Towards High-Fidelity Single-view Holistic Reconstruction of Indoor Scenes – *Supplementary Material* –

A Implementation Details

Room background estimation For background estimation, we use ResNet-18 to encode the global room feature, where the input scene image is resized to 256×256 and the output feature vector has a dimension of 256. As for local features, we adopt the same modified stacked hourglass network as PIFu [7], and the size of the input scene image is 484×648. We also use the same decoder as [7] to be our room reconstruction PIFu, but change the input dimension of the first layer. Different from [7], we use a positional encoding block like [10] with a frequency of 4, which makes the occupancy decoding achieve better performance.
Object reconstruction For object reconstruction, we first crop the target object using its 2D bounding box and resize it to 256×256 before feeding it into a ResNet-18 encoder to produce the object global feature. We also use this 2D bounding box to perform RoI Alignment [4] on the feature map of size $256 \times 64 \times 64$ encoded by the hourglass network. Network architectures for channel-wise attention and mask prediction are all MLPs with 3 and 4 layers, respectively. The architecture of the PIFu decoder is the same as the one for room background estimation. Attentional Channel Filtering module is constructed by two layers of convolution, followed by a global average pooling to generate a feature vector. Two more layers of MLP and a Sigmoid activation is applied to generate the channel-wise attentional weight. The mask segmentation module has four layers of convolution with kernel size 1, all layers is followed with ReLU activation except the last layer that uses Sigmoid activation. The whole object reconstruction pipeline is trained jointly. As our InstPIFu is trained with cross-category data, the shape decoder needs an extra category code like the MGN in Total3D [6]. The loss weight ratio of the occupancy loss to the mask loss is $1:1$ during training.
More Implementation Details We use a similar camera estimator, 2D and 3D detection networks as Total3D [6]. The global image encoders G and $G^{'}$ in Fig. 2 are both simple MLPs after a few pooling and convolution operations. For the local feature extractor and the PIFu decoders, we adopt the same structures as in [7]. Our InstPIFu is trained on 3D-FRONT, where $W_r = H_r = 64$ and $L_c = 256$. 3D points used to train InstPIFu are sampled around the mesh surface randomly as [7] and within the bounding box uniformly with a ratio of $1:1$. Training is conducted using the batch size of 16 for 100 epochs on two NVIDIA RTX3090Ti using 80 hours. The learning rate is initialized as 0.0001, and decayed by a factor of 0.2 in the 50^{th} and 80^{th} epochs.

B More Quantitative Comparisons

Coordinate system As we mention in Section 4.3, indoor shapes are recovered in canonical coordinate to ease the learning of reconstructing indoor objects with various poses and scales. Tab. S1 gives the results of object reconstruction in camera coordinate, which is worse, justifying the design of our method.

Category	bed	chair	sofa	table	desk	nightstand	cabinet	bookshelf	mean ↓ / ↑
$Ours_{cam}$	35.40 / 31.22	48.98 / 27.07	14.12 / 50.46	50.51 / 37.17	58.98 / 30.02	71.63 / 22.29	35.02 / 34.23	24.64 / 39.48	41.64 / 34.93
$Ours_{cano}$	**18.17 / 47.85**	**14.06 / 59.08**	**7.66 / 67.60**	**23.25 / 56.43**	**33.33 / 48.49**	**11.73 / 57.14**	**6.04 / 73.32**	**8.03 / 66.13**	**14.46 / 61.32**

Table S1. Quantitative comparisons of using different coordinate systems for object reconstruction on 3D-FUTURE (CD / F-Score). The values of CD are in units of 10^{-3}.

Effect of 2D detection accuracy We follow Total3D and Im3D to use GT 2D detection during training. There are misaligned bounding boxes when using off-shelf 2D detection while testing. Tab. S2 gives the reconstruction results with GT 2D detection and predicted 2D detection by Faster R-CNN trained on MS COCO dataset, and using predicted 2D bounding box causes minor performance drop.

Category	bed	chair	sofa	table	desk	nightstand	cabinet	bookshelf	mean ↓ / ↑
$Ours_{pred}$	19.60 / **47.99**	18.30 / 55.48	8.14 / 64.91	25.81 / 50.23	56.10 / 41.87	**10.16** / 46.70	8.59 / 64.35	**7.15 / 69.39**	15.35 / 58.21
$Ours_{gt}$	**18.17** / 47.85	**14.06 / 59.08**	**7.66 / 67.60**	**23.25 / 56.43**	**33.33 / 48.49**	11.73 / **57.14**	**6.04 / 73.32**	8.03 / 66.13	**14.46 / 61.32**

Table S2. Quantitative comparisons of using different 2D object proposals for the testing of object reconstruction network on 3D-FUTURE (CD / F-Score).

Evaluation on the Whole Scene We evaluate the reconstruction results of the whole scene including the background and instances of the foreground by CD on 500 images of test scenes of 3D-FRONT, where our approach gets 0.172 and the baseline gets 0.290.

C More Qualitative Comparisons

Background on 3D-FRONT More visual results of background on 3D-FRONT [2] is shown in Fig. S1 and Fig. S2.
3D-FUTURE More qualitative comparisons of indoor object reconstruction on 3D-FUTURE [3] are shown in Fig. S3.
3D-FRONT We present more scene reconstruction results on the testing set of 3D-FRONT [2] in Fig. S4.
SUN RGB-D We present more scene reconstruction results on the testing set of SUN RGB-D [8] in Fig. S5.

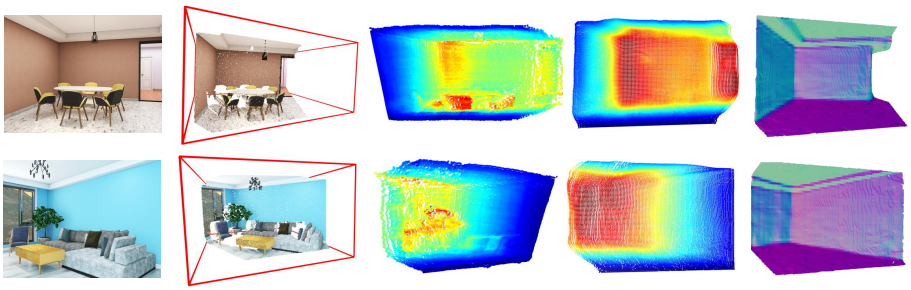

Fig. S1. Visual comparisons for background representation, from left to right: scene images, reconstructed 3D room bounding boxes [11], depth estimation for Factored3D [9] and Adabins [1], as well as our results.

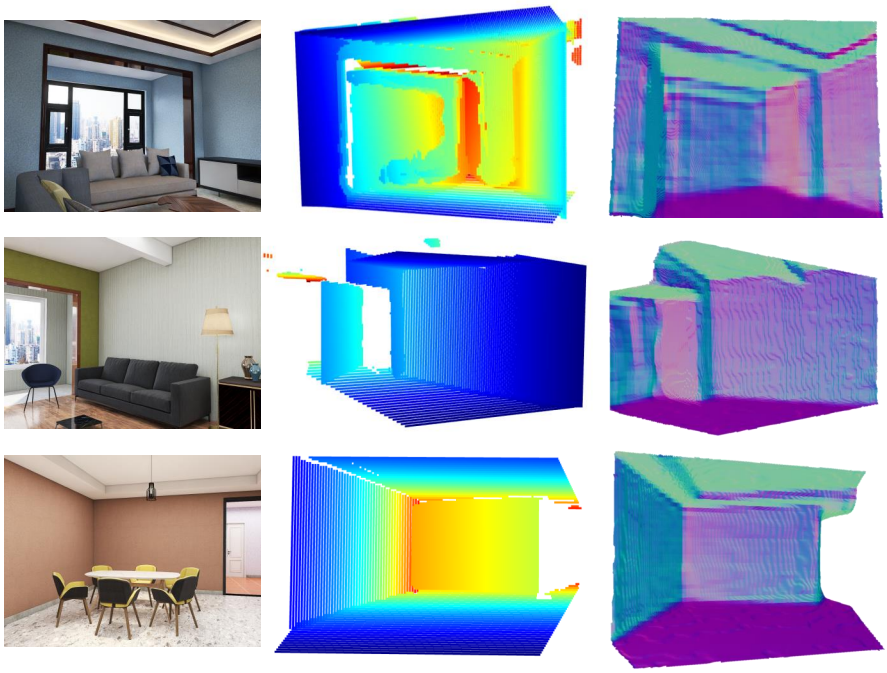

Fig. S2. Visual comparisons for background representation between PlaneRCNN [5] (middle) and ours (right).

Fig. S3. More qualitative comparisons of indoor object reconstruction on 3D-FUTURE. From left to right of every quintuplet: (1) Input image, results from (2) MGN [6], (3) LDIF [11], (4) Ours, and (5) Ground truth.

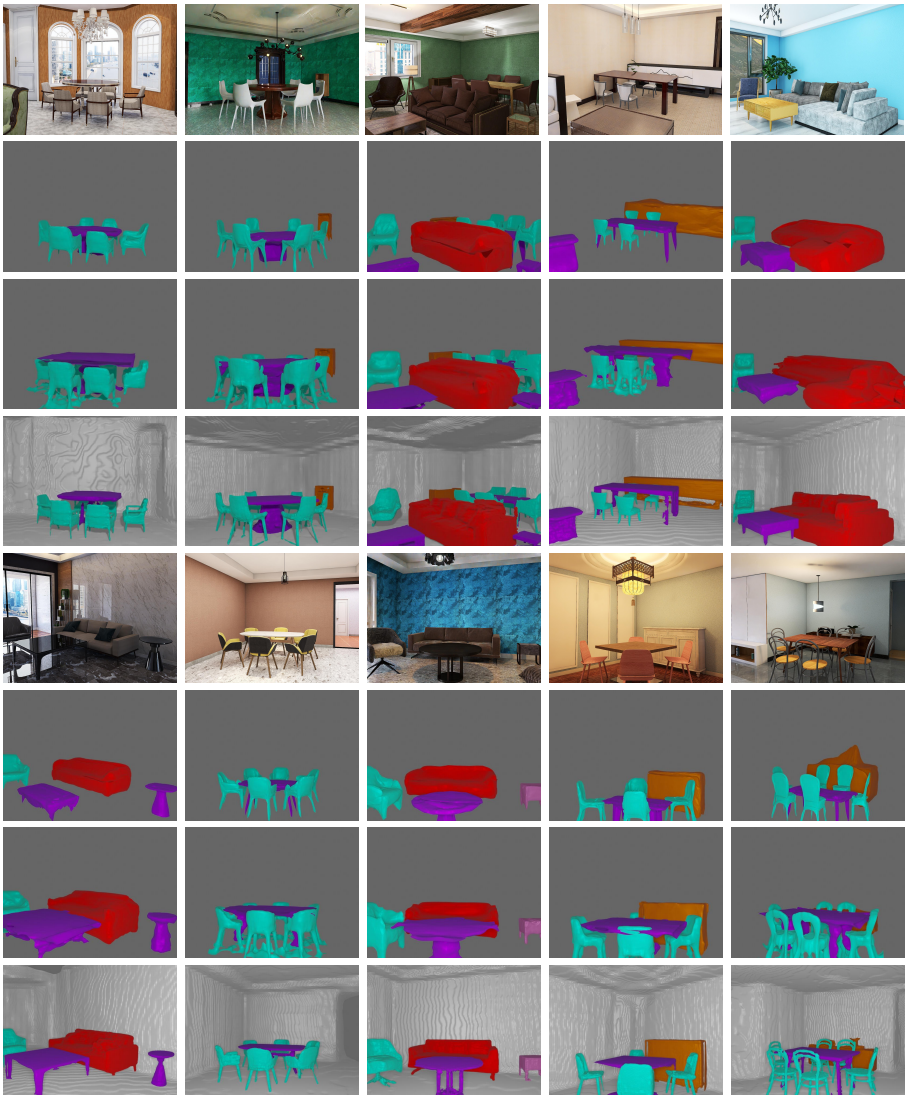

Fig. S4. More qualitative comparisons of holistic scene reconstruction on 3D-FRONT. From the first row to the last: the input image, scene reconstruction results of Total3D [6], Im3D [11], and Ours.

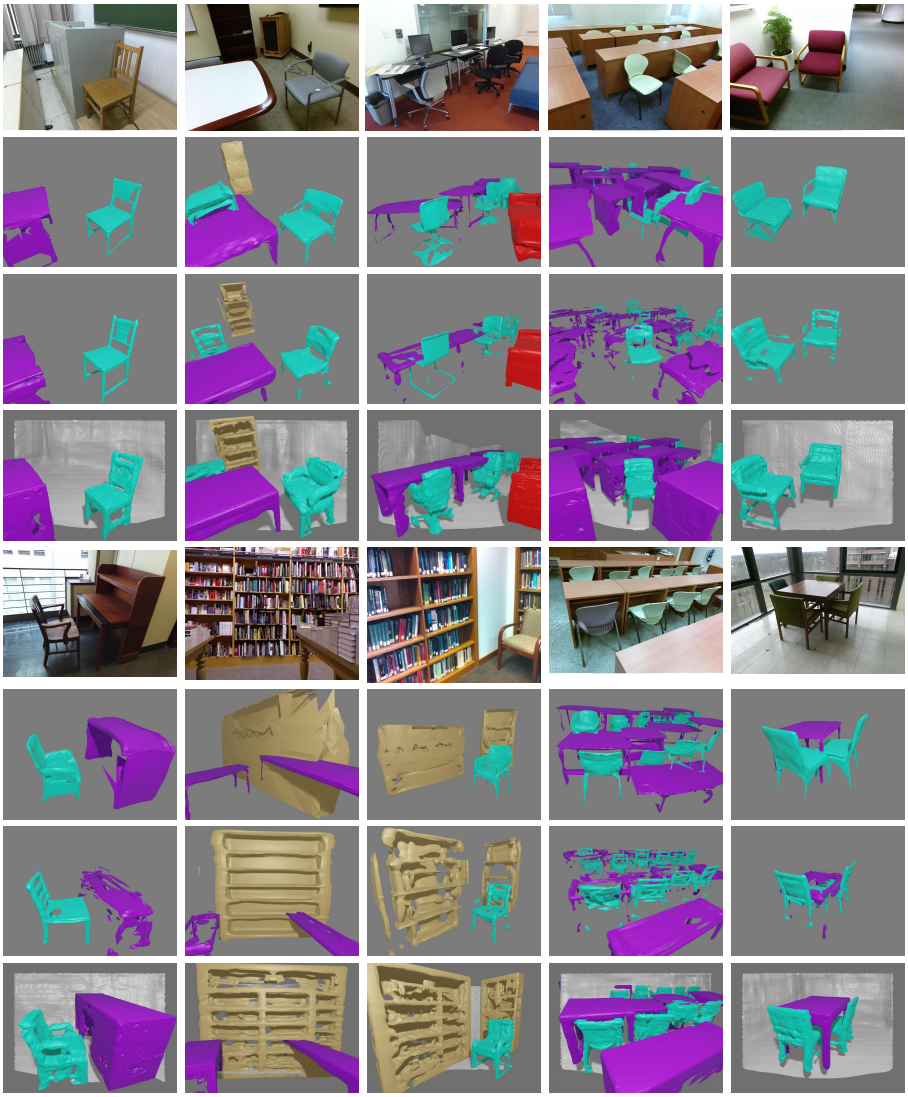

Fig. S5. More qualitative comparisons of holistic scene reconstruction on SUN RGB-D. From the first row to the last: the input image, scene reconstruction results of Total3D [6], Im3D [11], and Ours.

References

1. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4009–4018 (2021)
2. Fu, H., Cai, B., Gao, L., Zhang, L.X., Wang, J., Li, C., Zeng, Q., Sun, C., Jia, R., Zhao, B., et al.: 3d-front: 3d furnished rooms with layouts and semantics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10933–10942 (2021)
3. Fu, H., Jia, R., Gao, L., Gong, M., Zhao, B., Maybank, S., Tao, D.: 3d-future: 3d furniture shape with texture. International Journal of Computer Vision pp. 1–25 (2021)
4. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
5. Liu, C., Kim, K., Gu, J., Furukawa, Y., Kautz, J.: Planercnn: 3d plane detection and reconstruction from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4450–4459 (2019)
6. Nie, Y., Han, X., Guo, S., Zheng, Y., Chang, J., Zhang, J.J.: Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 55–64 (2020)
7. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2304–2314 (2019)
8. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 567–576 (2015)
9. Tulsiani, S., Gupta, S., Fouhey, D.F., Efros, A.A., Malik, J.: Factoring shape, pose, and layout from the 2d image of a 3d scene. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 302–310 (2018)
10. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2021)
11. Zhang, C., Cui, Z., Zhang, Y., Zeng, B., Pollefeys, M., Liu, S.: Holistic 3d scene understanding from a single image with implicit representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8833–8842 (2021)