Supplementary Material
3D-to-2D Distillation for Indoor Scene Parsing

Zhengzhe Liu[1] Xiaojuan Qi[2] Chi-Wing Fu[1]
[1]The Chinese University of Hong Kong [2]The University of Hong Kong
{zzliu,cwfu}@cse.cuhk.edu.hk xjqi@eee.hku.edu.hk

In this supplementary material, we show sparse 3D data projected onto associated 2D images (Section 1), quantitative (per-category IoU) and qualitative comparisons for unpaired 3D-to-2D distillation (Section 2), visual results on NYU-v2 to show the generalizability of our method (Section 3), the configurations in the depth prediction task (Section 4), and the network architecture of the discriminator in Semantic Aware Adversarial Loss (Section 5). The code can be found in `https://github.com/liuzhengzhe/3D-to-2D-Distillation-for-Indoor-Scene-Parsing`.

1. Visualization of the Projected 3D data to 2D Image

This section corresponds to Section 4.2 in the main paper. In this section, we show 3D data projected onto associated 2D images.

We use full 3D data for feature extraction, but after projecting all points to the 2D image domain, we save only the 3D point features in the 2D image grid at (x, y) pixel locations that are multiples of K (empirically set as 8) to reduce the I/O burden when training the 2D network. This is certainly a trade-off between the I/O burden and the performance.

As shown in Figure 1, we can see that the projected 3D data points are very sparse, even though they are dilated to increase the number of pixels occupied by each 3D point. On average, only 16.38% and 10.57% of 2D pixels have corresponding 3D points in ScanNet-v2 and in S3DIS, respectively.

2. Detailed Results and Visualization of the Unpaired 3D-2D distillation results

This section corresponds to Section 4.3 in the paper. Here, we present the per-category performance and quantitative results of our 3D-to-2D distillation on unpaired 2D-3D data. As shown in Table 1, the results of most categories can be improved using our method. Further, as Figure 2 shows, PSPNet (baseline) has some failure cases, especially when the 2D image cues are confusing or misleading. For example, as marked by the red boxes in Figure 2, the door in row 1 has an unusual blue color and the refrigerator in row 2 has a similar color and texture as the door. The bookshelf in row 3 is partially occluded and the chair in row 4 looks similar to the floor. In comparison, our 3D-to-2D distillation model improves PSPNet (baseline). This demonstrates that our approach enhances the 2D network to better leverage the geometrical information for resolving issues like occlusions, viewpoints, and texture, even without paired 2D-3D data.

3. Visualization of the Model Generalization Results on NYU-v2

This section corresponds to Section 4.6 in the main paper. Here, we show the visual results of the baseline and our model, both trained only on ScanNet-v2 and directly tested on NYU-v2. The results in Figure 3(d) manifest that our model can better segment the objects in the unseen NYU-v2 dataset by leveraging the distilled 3D information. The results demonstrate the generalization ability of our 3D-to-2D distillation model.

4. Details of Depth Prediction for 3D Information Evaluation

This section provides details for Section 4.5 in the main paper. In detail, the depth prediction network uses VGG [1] as the backbone. We use the first 2k images in the ScanNet-v2 training set for training and the first 200 images in the ScanNet-v2 validation set for testing. All models (including the baseline) are trained with SGD for 10 epochs. The initial learning rate is $5e^{-7}$ and decreased by $2.5e^{-8}$ for each epoch. Figure 4 shows two more examples that follow the style of Figure 7 in the main paper, showing that the depth information reconstructed from our network has better 3D structure than the depth information reconstructed from the baseline 2D network without our 3D-to-2D distillation.

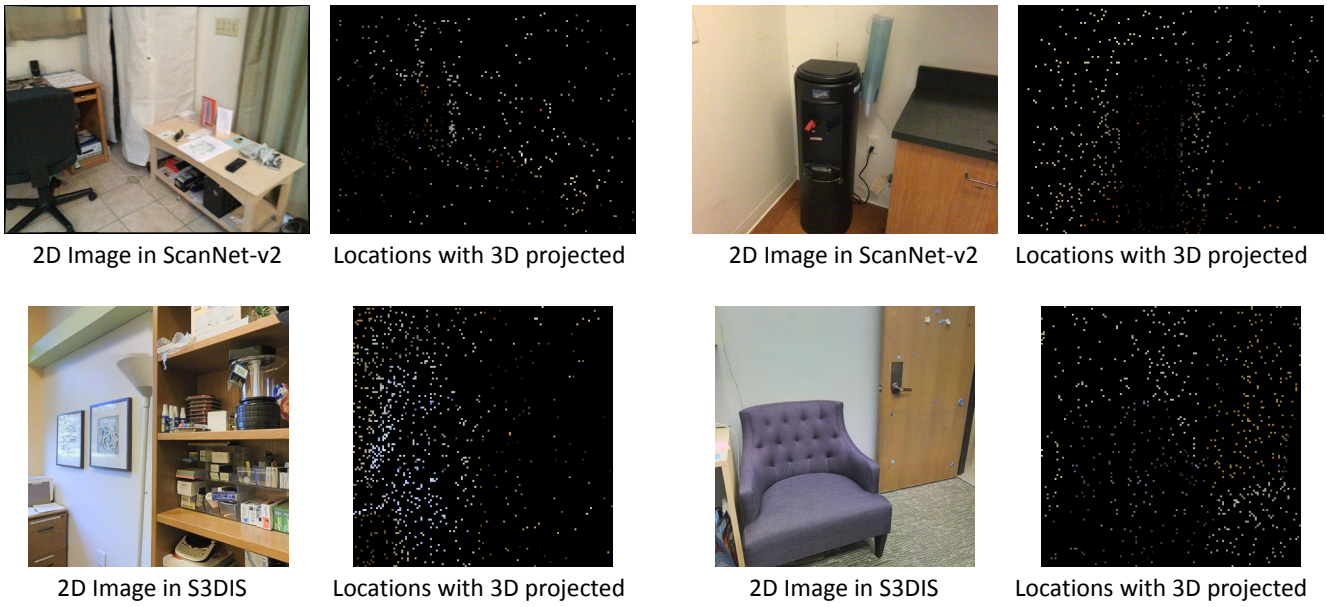


Figure 1. Projecting the sparse 3D data onto the associated 2D images for the sparse paired 2D-3D data in ScanNet-v2 (top row) and S3DIS (bottom row). We only perform the regression as Equation 1 on the pixels with projected 3D data.

Method	mIoU	wall picture	floor counter	cabinet desk	bed curtain	chair refrigerator	sofa shower curtain	table toilet	door sink	window bathtub	bookshelf other-furniture
PSPNet-50 [3]	49.50	79.06 62.70	84.81 57.30	61.62 15.37	69.73 48.74	60.13 36.86	60.53 14.25	41.10 62.65	36.28 49.35	54.52 24.59	54.53 16.03
PSPNet-50 + **Our Unpaired** 3D-to-2D Distillation	**51.70**	80.11 64.28	83.28 58.80	60.96 17.63	71.57 49.12	61.19 45.84	63.82 22.43	40.16 67.67	39.29 53.38	55.48 25.00	56.22 17.61

Table 1. Unpaired 3D-to-2D distillation results on the ScanNet-v2 validation set. In the 2nd column from the left, we compare the overall performance (mIoU) of the baseline (PSPNet-50) vs. our full model (bottom), and then in the subsequent ten columns, we compare the per-category performance (20 categories in total), where in each of these columns, we show results for two categories in each cell. With almost negligible extra effort, our approach can improve the results of most of the categories, and relatively improve the mean IoU by almost 5%, *i.e.*, from 49.50 to 51.70, by means of training the network with unpaired 2D-3D data.

5. Network Architecture of the Discriminator of Semantic-Aware Adversarial Loss

The architecture of discriminator D^c is composed of 6 fully-connected layers with channels 64, 32, 16, 8, 4, and 1, respectively. Each of the first five is followed by a Leaky-ReLU [2] as the activation function, and the last one is followed by a Sigmoid operation to modulate the output within the interval $(0, 1)$, indicating the confidence of whether the input feature vector is from the 2D network or 3D network. D^c is trained with all the feature vectors belonging to category c and the discriminators for different categories are optimized individually without sharing weights. Each discriminator only has 0.009M parameter, and its input is an $8\times$ down-sampled feature map with 96 channels, the computation is 63.9M MAC for all of them, and all discriminators are trained simultaneously.

References

[1] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[2] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[3] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.

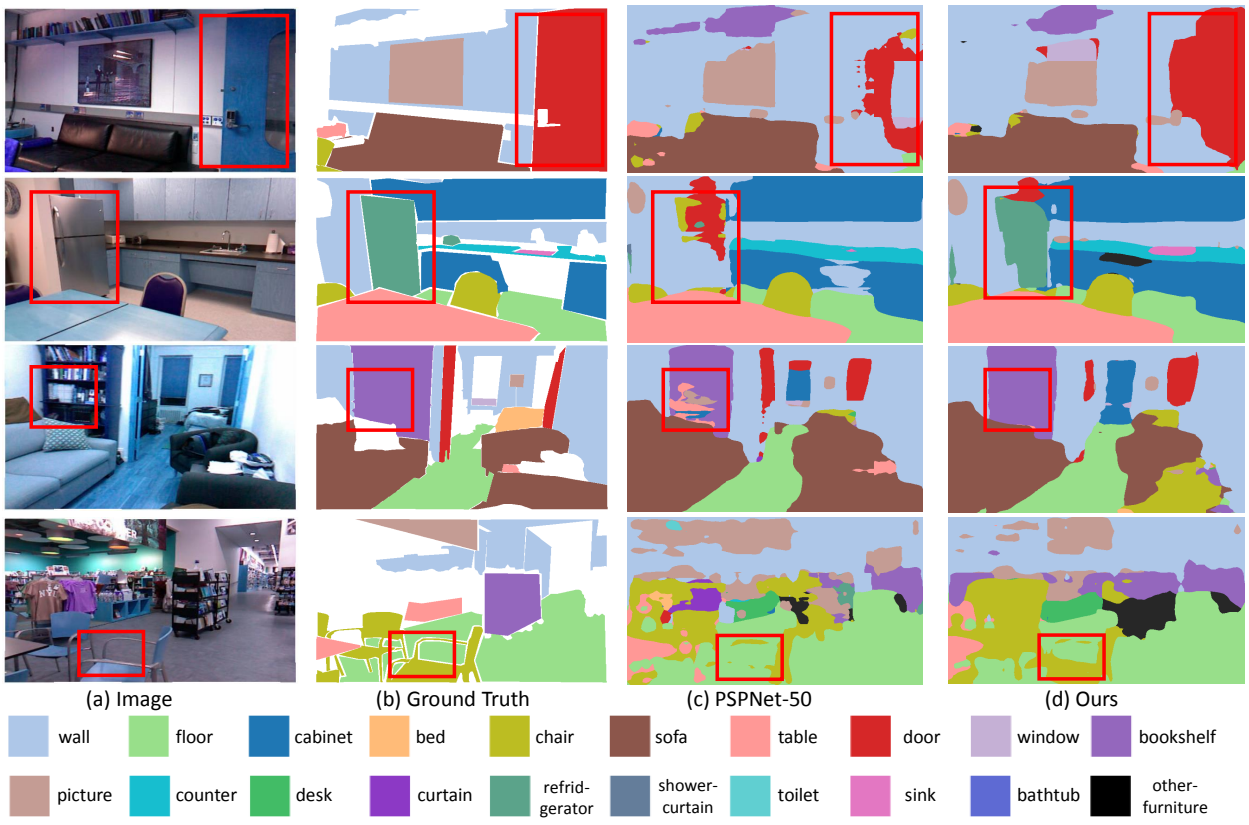


Figure 2. Visual comparison on the NYU-v2 data, with and without using our 3D-to-2D distillation method. Comparing the results shown on the rightmost two columns, we can see that our 3D-to-2D distillation model better predicts the door, the refrigerator, the bookshelf, and the chair. Our model makes these predictions by leveraging the 3D information distilled from the "ScanNet-v2" 3D scenes, which are *not paired* with the NYU-v2 image data shown above.

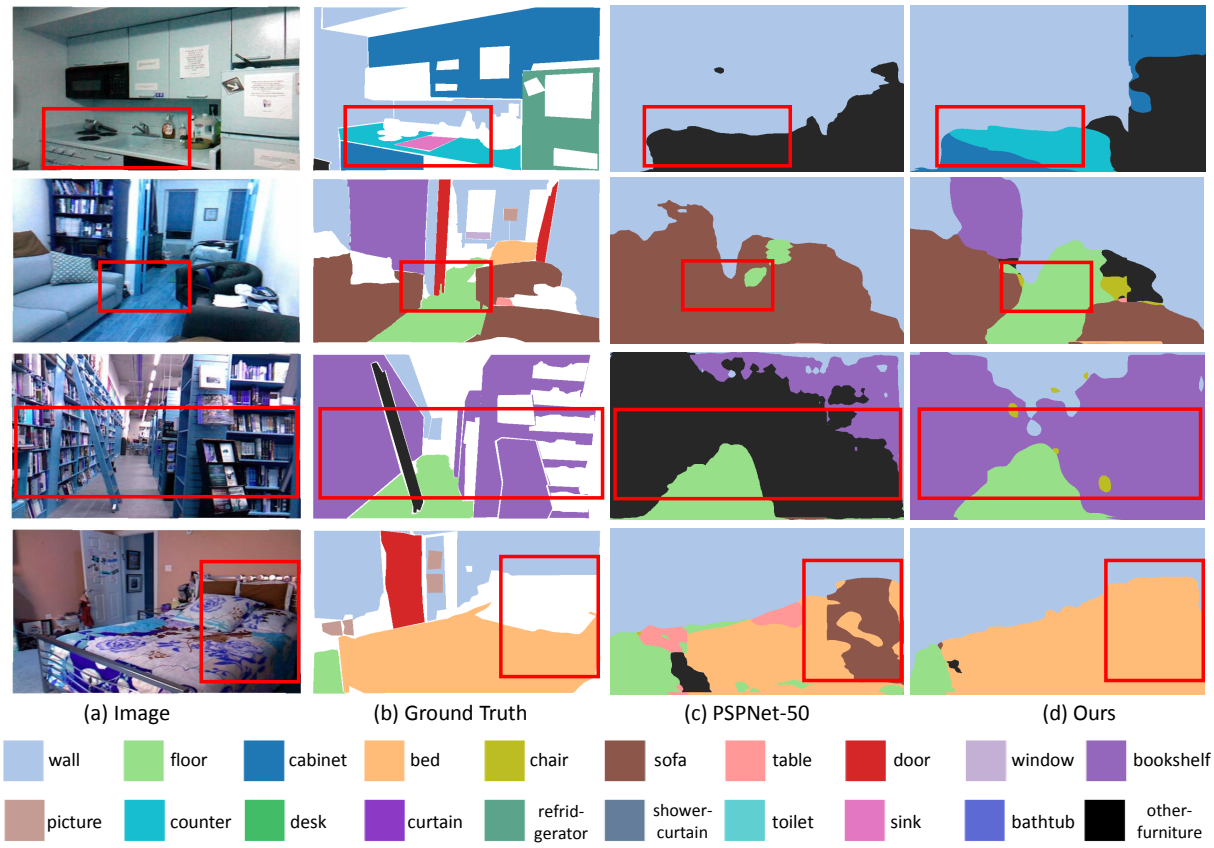
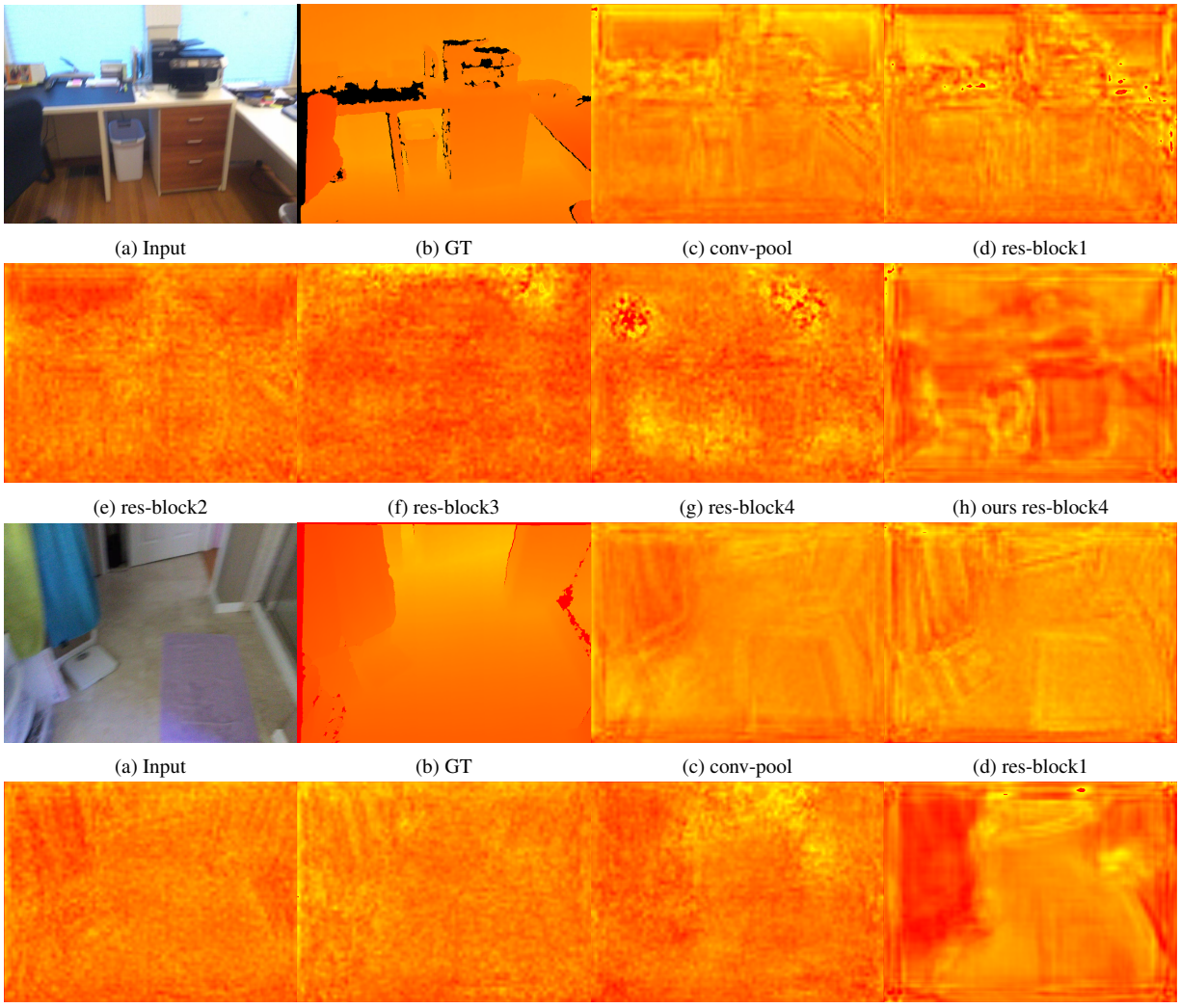


Figure 3. Visualization of the results that demonstrate the generalizability of our model. Both models, *i.e.*, (c) & (d), are trained on ScanNet-v2 and tested on NYU-v2. Our model (d) with the 3D-to-2D distillation can better predict the cabinet, counter, bookshelf, floor, and bed (see the red boxes) by leveraging the distilled 3D information. Quantitative comparison results can be found in the main paper.



(a) Input (b) GT (c) conv-pool (d) res-block1

(e) res-block2 (f) res-block3 (g) res-block4 (h) ours res-block4

(a) Input (b) GT (c) conv-pool (d) res-block1

(e) res-block2 (f) res-block3 (g) res-block4 (h) ours res-block4

Figure 4. Depth reconstructed from the original 2D network as baseline (c-g) and ours (h) vs. the ground truth (GT) shown in (b). (c-g) are from the feature maps of PSPNet-50, whereas (h) is from the "res-block-4" of the network with our 3D-to-2D distillation.