Physically-Based Rendering for Indoor Scene Understanding Using Convolutional Neural Networks

SUPPLEMENTAL MATERIAL

Yinda Zhang Shuran Song Ersin Yumer Manolis Savva
Joon-Young Lee Hailin Jin Thomas Funkhouser

Princeton University Adobe Research

1. Dataset Analysis

1.1. Semantic Segmentation

Our synthetic dataset contains on average 11 48 objects per image, and 54 90% of the pixels are covered by objects, i.e. not wall, floor, or ceiling. On the contrary, NYUv2 contains 24 20 objects per image, and 68 17% of the pixels are covered by objects. Fewer number of instance and object-covered pixels suggests that the real scene is more cluttered containing more objects, and probably our synthetic camera should move closer to the objects to have a zoomed in view.

1.2. Distribution of Surface Normal

Figure 1 shows the distribution of surface normal for all pixels in our synthetic data (the LEFT column) and NYUv2 (the RIGHT column) respectively. The normal distribution is visualized in a panorama, with x axis corresponding to angle in horizontal plane from [ ], and y axis corresponding to the vertical angle from [2 2]. The normal is calculated in camera coordinates, where z- is gravity direction, x+ points to the right-hand side, and y+ points to the front of the camera. We also show the distribution of normal direction on foreground (pixels belong to an object) and background (belong to wall, floor, or ceiling) area respectively on the 2nd and 3rd row. We can see that the overall and foreground distribution of synthetic data is similar to that of the NYUv2 dataset. However, the background distribution is different, because the vertical tilted angle is fixed such that the normal direction of floor or ceiling are all the same (two highlighted single dots) and the normal of wall falls in a great circle on the panorama.

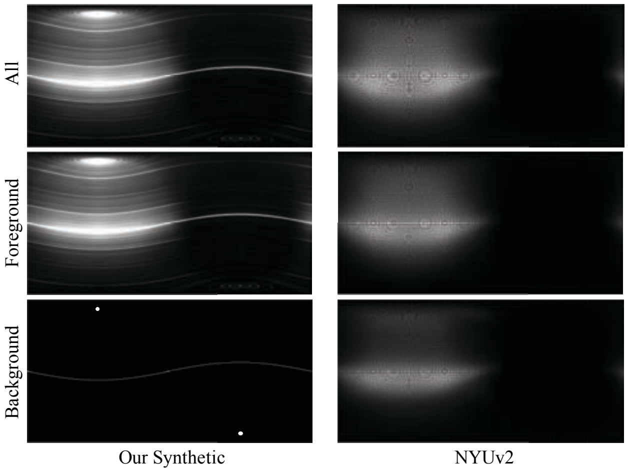


Figure 1. Surface normal distribution of our synthetic dataset and NYUv2. The normal distribution is visualized in a panorama, with x axis corresponding to [ ], and y axis corresponding to [2 2]. The normal is calculated in camera coordinates, where z- is gravity direction, x+ points to right-hand side, and y+ points to the front of the camera. There are two single dots on the background distribution of our synthetic data highlighted for visualization purpose.

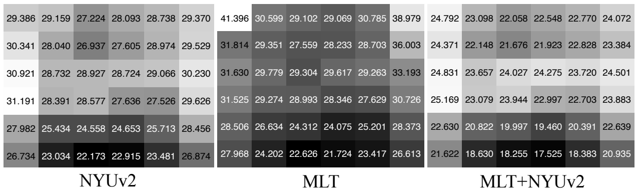
2. Additional Results

2.1. Normal Prediction

2.1.1 Quantitative Analysis

Figure 2 shows the angle error for pixels within each sub-region of the images, i.e. error along x and y axis of the camera coordinates mentioned above. The image dimension (640 480) is divided into 6 6 sub-regions. The number on each sub-region shows the mean of angle error, and darker intensity indicates lower error. "NYUv2" is the model directly trained on NYUv2. "MLT" is model pre-

indicates equal contributions.



29.386	29.159	27.224	28.093	28.738	29.370
30.341	28.040	26.937	27.605	28.974	29.529
30.921	28.732	28.927	28.724	29.066	30.230
31.191	28.391	28.577	27.636	27.526	29.626
27.982	25.434	24.558	24.653	25.713	28.456
26.734	23.034	22.173	22.915	23.481	26.874

NYUv2

41.396	30.599	29.102	29.069	30.785	38.979
31.814	29.351	27.559	28.233	28.703	36.003
31.630	29.779	29.304	29.617	29.263	33.193
31.525	29.274	28.993	28.346	27.629	30.726
28.506	26.634	24.312	24.075	25.201	28.373
27.968	24.202	22.626	21.724	23.417	26.613

MLT

24.792	23.098	22.058	22.548	22.770	24.072
24.371	22.148	21.676	21.923	22.828	23.384
24.831	23.657	24.027	24.275	23.720	24.501
25.169	23.079	23.944	22.997	22.703	23.883
22.630	20.822	19.997	19.460	20.391	22.639
21.622	18.630	18.255	17.525	18.383	20.935

MLT+NYUv2

Figure 2. Surface normal estimation error of different sub-area in image. The image dimension (640 480) is divided into 6 6 sub-regions. The number on each sub-region shows the mean of angle error, and darker intensity indicates lower error. "NYUv2" is the model directly trained on NYUv2. "MLT" is model pretrained on our synthetic data. "MLT+NYUv2" is the "MLT" model further finetuned on NYUv2.

trained on our synthetic data. "MLT+NYUv2" is the "MLT" model further finetuned on NYUv2. It is clear to see that all of the models works better in the mid-lower part of the image, which is mostly occupied by floor or top of the furniture, e.g. table, bed, that shows upward normal direction. The area near left and right boundary of the image shows comparatively worse performance.

Figure 3 shows the angle error with regard to the depth of the pixel, i.e. error along the z axis of the camera coordinates. As we can see, the error is the smallest for pixels with depth in range of [2 3], and keeps increasing when the points are further away from the camera. This indicates that pixels far away from camera shows less evidence of local geometry in color image. On the other hand, as the noise of depth is proportional to depth for most of the depth sensor, the noise in the ground truth may also contribute to the error.

Table 1 shows the performance of different models on pixels from different semantic area. We can see that the error on the foreground area which consists of objects is significantly larger than the error on the background area covered by wall, ceiling, and floor. It is consistent with the observation that foreground area containing various of objects exhibits more diverse and rapidly changing surface normal, which is hard to predict. However, the error on the background area is still comparatively big, which is a bit of surprising as the area mostly consists of large plane surfaces that are easy to deal with. We hypothesize that the noise in the ground truth contributes to the error of both foreground and background area, which is more visible to the later one.

2.1.2 Additional Visual Results

We provide more results of surface normal estimation in Figure 4 and Figure 5. The 1st and 2nd column show input images and ground truth normal converted from the depth image. The 3rd to 5th column show the results of the model directly trained on NYUv2, pretrained on MLT-IL/OL rendering, and finetuned on NYUv2 after pretraining.

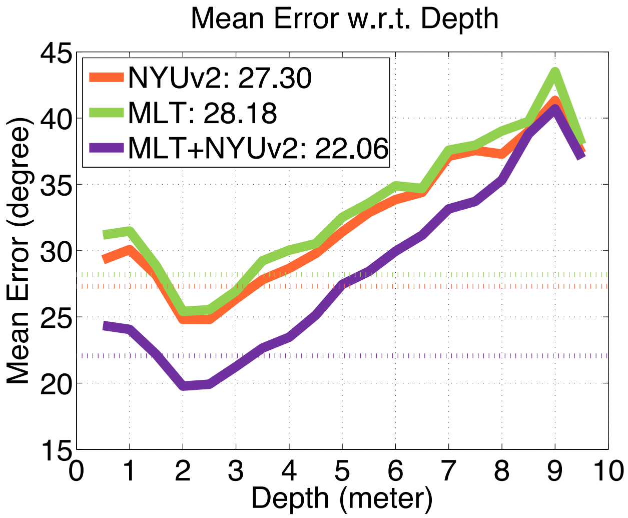


Figure 3. Surface normal estimation error w.r.t. depth. The number in the legend shows the average of overall error for each method respectively. The dashed line indicates these values in the figure. The performance is better if the curves and numbers are lower.

Model	Area	Mean ( )	Median ( )	11.25	22.5	30
NYUv2	F	29.26	23.48	22.54	48.14	61.03
	B	24.95	18.27	32.81	57.98	69.15
MLT	F	29.37	23.78	22.38	47.67	60.08
	B	26.76	19.28	31.33	55.75	66.33
MLT +NYUv2	F	24.17	17.29	33.42	60.48	71.50
	B	19.54	12.15	47.01	71.75	79.74

Table 1. Surface normal estimation error for fore/background area. For each model, we provide the performance for pixels on either objects ("F") or background ("B"), i.e. wall, floor, or ceiling.

2.2. Semantic Segmentation

Table 2 shows the per-class semantic segmentation results. Table 3 shows the object mapping from our synthetic dataset 84 category to NYUv2 40 category. Figure 6 shows additional visual results from semantic segmentation task.

2.3. Boundary Edge Prediction

Figure 7 shows additional visual results of the boundary detection. First column is the input color images, second to fourth columns are the results of the model, after initialized with weights learned from ImageNet, (2) directly trained on NYUv2, (3) pretrained on MLT-IL/OL rendering, and (4) pretrained on MLT-IL/OL rendering followed by finetuning on NYUv2. The last column is the ground truth overlaid with the difference between the model w/wo pretraining on our MLT-IL/OL. Red pixels denote enhanced, and green pixels denote suppressed edges as object boundary by the model with pretraining. We can see that edges within objects or on the background are successfully suppressed.

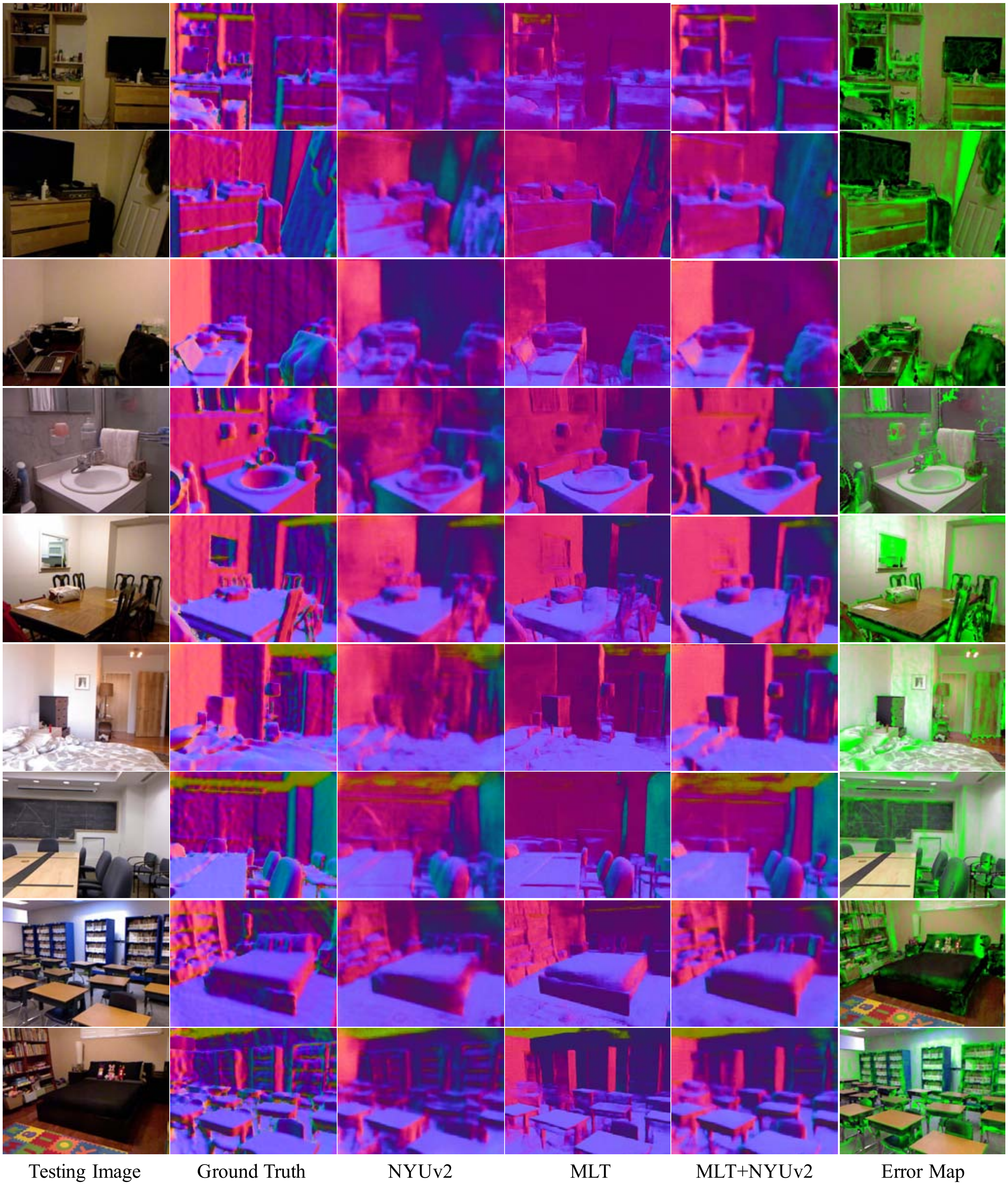


Figure 4. Visualization of surface normal estimation on NYUv2 testing images. The 1st and 2nd column show input images and ground truth normal converted from the depth image. The 3rd to 5th column show the results of the model directly trained on NYUv2, pretrained on MLT-IL/OL rendering, and finetuned on NYUv2 after pretraining.

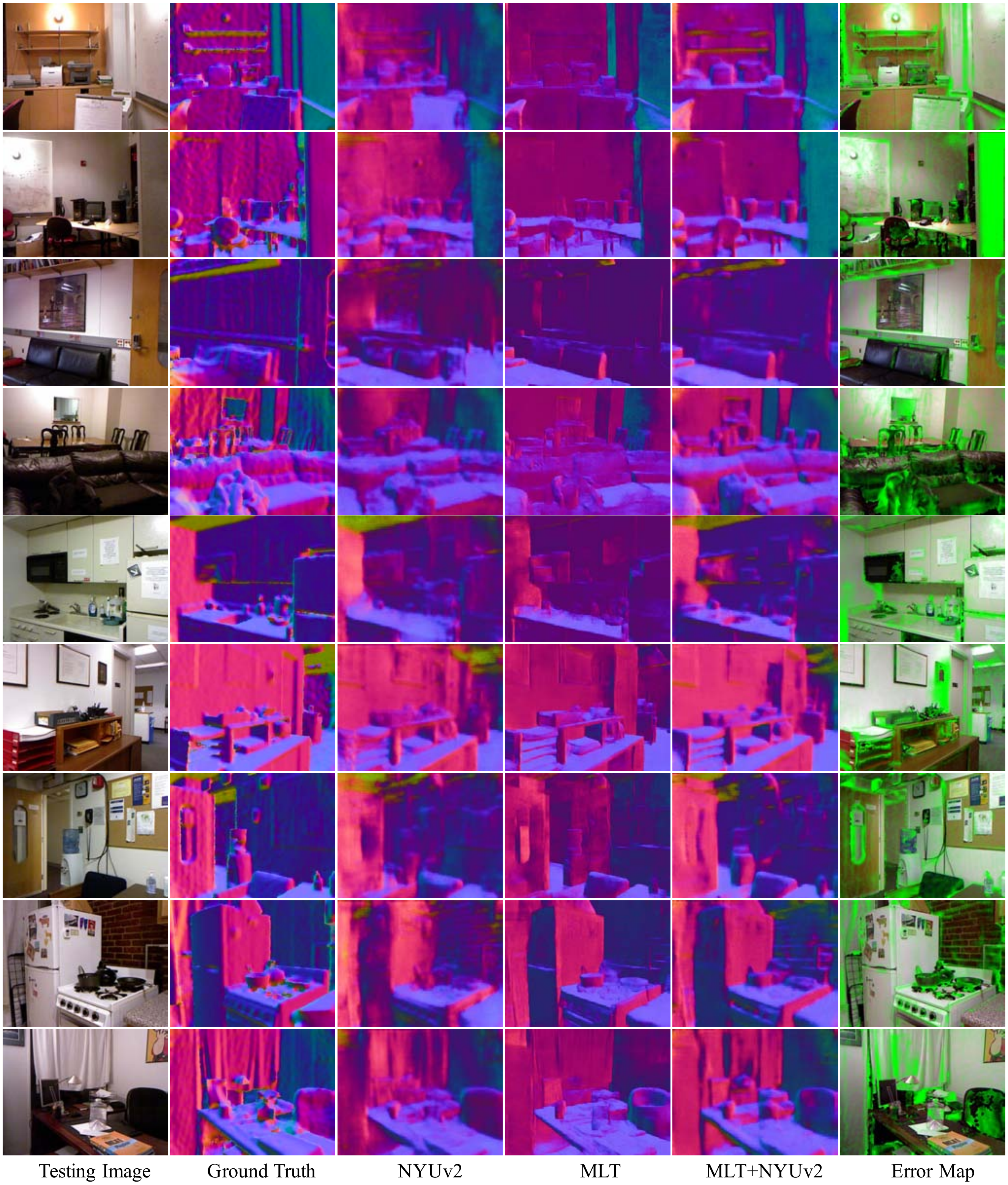


Figure 5. Visualization of surface normal estimation on NYUv2 testing images. The 1st and 2nd column show input images and ground truth normal converted from the depth image. The 3rd to 5th column show the results of the model directly trained on NYUv2, pretrained on MLT-IL/OL rendering, and finetuned on NYUv2 after pretraining.

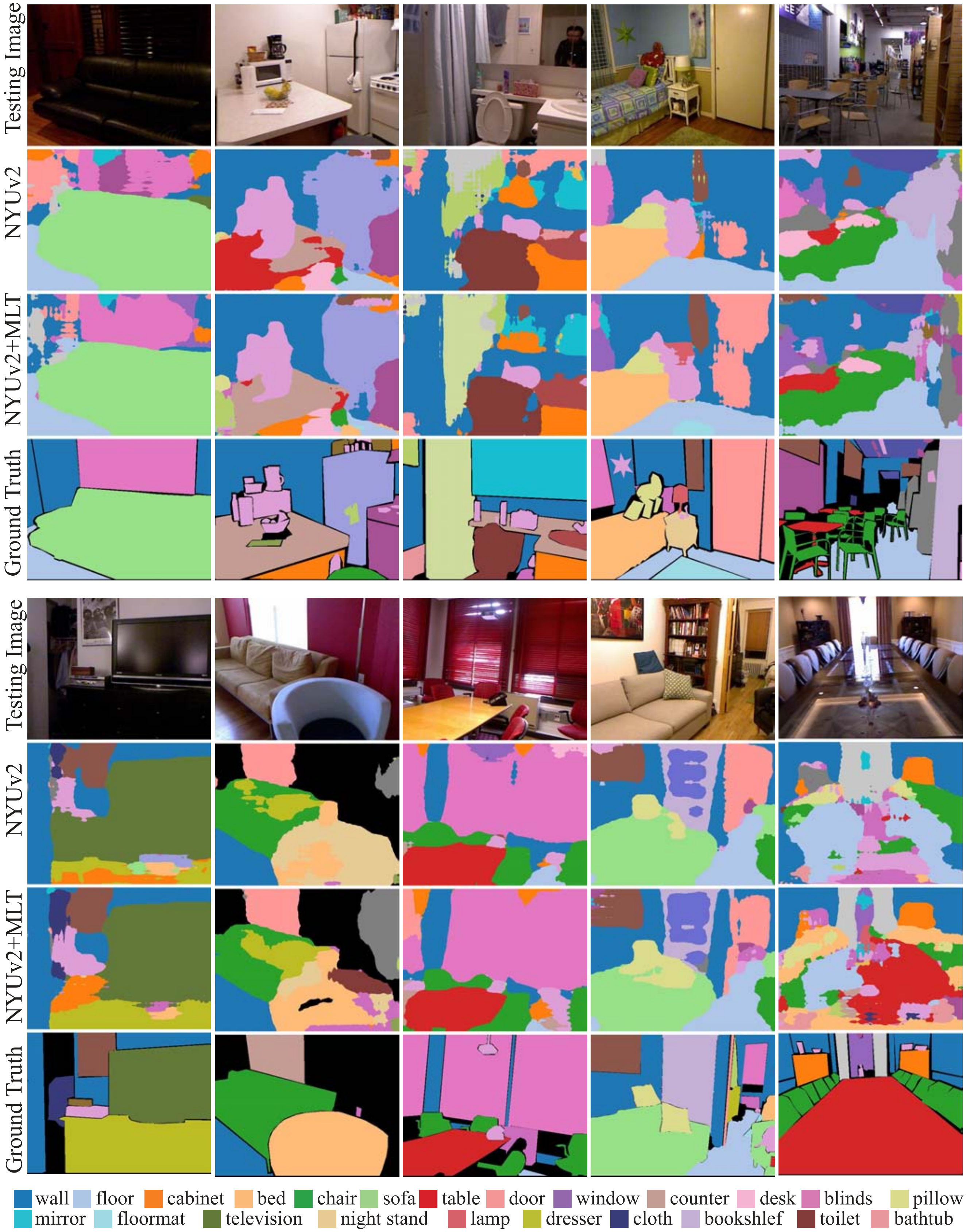


Figure 6. Visualization of semantic segmentation result on NYUv2 testing images.

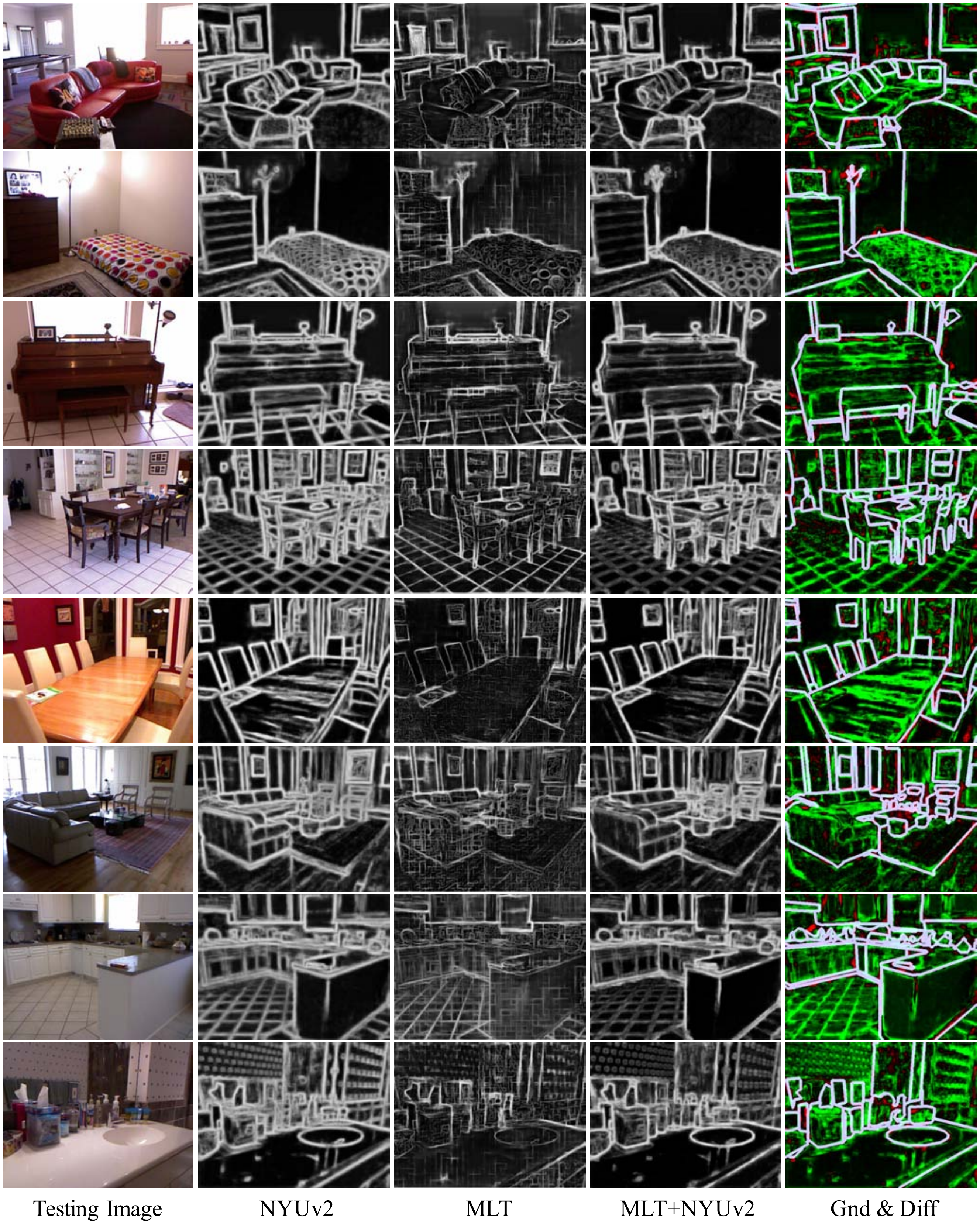


Figure 7. Visualization of object boundary detection on NYUv2 testing images. First column is the input color images, second to fourth columns are the results of the model, after initialized with weights learned from ImageNet, (2) directly trained on NYUv2, (3) pretrained on MLT-IL/OL rendering, and (4) pretrained on MLT-IL/OL rendering followed by finetuning on NYUv2. The last column is the ground truth overlaid with the difference between the model w/wo pretraining on our MLT-IL/OL. Red pixels denote enhanced, and green pixels denote suppressed edges as object boundary by the model with pretraining.

	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	blinds	desk	shelves	curtain	dresser	pillow	mirror	floor mat
ImageNet+NYU	65.7	71.1	48.4	53.4	43.6	46.0	31.6	25.2	38.7	36.8	44.8	40.7	46.1	14.5	7.4	33.2	26.1	28.7	11.0	24.2
ImageNet+MLT	51.1	47.9	4.1	23.4	23.2	19.6	14.1	9.0	11.1	0.0	7.0	8.8	32.2	5.2	2.3	14.6	3.4	0.0	0.0	11.8
ImageNet+MLT+NYU	67.1	72.5	46.9	53.8	45.5	45.3	32.2	26.5	40.2	32.7	46.9	41.6	51.9	14.8	7.0	37.0	31.3	30.1	14.9	28.7
ImageNet+OPNGL	25.6	13.3	2.0	16.0	6.9	10.4	3.4	0.4	0.0	0.0	3.3	3.4	1.0	0.0	0.5	7.8	0.0	0.0	0.0	6.0
ImageNet+OPNGL+NYU	66.6	72.8	48.2	52.7	43.7	46.3	31.5	22.5	37.4	35.1	47.2	42.4	44.6	14.7	6.8	31.4	35.6	31.0	17.0	25.4
clothes	ceiling	books	refridgerator	television	paper	towel	shower curtain	box	whiteboard	person	night stand	toilet	sink	lamp	bathtub	bag	otherstructure	otherfurniture	otherprop	mean
15.5	46.7	25.9	30.7	36.0	21.1	21.7	8.6	6.8	27.5	50.2	21.9	56.2	39.9	29.1	33.0	6.5	17.1	9.1	29.4	31.7
0.0	10.9	0.0	2.5	5.7	0.0	0.0	0.0	0.0	0.0	12.2	6.1	8.1	7.1	14.4	15.3	0.0	1.0	0.0	11.3	9.6
19.9	47.3	25.7	31.7	37.4	23.4	23.5	10.8	5.7	36.1	54.8	28.3	53.1	39.5	32.8	29.9	8.8	16.9	8.8	28.4	33.2
0.0	0.0	0.0	0.0	2.2	0.0	0.0	0.0	0.0	0.0	3.5	0.0	1.9	8.0	7.3	3.6	0.0	0.5	0.0	0.6	3.2
18.2	46.4	24.2	32.2	37.7	21.2	23.3	6.9	6.0	40.1	51.4	24.9	55.4	42.7	29.4	35.4	10.5	16.6	8.9	28.9	32.8

Table 2. Semantic segmentation performance for NYUv2 40 categories. For each semantic category, we show the IOU accuracy of models w/wo pretraining on synthetic data with different rendering qualities.

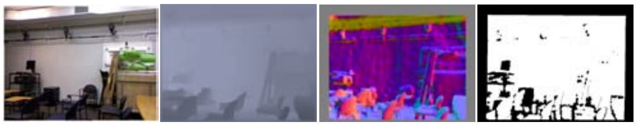


Figure 8. Example of surface normal ground truth. The surface normal is converted from depth map, which might be noisy due to the limit of sensor technology. The valid map indicates if the normal on each pixel is reliable. Only valid pixels are used for training and testing.

3. Ground Truth for Surface Normals

For the results presented in the paper, we use the ground truth provided by Eigen *et al*. [2] on their project webpage. The ground truth is computed at each pixel by fitting a least squares plane, using the code released by Silberman *et al*. [3]. Given a pixel location, they first sample 3D points from 18 18 nearby region, and form them into a matrix of $A = N \quad 3$. The normal for the pixel is then computed as the eigenvector of $A^T A$ corresponding to the smallest eigenvalue. The confidence of this estimated normal is defined as $1 \quad {}_1 \ sigma_2$, where ${}_1$ is the smallest, and ${}_2$ is the second smallest eigenvalue of $A^T A$. At training time, we only compute loss on valid pixels, such that invalid pixels always have a zero loss and hence do not propagate any gradient back. At test time, only the valid pixels are evaluated.

The "ground truth" normals computed in this way are quite noisy, due to noise in the depth sensor. To evaluate the effect of this noise, and to provide results with respect to normals estimated more robustly, we fit planes to each of the area labelled as either wall, ceiling, and floor, and replace the surface normal of these area with the normal of the fit plane. Figure 9 shows an example of the ground truth before and after plane fitting. As an additional experiment beyond the ones described in the paper, we evaluate MLT+NYUv2, NYUv2, and MLT models presented in the paper on this new ground truth. We find that they achieve mean angle errors of 23 12, 28 18, and 28 28, respectively, compared to the 22 06, 27 30, and 28 59 on the original ground truth. We can see that only MLT, the model pretrained on synthetic data, achieves comparatively better performance. Table 4 shows the evaluation of each model on background area on the original and plane fitting ground truth. Again, the MLT model shows the most improvement, and performs even better than the model directly trained on NYUv2. This indicates that the model pretrained on synthetic predicts cleaner and more accurate background geometries than ones trained on the noisy ground truth.

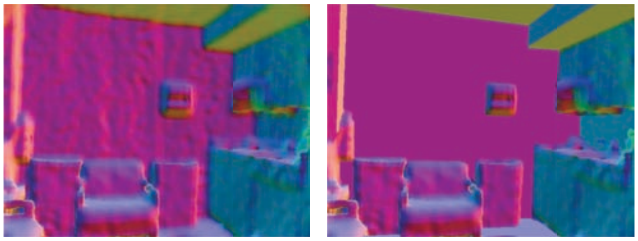


Figure 9. Surface normal ground truth before and after plane fitting for wall, ceiling, and floor.

4. Object Boundary Detection Network

We adopt network proposed in Xie *et al*. [4]. The network is a trimmed VGG-16, where only first 5 convolution layers are used. An intermediate output layer is added to each convolution stage before pooling, which results in 5 intermediate outputs with stride 1, 2, 4, 8, and 16 respectively. Their final output is the fusion of these 5 intermediate outputs.

We use their code, to replicate their results. The VGG-16 layers are initialized with the pretrained model on ImageNet. In their original setting for training on BSDS500 [1], the learning rate is initially set as $1 \quad 10^{\ 6}$ and reduced to 10% after each $10K$ iterations. The momentum is 0 9, and the weight decay is $2 \quad 10^{\ 4}$.

However, this training prescription does not apply to NYUv2. The loss goes out of range and training fails, because the NYUv2 provides larger images with more pixels and the loss accumulates significantly more error from all pixels. To deal with this problem, when training on NYUv2, we reduce the initial learning rate to $2 \quad 10^{\ 7}$. Empirically, this learning rate keeps the total loss in range, and is large enough to finetune the model.

Ousrs 84 class	NYUv2	Ousrs	NYUv2
ac	otherprop	kitchenware	otherprop
arch	door	mailbox	otherprop
armchair	chair	mirror	mirror
baby_bed	bed	music	otherprop
bar	otherfurniture	office_chairs	chair
bathroom_stuff	otherprop	ottoman	otherprop
bathtub	bathtub	outdoor_lamp	lamp
bench_chair	chair	outdoor_rest	chair
bookshelf	bookshelf	outdoor_spring	otherprop
bunker_bed	bed	paintings	picture
candel	lamp	partitions	otherstructure
car	otherprop	people	people
chair	chair	pets	otherprop
chandelier	lamp	pillow	otherprop
clock	otherprop	plants	otherprop
closets _wardrobes_cabinets	cabinet	pool	otherprop
cloth	clothes	recreation	otherprop
coffee_table	table	rug	floormat
column	wall	safe	otherprop
computer	television	shelves	shelves
curtain	curtain	shoes	otherprop
desk	desk	shoes_cabinet	cabinet
dinning_table	table	shower	shower curtain
door	door	single_bed	bed
double_bed	bed	sofa	sofa
dresser	dresser	stair	otherstructure
dressing_table	table	stand	night stand
fan	otherprop	switch	otherprop
fences_gate	otherprop	table_and_chair	table
figurines	otherprop	table_lamp	lamp
fireplaces	otherstructure	toilet	toilet
floor_lamps	lamp	toys	otherprop
fridges	refridgerator	trash_can	otherfurniture
gym	otherprop	tripole	otherprop
hangers	otherprop	tv_bench	cabinet
hanging _kitchen_cabinet	cabinet	tvs	television
heater	otherprop	vases	otherprop
household_appliance	otherprop	wall_lamp	lamp
idk	otherprop	wash_basins	sink
kitchen_appliance	otherprop	whitebroad	whiteboard
kitchen_cabinet	cabinet	windows	window
kitchen_set	otherprop	workplace	desk

Table 3. Class mapping from our synthetic dataset 84 category to NYUv2 40 category.

Model	GT	Mean (°) ↓	Median (°) ↓	11.25 ↑	22.5 ↑	30 ↑
NYUv2	Ori	24.95	18.27	32.81	57.98	69.15
	Fit	26.90	18.94	34.02	55.81	65.95
MLT	Ori	26.76	19.28	31.33	55.75	66.33
	Fit	26.87	17.47	36.76	57.80	66.54
MLT +NYUv2	Ori	19.54	12.15	47.01	71.75	79.74
	Fit	21.85	11.56	49.19	68.10	75.12

Table 4. Comparison of performance on background region that is either wall, floor, and ceiling. "Ori" represents the original ground truth. "Fit" is the ground truth with plane fitting.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):898–916, May 2011. 7

[2] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015. 7

[3] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012. 7

[4] S. Xie and Z. Tu. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1395–1403, 2015. 7