Supplementary Materials: Learning to Recover 3D Scene Shape from a Single Image

1. Datasets

1.1. Datasets for Training

To train a robust model, we use a variety of data sources, each with its own unique properties:

- Taskonomy [21] contains high-quality RGBD data captured by a LiDAR scanner. We sampled around 114K RGBD pairs for training.
- DIML [10] contains calibrated stereo images. We use the GA-Net [22] method to compute the disparity for supervision. We sampled around 121K RGBD pairs for training.
- 3D Ken Burns [13] contains synthetic data with ground truth depth. We sampled around 51K RGBD pairs for training.
- Holopix50K [8] contains diverse uncalibrated web stereo images. Following [17], we use FlowNet [9] to compute the relative depth (inverse depth) data for training.
- HRWSI [18] contains diverse uncalibrated web stereo images. We use the entire dataset, consisting of 20K RGBD images.

1.2. Datasets Used in Testing

To evaluate the generalizability of our method, we test our depth model on a range of datasets:

- NYU [15] consists of mostly indoor RGBD images where the depth is captured by a Kinect sensor. We test our method on the official test set, which contains 654 images.
- KITTI [7] consists of street scenes, with sparse metric depth captured by a LiDAR sensor. We use the standard test set (652 images) of the Eigen split.
- ScanNet [6] contains similar data to NYU, indoor scenes captured by a Kinect. We randomly sampled 700 images from the official validation set for testing.
- DIODE [16] contains high-quality LiDAR-generated depth maps of both indoor and outdoor scenes. We use the whole validation set (771 images) for testing.
- ETH3D [14] consists of outdoor scenes whose depth is captured by a LiDAR sensor. We sampled 431 images from it for testing.
- Sintel [1] is a synthetic dataset, mostly with outdoor scenes. We collected 641 images from it for testing.
- OASIS [5] is a diverse dataset consisting of images in the wild, with ground truth depth annotations by humans. It contains both sparse relative depth labels (similar to DIW [3]), and some planar regions. We test on the entire validation set, containing 10K images.
- YouTube3D [4] consists of in-the-wild videos that are reconstructed using structure from motion, with the sparse reconstructed points as supervision. We randomly sampled 58K images from the whole dataset for testing.
- RedWeb [17] consists of in-the-wild stereo images, with disparity labels derived from an optical flow matching algorithm. We use 3.6K images to evaluate the WHDR error, and we randomly sampled 5K points pairs on each image.
- iBims-1 [11] is an indoor-scene dataset, which consists of 100 high-quality images captured by a LiDAR sensor. We use the whole dataset for evaluating edge and plane quality.

We will release a list of all images used for testing to facilitate reproducibility.

2. Details for Depth Prediction Model and Training.

We use the depth prediction model proposed by Xian *et al*. [18]. We follow [20] and combine the multi-source training data by evenly sampling from all sources per batch. As HRWSI and Holopix50K are both web stereo data,

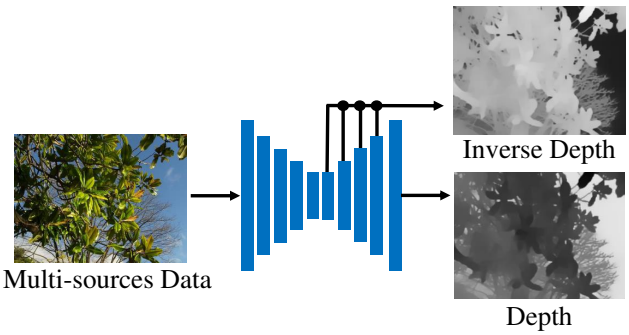


Figure 1: The network architecture for the DPM. The network has two output branches. The decoder outputs the depth map, while the auxiliary path outputs the inverse depth. Different losses are enforced on these two branches.

we merge them together. Therefore, there are four different data sources, i.e. high-quality Taskonomy, synthetic 3D Ken Burn, middle-quality DIML, and low-quality Holopix50K and HRWSI. For example, if the batch size is 8, we sample 2 images from each of the four sources. Furthermore, as the ground truth depth quality varies between data sources, we enforce different losses for them.

For the web-stereo data, such as Holopix50K [8] and HRWSI [18], as their inverse depths have unknown scale and shift, these inverse depths cannot be used to compute the affine-invariant depth (up to an unknown scale and shift to the metric depth). The pixel-wise regression loss and geometry loss cannot be applied for such data. Therefore, during training, we only enforce the ranking loss [17] on them.

For the middle-quality calibrated stereo data, such as DIML [10], we enforce the proposed image-level normalized regression loss, multi-scale gradient loss and ranking loss. As the recovered disparities contain much noise in local regions, enforcing the pair-wise normal regression loss on noisy edges will cause many artifacts. Therefore, we enforce the pair-wise normal regression loss only on planar regions for this data.

For the high-quality data, such as Taskonomy [21] and synthetic 3D Ken Burns [13], accurate edges and planes can be located. Therefore, we apply the pair-wise normal regression loss, ranking loss, and multi-scale gradient loss for this data.

Furthermore, we follow [12] and add a light-weight auxiliary path on the decoder. The auxiliary outputs the inverse depth and the main branch (decoder) outputs the depth. For the auxiliary path, we enforce the ranking loss, image-level normalized regression loss in the inverse depth space on all data. The network is illustrated in Fig. 1.

3. Sampling Strategy for Pairwise Normal Loss

We enforce the pairwise normal regression loss on Taskonomy and DIML data. As DIML is more noisy than Taskonomy, we only enforce the normal regression loss on the planar regions, such as pavements and roads, whereas for Taskonomy, we sample points on edges and on planar regions. We use the local least squares fitting method [19] to compute the surface normal from the depth map.

For edges, we follow the method of Xian *et al*. [18], which we describe here. The first step is to locate image edges. At each edge point, we then sample pairs of points on both sides of the edge, i.e. $\mathscr{P} = \{(P_A, P_B)_i | i = 0, ..., n\}$. The ground truth normals for these points are $\mathscr{N}^* = \{(\boldsymbol{n}_A^*, \boldsymbol{n}_B^*)_i | i = 0, ..., n\}$, while the predicted normals are $\mathscr{N} = \{(\boldsymbol{n}_A, \boldsymbol{n}_B)_i | i = 0, ..., n\}$. To locate the object boundaries and planes folders, where the normals changes significantly, we set the angle difference of two normals greater than $arccos(0.3)$. To balance the samples, we also get some negative samples, where the angle difference is smaller than $arccos(0.95)$ and they are also detected as edges. The sampling method is illustrated as follow.

$$\mathscr{S}_1 = \{\boldsymbol{n}_A^* \cdot \boldsymbol{n}_B^* > 0.95, \boldsymbol{n}_A^* \cdot \boldsymbol{n}_B^* < 0.3 | (\boldsymbol{n}_A^*, \boldsymbol{n}_B^*)_i \in \mathscr{N}^*\} \tag{1}$$

For planes, on DIML, we use [2] to segment the roads, which we assume to be planar regions. On Taskonmy, we locate planes by finding regions with the same normal. On each detected plane, we sample 5000 paired points on average. Finally, we combine both sets of paired points and enforce the normal regression loss on them, see E.q. 4 in our main paper.

4. Illustration of the Reconstructed Point Cloud

We illustrate some examples of the reconstructed 3D point cloud from our proposed approach in Fig. 2. All these data are unseen during training. This shows that our method demonstrates good generalizability on in-the-wild scenes and can recover realistic shape of a wide range of scenes.

5. Illustration of Depth Prediction In the Wild

We illustrate examples of our single image depth prediction results in Fig. 3. The images are randomly sampled from DIW and OASIS, which are unseen during training. On these diverse scenes, our method predicts reasonably accurate depth maps, in terms of global structure and local details.

References

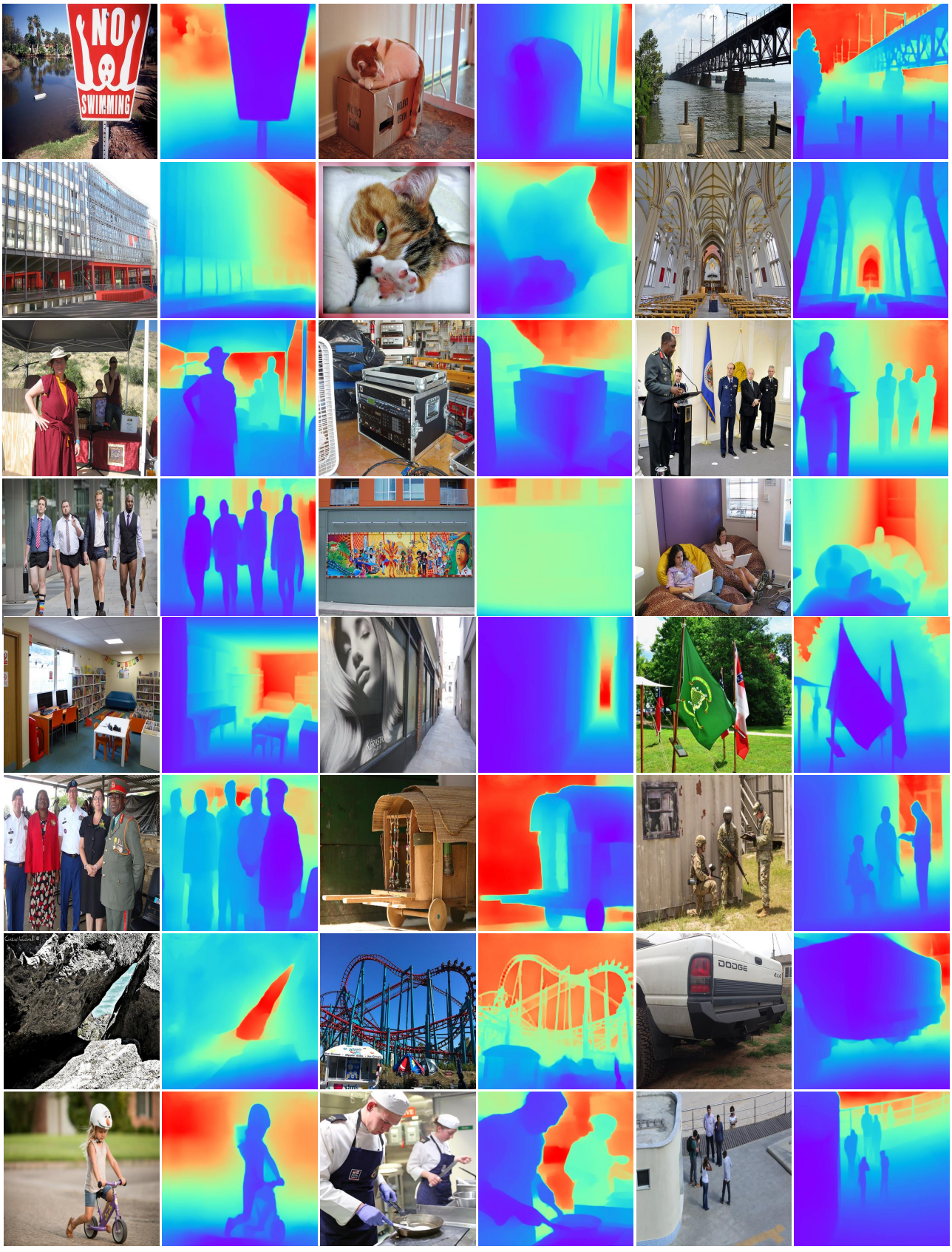
[1] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In





Figure 2: Point Cloud Illustration. The first column shows the input images. The remaining columns show the point cloud recovered from our proposed approach from the left, right, and top respectively.


NO
SWIMMING


NO
SWIMMING
DODGE


Estrutural
KeyBank

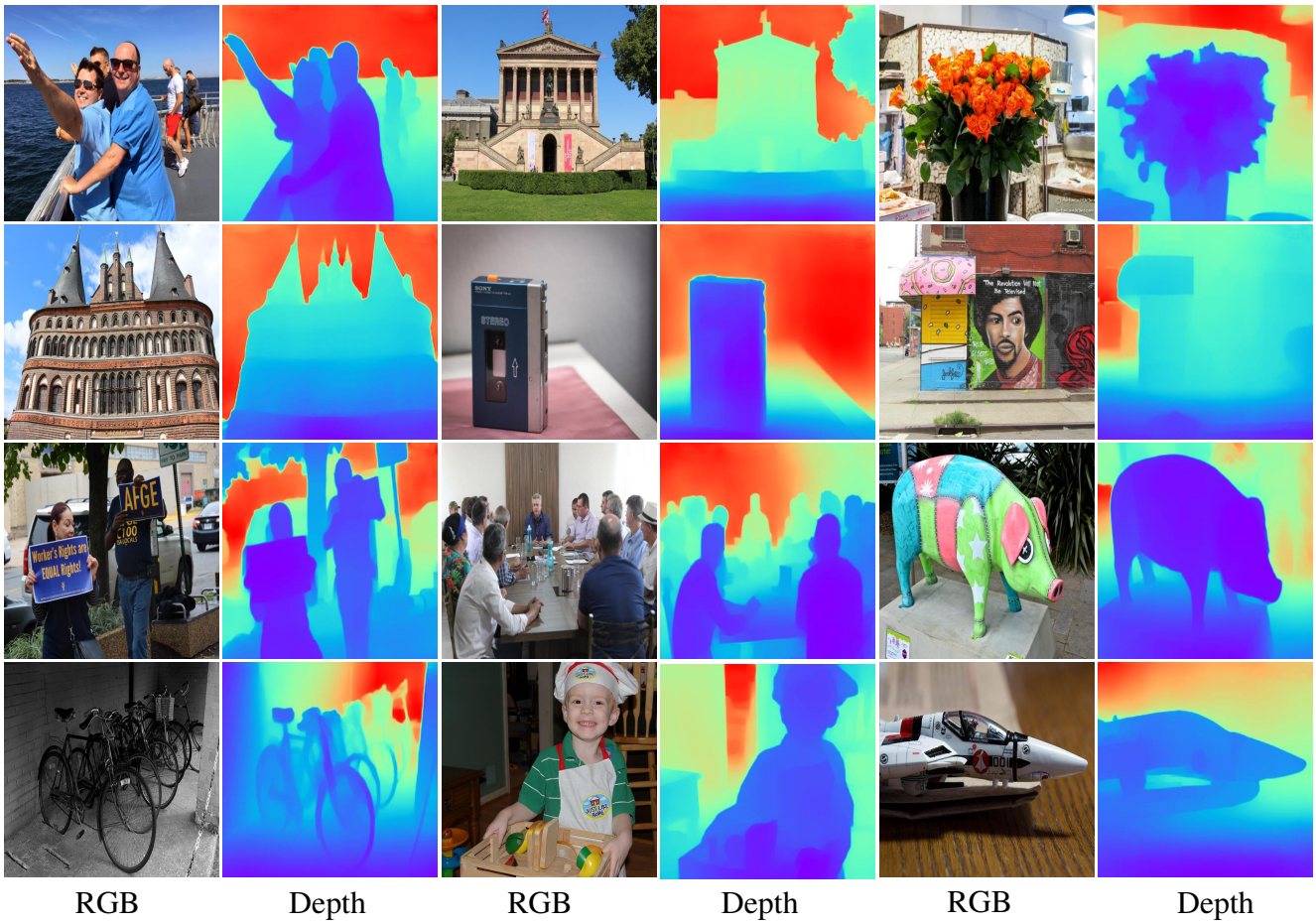


Figure 3: Examples of depths on in-the-wild scenes. Purple indicates closer regions whereas red indicates farther regions.

Proc. Eur. Conf. Comp. Vis., pages 611–625. Springer, 2012. 1

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. Eur. Conf. Comp. Vis.*, 2018. 2

[3] 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

[4] 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. 1

[5] 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

[6] 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. 1

[7] 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. 1

[8] 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. 1, 2

[9] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017. 1

[10] 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. 1, 2

[11] 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. 1

[12] Yifan Liu, Bohan Zhuang, Chunhua Shen, Hao Chen, and Wei Yin. Training compact neural networks via auxiliary overparameterization. *arXiv: Comp. Res. Repository*, page 1909.02214, 2019. 2

[13] 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. 1, 2

[14] 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. 1

[15] 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. 1

[16] 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. 1

[17] 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

[18] 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

[19] 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. 2

[20] 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

[21] 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. 1, 2

[22] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 185–194, 2019. 1