DnD: Dense Depth Estimation in Crowded Indoor Dynamic Scenes Supplementary Material

Dongki Jung [*1] Jaehoon Choi [*1,2] Yonghan Lee[1] Deokhwa Kim[1] Changick Kim[3]
Dinesh Manocha[2] Donghwan Lee[1]
[1]NAVER LABS [2]University of Maryland [3]KAIST

A. Datasets

We provide a brief overview of the NAVERLABS Indoor Localization dataset [4, 1] and the specific properties that make it desirable as a dense depth estimation in crowded indoor dynamic scenes. The NAVERLABS dataset consists of various video sequences captured on five different floors in a department store and a metro station. Among these places, we selected two places, the metro station B1 (MS) and the department store B1 (Dept), to build a benchmark for depth estimation in crowded indoor dynamic scenes. The MS dataset is one of the most crowded dataset in the NAVERLABS dataset. We also chose the Dept dataset because it has totally different environments. In the NAVERLABS dataset, a mapping robot utilizes two 16-channel LiDAR sensors, six RGB cameras and four smartphone cameras. We only utilized images collected from six RGB cameras. LiDAR sensors are only used for generating groundtruth depth maps for evaluation. As the NAVERLABS dataset consists of the split of 60735 images for training and 835 images for testing in MS, and 25083 images for training and 443 images for testing in Dept, we adopt this dataset for dense depth estimation in crowded indoor dynamic scenes.

Dataset	Environment	Location	**Dynamic**	Video	**# Images**	**Metric**	Annotation
(a) Indoor datasets							
NYUv2 [29]	Indoor	Small Rooms		✓	407K	✓	RGB-D
ScanNet [6]	Indoor	Small Rooms		✓	2.5M	✓	RGB-D
Stanford [2]	Indoor	Small Rooms		✓	72K	✓	Laser
Matterport3D [3]	Indoor	Small Rooms		✓	194K	✓	Laser
7 scenes [28]	Indoor	Small Rooms		✓	26K	✓	Laser
InLoc [31]	Indoor	Univ. bldg.		✓	10K	✓	Laser
Baidu [30]	Indoor	Mall	✓	✓	682	✓	Laser
(b) Internet Image collections							
DIW [5]	Indoor & Outdoor	Diverse	✓		495K		Ordinal
MegaDepth [17]	Outdoor	Diverse			130K		SfM
ReDWeb [33]	Indoor & Outdoor	Diverse	✓	✓	3600		Stereo
WSVD [32]	Indoor & Outdoor	Diverse	✓	✓	150K		Stereo
MC [16]	Indoor & Outdoor	Diverse	✓	✓	136K		
(c) **NAVERLABS**							
Dept [4, 1]	Indoor	Mall	✓	✓	25K	✓	Laser
MS [4, 1]	Indoor	Mall & Turnstiles	✓	✓	60K	✓	Laser

Table 1: (a) Most large-scale indoor datasets do not have dynamic scenes except for Baidu [30]. (b) All datasets collected from the internet do not contain metric 3D models. To train depth estimation algorithms in crowded indoor dynamic scenes, a dataset must contain three properties: dynamic, metric and large amount of images. Different from previous datasets (a) and (b), NAVERLABS contains large amounts of dynamic scenes with the metric 3D model in indoor environments.

*These two authors contributed equally.

Dense depth estimation in crowded indoor dynamic scenes is a significant task for Robotics and AR applications. To the best of our knowledge, however, it is difficult to find large-scale public datasets for dense depth estimation in crowded indoor dynamic scenes. In Table 1 (a), most indoor datasets [29, 6, 2, 3, 28, 31, 30] are captured from small collections of room, office, and university buildings. They cover only a restricted scale of spaces and have similar design and internal structures. Additionally, existing indoor datasets do not provide scenes containing multiple dynamic objects such as moving people except for the Baidu dataset [30]. The Baidu dataset [30] collected from a shopping mall has dynamic scenes with 3.75% crowd density. However, Baidu contains only 689 images for training. In Table 1 (b), other works [5, 17, 33, 32, 16] explore the use of internet photo collections to train the depth estimation models for dynamic scenes. These datasets contain a large number of dynamic scenes, but they do not provide metric depth. They often have relative depth obtained by either stereo matching or COLMAP [26, 27], or ordinal depth relation from manual annotations. Thus, we are not able to evaluate our algorithms using metric depth maps for internet images.

B. Evaluation Metrics

The depth value is represented in the metric scale (m). Following [7], we evaluate our method using the following metrics: absolute relative difference (Abs Rel), square relative difference (Sq Rel), root mean squared error (RMSE), RMSE in logarithmic scale (RMSE log), and δ_i meaning the percentage of predicted pixels for which the relative error is less than a threshold i.

C. Implementation Details

We implement our method in PyTorch [20] and conduct all experiments on a V100 GPU. We train our network with a batch size of 8 images with size 1024×768 for 20 epochs. We use the following set of weights for each loss term in the loss function: $\lambda_d = 0.001$, $\lambda_{ph} = 1$, $\lambda_s = 0.3$, $\lambda_f = 0.1$, and $\lambda_n = 0.001$. We utilize the Adam [13] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We use an initial learning rate of 10^{-4} for the first 10 epochs and halve it for the remaining 10 epochs. We use human masks to eliminate the points in human regions to reduce the noise on projected depth maps.

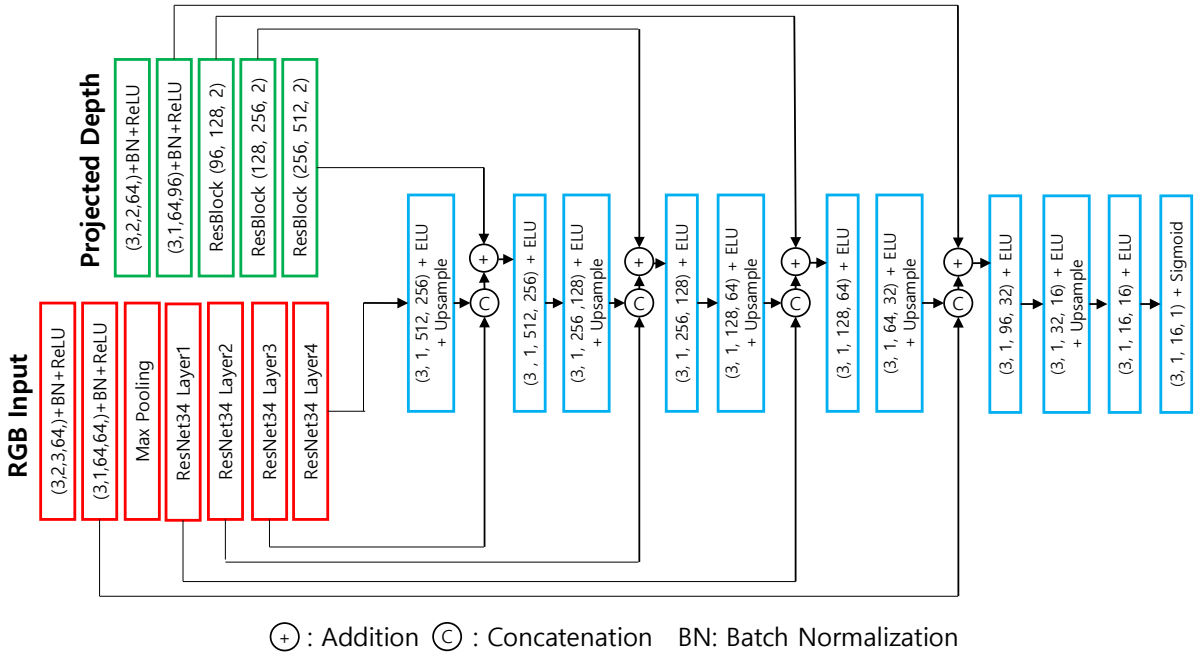

Figure 1: The network architecture of our proposed method. The red block, green block, and blue block are an RGB encoder, a depth encoder, and a decoder. We concatenate a projected depth map and a binary mask, which indicates where the projected depth values exist. The convolutional filter is defined as (filter size, stride size, input channel, output channel). The RGB encoder adopts a ResNet 34 [11] backbone pretrained with ImageNet [24], and the depth encoder consists of two convolutional blocks and three residual blocks.

D. DnD with Localization

DnD network requires a monocular image and the corresponding reprojected depth map derived from the 3D model with the camera position. In order to verify the applicability, it is necessary to demonstrate both pose estimation and depth map projection about the new test images. Thus, we implement the whole pipeline of DnD, including mapping and localization process using Kapture [19], which is the visual localization toolbox. As a mapping procedure, the sparse 3D model is reconstructed via Structure-from-Motion (SfM) [26] in the set of training images. When a test image is given, image retrieval (AP-GeM [22]) is performed to obtain top k ranked images in the database of the 3D model, which can consider covisibility with the images and estimate a more accurate camera pose. The local feature extractor (R2D2 [23]) produces 2D-2D matches between the test image and top-ranked database images, resulting in 2D-3D matches between the test image and the 3D model. Perspective-n-Point (PnP [15]) problem with Random Sample Consensus (RANSAC [9]) is solved in these matching points and consequently yields the predicted camera pose. With the visible 3D points from the test image and the estimated pose, we can obtain the reprojected depth map and activate our DnD framework. We assign 5 for the top k, and the results of the MS dataset are reported in Fig. 2 and Table 2.

Method	3D Model	F+B / F (Lower is better)				F+B / F (Higher is better)		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
DnD	SfM + MVS	0.189 / 0.240	0.20 / 0.16	1.76 / 2.44	0.084 / 0.133	0.806 / 0.677	0.881 / 0.798	0.919 / 0.856
DnD	SfM	0.194 / 0.249	0.17 / 0.15	1.85 / 2.56	0.092 / 0.141	0.774 / 0.638	0.874 / 0.785	0.918 / 0.850

Table 2: The performance with or without Kapture [19] in the MS dataset. The upper line in the table shows the results of the camera poses, which are used in reconstruction of the 3D model (SfM + MVS). The lower line in the table indicates the performance of the whole pipeline including visual localization under the sparse 3D model from SfM (without MVS). Although the SfM point clouds are highly sparse and visual localization methods has pose uncertainty, the comparable results show the robustness of our framework.

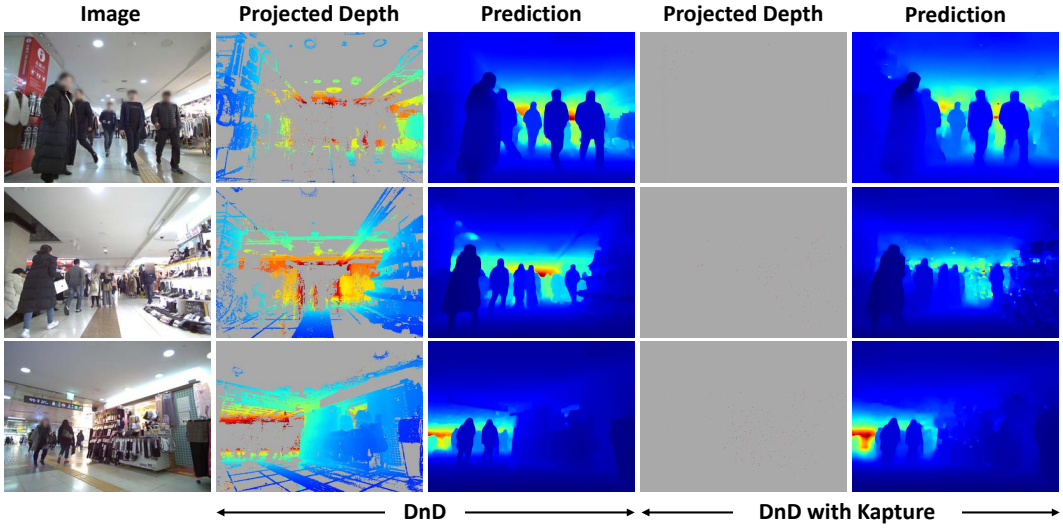

Figure 2: DnD (2-3 columns) estimates depth from the image and the projected depth map, which is derived from the 3D model. Since DnD with Kapture [19] (4-5 columns) exploit the sparse 3D model from SfM, the projected depth maps are also highly sparse, and the background details of the estimated depth maps decrease. However, the overall qualities, especially moving people, are predicted robustly. (low high; grey means empty depth values.)

E. Robustness to Orientation Errors

We assume a scenario that the camera pose is inaccurate and projected depth maps are not aligned well with the current image. This scenario is common in real-world applications because the visual localization system often computes uncertain poses, especially in dynamic scenes where the moving people cover a large part of the image. Therefore, we validate that our method is robust to orientation errors caused by visual localization algorithms. We add noises α to ground truth poses and then project depth maps from the incomplete 3D model by using noisy poses. α is set to 2°, 5°, and 10°. Following the [25],

we report the quantitative evaluation with different noises for pose values. We perform this experiment on the MS dataset because it is highly crowded than the Dept dataset.

Method	Orientation Error	F+B / F (Lower is better)				F+B / F (Higher is better)		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
DnD	-	0.189 / 0.240	0.20 / 0.16	1.76 / 2.44	0.084 / 0.133	0.806 / 0.677	0.881 / 0.798	0.919 / 0.856
DnD	2°	0.205 / 0.246	0.24 / 0.17	1.82 / 2.46	0.089 / 0.135	0.786 / 0.666	0.876 / 0.796	0.917 / 0.855
	5°	0.207 / 0.246	0.23 / 0.17	1.83 / 2.46	0.090 / 0.135	0.779 / 0.666	0.875 / 0.796	0.917 / 0.855
	10°	0.211 / 0.245	0.24 / 0.17	1.84 / 2.46	0.091 / 0.135	0.773 / 0.666	0.873 / 0.795	0.916 / 0.855

Table 3: Experimental evaluations of our proposed method's robustness to orientation errors.

F. Qualitative Results of Depth Prediction Corresponding Sparse Depth Points

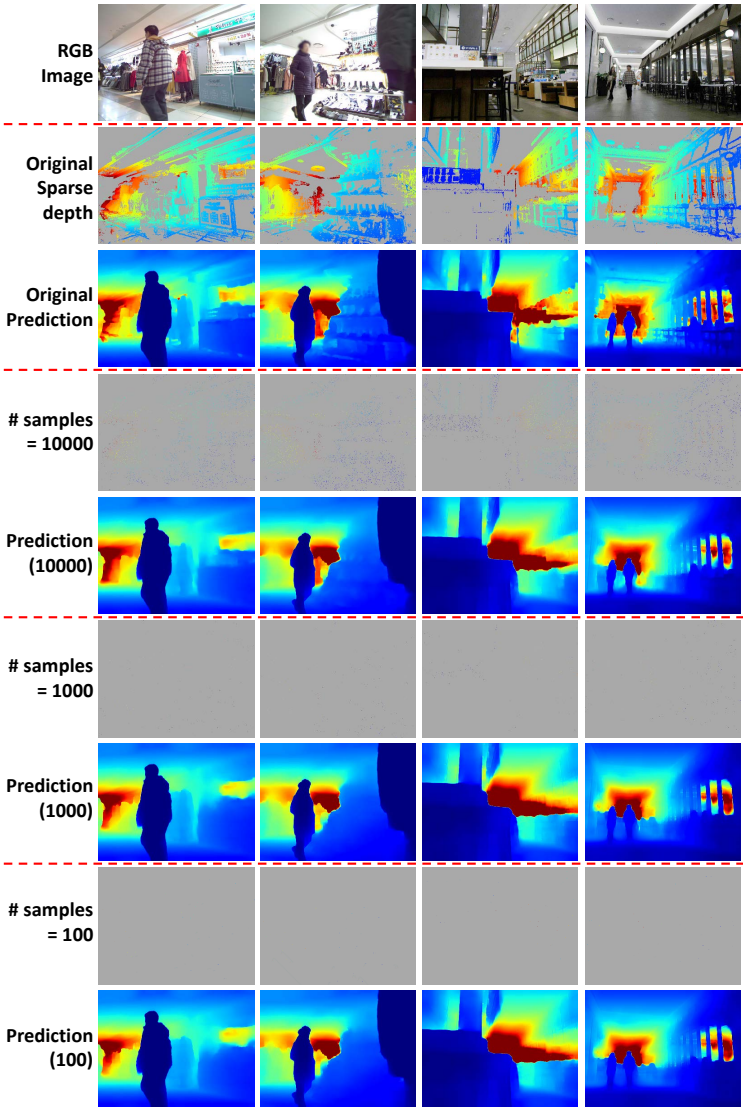

Figure 3: Qualitative results of depth prediction corresponding to the number of uniform-sampled input depth points (refer the Fig. 6 (a) in the manuscript). The first and the second columns show the results of the MS dataset, and the third and the fourth columns indicate the Dept dataset. As the number of sampled points decreases, some details of the static background disappear. However, the overall quality of depth estimation results is maintained robustly. All color maps use the jet color map (low high; grey means empty depth values.)

G. Ablation Study

Method	F+B / F (Lower is better)				F+B / F (Higher is better)		
	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
DnD (L_{ph} only)	0.289 / 0.388	0.47 / 0.60	2.60 / 3.24	0.109 / 0.148	0.663 / 0.564	0.837 / 0.759	0.911 / 0.847
DnD w/o FSC, NSC	0.240 / 0.355	0.42 / 0.71	2.51 / 3.63	0.091 / 0.133	0.741 / 0.642	0.877 / 0.806	0.929 / 0.868
DnD w/o NSC	0.226 / 0.335	0.37 / 0.72	2.39 / 3.19	0.087 / 0.130	0.753 / 0.635	0.885 / 0.810	0.930 / 0.878
DnD w/o FSC	0.230 / 0.272	**0.31** / 0.31	2.44 / 3.15	0.092 / 0.127	0.730 / 0.644	0.871 / 0.814	0.927 / 0.880
DnD (full)	**0.213 / 0.250**	0.32 / **0.30**	**2.36 / 3.04**	**0.084 / 0.116**	**0.761 / 0.707**	**0.889 / 0.836**	**0.932 / 0.886**

Table 4: Contributions of our proposed modules to the evaluation results on the Dept dataset. F means the evaluation results on depth values in the human regions, and F+B indicates the evaluation results on depth values over the entire scene. The median scaling is applied to DnD (L_{ph} only) for absolute scale depth prediction.

H. Qualitative Results of Depth Completion Algorithm

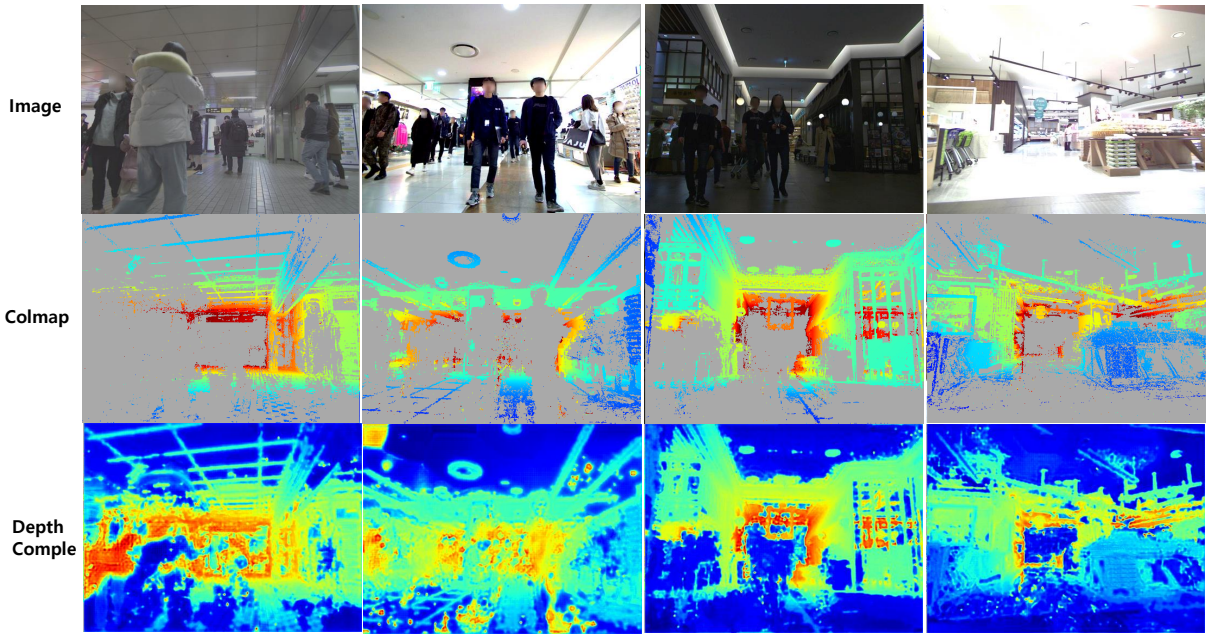

Figure 4: As we metioned in the manuscript, we show the qualitative results of DepthComple [18]. We adopt an early-fusion encoder-decoder network from [18] combined with normalized convolution layers [8]. We use projected depth maps as ground truth for supervised training. The depth completion network fails to fill the empty regions in COLMAP results. Also, it fails to produce estimated depth in human regions with accurate depth values. Since most depth completion models show poor performance, we did not add these experimental results in the manuscript. All color maps use the jet color map (low high; grey means empty depth values.)

I. Visualization of Intermediate Results for Our Two Novel Consraints

Figure 5: Examples of the optical flow in the MS dataset. From top to bottom, each row shows the optical flows estimated by FlowNet2.0 [12], the temporally adjacent images $I_{t'}$, the warped images I'_t, and the current images I_t. Both I_t and $I_{t'}$ are used as the input for our proposed training method, and the warped images I'_t are projected from the image coordinates of $I_{t'}$ to I_t for flow-guided shape constraint (for more details, see Section 3.3 in the manuscript).

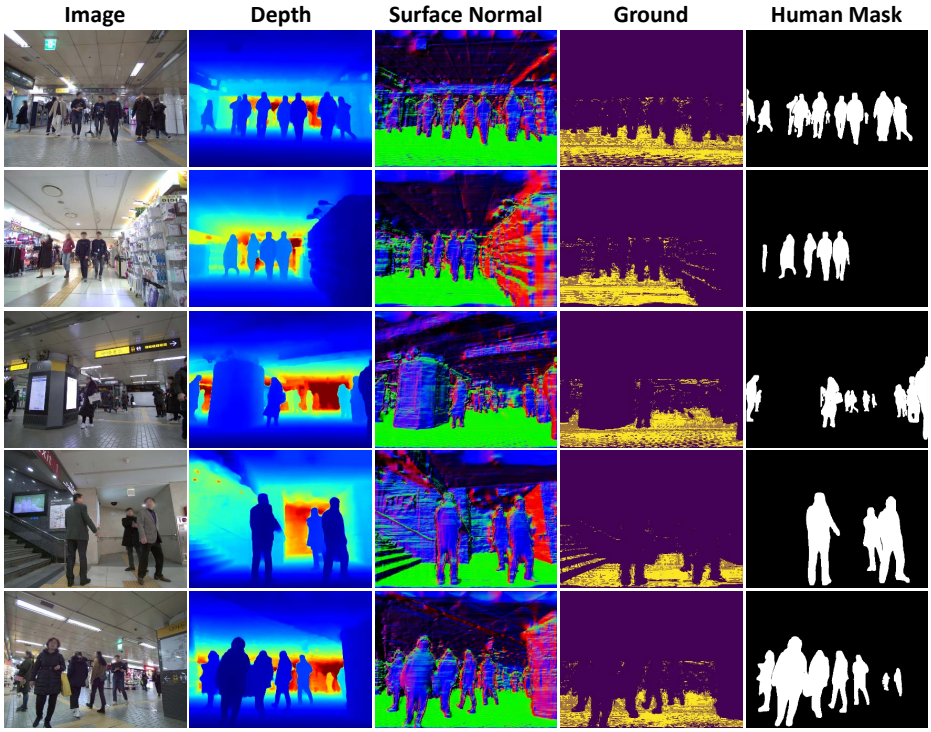

Figure 6: The intermediate results for the normal-guided scale constraints (for more details, see Section 3.4 in the manuscript). We can obtain the normal directions from the estimated depth maps and find the ground regions. Human masks are estimated by Mask R-CNN [10]. In training, the sampled pixels in each of the people are constrained by the estimated depth values at the human's ground contact point. All color maps use the jet color map (low high; grey means empty depth values.)

J. Results on NYUv2 and TUM RGB-D

Figure 7 and 8 show the qualitative results of NYUv2 and TUM RGB-D, respectively. We acquire the sparse depth map by sparsifying the groundtruth depth map in order to obtain the metric scale. Assuming the similar situation that the reprojected depth maps are derived from the 3D model, we only sample the depth values in the position of SIFT features for the corresponding images. In TUM, we used the ground truth camera pose provided by the dataset. In NYUv2, we used R2D2 [23] to obtain correspondences by using FLANN-based search algorithm and estimate the relative pose via the Perspective-n-Point (PnP [15]).

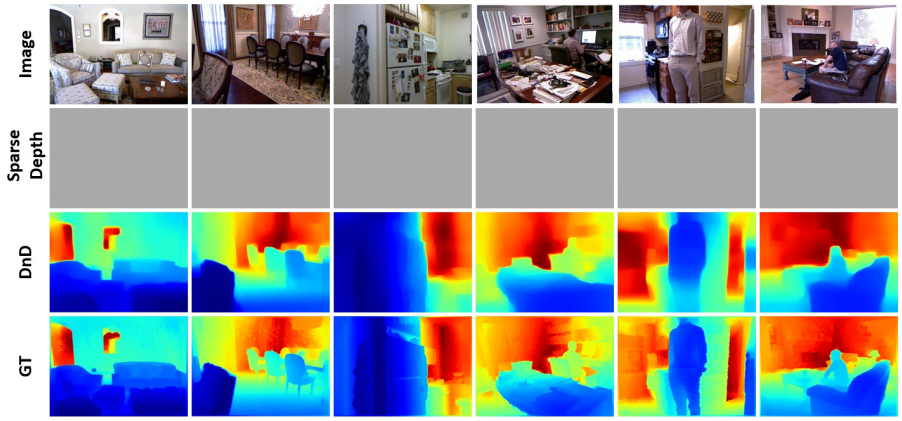

Figure 7: The qualitative results of the NYUv2 dataset.

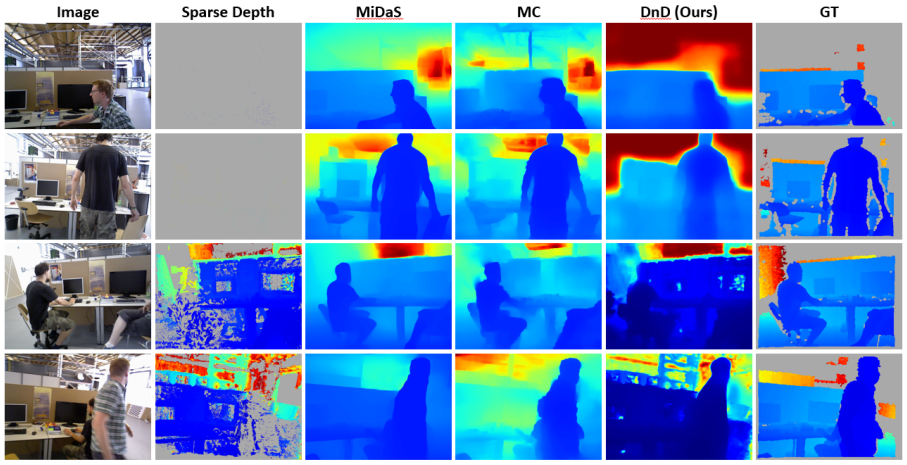

Figure 8: The qualitative results of the TUM RGB-D dataset. The 1st and 2nd raw images result from the sparse depth maps, which are the sampled values of the Kinect with SIFT features. The 3rd and 4th raw images indicate the results for the projection of the scale ambiguous 3D model. Although there is no significant improvement in the visual quality compared to MiDaS [21] and Mannequin (MC) [16], DnD is able to estimate more accurate depth values in the numerical results since our method is based on the metric depth.

Since the dynamic objects category of the TUM RGB-D dataset is a small-scale dataset, this dataset is not suitable for self-supervised training method like DnD. As shown in Fig. 8, DnD does not show improved performance on qualitative results compared to MiDaS and Mannequin (MC). Both MiDaS and MC are based on supervised learning which require dense ground truth depth maps across different environments. These works trained their model on diverse and large-scale datasets. Specifically, MC used 136K image-depth pairs from Mannequin challenge videos, and MiDaS was trained on 10 different large-scale datasets [1]. However, DnD was trained on either MS or Dept datasets (25K or 60K images) without ground truth depth maps. Therefore, DnD does not show better generalization ability to produce sharp depth maps. Instead, since we focus on building a novel method to estimate metric depth maps, our method can show consistently better performance on quantitative results on TUM RGB-D dataset.

[1] https://github.com/intel-isl/MiDaS

K. Additional Qualitative Comparisons

In this section, we provide more additional qualitative comparisons (See Figure 4. in the manuscript).

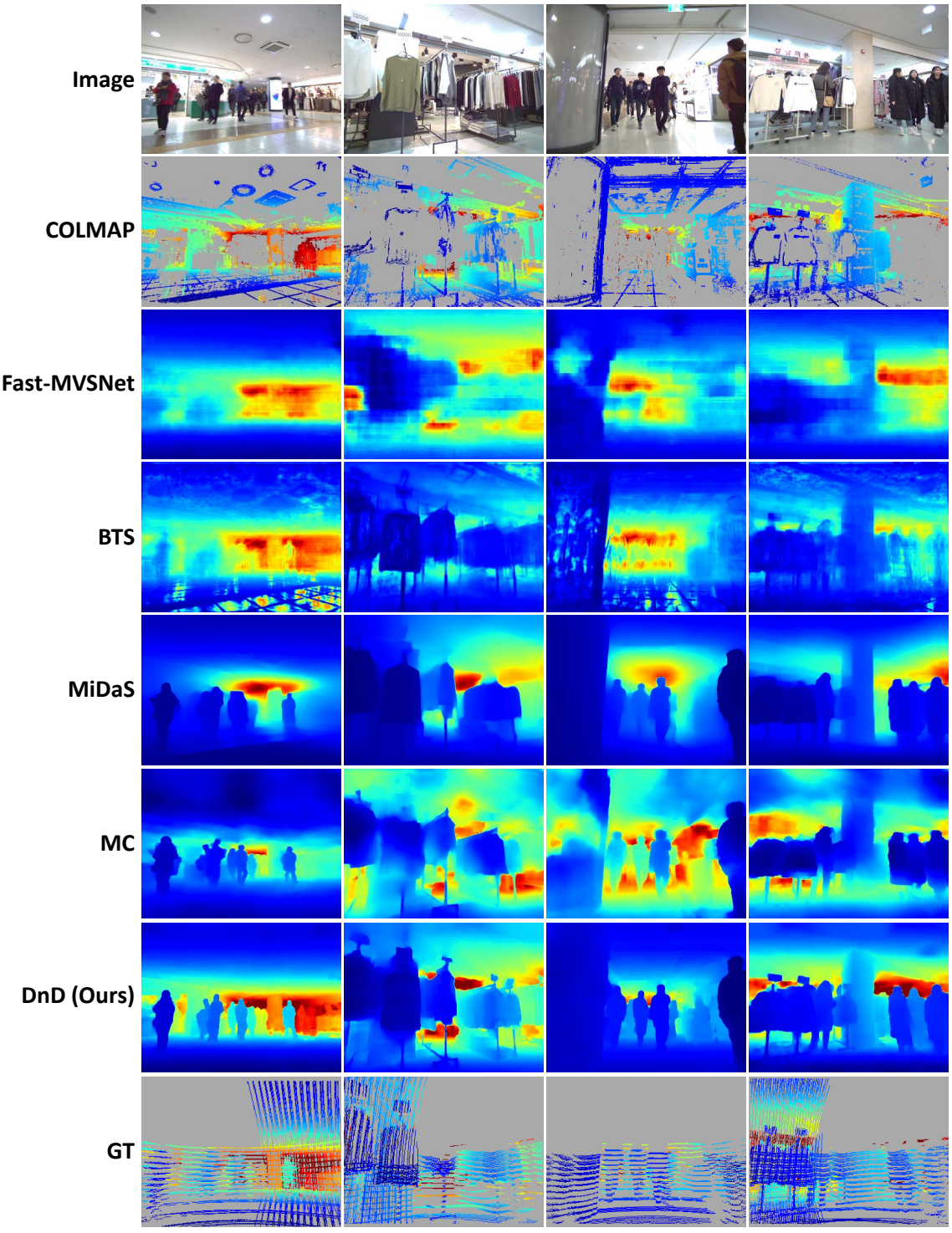

Figure 9: Qualitative comparisons with the state-of-the-art depth estimation algorithms in the MS dataset. From the third row to bottom, the results of Fast-MVSNet [34], BTS [14], MiDaS [21], Mannequin (MC) [16], and DnD are shown. All color maps use the jet color map (low high; grey means empty depth values.)

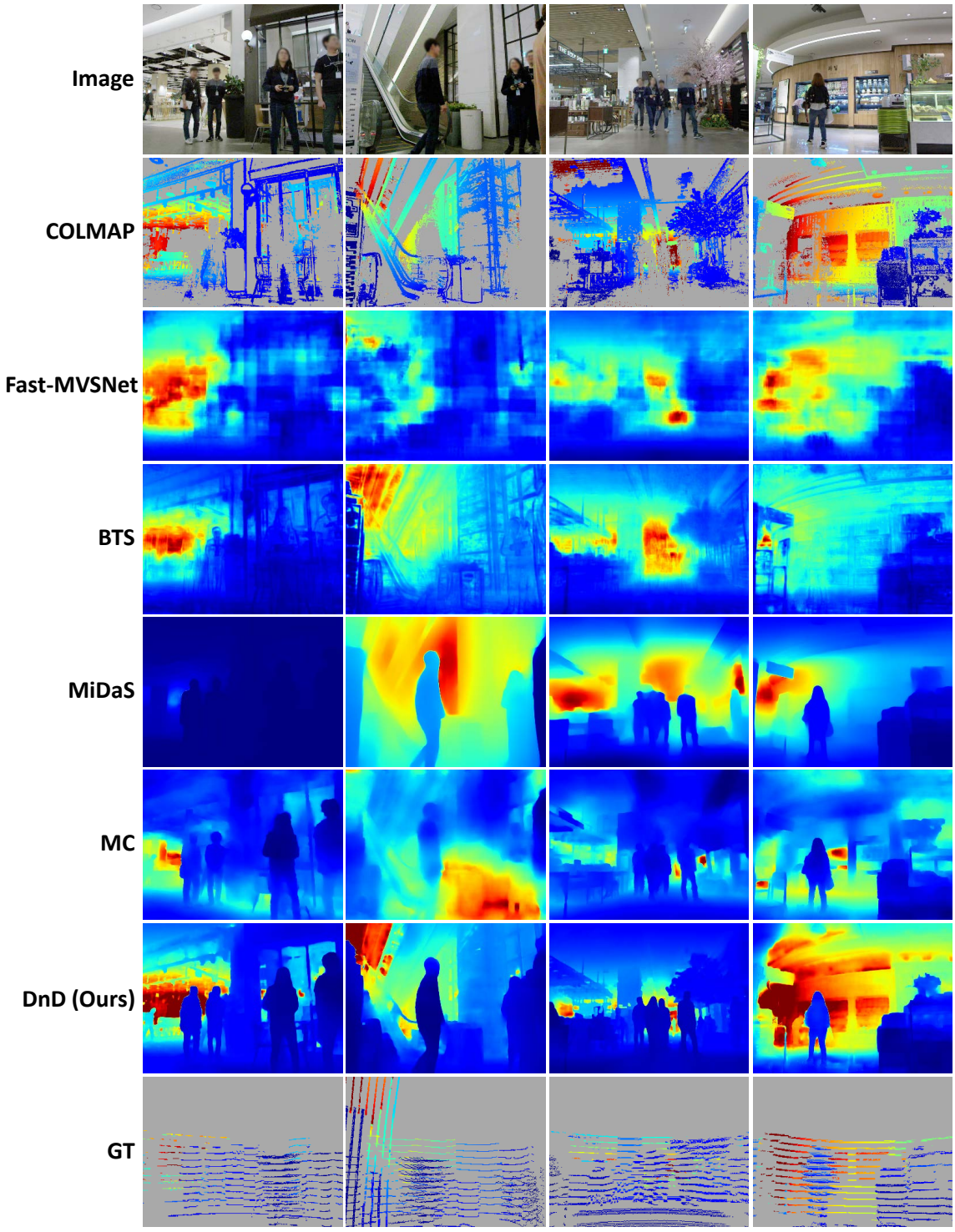

Figure 10: Qualitative comparisons with the state-of-the-art depth estimation algorithms in the Dept dataset. From the third row to bottom, the results of Fast-MVSNet [34], BTS [14], MiDaS [21], Mannequin (MC) [16], and DnD are shown. All color maps use the jet color map (low high; grey means empty depth values.)

References

[1] Navi: Large-scale localization datasets in crowded indoor spaces, 2020. Manuscript submitted for publication. 1

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

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1, 2

[4] Min Young Chang, Suyong Yeon, Soohyun Ryu, and Donghwan Lee. Spoxelnet: Spherical voxel-based deep place recognition for 3d point clouds of crowded indoor spaces. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020. 1

[5] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *Advances in neural information processing systems*, pages 730–738, 2016. 1, 2

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 1, 2

[7] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014. 2

[8] Abdelrahman Eldesokey, Michael Felsberg, and Fahad Shahbaz Khan. Confidence propagation through cnns for guided sparse depth regression. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 5

[9] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[12] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 6

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[14] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 8, 9

[15] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2):155, 2009. 3, 7

[16] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4521–4530, 2019. 1, 2, 7, 8, 9

[17] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018. 1, 2

[18] Fangchang Ma, Guilherme Venturelli Cavalheiro, and Sertac Karaman. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3288–3295. IEEE, 2019. 5

[19] Gabriela Csurka Yohann Cabon Torsten Sattler Noé Pion, Martin Humenberger. Benchmarking image retrieval for visual localization. In *International Conference on 3D Vision*, 2020. 3

[20] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 2

[21] 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 Transactions on Pattern Analysis and Machine Intelligence*, 2020. 7, 8, 9

[22] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5107–5116, 2019. 3

[23] Jerome Revaud, Philippe Weinzaepfel, César Roberto de Souza, and Martin Humenberger. R2D2: repeatable and reliable detector and descriptor. In *NeurIPS*, 2019. 3, 7

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 2

[25] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019. 3

[26] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. 2, 3

[27] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016. 2

[28] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2937, 2013. 1, 2

[29] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 1, 2

[30] Xun Sun, Yuanfan Xie, Pei Luo, and Liang Wang. A dataset for benchmarking image-based localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7436–7444, 2017. 1, 2

[31] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7199–7209, 2018. 1, 2

[32] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. In *2019 International Conference on 3D Vision (3DV)*, pages 348–357. IEEE, 2019. 1, 2

[33] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 311–320, 2018. 1, 2

[34] Zehao Yu and Shenghua Gao. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1949–1958, 2020. 8, 9