*. PhD thesis, MIT, 1963. 3

[38] 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 *ICCV*, 2019. 2, 3, 1

[39] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *CVPR*, 2020. 1, 2, 3

[40] Shuwei Shao, Zhongcai Pei, Weihai Chen, Xingming Wu, and Zhengguo Li. NDDepth: Normal-distance assisted monocular depth estimation. In *ICCV*, 2023. 8

[41] Yichen Sheng, Yifan Liu, Jianming Zhang, Wei Yin, A Cengiz Oztireli, He Zhang, Zhe Lin, Eli Shechtman, and Bedrich Benes. Controllable shadow generation using pixel height maps. In *ECCV*, 2022. 2, 3

[42] Yichen Sheng, Jianming Zhang, Julien Philip, Yannick Hold-Geoffroy, Xin Sun, He Zhang, Lu Ling, and Bedrich Benes. PixHt-Lab: Pixel height based light effect generation for image compositing. In *CVPR*, 2023. 3

[43] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3D: High-fidelity 3D creation from a single image with diffusion prior. In *ICCV*, 2023. 2, 3

[44] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. In *3DV*, 2019. 2

[45] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2D diffusion models for 3D generation. In *CVPR*, 2023. 7

[46] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2Mesh: Generating 3D mesh models from single RGB images. In *ECCV*, 2018. 2, 3

[47] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *CVPR*, 2024. 8

[48] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021. 3, 1

[49] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *CVMJ*, 2022. 3, 4

[50] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. ProlificDreamer: High-fidelity and diverse text-to-3D generation with variational score distillation. In *NeurIPS*, 2023. 3

[51] Scott Workman, Connor Greenwell, Menghua Zhai, Ryan Baltenberger, and Nathan Jacobs. Deepfocal: A method for direct focal length estimation. In *ICIP*, 2015. 3

[52] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3D reconstruction. In *CVPR*, 2023. 3

[53] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T Freeman, and Joshua B Tenenbaum. Learning shape priors for single-view 3D completion and reconstruction. In *ECCV*, 2018. 2

[54] 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 *CVPR*, 2018. 2

[55] Wenqi Xian, Zhengqi Li, Matthew Fisher, Jonathan Eisenmann, Eli Shechtman, and Noah Snavely. UprightNet: Geometry-aware camera orientation estimation from single images. In *ICCV*, 2019. 3

[56] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 4, 1

[57] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. NeuralLift-360: Lifting an in-the-wild 2D photo to a 3D object with 360° views. In *CVPR*, 2023. 1, 2, 3

[58] Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *CVPR*, 2021. 3

[59] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3D reconstruction of generic objects in hands. In *CVPR*, 2022. 3

[60] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *ICCV*, 2019. 2

[61] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3D scene shape from a single image. In *CVPR*, 2021. 1, 2, 4, 5, 6, 7, 8

[62] Zhengyou Zhang. A flexible new technique for camera calibration. *TPAMI*, 2000. 3