OpenRooms: An Open Framework for Photorealistic Indoor Scene Datasets

Zhengqin Li[1] Ting-Wei Yu[1] Shen Sang[1] Sarah Wang[1] Meng Song[1] Yuhan Liu[1] Yu-Ying Yeh[1]
Rui Zhu[1] Nitesh Gundavarapu[1] Jia Shi[1] Sai Bi[1] Hong-Xing Yu[1] Zexiang Xu[2]
Kalyan Sunkavalli[2] Miloš Hašan[2] Ravi Ramamoorthi[1] Manmohan Chandraker[1]

[1]UC San Diego [2]Adobe Research

Abstract

We propose a novel framework for creating large-scale photorealistic datasets of indoor scenes, with ground truth geometry, material, lighting and semantics. Our goal is to make the dataset creation process widely accessible, transforming scans into photorealistic datasets with high-quality ground truth for appearance, layout, semantic labels, high quality spatially-varying BRDF and complex lighting, including direct, indirect and visibility components. This enables important applications in inverse rendering, scene understanding and robotics. We show that deep networks trained on the proposed dataset achieve competitive performance for shape, material and lighting estimation on real images, enabling photorealistic augmented reality applications, such as object insertion and material editing. We also show our semantic labels may be used for segmentation and multi-task learning. Finally, we demonstrate that our framework may also be integrated with physics engines, to create virtual robotics environments with unique ground truth such as friction coefficients and correspondence to real scenes. The dataset and all the tools to create such datasets will be made publicly available.[1]

1. Introduction

Indoor scenes represent important environments for visual perception and scene understanding, for applications such as augmented reality and robotics. However, their appearance is a complex function of multiple factors such as shape, material and lighting, and demonstrates phenomena like significant occlusions, shadows, interreflections and large spatial variations in lighting. Reasoning about these underlying, entangled factors requires large-scale high-quality ground truth, which remains hard to acquire. While ground truth geometry can be captured using a 3D scanner, it is extremely challenging (if not nearly impossible) to accurately acquire the complex spatially-varying material and lighting of indoor scenes. An alternative is to consider synthetic datasets, but large-scale synthetic datasets of indoor scenes with plausible geometry, materials and lighting are also non-trivial to create.

This paper presents OpenRooms, a framework for synthesizing photorealistic indoor scenes, with broad applicability across computer vision, graphics and robotics. It has several advantages over prior works, summarized in Table 1. First, rather than use artist-created scenes and assets, we ascribe high-quality material and lighting to RGBD scans of real indoor scenes. Beyond just the data, we provide all the tools necessary to accomplish this, allowing any researcher to inexpensively create such datasets. While prior works can align CAD models to scanned point clouds [5, 26, 6], they do not explore how to assign materials and lighting appropriately to build a large-scale photorealistic dataset. Second, we provide extensive high-quality ground truth for complex light transport that is unmatched in prior works. Our material is represented by a spatially-varying microfacet bidirectional reflectance distribution function (SVBRDF), and our lighting includes windows, environment maps and area lights, along with their per-pixel spatially-varying effects to account for visibility, shadows and inter-reflections. Third, we render photorealistic images with our data and tools, which include a custom GPU-accelerated physically-based renderer.

We create an instance of such a dataset by building on existing repositories: 3D scans from ScanNet [16], CAD model alignment [5], reflectance [1] and illumination [23, 24]. The resulting dataset contains over 100K HDR images with ground-truth depths, normals, spatially-varying BRDF and light sources, along with per-pixel spatially-varying lighting and visibility masks for every light source. We also provide per-pixel semantic labels. Besides being publicly available, the dataset can be significantly extended through future community efforts based on our tools. We also demonstrate applicability of our method to other choices for material [4] and geometry [48].

We believe that our effort will significantly accelerate research in multiple areas. Inverse rendering tasks are directly related, including single-view [17] and multi-view [55] depth prediction, intrinsic decomposition [33, 11], material classification [10] and lighting estimation [20, 21, 32].

[1]Webpage: https://ucsd-openrooms.github.io/

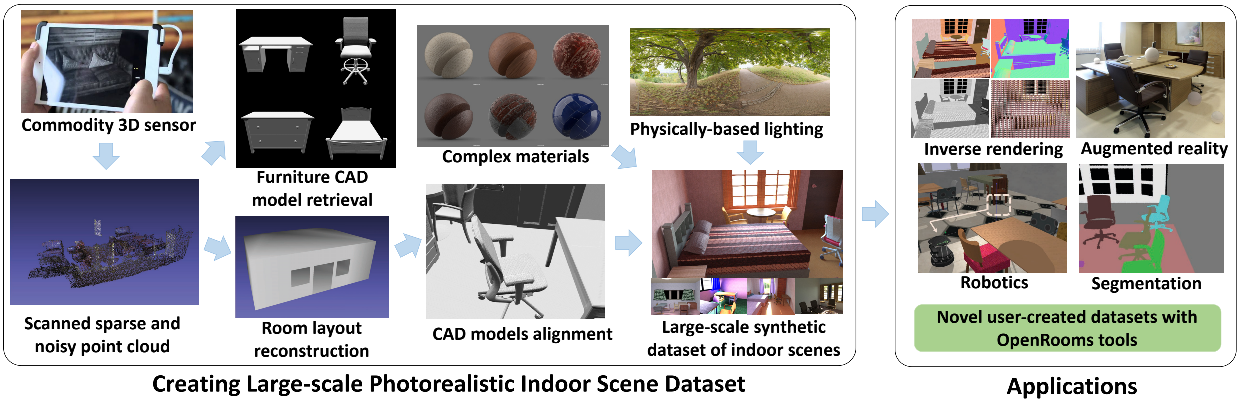

Figure 1: Our framework for creating a synthetic dataset of complex indoor scenes with ground truth shape, SVBRDF and SV-lighting, along with the resulting applications. Given possibly noisy scans acquired with a commodity 3D sensor, we generate consistent layouts for room and furniture. We ascribe per-pixel ground truth for material in the form of high-quality SVBRDF and for lighting as spatially-varying physically-based representations. We render a large-scale dataset of images associated with this ground truth, which can be used to train deep networks for inverse rendering and semantic segmentation. We further motivate applications for augmented reality and robotics, while suggesting that the open source tools we make available can be used by the community to create other large-scale datasets too.

To demonstrate the efficacy of the dataset, we train a state-of-the-art inverse rendering network and achieve accurate results on real images. We also demonstrate that OpenRooms may be used for training semantic segmentation networks [60, 15], as well as multi-task learning to jointly estimate shape, material and semantics. Our high-quality and extensive ground truth may help better understand complex light transport in indoor scenes and enable new applications in photorealistic augmented reality, where we demonstrate object insertion, material editing light source detection as examples, and may include light editing in the future.

Studies in robotics may also benefit by using our ground truth to enhance existing simulation environments [52, 43, 53, 37]. We demonstrate this possibility by combining OpenRooms assets with the PyBullet engine [3] and mapping our SVBRDFs to friction coefficients, to motivate navigation and rearrangement under different material and lighting. We also note that OpenRooms allows a one-to-one correspondence between real videos and simulations, which can be valuable for sim-to-real transfer [27].

In Figure 1 we illustrate the OpenRooms framework for creating large-scale, high-quality synthetic indoor datasets from commodity RGBD sensor scans and demonstrate some of the applications that our work enables.

2. Related Work

Indoor scene datasets. The importance of indoor scene reconstruction and understanding has led to a number of real datasets [46, 16, 13, 53, 50]. While they are by nature photorealistic, they only capture some scene information (usually images, geometry and semantic labels). However, we are interested in studying geometry, reflectance and illumination, where the latter two are particularly challenging to acquire in real datasets. Synthetic datasets provide an alternative [38, 49, 31], but prior ones are limited with respect to rendering arbitrary data [31], scene layout [38], material [49], or baselines [42], as summarized in Table 1.

Several methods build 3D models for indoor scenes from a single image [26] or scans [5, 6, 12, 14]. However, our focus is beyond geometry, to assign real-world materials and lighting to create photorealistic scenes. To the best of our knowledge, the only existing dataset with complex materials and spatially-varying lighting annotations is from Li et al. [32], but is built on artist-created assets that are not publicly available [49]. We instead create photorealistic indoor scene datasets that start with 3D scans to provide high-quality ground truth for geometry, reflectance and lighting.

Several indoor virtual environments have also been proposed for robotics and embodied vision [52, 43, 53, 37, 30]. Our work is complementary, where our photorealistic ground truth and suite of tools could be used to enhance existing virtual environments and conduct new types of studies. In Sec. 4.3, we seek to motivate such adoption by illustrating integration with a physics engine and computing ground truth for friction coefficients.

Inverse rendering for indoor scenes. Indoor scene inverse rendering seeks to reconstruct geometry, reflectance and lighting from (in our case, monocular) RGB images. Estimating geometry, in the form of scene depth or surface normals, has been widely studied [17, 7, 55, 36]. Most scene material estimation methods either recognize material classes [10] or only reconstruct diffuse albedo [33, 8, 29]. Scaling these methods to real-world images requires scene datasets with complex physically-based materials. Li et al. [32] augment a proprietary dataset [49] with ground-truth SVBRDF annotations to train a physically-motivated network. We demonstrate comparable inverse rendering performance using their network, but trained on Open-

Dataset	Available annotations						Publicly available assets				Corresponding real images and scenes
	Geometry	Material	Lighting			Segmentation	Images	CAD	Baseline	Tool	
			Light sources	Per-pixel	Visibility						
PBRS [59]	✓	diffuse	✗	shading	✗	✓	✗	✗	✓	✓	✗
Scenenet [38]	✓	✗	✗	✗	✗	✓	✓	✓	✗	✓	✗
CGIntrinsic [33]	✗	diffuse	✗	shading	✗	✗	✓	✗	✓	✓	✗
InteriorNet [31]	✓	diffuse	✗	shading	✗	✓	✓	✗	✓	✓	✗
CG-PBR [44]	✓	phong	✗	shading	✗	✗	✗	✗	✗	✓	✗
InvIndoor [32]	✓	microfacet	✗	envmap	✗	✗	✗	✗	✓	✓	✗
3D-Future [18]	✓	✗	✗	✗	✗	✓	✓	✓	✗	✗	✗
AI2-THOR [30]	✓	✓	✓	✗	✗	✓	✓	✓	✗	✗	✓
Structure3D [61]	✓	✗	✗	shading	✗	✓	✓	✗	✗	✓	✗
Hypersim [42]	✓	diffuse	✗	highlight	✗	✓	✓	✗	✗	✓	✗
OpenRooms	✓	microfacet	✓	envmap	✓	✓	✓	✓	✓	✓	✓

Table 1: OpenRooms is distinct in providing extensive ground truth for photorealism (especially material and lighting), with publicly available assets and tools. The tools in OpenRooms framework allow generating synthetic counterparts of real scenes, with high-quality ground truth.

Rooms, developed using publicly available assets.

Previous indoor scene lighting estimation methods only predict shading (which entangles geometry and lighting) [33], require RGBD inputs [8], or rely on hand-crafted heuristics [28, 29]. More recently, deep network-based lighting estimation methods have shown great progress for estimating both global [20, 19] and spatially-varying lighting [21, 47, 32] from single RGB images. The latter set of methods largely rely on proprietary synthetic data to generate spatially-varying lighting annotations; we demonstrate comparable performance by training on our dataset.

3. Building a Photorealistic Indoor Dataset

We now describe our framework for building a synthetic dataset of complex indoor scenes. We demonstrate this using ScanNet, a large-scale repository of real indoor scans [16], but our work is also applicable to other datasets [48, 25], as shown in the supplementary. We briefly describe the geometry creation, while focusing on our principal novelties of photorealistic material and lighting.

3.1. Creating CAD Models from 3D Scans

While recent methods such as [6] are possible alternatives, we demonstrate our dataset creation example utilizing existing labels in ScanNet and initial CAD alignment [5] to create the ground truth geometry robustly.

Reconstructing the room layout We fuse the depth maps from different views of a scene to obtain a single point cloud. We design a UI for fast layout annotation (Fig. 3), which projects the 3D point cloud to the floor plane and a polygon may be selected for the layout. While the annotation needs less than a minute per scene, we also train a Floor-SP network [14] on these annotations that users may employ for their own scenes (shown in the supplementary). Next we use RANSAC to determine the horizontal floor plane. Since ScanNet views generally do not cover the ceiling, we assign a constant room height of 3 meters.

Windows and doors Special consideration is needed for doors and windows as they are important illuminants in indoor scenes. We project the 3D points labeled as doors and windows to the closest wall, then divide the wall into segments and merge connected segments with sufficient number of points, to which a ShapeNet CAD model is assigned.

Consistent furniture placement We use initial poses from Scan2CAD [5] to align CAD models with furniture instances. We do not require appearances to closely match the input images, but generate plausible layouts and shapes with as much automation as possible. Our tool automatically moves bounding boxes for furniture perpendicular to the floors and walls to resolve floating objects and intersections. Such geometric consistency is important since our dataset may also be used for tasks such as navigation.

Semantic labels Given our geometry ground truth, it is straightforward to obtain labels for semantic and instance segmentation based on PartNet annotations, as shown in Fig. 4. We demonstate in experiments that our labels can be used to train single and multi-task deep networks.

3.2. Assigning Complex Materials to Indoor Scenes

One of the major contributions of our dataset is ground-truth annotation of complex material parameters for indoor scenes. Previous works typically provide material annotations as simple diffuse or Phong reflectance [49, 45], while we provide a physically-based microfacet SVBRDF.

Assigning materials to ShapeNet Many ShapeNet CAD models do not have texture coordinates, so we use Blender's [2] cube projection UV mapping to compute texture coordinates for them automatically. Inspired by Photoshape [41], we split CAD models into semantically meaningful parts and assign a material to each part. While Photoshape does this for only chairs, we do so for all furniture types in indoor scenes, using the semantically meaningful part segmentation of 24 categories of models provided by PartNet [40].

Material annotation UI We design a custom UI tool to annotate material category for each part, as shown in Fig. 3. It allows merging over-segmented parts which should be assigned the same material. To allow material annotation, we group 1,078 SVBRDFs into 9 categories based on their appearances, similar to [34, 32], as shown in Fig. 3. Annotators label a material category for each part, with a specific material sampled randomly from the category. While we do not pursue mimicking input appearances, we do seek that photorealism and semantics be respected in the dataset. Experiments show that our dataset created following the above choices enables state-of-the-art inverse rendering performances. Note our distinction from domain randomization,

Figure 2: Images from ScanNet and our corresponding synthetic scene layouts rendered with different materials, different lighting, and different views selected by our algorithm. A video is included in the supplementary. The third row shows the same scene as the second one, but rendered with freely available Substance Share materials [4] instead of the public but non-free Adobe Stock materials [1].

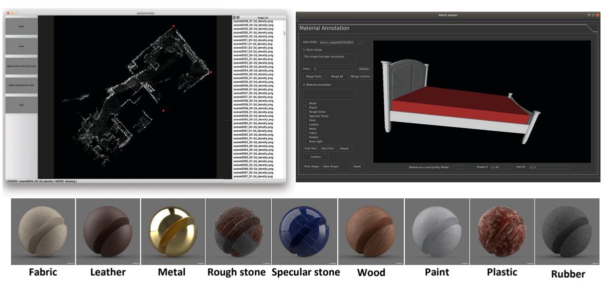

Figure 3: UIs for annotating room layout (Left top) and material category (Right top). (Bottom) Material examples from each category. Please zoom in for better visualization.

since arbitrary choices for material and lighting might not allow generalization on real scenes for extremely ill-posed problems like material and lighting estimation. Our tools and the annotations will be released for future research.

3.3. Ground Truth Lighting for Indoor Scenes

Lighting plays one of the most important roles in image formation. However, prior datasets usually only provide diffuse shading as their lighting representation [33, 59]. Recent work provides per-pixel environment maps by rendering the incoming radiance at every surface point in the camera frustum [32], which allows modeling shadows and specular highlights, but not the complex interactions among global light sources, scene geometry, materials and local lighting. On the contrary, OpenRooms provides extra supervision for visible and invisible light sources, the contribution of each individual light source to the local lighting, direct and indirect lighting, as well as visibility. Such rich supervision may help better understand the complex light transport in indoor scenes and enable new applications such as editing of light sources and dynamic scenes.

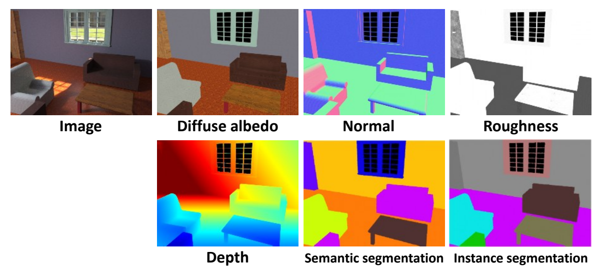

Figure 4: One of our rendered images with ground-truth geometry, spatially-varying material and segmentation labels.

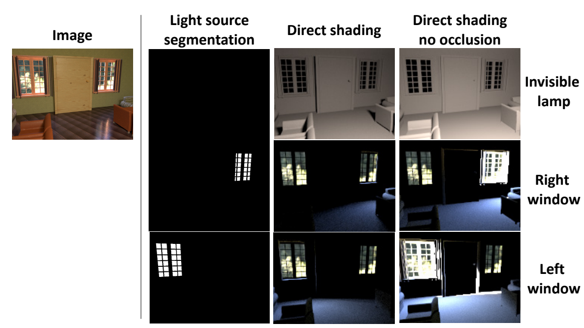

Figure 5: Our ground-truth light source annotations. From left to right: input and for each light source, its instance segmentation, and direct shading with and without occlusion. Our annotations reveal rich information about light transport in indoor scenes.

Light sources We model two types of light sources in OpenRooms—windows and lamps—and we provide ground-truth annotations for them. The annotations include instance segmentation masks for visible light sources and a consistent parameterized representation for both visible and invisible light sources. More specifically, for each window, we model its geometry using a rectangular plane and the lighting coming through the window using an environment map rendered at its center. We represent each lamp as a 3D bounding box following the standard area light model. We

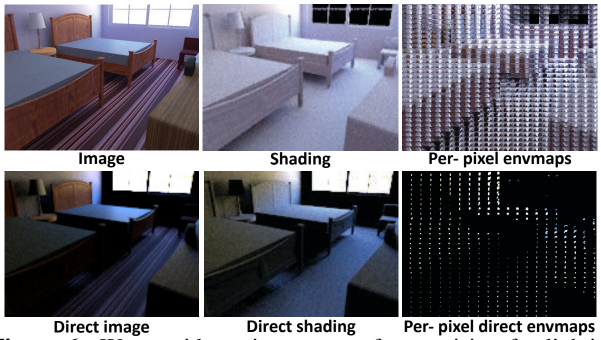

Figure 6: We provide various types of supervision for lighting analysis of indoor scenes, including per-pixel environment maps with only direct illumination, or including indirect illumination.

visualize our light source annotations in Figure 5. Our light source representation has clear physical meaning and can model the full physics of image formation in indoor scenes.

Light source colors For environment maps, we use 414 high-resolution HDR panoramas of natural outdoor scenes, from [24] and [23]. For indoor lamps, unlike previous synthetic datasets that randomly sample the spectrum of area lights [32, 59, 33], we follow a physically plausible black-body model to determine the spectrum of the light source by its temperature, chosen between 4000K to 8000K.

Per-pixel lighting Additionally, as in prior works [32, 33, 59], we render per-pixel environment maps and shading as a spatially-varying lighting representation. However, we render both with direct, as well as combined direct and indirect illumination. This will help to separately analyze the direct contribution from light sources and indirect reflections from the indoor scene. We visualize an example in Figure 6.

Per-light direct shading and visibility In order to understand complex light transport in indoor scenes, we also provide the separate contribution of every individual light source and its visibility map. For each image, we render the direct shading of each light source, with and without considering the occlusion term, by turning on only that particular light source. The visibility map can be computed as the ratio of the two direct shading images. We visualize these annotations in Figures 5 and 6. These will allow new challenging light editing tasks not possible with prior datasets, such as turning on and off a light or opening a window.

3.4. Rendering with a Physically-based Renderer

To minimize the domain gap between synthetic and real data, we modify the physically-based GPU-accelerated renderer from our prior work [32] to support ground-truth per-light contribution and fast rendering of per-pixel environment map. Our renderer models complex light transport up to 7 bounces of inter-reflection.

View selection ScanNet provides the camera pose of each RGBD image. However, their distribution is biased towards views close to the scene geometry, to optimize scanning.

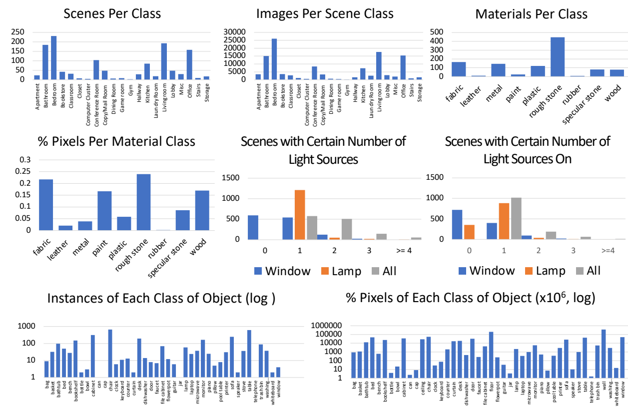

Figure 7: Dataset statistics for scene categories, images, materials, lighting and semantic labels (please zoom for viewing).

On the contrary, we prefer views covering larger regions, matching typical human viewing conditions. To achieve this, we first sample different views along the wall, facing the center of the room. For each view, we render its depth and normal maps. Let d_p and $\hat{\mathbf{n}}_p$ be the depth and normal of pixel p, $\mathbf{Grad}(\hat{\mathbf{n}}_p)$ be the sum of absolute gradients of the normal in the three channels. We choose the view based on computing a score defined as

$$\sum_{p \in \mathcal{P}} \mathbf{Grad}(\hat{\mathbf{n}}_p) + 0.3 \sum_{p \in \mathcal{P}} \log(d_p + 1). \tag{1}$$

Views with higher scores are used to create the dataset. An example of our view selection results is shown in Figure 2 (bottom right). Details are included in the supplementary.

Other renderers While our renderer will be publicly released, our assets (geometry, material maps, lights) are in a standard graphics format that could be used in other rendering environments. For example, common real-time rasterization engines like Unity or Unreal could be used for applications (such as robotics) which prefer real-time performance and do not require fully accurate global illumination. Furthermore, our per-pixel spatially-varying lighting maps could be used as high-quality precomputed lighting probes for photorealistic real-time rendering [39].

3.5. OpenRooms Dataset Statistics

Scene, image, semantic label distribution We pick 1,287 of the 1,506 ScanNet scenes to instantiate our dataset, discarding those which cover very small portions of rooms. We randomly choose 1,178 scenes for training and 109 scenes for validation. For each scene, we choose views using our view selection method. For each rendered image, we render two others with different materials and lighting, as shown in Fig. 2 (bottom-left). We render 118,233 HDR images at 480×640 resolution, with 108,159 in the training set and 10,074 in the validation set. We render semantic labels of all 44 classes of CAD models in OpenRooms.

The distributions of scene categories and images, number of objects per class and the percentage of pixels per class are summarized in Figure 7. Note that the class distribution follows that of real scans in ScanNet indoor scenes.

Material distribution We use 1,075 SVBRDFs from [1] to build OpenRooms, corresponding to the 9 categories shown in Fig. 3. The number of materials per-category and their pixel distributions are summarized in Fig. 7.

Lighting distribution Figure 7 shows the distribution of the two types of light sources (windows and lamps). Each image has at least one light source "on" for rendering. For all the 118K images, we render spatially-varying environment maps and shading, with direct illumination only and with combined direct and indirect illumination. Moreover, we provide a parameterized representation for every visible and invisible light source, as well as render their individual direct shading contribution and visibility map. Compared to all prior works, OpenRooms provides significantly more extensive and detailed supervision for complex lighting, which may allow new applications such as light source detection and editing.

Asset cost Almost all the assets used for creating our dataset are publicly available and free for research use. The only non-free (but also publicly available) assets are the raw material maps from Adobe Stock [1] that cost less than US$500, while the material parameters annotated with our scenes are freely available. Note that photorealistic appearances may also be achieved using our tools with freely available materials, such as Substance Share [4] in Fig. 2.

Dataset creation time It takes 30s to annotate one scene layout and 1 minute to label materials for one object, leading to 64 hours for labeling the whole dataset, which was accomplished by students with knowledge of computer vision. Almost all rendering time is spent to render images and spatially-varying per-pixel environment maps, which takes 600s and 100s per image, respectively, for our custom renderer on a single 2080Ti GPU. In principle, we can render the dataset in 1 month using 40 GPUs.

4. Applications

4.1. Inverse Rendering

We verify the effectiveness for inverse rendering by testing networks trained on our dataset on various benchmarks, where both quantitative and qualitative results show good generalization to real images. We use a state-of-the-art network architecture for inverse rendering in indoor scenes that handles spatially-varying material and lighting [32]. Please refer to the supplementary material for more details.

Intrinsic decomposition. We compare our intrinsic decomposition results with 3 previous approaches. The qualitative comparison is shown in Fig. 8 while quantitative results are in Table 2, which are comparable to prior state-of-the-art based on artist-created SUNCG dataset [49].

Figure 8: Comparisons with previous state-of-the-art on intrinsic decomposition (albedo prediction shown).

	Training	WHDR↓
Ours	Ours + IIW	16.4
Li18[33]	CGI + IIW	17.5
Sen.19[44]	CGP + IIW	16.7
Li20[32]	CGM + IIW	15.9

Table 2: Intrinsic decomposition on IIW [9].

Method	Mean(°)↓	Med.(°)↓	Depth↓
Ours	25.3	18.0	0.171
Li20[32]	24.1	17.3	0.184
Sen.19[44]	21.1	16.9	–
Zhang17[59]	21.7	14.8	–

Table 3: Normal and depth predictions on NYU dataset [46].

Test on	OpenRoom		NYU2			
Train on OR/NYU2	Yes/ No		No/ Yes		Yes/ Yes	
	bbox	seg	bbox	seg	bbox	seg
AP(0.5:0.95)	80.2	70.1	17.1	15.3	23.5	21.6
AP-windows	85.8	63.2	11.9	12.7	20.5	20.6
AP-lamp	74.7	76.9	22.2	18.0	26.6	22.7

Table 4: Bounding box regression and mask AP on OpenRooms and NYU2 [46] for light source (windows and lamps) detection.

Depth and normal estimation. We evaluate the normal and depth estimation on the NYU dataset. The quantitative evaluation is in Table 3. We perform slightly worse than Li et al.'s dataset, possibly because their SUNCG-based dataset has more diverse and complex geometry compared to our ShapeNet-based furnitures.

Light source detection We use a ResNeXt101 [54] and FPN [35] pretrained model from Detectron2 [51] to train an instance segmentation network for light source detection (windows and lamps). We evaluate on OpenRooms and NYUv2 [46]. As shown in Tab. 4 and Fig. 10, training on OpenRooms boosts accuracy on NYUv2 testing by around 5%, for both bounding box regression and segmentation.

Per-pixel lighting estimation The above network also predicts per-pixel spatially-varying lighting, with qualitative results shown in Fig. 9 and quantitative results in supplementary. Note that we also provide ground truth for per-pixel direct lighting, shading and visibility, which are not predicted by our network but may be useful for studies in light transport, editing and augmented reality.

Semantic segmentation We use DeepLabV3 [15] and PSPNet(50) [60] to pre-train semantic segmentation models

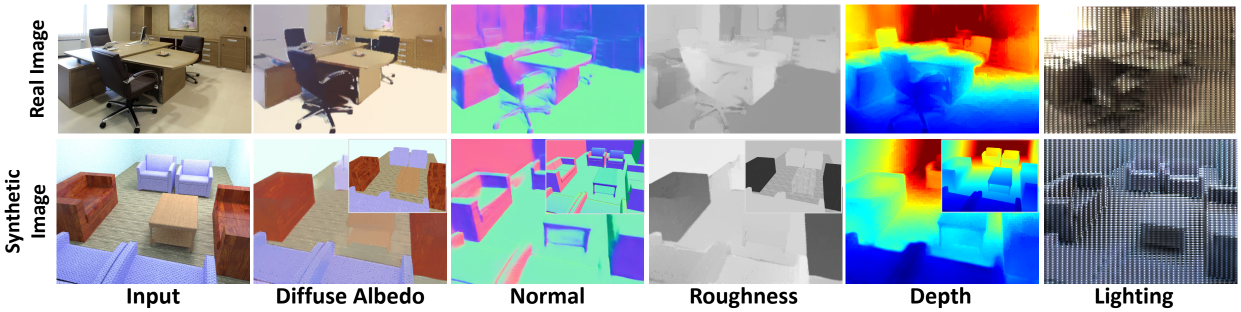

Figure 9: Inverse rendering results on a real example and a synthetic example. The insets in the bottom row are the ground truth.

Figure 10: Light source detection on OpenRooms (OR) and NYUv2 [46]. Windows are better detected with OR training.

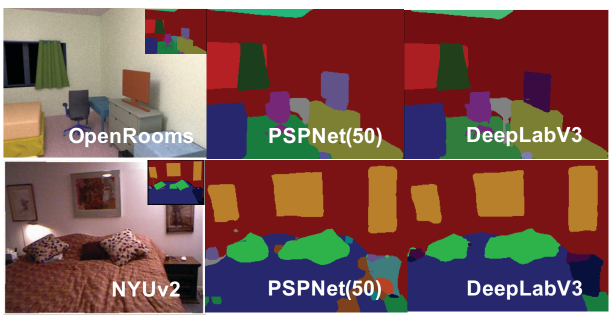

Figure 11: Semantic segmentation on OpenRooms and NYUv2 [46] using PSPNet(50) [60] and DeepLabV3 [15].

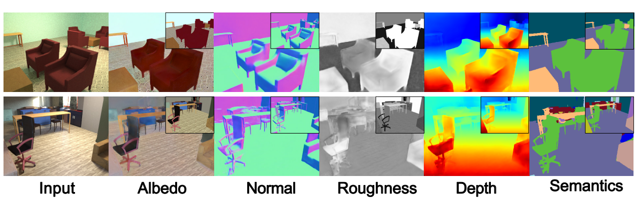

Figure 12: Multi-task estimation on OpenRooms.

	PSPNet(50) [60]				DeepLabV3 [15]			
	mIoU		mAcc		mIoU		mAcc	
	10K	50K	10K	50K	10K	50K	10K	50K
IN	41.1	41.2	53.3	53.4	41.7	42.2	53.6	54.4
OR	40.8	41.1	53.0	52.5	42.5	42.9	54.5	55.1

Table 5: Semantic segmentation trained on OpenRoom (OR) and InteriorNet (IN) [31] and fine-tuned on NYUv2 [46] with PSP-Net(50) and DeepLabV3, using different number of images.

on OpenRooms, then finetune and evaluate on NYUv2 [46] with 40 labels [22]. We also compare the results pre-trained on InteriorNet [31] with the same number of training images. As shown in Tab. 5 and Fig. 11, results are comparable for the two models and register improvements with greater number of images for the two pre-training datasets.

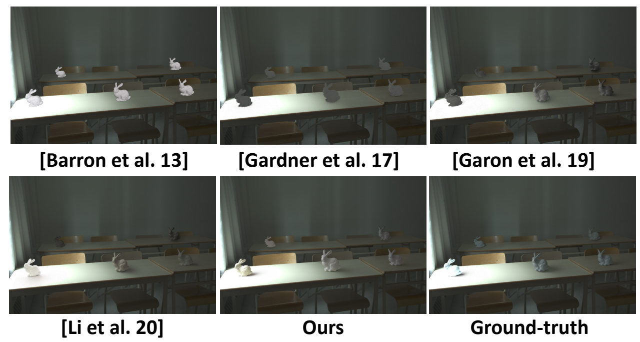

Figure 13: Object insertion on a real benchmark dataset [21]. Our dataset leads to photorealistic insertion results comparable to state-of-the-art [32][21]. Please zoom in for more details.

Figure 14: Material editing in real images. Note that the network trained on our dataset handles specular effects and spatially-varying lighting well.

Multi-task estimation An advantage of OpenRooms is the ground truth available for a range of both inverse rendering and semantic tasks. This may be useful for learning correlations among different vision tasks, and therefore can be of great interest to researchers in multi-task and transfer learning. As an illustration, we add a simple segmentation head to the inverse rendering network described above. Qualitative results are shown in Figure 12. Quantitative results are shown in the supplemenatry. We hope such data will motivate and be useful for studies in multi-task learning, such as [56, 57].

4.2. Applications to Augmented Reality

Object insertion Photorealistic virtual object insertion in augmented reality requires high-quality estimation of geometry, material and lighting. We test our inverse network

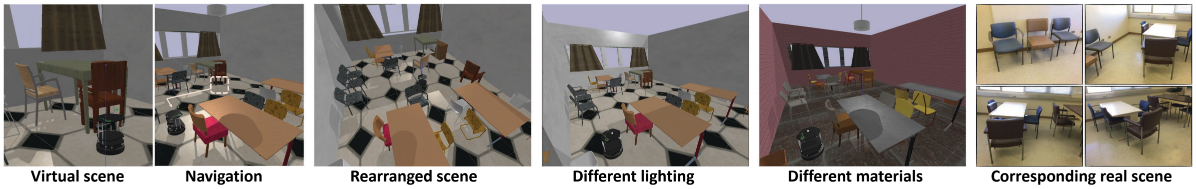

Figure 15: OpenRooms is integrated with a physics engine to create virtual scenes for robotics, potentially enabling studies for navigation and rearrangement across varying lighting and material, with possible correspondence to real scenes.

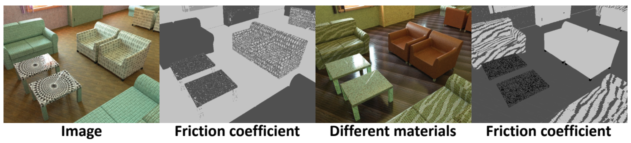

Figure 16: Ground-truth friction coefficients for the same scene with different materials. Specular materials tend to have lower coefficients of friction (darker).

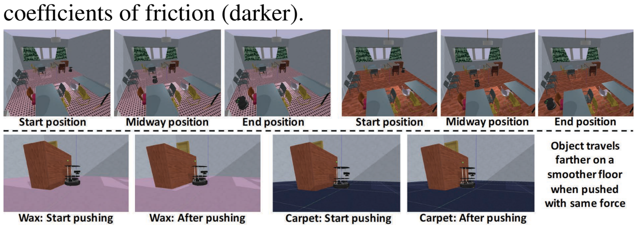

Figure 17: OpenRooms enables novel studies in navigation and rearrangement with material and lighting variations.

	Barron13 [8]	Gardner17 [20]	Garon19 [21]	Li20 [32]
Ours vs.	88.19%	66.16%	56.53%	54.77%

Table 6: User study on object insertion indicating the % of pairwise comparisons where human annotators thought we outperformed an alternative method; we outperform all prior methods. More details and comparisons are in supplementary.

on the dataset from [21], which contains around 80 ground-truth spatially-varying light probes. As shown in Fig. 13, our network outperforms those methods that cannot handle spatially-varying or high-frequency lighting well. It even generates more consistent lighting compared to [32] which is trained on a SUNCG-based dataset, probably because our dataset has more diverse outdoor lighting and handles indoor lighting in a physically meaningful way. The quantitative user study in Table 6 also suggests that a network trained on our dataset performs better on object insertion.

Material editing We illustrate replacement of the material of a planar surface in Fig. 14 using the method of [32]. We note that spatially-varying lighting effects and specularity are handled quite well, with results comparable to [32], even though our dataset is created from noisy scans acquired with a commodity sensor.

4.3. Applications to Robotics and Embodied Vision

To facilitate research in robotics and embodied AI, OpenRooms supports transforming a rich 3D indoor scene model into an interactive environment, with realistic physical simulation through PyBullet [3]. A URDF file describe physical properties, such as mass and friction coefficients, for CAD models. This feature of OpenRooms establishes direct connections between appearance and physical properties of the environment, to provide a learning testbed for a range of topics including physics understanding from perception and policy generalization across environment and configuration changes. As an example, Fig. 15 shows a classroom scenario where a robot is inserted into the scene and may perform a navigation task. Furniture in the scene can be rearranged, while the lighting and material properties can also be changed. In Fig. 17, we show navigation and rearrangement where different frictions of coefficient for the same scene lead to different pushing outcomes (see supplementary for details). Since we create the scene from scans, correspondence is available to real scenes, which may be useful for sim-to-real transfer studies [27].

Ground truth for friction coefficients We use our albedo and roughness ground truth to render reflectance disks through a virtual equivalent of the acquisition in [58], then do a nearest neighbor search to compute the friction coefficients. Examples of per-pixel friction coefficients are in Fig. 16, where specular materials have lower friction coefficients. More details are included in the supplementary.

5. Conclusion and Future Work

We have proposed methods that enable user-generated photorealistic datasets for complex indoor scenes, starting from existing public repositories of 3D scans, shapes and materials. We illustrate the process on over 1000 indoor scenes from ScanNet. In contrast to prior works, we provide high-quality ground truth for complex materials and spatially-varying lighting, including direct and indirect illumination, light sources, per-pixel environment maps and visibility. We demonstrate that inverse rendering and segmentation networks can be trained on OpenRooms, towards augmented reality applications like object insertion and material editing. We also show our dataset can be integrated with physics engines and provide friction coefficients, which suggest interesting future studies in navigation, rearrangement and sim-to-real transfer. Our dataset and all tools used for its creation will be publicly released.

Please refer to the **supplementary material** for further details, extensive experimental results and videos.

Acknowledgments: We thank NSF CAREER 1751365, a Google Award, generous gifts from Adobe, NSF CHASE-CI, ONR N000142012529, N000141912293, NSF 1703957, the Ronald L. Graham Chair and UCSD Center for Visual Computing.

References

[1] Adobe Stock. https://stock.adobe.com/3d-assets.

[2] Blender. http://www.blender.org.

[3] Bullet Real-Time Physics Simulation. https://pybullet.org.

[4] Substance Share. https://share.substance3d.com/.

[5] Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X Chang, and Matthias Nießner. Scan2CAD: Learning CAD model alignment in RGB-D scans. In *Proc. CVPR*, 2019.

[6] Armen Avetisyan, Tatiana Khanova, Christopher Choy, Denver Dash, Angela Dai, and Matthias Nießner. SceneCAD: Predicting object alignments and layouts in RGB-D scans. In *Proc. ECCV*, 2020.

[7] Aayush Bansal, Bryan Russell, and Abhinav Gupta. Marr revisited: 2D-3D model alignment via surface normal prediction. In *Proc. CVPR*, 2016.

[8] Jonathan T Barron and Jitendra Malik. Intrinsic scene properties from a single RGB-D image. In *Proc. CVPR*, 2013.

[9] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Trans. Graphics*, 33(4):159, 2014.

[10] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proc. CVPR*, 2015.

[11] Sai Bi, Nima Khademi Kalantari, and Ravi Ramamoorthi. Deep hybrid real and synthetic training for intrinsic decomposition. In *Proc. EGSR*, 2018.

[12] Ricardo Cabral and Yasutaka Furukawa. Piecewise planar and compact floorplan reconstruction from images. In *Proc. CVPR*, 2014.

[13] 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. 2017.

[14] Jiacheng Chen, Chen Liu, Jiaye Wu, and Yasutaka Furukawa. Floor-SP: Inverse CAD for floorplans by sequential room-wise shortest path. In *Proc. ICCV*, 2019.

[15] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[16] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. CVPR*, 2017.

[17] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proc. ICCV*, 2015.

[18] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *arXiv preprint arXiv:2009.09633*, 2020.

[19] Marc-André Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-François Lalonde. Deep parametric indoor lighting estimation. In *Proc. ICCV*, 2019.

[20] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Trans. Graphics*, 9(4), 2017.

[21] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-François Lalonde. Fast spatially-varying indoor lighting estimation. In *Proc. CVPR*, 2019.

[22] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *Proc. CVPR*, 2013.

[23] HDRI HAVEN. https://hdrihaven.com/.

[24] Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. Deep sky modeling for single image outdoor lighting estimation. In *Proc. CVPR*, 2019.

[25] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. SceneNN: A scene meshes dataset with annotations. In *Proc. 3DV*, 2016.

[26] Hamid Izadinia, Qi Shan, and Steven M Seitz. Im2cad. In *Proc. CVPR*, 2017.

[27] Abhishek Kadian, Joanne Truong, Aaron Gokaslan, Alexander Clegg, Erik Wijmans, Stefan Lee, Manolis Savva, Sonia Chernova, and Dhruv Batra. Sim2real predictivity: Does evaluation in simulation predict real-world performance? *IEEE Robotics and Automation Letters (RA-L)*, 2020.

[28] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. *ACM Trans. Graphics*, 30(6):1, 2011.

[29] Kevin Karsch, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, Hailin Jin, Rafael Fonte, Michael Sittig, and David Forsyth. Automatic scene inference for 3d object compositing. *ACM Trans. Graphics*, pages 32:1–32:15, 2014.

[30] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An interactive 3d environment for visual AI. *arXiv preprint arXiv:1712.05474*, 2017.

[31] Wenbin Li, Sajad Saeedi, John McCormac, Ronald Clark, Dimos Tzoumanikas, Qing Ye, Yuzhong Huang, Rui Tang, and Stefan Leutenegger. Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. In *Proc. BMVC*, 2018.

[32] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and SVBRDF from a single image. In *Proc. CVPR*, 2020.

[33] Zhengqi Li and Noah Snavely. CGIntrinsics: Better intrinsic image decomposition through physically-based rendering. In *Proc. ECCV*, 2018.

[34] Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. Materials for masses: SVBRDF acquisition with a single mobile phone image. In *Proc. ECCV*, 2018.

[35] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proc. CVPR*, 2017.

[36] Chen Liu, Jimei Yang, Duygu Ceylan, Ersin Yumer, and Yasutaka Furukawa. Planenet: Piece-wise planar reconstruction from a single rgb image. In *Proc. CVPR*, 2018.

[37] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proc. ICCV*, 2019.

[38] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J.Davison. SceneNet RGB-D: Can 5M synthetic images beat generic Imagenet pre-training on indoor segmentation? In *Proc. ICCV*, 2017.

[39] Morgan McGuire, Michael Mara, Derek Nowrouzezahrai, and David Luebke. Real-time global illumination using precomputed light field probes. In *Proc. ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, 2017.

[40] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *Proc. CVPR*, June 2019.

[41] Keunhong Park, Konstantinos Rematas, Ali Farhadi, and Steven M Seitz. Photoshape: photorealistic materials for large-scale shape collections. *ACM Trans. Graphics*, 37(6):192, 2019.

[42] Mike Roberts and Nathan Paczan. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. *arXiv preprint arXiv:2011.02523*, 2020.

[43] Manolis Savva, Angel X. Chang, Alexey Dosovitskiy, Thomas Funkhouser, and Vladlen Koltun. MINOS: Multimodal indoor simulator for navigation in complex environments. *arXiv:1712.03931*, 2017.

[44] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *Proc. ICCV*, 2019.

[45] Jian Shi, Yue Dong, Hao Su, and X Yu Stella. Learning non-Lambertian object intrinsics across ShapeNet categories. In *Proc. CVPR*, 2017.

[46] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proc. ECCV*, 2012.

[47] Shuran Song and Thomas Funkhouser. Neural illumination: Lighting prediction for indoor environments. In *Proc. CVPR*, pages 6918–6926, June 2019.

[48] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proc. CVPR*, 2015.

[49] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proc. ICCV*, 2017.

[50] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[51] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[52] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. Building generalizable agents with a realistic and rich 3d environment. *arXiv preprint arXiv:1801.02209*, 2018.

[53] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: real-world perception for embodied agents. In *Proc. CVPR*, 2018.

[54] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proc. CVPR*, 2017.

[55] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proc. ECCV*, 2018.

[56] A. R. Zamir, A. Sax, N. Cheerla, R. Suri, Z. Cao, J. Malik, and L. J. Guibas. Robust learning through cross-task consistency. In *Proc. CVPR*, 2020.

[57] A. R. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In *Proc. CVPR*, 2018.

[58] Hang Zhang, Kristin Dana, and Ko Nishino. Friction from reflectance: Deep reflectance codes for predicting physical surface properties from one-shot in-field reflectance. In *Proc. ECCV*, 2016.

[59] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *Proc. CVPR*, 2017.

[60] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proc. CVPR*, 2017.

[61] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. *Proc. ECCV*, 2020.