Supplementary Material: Factorized Inverse Path Tracing for Efficient and Accurate Material-Lighting Estimation

Positional encoding Hash grid Ground truth

Figure 1: **Qualitative comparison of different input encoding** shows a hash grid can better model the detailed texture on the floor.

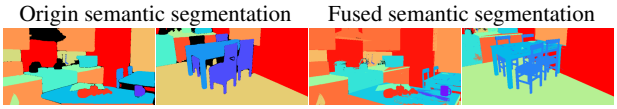

Figure 2: **Fusing segmentation on raw images** onto the mesh produces multi-view consistent segmentation.

Supplementary Material Overview

In Sec. A, we provide the details of our pipeline implementation and data pre-processing.

In Sec. B, we present additional details of our experiments, including: (1) the setup of real world scenes (Sec. B.1); (2) detailed ablation study (Sec. B.2); (2) additional results (Sec. B.3); and (4) the schemes for evaluating the baselines (Sec. B.4).

A. Implementation Details

We implement our method in PyTorch [12] and Mitsuba 3 [5]. The diffuse and specular shadings in Eq. 6 are path-traced and denoised by the OptiX denoiser [11], where we use 128 samples per pixel for diffuse shadings and 64 for specular shadings. Importance sampling of the BRDF is applied for shading initialization (stage 1), and multiple importance sampling is applied for shading refinement (stage 3). For each round of BRDF-emission mask estimation (stage 2), the optimization is run over the entire training set for 2 epochs using Adam [6] optimizer with a learning rate of 1e-3 and a batch size of 8,192. Stage 2 and 3 are repeated twice after stage 1, and all the experiments are run on a single 3090Ti GPU.

Network architecture. The BRDF network MLP_{brdf} has 2 hidden layers of size 64, and its hash encoding [10] has 32 levels and 19 $\log_2$ hash map size with other parameters set to their recommended defaults. For emission mask network MLP_{emit}, we use positional encoding [9] with 10 frequency bands, 6 hidden layers of size 128, and one residual connection in the middle. Hash encoding is preferred for the BRDF network as albedo usually demonstrates high frequency pattern, which can be more efficiently modeled by a hash grid (Fig. 1). Both networks use ReLU activation between the intermediate layers.

Semantic segmentation acquisition. To obtain semantic segmentation, we use Mask2Former [3] pre-trained on the COCO dataset [8] with Swin-L backbone. The input images are firstly tone-mapped with $\gamma = 1/2.2$ then clipped to be in the range $[0, 1]$. Given segmentation from multi-view images, we fuse them onto the mesh and let each mesh triangle take the segmentation ID with the maximum occurrence (Fig. 2).

Geometry acquisition with MonoSDF [17]. We adapt the original code from MonoSDF in the default configuration for ScanNet with Multi-Resolutional Feature Grids architecture and the following changes: (1) instead of having all rays coming from one image in each training iteration, we randomly sample over all training pixels, which is empirically found to yield more stable convergence on noisy inputs especially for real world images; (2) input images are changed from SDR to HDR to be in the same format as our model input; accordingly, output activation of MLP is changed to ReLU, and re-rendering loss is changed to L1 loss on tone-mapped outputs and labels. Considering MonoSDF does not incorporate an outlier rejection algorithm, we employ a two-step training strategy to deal with the bad camera poses. We first train for one epoch to acquire a rough mesh and reproject the mesh onto all frames. Frames with significant misalignment are then rejected and the model is re-trained. To extract the mesh, we employ Marching Cubes with a grid size of 512. In total, the entire process takes around 1 day per-scene.

B. Experiment Details

B.1. Real world scene capture and relighting

Need for acquiring new real world data. Existing datasets that provide multi-view HDR images and camera poses of real world scenes may include: Replica [14] , Matterport3D [2], and sample scenes from FVP [13]. However, each dataset has their own limitations that prohibit usage

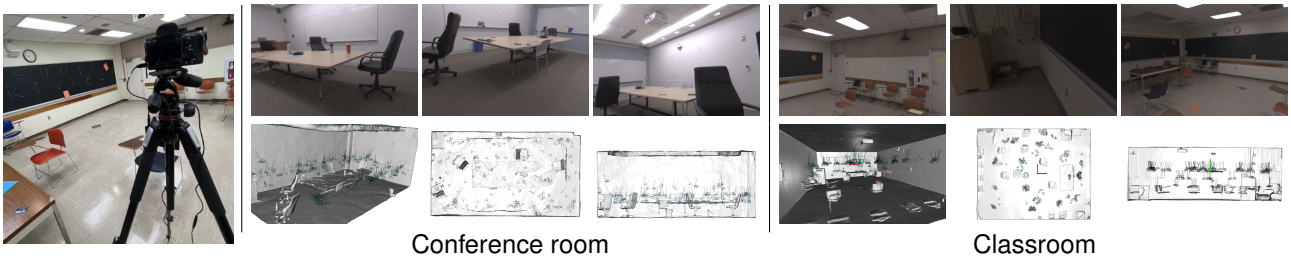

Figure 3: **The capture setting (left) and observations of the real world scenes (middle and right).** We present two real world scenes (Conference room and Classroom) with samples of captured images, reconstructed geometries in 3 views, and all camera poses.

in our evaluation. Specifically, HDR images from FVP do not employ exposure bracketing, which results in overexposed emission that is not applicable to our physically-based light transport modeling. HDR images from Replica are not publicly available, thus view-dependent effects cannot be observed. For Matterport3D, the captured images exhibit artifacts including camera glare and problematic tone-mapping.

Therefore, we capture a few scenes as proof of concept of our method, including a conference room scene presented in the main paper and an additional classroom scene. Fig. 3 demonstrates our capture setting. We mount a Sony A7M3 full-frame camera on a tripod and use a remote control shutter release to capture images with exposure bracketing of 5 steps 1EV each or 5 steps 2EV each depending on the dynamic range of the room. We take images from multiple locations of the room, starting roughly with a direction towards the room center, then randomizing yaw angles between $-60°$ to $60°$, pitch angles between $-45°$ to $45°$, with minimal roll. The camera height is sampled between $0.5m$ to $2.5m$. For HDR reconstruction, we process the captured RAW images with black level subtraction, demosaicing, devignetting, and undistortion. The recovered images are assumed to follow linear camera response and are combined using a hat function similar to Debevec *et al*. [4].

Reference relighting of Classroom. As is shown in Fig. 4, lights in the Classroom can be switched between front and rear light modes. We choose the rear lights as original lighting for the main capture, and take a few additional photos with only front lights on as reference for relighting. Given BRDF-emission estimation from the main capture, we relight the scene by turning the estimated emission off and insert simple novel emitters to roughly match the front lights in their actual locations (see demonstration in Fig. 4, bottom). Considering it is not possible to have the manually inserted novel emitters to perfectly match the actual complex front lights, we treat the reference relighting photos only as pseudo-ground truth.

B.2. Ablation study details

Fig. 5 shows the effect of different training strategies on BRDF-emission estimation as discussed in Sec. 5.4. To demonstrate the impact of noisy inputs, Fig. 6 shows the quality of estimated geometry and semantic segmentation with respect to their ground truth together with the corresponding reconstruction results. It can be seen that surface roughness for regions with weak highlights can be very sensitive to inputs, while emission and material reflectance estimation are robust as long as the noise stays in a reasonable range.

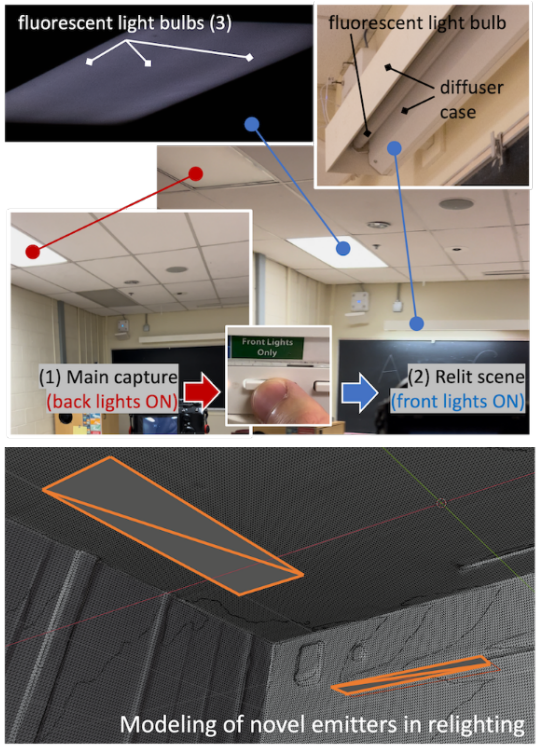

Figure 4: **Relighting of the Classroom scene.** The upper figure shows (1) the lighting for main captures with only back lights turned on, and (2) the relit scene with only front lights on (as reference for our relighting experiments). The lower figure shows our inserted area emitters as approximation of the actual front lights.

Failure cases. As discussed in the limitation section (Sec. 6), broken geometry can lead to large artifacts in our BRDF-emission reconstruction. A dormitory scene capture

Figure 5: **Qualitative comparison of different training strategies** shows all of the strategies are necessary for efficient and accurate BRDF-emission estimation. The insets are the ground truth.

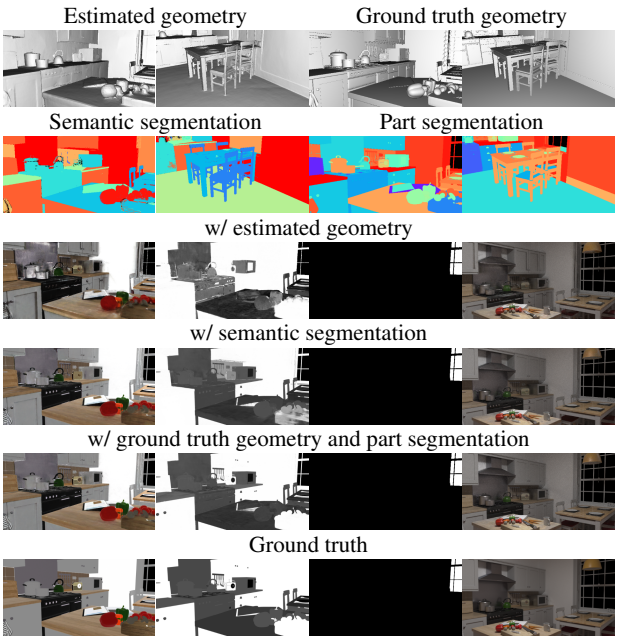

Figure 6: **Sensitivity analysis on training inputs.** Imperfect geometry and the usage of semantic segmentation instead of fine-grained part segmentation (row 1-2) can be acceptable for our BRDF-emission estimation (row 3-4, column 1-2). Ambiguity in roughness increases as geometry is imperfect or coarser segmentation is used (row 3-4, column 2), but they do not significantly affect applications like relighting (row 3-4, column 4).

is shown in Fig. 7 to demonstrate the problem, where the front face of the reconstructed wall cabinet fails to align with the actual geometry (because of insufficient view coverage), causing the shadow boundary to be baked into the reflectance map. Meanwhile, geometry of the lamp on the ceiling fan is partly missing, which causes the emission to be incorrectly projected to the background wall and cabinet surface, creating bright artifacts on the reflectance map and phantom emitters on the wall.

B.3. Additional results

In Fig. 8, 9, we show the per-scene qualitative comparison of estimated BRDF and emission for all methods on synthetic dataset, and we compare the view synthesis and relighting results in Fig. 10, 11. In Fig. 12 and Fig. 13, we provide evaluation on additional views of our real world captures.

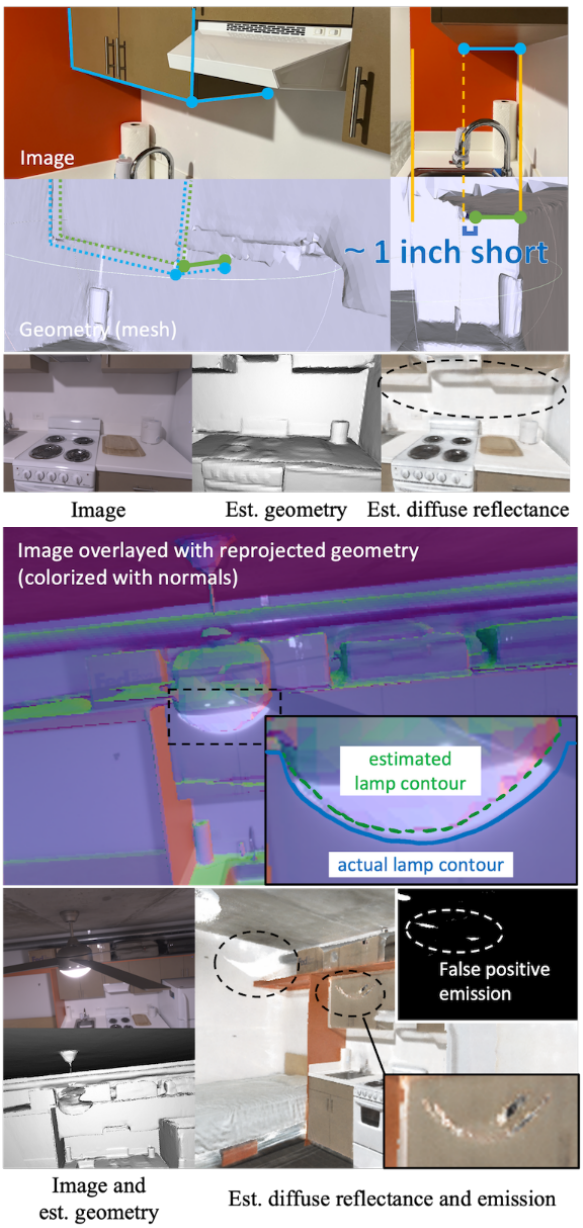

Figure 7: **Failure cases of our method due to bad geometry.** Top: indented cupboard (actual boundary in blue; estimated in green which has a geometry error of around 1 inch) results in incorrect light-surface intersection and boundary artifacts on the diffuse reflectance map (circled). Bottom: emission and bright artifacts (for diffuse reflectance) get erroneously baked onto the wall and cupboard (circled) because of the missing geometry on the lamp.

B.4. Evaluation scheme of baseline methods

For FVP [13], We use its original code with the following adaptations: FVP relies on thresholding RGB values to locate emitters, so we pick the threshold that separates emitters from the rest of the scene in our images. It also involves a step to manually set the exposure of each overexposed

emitter, which in our adaptation is provided as the median radiance within each emitter. In relighting, FVP assumes the maximum radiance of novel emitters to be 1 so as to yield shading in SDR for input into its network. Afterwards, the exposure of novel emitters can be set to arbitrary numbers in FVP's GUI. We follow the strategy but set exposure as our desired radiance values for novel emitters, so that relighting results from FVP can be directly comparable.

For evaluation of IPT [1] and MILO [16], since their code is not available, and a re-implementation requires careful design choices, we depend on results provided by the authors of MILO and IPT on our data, where it was possible for them to evaluate. Because MILO and FVP use texture-based representations, geometry is remeshed to prevent artifacts like UV seam and bleeding, which gives equivalent quality in most of the cases except for thin structures like disks on the kitchen table.

For Li22 [7], instead of using predicted depth, we directly back-project ground truth geometry to obtain a depth image as its input. On real scenes, considering the method is based on single-view input and does not allow rerendering to novel views under different lighting, we directly feed the reference relighting image as the input, replace all estimated emitters by our novel eimtters (Sec. B.1), and re-render the scene with estimated materials using its neural rendering pipeline.

References

[1] Dejan Azinovic, Tzu-Mao Li, Anton Kaplanyan, and Matthias Nießner. Inverse path tracing for joint material and lighting estimation. In *CVPR*, 2019. 4, 5, 6, 7, 8

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *3DV*, pages 667–676, 2018. 1

[3] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 1

[4] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH 2008 classes*, pages 1–10, 2008. 2

[5] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. Mitsuba 3 renderer, 2022. https://mitsuba-renderer.org. 1

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[7] Zhengqin Li, Jia Shi, Sai Bi, Rui Zhu, Kalyan Sunkavalli, Miloš Hašan, Zexiang Xu, Ravi Ramamoorthi, and Manmohan Chandraker. Physically-based editing of indoor scene lighting from a single image. In *ECCV*, 2022. 4, 5, 6, 8, 9

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 1

[9] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1

[10] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. In *ACM Transactions on Graphics (ToG)*, volume 41, pages 1–15, 2022. 1

[11] Steven G Parker, James Bigler, Andreas Dietrich, Heiko Friedrich, Jared Hoberock, David Luebke, David McAllister, Morgan McGuire, Keith Morley, Austin Robison, et al. Optix: a general purpose ray tracing engine. In *Acm transactions on graphics (tog)*, volume 29, pages 1–13, 2010. 1

[12] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, volume 32, 2019. 1

[13] Julien Philip, Sébastien Morgenthaler, Michaël Gharbi, and George Drettakis. Free-viewpoint indoor neural relighting from multi-view stereo. In *ACM Transactions on Graphics (TOG)*, volume 40, pages 1–18, 2021. 1, 3, 7, 8, 9

[14] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 1

[15] Yao Yao, Jingyang Zhang, Jingbo Liu, Yihang Qu, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Neilf: Neural incident light field for physically-based material estimation. In *ECCV*, pages 700–716, 2022. 5, 6

[16] Bohan Yu, Siqi Yang, Xuanning Cui, Siyan Dong, Baoquan Chen, and Boxin Shi. Milo: Multi-bounce inverse rendering for indoor scene with light-emitting objects. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 4, 5, 6, 7, 8, 9

[17] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. In *NeurIPS*, 2022. 1

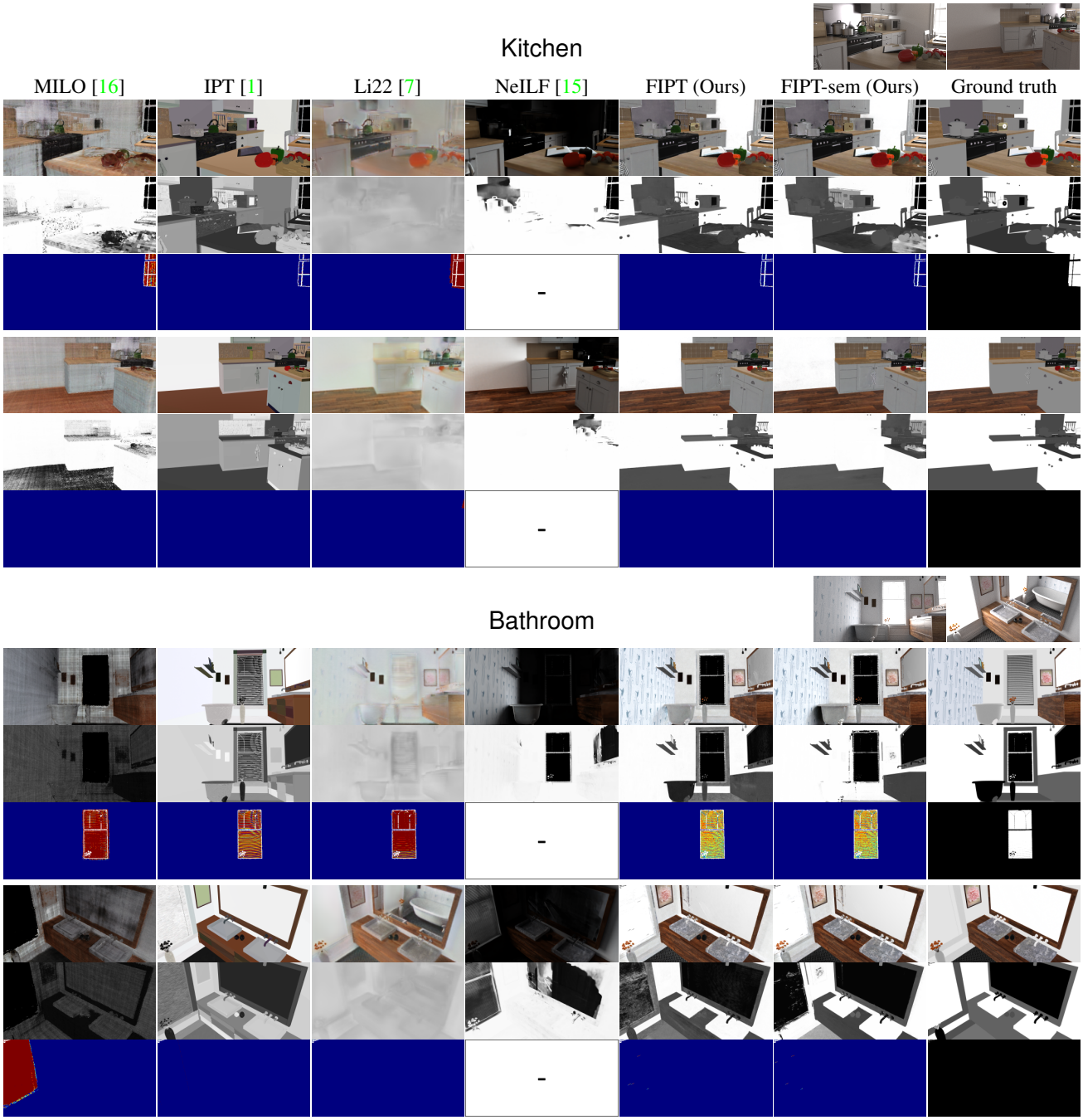

Figure 8: **BRDF and emission estimation on synthetic Kitchen and Bathroom for all methods.** Input views are shown in the upper-right corner of each scene.

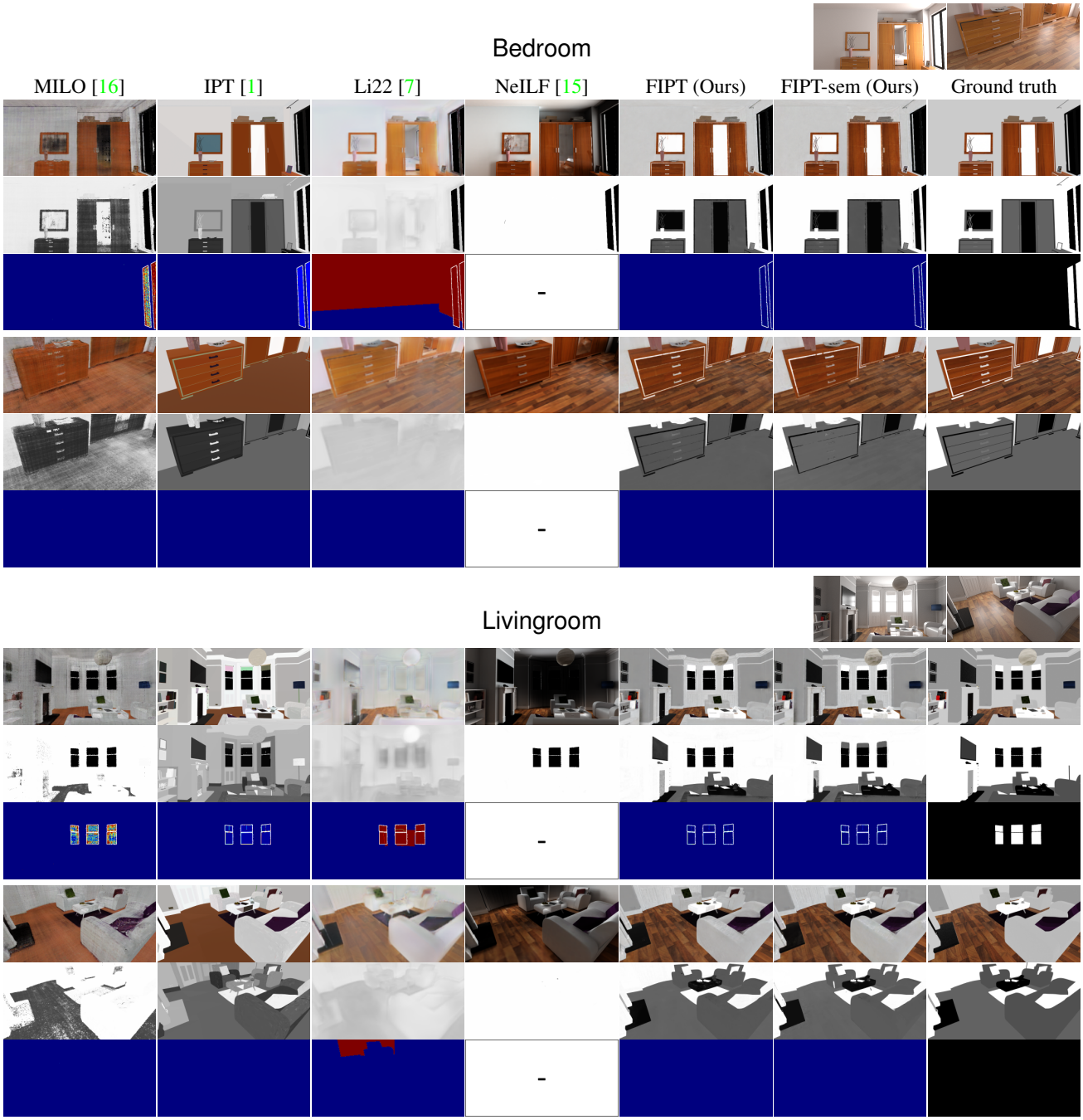

Figure 9: **BRDF and emission estimation results on synthetic Bedroom and Livingroom for all methods**, showing 2 views per-scene. Input views are shown in the upper-right corner of each scene.

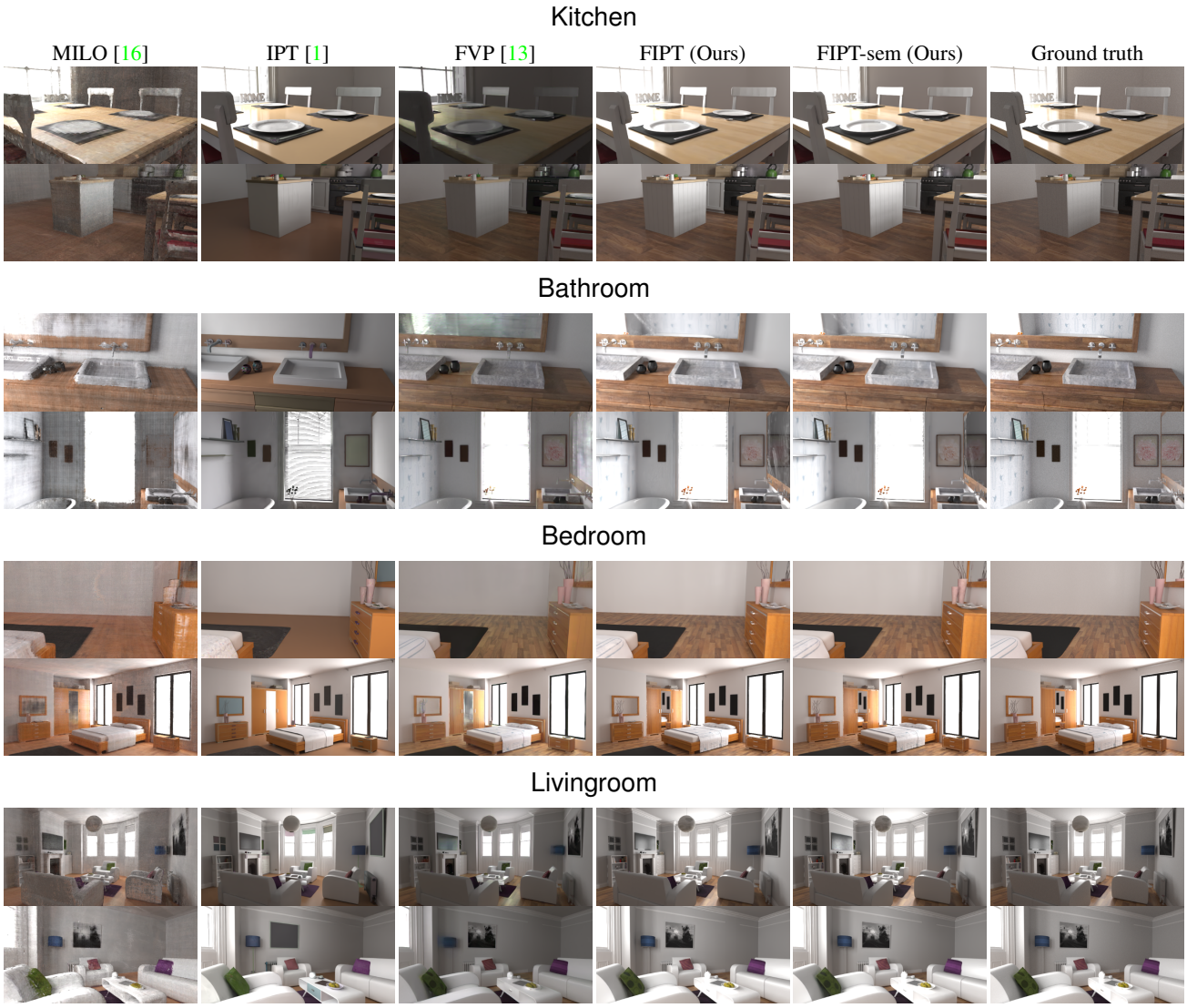

Figure 10: **View synthesis results on synthetic scenes for all methods**, showing 2 views per-scene.

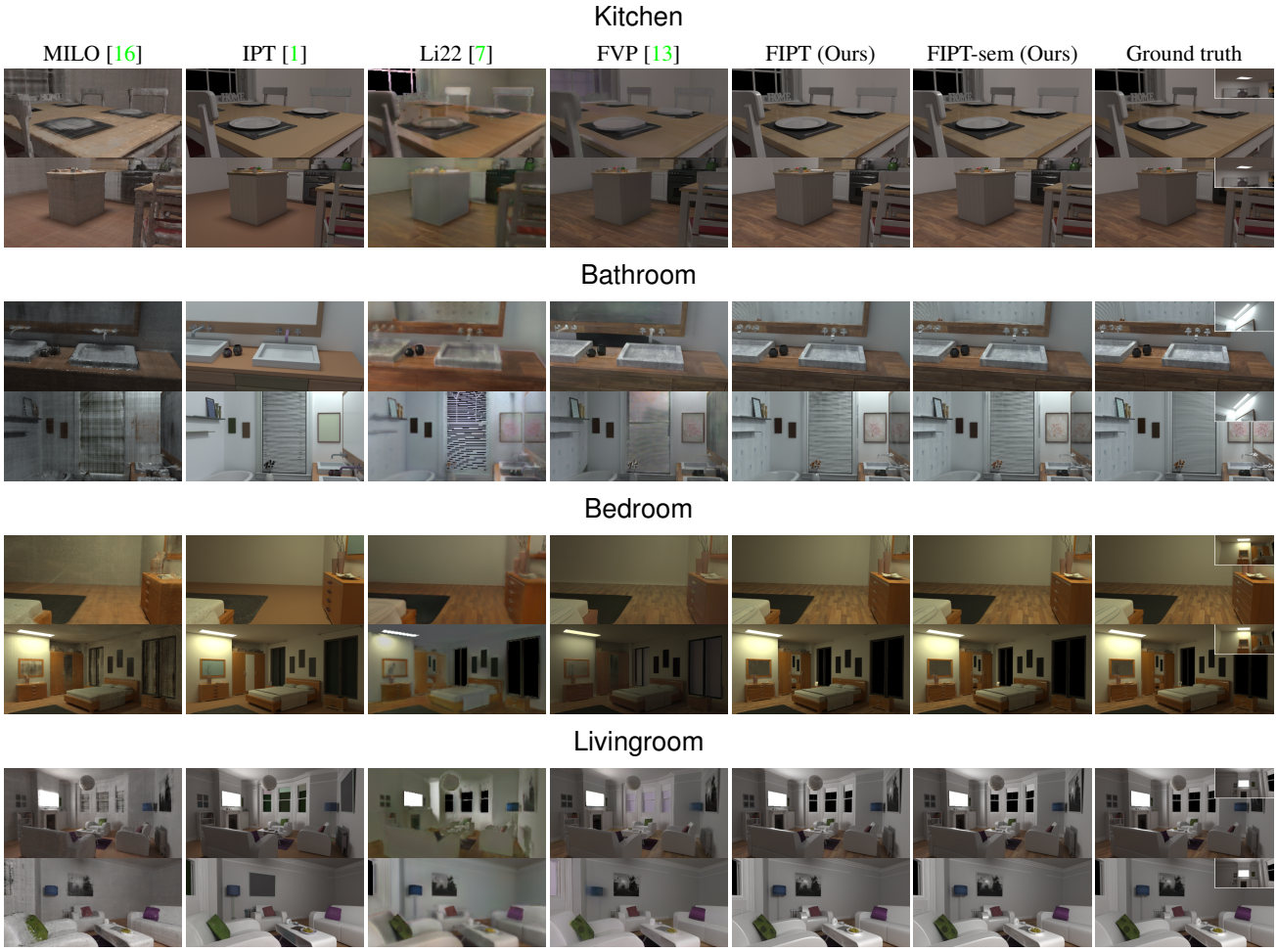

Figure 11: **Relighting results on synthetic scenes for all methods**, showing 2 views per-scene.

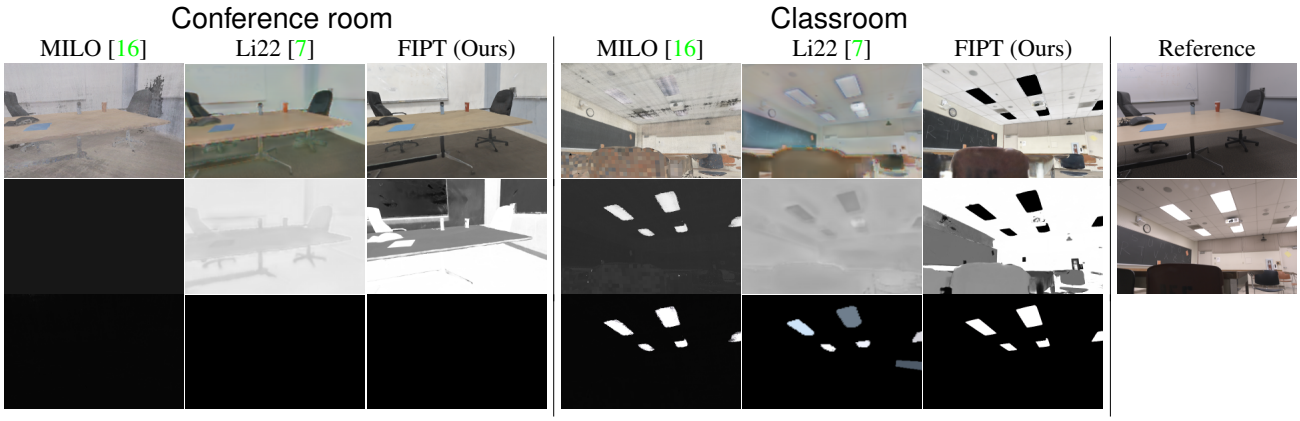

Figure 12: **BRDF and emission estimation on real scenes**, showing 1 additional view per-scene besides views shown in the main paper.

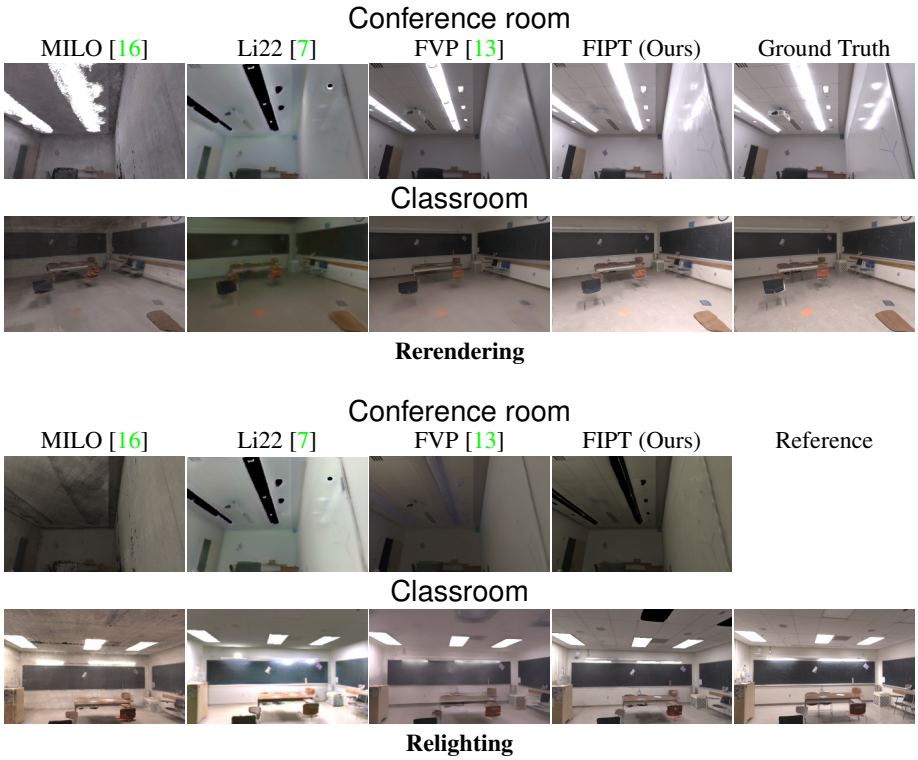

Figure 13: **Rerendering and relighting on real scenes,** showing 1 additional view per-scene for each task besides views shown in the main paper. Top two rows show the rerendering with original lighting (Conference room: all ceiling lamps on; Classroom: rear lights on and fronts lights off). Bottom two rows show the relighting under novel light with relit Classroom also included as pseudo-ground truth (with rear lights off, and front lights on).