PhotoScene: Photorealistic Material and Lighting Transfer for Indoor Scenes Supplementary Material

Yu-Ying Yeh[1] Zhengqin Li[1] Yannick Hold-Geoffroy[2] Rui Zhu[1] Zexiang Xu[2]
Miloš Hašan[2] Kalyan Sunkavalli[2] Manmohan Chandraker[1]

[1]University of California, San Diego [2]Adobe Research

1. Ablation Study

We provide ablation study on our entire framework, by removing each component one at a time. We visualize the final results and compute RMSE on the optimized region for each material part with one scene from Photo-to-Manual dataset as shown in Figure 1 and with 18 randomly selected scenes over 171 materials from ScanNet-to-OpenRooms dataset in Table 1. We demonstrate the results a) without warping, b) with random graph selection, c) with material classifier graph selection, d) with stat loss only, e) with VGG loss only, f) without UV transformation parameters, and g) without material reoptimization, as well as our full framework and the baselines mentioned in the main paper.

Without warping, the mask and UV map cannot correctly fetch accurate material regions for optimization. This might lead to wrong material portions being considered due to misalignment so that the overall color and the pattern are not accurate. If we choose procedural graphs randomly from the entire collection or conditioned on a material super-class, the results do not have similar patterns as each procedural graph represents a distinct type of material (e.g. wood, homogeneous, ... etc.). With only statistics loss, the spatially-varying patterns become unconstrained and only match color statistics without considering spatial structures. The UV parameters cannot be estimated correctly and the statistics loss does not contain structure information. With only VGG loss, the results have similar spatial structures but are not guaranteed to have similar color to the reference photo without statistics loss. Without optimization of UV transformation, the orientation and scale of the textures are not guaranteed to be consistent to the reference photo. Note that even though our full method has slightly higher RMSE than (f), its qualitative superiority is not reflected in the metric since our optimization objectives are to align the pixel statistics and masked VGG features rather than per-pixel appearances. Without material re-optimization, sometimes the initial albedo colors have lighting baked-in, resulting in mismatched color under globally-consistent lighting. It is possible to get lower RMSE values with worse UV parameters. In sum, our full framework generates more similar appearances to the photo by considering all the components.

2. More Results

We demonstrate more results on ScanNet-to-OpenRooms, Photos-to-Manual, SUN-RGBD-to-Total3D material and lighting transfer results with novel view and relighting results in Figure 3, 4, 5, 6 and supplementary videos.

We also provide results with panoptic labels predicted by MaskFormer [1] instead of ground truth labels for ScanNet-to-Openrooms, and compute the results with baselines with randomly selected 62 scenes and over more than 521 materials, as shown in Table 2. The RMSE errors are slightly higher when using panoptic predictions, but still lower than baseline methods with panoptic ground truths. This demonstrate that our method is robust to imperfect input mask and outperform baseline methods regardless the input masks.

3. Additional Details for Proposed Method

3.1. Initialization and Alignment (Sec. 3.1)

Consensus-aware view selection. When a video sequence is available as input, we subsample views that are at least 30° or $1m$ apart, then choose the optimal view among them for optimizing each material part. We choose the best view based on three criteria – coverage, field-of-view and consensus. We expect good material transfer from those input images where a substantial number of pixels from the material part are observed. To ensure they occupy a favorable field-of-view, we weigh the number of pixels with a Gaussian, G, centered at the middle of the image and with variance one-fourth of the image dimensions. Finally, the goodness of a material part in a given view is also determined by the number of other views, n_i, where material estimates are in consensus, which is determined as the L2-norm of the mean and standard deviations of the per-pixel albedo and roughness predictions from InvRenderNet. We choose the view with the highest value of $n_i \cdot \sum(G \odot M_{photo})$ as the one to use for material

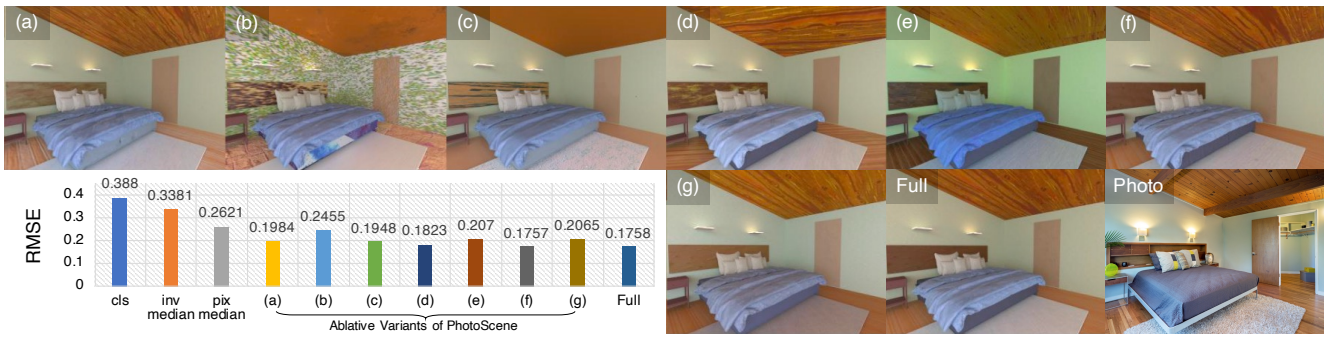

Figure 1. Ablation study on our entire framework with one selected scene from Photos-to-Manual. We compare the results by removing different modules from our full framework: a) without warping, b) with random graph selection, c) with material classifier graph selection, d) with stat loss only, e) with vgg loss only, f) without UV transformation parameters, and g) without material reoptimization, as well as our full framework and the baselines mentioned in the main paper.

	Baseline Methods			Ablative Variants of PhotoScene							PhotoScene
	Classifier	InvRend. Med.	Pixel Med.	(a)	(b)	(c)	(d)	(e)	(f)	(g)	Full
RMSE	0.448	0.381	0.314	0.250	0.255	0.251	0.249	0.326	0.243	0.272	0.244

Table 1. Similarity evaluation between rendering results and reference photo on 18 selected scenes of the ScanNet-to-OpenRooms dataset, for baseline methods, various ablations and the full version of the proposed PhotoScene approach.

	Classifier	InvRend. Med.	Pixel Med.	Ours
RMSE (GT Mask)	0.453	0.337	0.342	0.259
RMSE (Pred. Mask)	0.467	0.373	0.354	0.285

Table 2. Similarity evaluation between baselines rendering results and reference photo with ScanNet-to-OpenRooms dataset using ground truth panoptic labels versus predictions from MaskFormer [1].

transfer.

More details on material part mask and mapping. We regard material part segmentation as non-trivial, since material parts are ambiguous, e.g. table legs can be treated as separated parts or same part as the entire table. We found that the instance-based segmentation from MaskFormer already provides robust candidates which can later be refined by the mapping and alignment with geometry mask M_{geo}. Again, we can always provide better segmentation from manually labeling or existing dataset.

When MaskFormer does not detect a valid mask or 3D shapes have too small parts or highly different geometries from the image, we cannot find a large enough mask. We determine these failure situations by setting a threshold on the number of valid pixels inside a mask which can be used for optimization, and simply compute median values on the valid pixels, or on geometry mask M_{geo} if no valid pixels at all. To be specific, we first compute a per-pixel weight map W_{aln} by the dot product between aligned normal from InvRenderNet N^{inv} and normal from geometry N_{geo} and then define the valid pixels by computing the number of pixels with the above dot product larger than 0.95 as J and only run our optimization if $J \geq 500$, otherwise, we compute median for small masks where $J < 500$. If there is no mask candidate being matched by IoU, we simply use M_{geo} to compute median.

The weight map W_{aln} is also multiplied with M_{aln} to obtain a weighted mask when computing mask-based losses during optimization.

Alignment and warping. Let M_{geo} be the 2D material part mask rendered from geometry under corresponding views and M_{photo} be the mask for the reference photo. We first decompose M_{geo} and M_{photo} into sub-masks $\{M^i_{geo}\}$ and $\{M^j_{photo}\}$ which represents a single instance (if there are multiple instances), and search for matching instance pairs by the highest mIoU values on *soft* instance submasks. If semantic labels for both photo and geometry are available, we can use it to reduce the sub-masks by selecting corresponding semantics. Here *soft* means we apply a Gaussian filter on the instance submask with mean set as the center of mask bounding box and standard deviation set as half of width and height of bounding box, respectively.

After finding the matching pairs of part instances, we need to find the warping relationship between M_{geo} and M_{photo} so that we can warp UV_{geo} to $\widehat{UV}_{geo} \approx UV_{photo}$, which is used to sample material parameters from UV space to image space in our material part-based differentiable rendering module. We formulate the warping as scaling and translation from bounding box B_{geo} of M_{geo} to bounding box B_{photo}

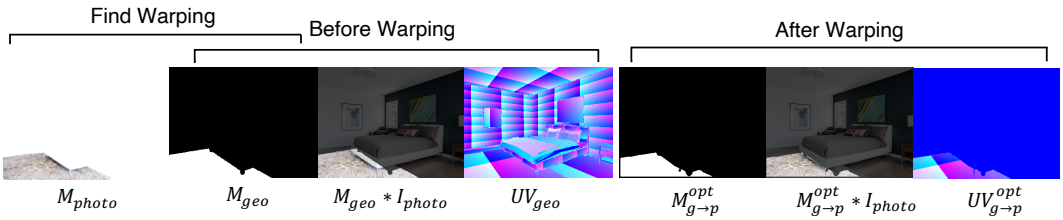

Figure 2. Example of part segmentation matching and UV warping between geometry and input image.

of M_{photo} to avoid unnecessary rotations. Let c_g and l_g be the center and size of B_{geo} and c_p and l_p be the center and size of B_{photo}. While c_g, c_p and l_g, l_p can be computed by minimum and maximum pixel locations in x and y directions of M_{geo} and M_{photo}, we can further find optimal c_g^* and l_g^* by optimizing intersection-over-union:

$$\max_{c_g, l_g} \frac{M_{photo} \cap \widehat{M}_{geo}}{M_{photo} \cup \widehat{M}_{geo}}, \tag{1}$$

to ensure higher percentage of overlap between $\widehat{M}_{geo}$ (the warped M_{geo}) and M_{photo}.

We warp the UV map $\widehat{UV}_{geo}$ by

$$x_s^* = (x_t - c_p)/l_p * l_g^* + c_g^*, \tag{2}$$

$$\widehat{UV}_{geo}(x_t) = UV_{geo}(x_s^*) \approx UV_{photo}(x_t). \tag{3}$$

Finally, we derive the warped material part mask $M^{opt}_{g\to p}$ and UV map $UV^{opt}_{g\to p}$ for optimization by overlapping regions after warping:

$$M^{opt}_{g\to p} = \widehat{M}_{geo} * M_{photo}, \quad \widehat{M}_{geo}(x_t) = M_{geo}(x_s^*), \tag{4}$$

$$UV^{opt}_{g\to p} = \widehat{UV}_{geo} * M^{opt}_{g\to p}. \tag{5}$$

Please see Figure 2 for an illustration. In the material optimization stage, M_{aln} refers to $M^{opt}_{g\to p}$.

With the improved view-consistent representation of light sources, we re-optimize the materials to achieve more accurate appearance in the material reoptimization stage. However, the per-pixel lighting bakes-in the geometry in certain views, which necessitates all inputs to be aligned with the geometry. So, we warp the reference photo I_{photo} and the material part mask M_{photo} to match the geometry. We again define bounding box parameters l_p and c_p to compute the warped $\widehat{M}_{photo}$ from M_{photo} to match M_{geo}:

$$x_s^* = (x_t - c_g)/l_g * l_p + c_p, \tag{6}$$

$$\widehat{M}_{photo}(x_t) = M_{photo}(x_s^*) \approx M_{geo}(x_t), \tag{7}$$

$$M^{opt}_{p\to g} = \widehat{M}_{photo} * M_{geo}, \quad I^{opt}_{p\to g} = \widehat{I}_{photo} * M^{opt}_{p\to g}. \tag{8}$$

Therefore, M_{aln} refers to $M^{opt}_{p\to g}$ in the material reoptimization stage.

4. Additional Details for Experiments

4.1. Material Classifier Implementation (Sec. 4.3)

The material classification model is based on ResNet-18 [2] backbone. We represent 2D convolution by *Conv2D*(C, K, S, P) where C is the output channels, K is the kernel size, S is stride and P is padding. Other operations in the model include *BN* for 2D batch normalization, ReLU, and Maxpool(K, S) for 2D max-pooling of kernel size K and stride S. The model takes the concatenation of the image and a binary mask of size 240×320×4 as input, followed by *Conv2D*(64, 7, 2, 3), BN, ReLU, Maxpool(3, 2), and modules *conv2.x*, *conv3.x*, *conv4.x*, *conv5.x* from ResNet-18, and 2D average-pooling, resulting by a feature vector of dimension 512. With the feature vector as input, a fully-connected (FC) layer classifies over 886 bins of materials and another FC layer classifies over 9 super-classes. A standard cross-entropy loss is used for the classification heads.

4.2. User Study Details (Sec. 4.3)

There are 60 random AMT users; each is asked to make a different binary comparison for each of 20 scenes *without* pre-training on our task. In each comparison, we ask users to choose the better set of multi-view renderings of transfer results between ours and one randomly sampled baseline from {cls, inv med, pix med}. Two options are randomly placed while the input photo is in the middle. Each comparison is evaluated by 20 different users.

5. Potential Negative Impacts

Our approach can synthesize high-quality digital counterparts of real scenes which may be rendered to create photorealistic images. Using a physically-based material prior also allows the ability to edit properties of these images by generating plausible new materials for specific regions or objects, which may be used for potentially harmful purposes. An avenue to overcome this negative impact might be further research in digital watermarks such as [3] for materials generated through our material priors, embedded in a manner that allows them to persist in an identifiable way through the rendering process.

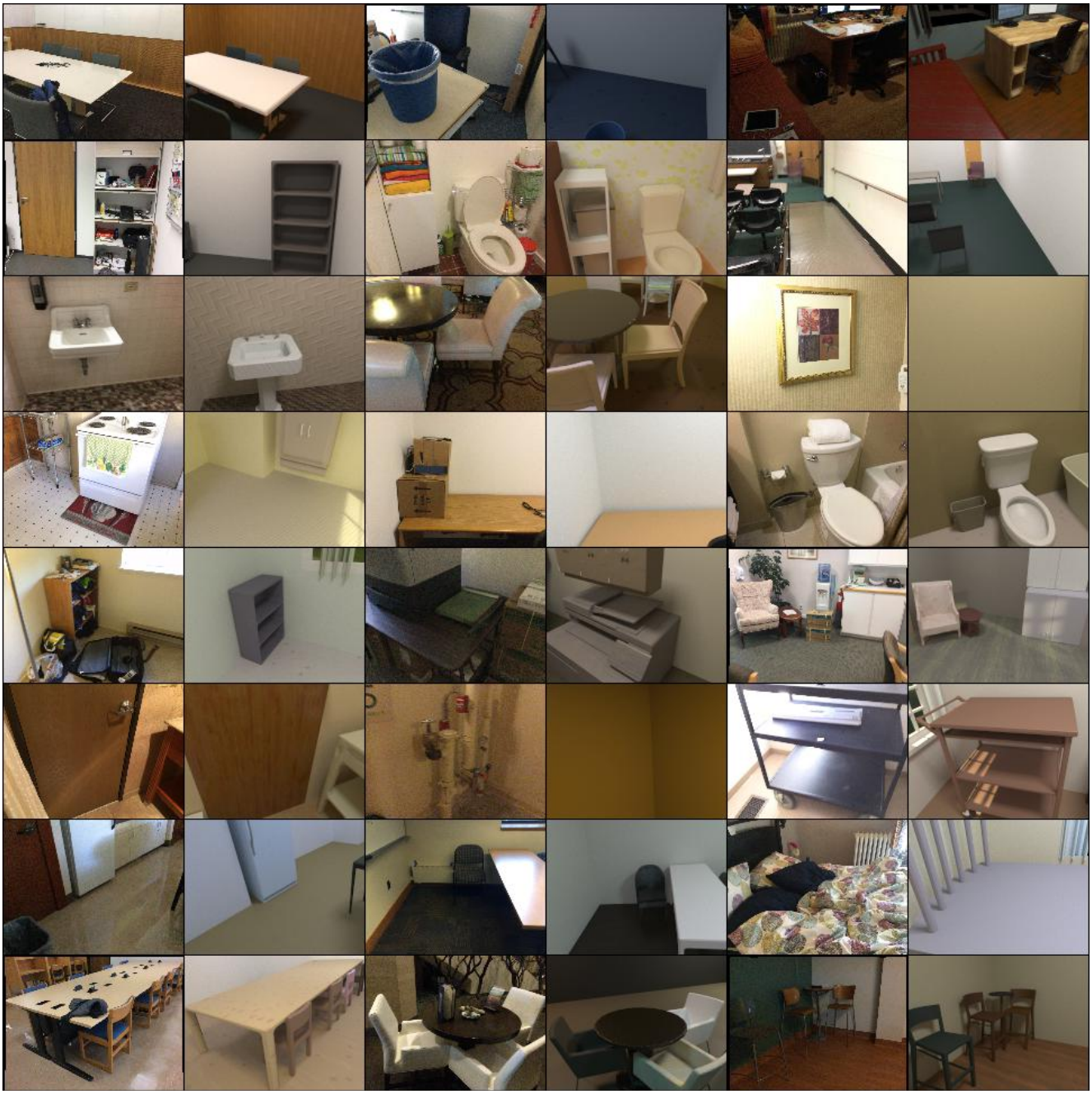

Figure 3. More results of material and lighting transfer with ScanNet-to-OpenRooms dataset.

References

[1] Bowen Cheng, Alexander G Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *arXiv preprint arXiv:2107.06278*, 2021. 1, 2

[2] 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. 3

[3] Matthew Tancik, Ben Mildenhall, and Ren Ng. StegaStamp: Invisible hyperlinks in physical photographs. In *CVPR*, 2020. 3

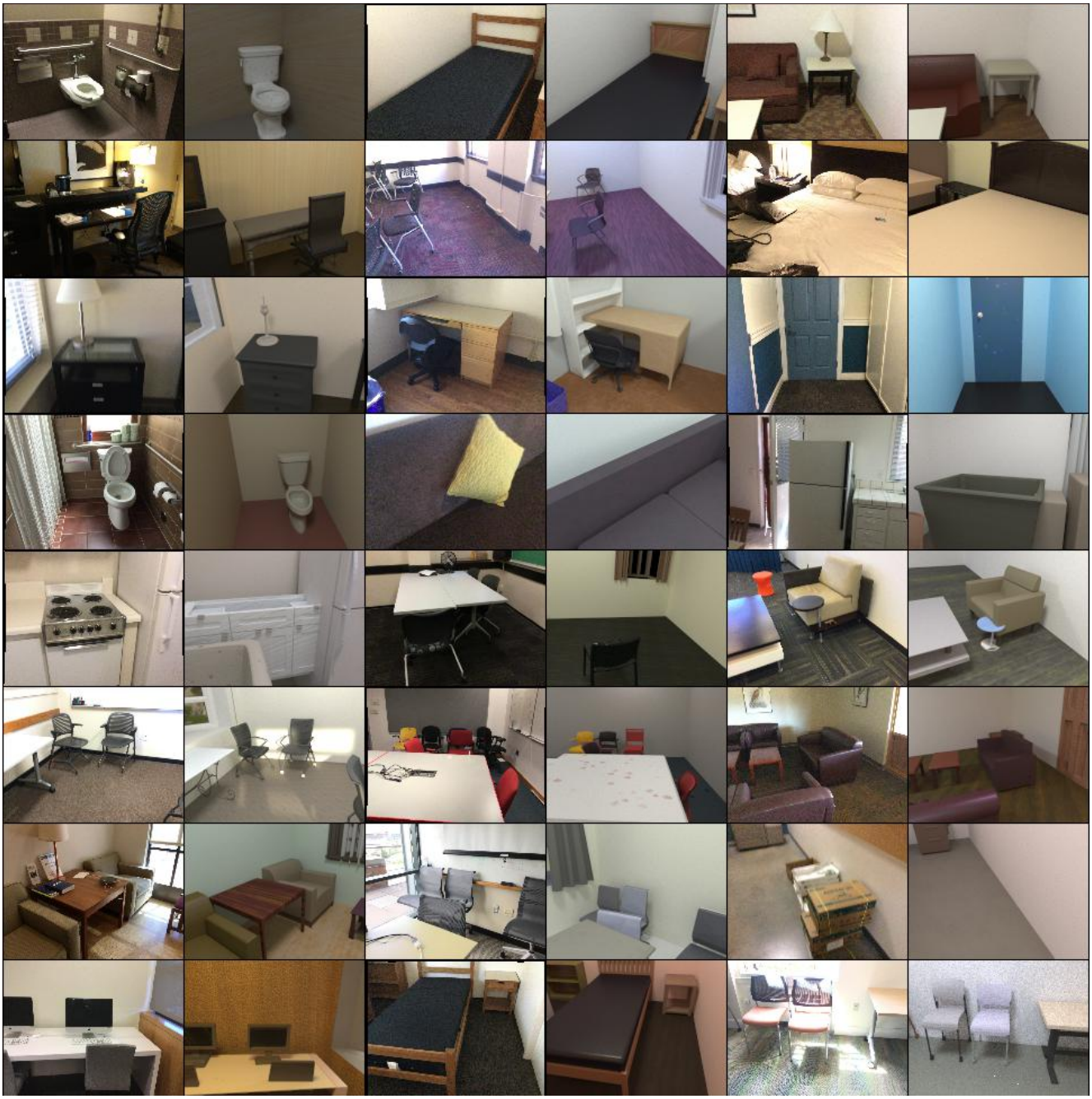

Figure 4. More results of material and lighting transfer with ScanNet-to-OpenRooms dataset.

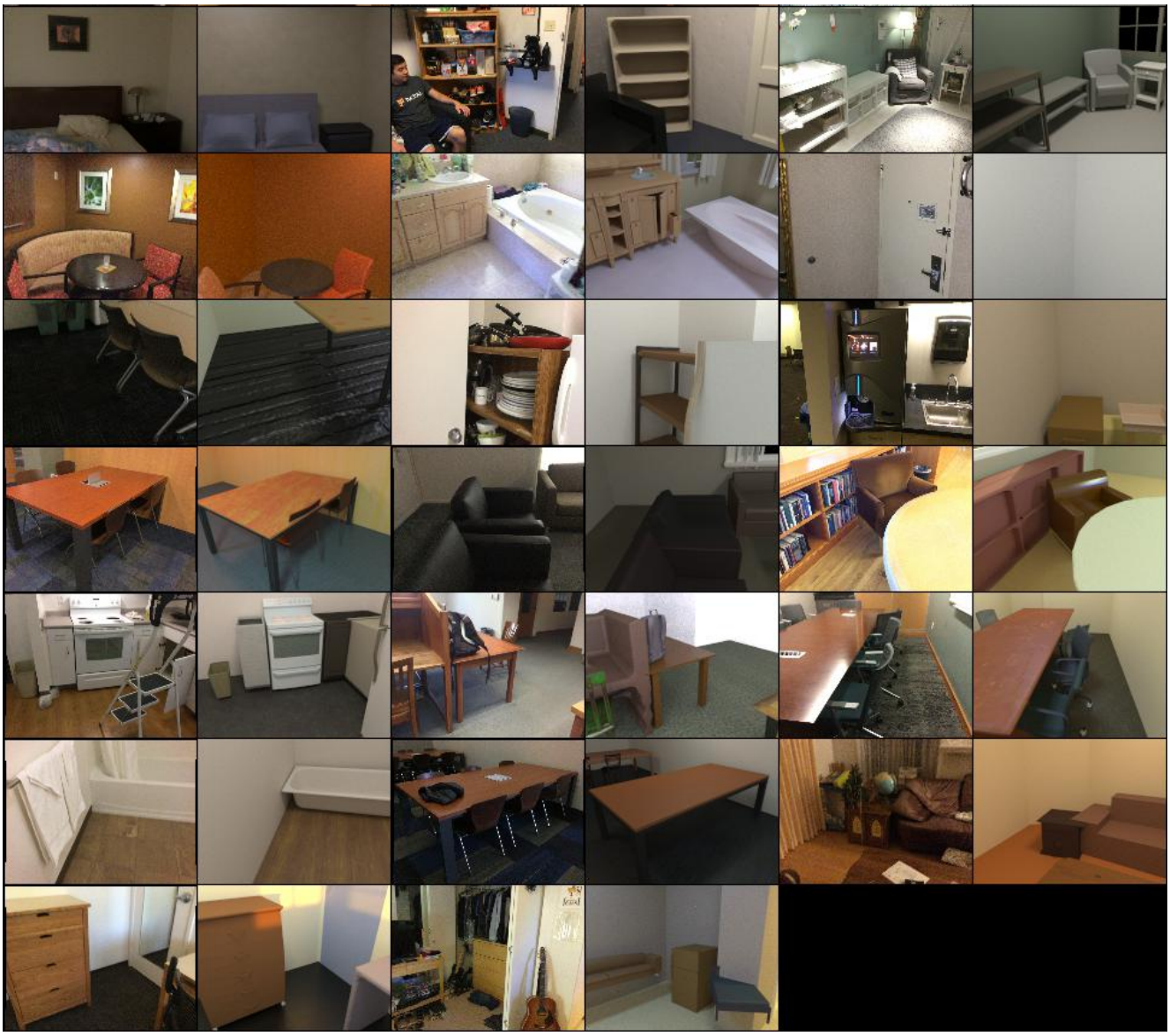

Figure 5. More results of material and lighting transfer with ScanNet-to-OpenRooms dataset.

Figure 6. More results of material and lighting transfer with SUN-RGBD-to-Total3D dataset.