

Intrinsic Image Fusion for Multi-View 3D Material Reconstruction

Supplementary Material

A. Additional Implementational Details

Dataset Details. We specify the scale of the test set in the inset table. We calculate PBR decomposition metrics against the ground-truth on all input views and report averaged results for the 4 synthetic scenes.

Number of parameters. Our parametric texture representations uses a base texture and a set of per-object affine transformations. The base texture uses an neural hashgrid [19] with 32 levels, 2 features per level and 2^{19} hashmap size to store albedo, rough, metallic means and scales, which amounts to 28 M parameters. These base texture parameters are optimized only during the aggregation phase. During the inverse path tracing, we only need to optimize the per-object affine transformations, which amounts to $O \times 3 \times 4$, where O is the number of 3D objects, giving a total of 1092 parameters for the kitchen scene. Optimizing on such a low-dimensional manifold makes the inverse path tracing more constrained, making it more robust against the rendering noise.

Real data pre-processing. Our method relies on instance segmentation, which we obtain in a pre-processing step for the ScanNet++ [28] scenes. We use SAM2 [22] to estimate a per-image segmentation. To aggregate the segmentations, we use MaskClustering [26]. This way, we get a per-face instance id. Then, we rasterize the mesh into all the views and render the instance ids to get per-pixel instances. These images are only used for evaluating the baseline IRIS [18].

Real-world lighting representation. To account for missing geometry and emission coming from outside of the scene, we additionally define an environment map of resolution 16×32 . Similarly to the mesh emission, we filter the potentially emissive environment map pixels by thresholding their aggregated observed radiance values ($t = 0.85$).

Lighting optimization. Our lighting optimization follows FIPT [25] and has four steps. First, we initialize the emissive triangles by filtering the aggregated observed radiance with a threshold of $t = 0.99$. Second, we optimize for the emission values using frozen base material textures with inverse path tracing using 128 samples per pixel with a single bounce. We use SGD optimizer for 1 epoch and batch size 8192 rays with initial learning rate of $lr = 1e + 2$. After 1000 iterations, we prune all the emitters, which has an

Dataset	Scene	Views
FIPT [25]	Kitchen	202
FIPT [25]	Bedroom	208
FIPT [25]	Livingroom	213
FIPT [25]	Bathroom	109
FIPT [25]	Conferenceroom	191
FIPT [25]	Classroom	278
ScanNet++ [28]	2a1b555966	349
ScanNet++ [28]	651dc6b4f1	64
ScanNet++ [28]	a003a6585e	106

Table 4. **Dataset Details.** We summarize the used scenes.

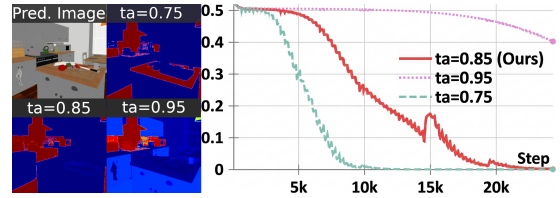


Figure 9. **Temperature Ablation.** Lower temperature values motivate quick convergence to hard assignment, but can get stuck in local optima, while too high value keeps soft assignment leading to oversmoothing. We use a value in-between to balance.

intensity lower than 5% of the overall maximum intensity. This stage takes 35 minutes using 20 GB of VRAM on the kitchen scene. Then, we cache the light transport into diffuse ($spp = 256$) and specular shading maps ($spp = 128$). This stage takes 5 minutes on the kitchen scene. Finally, we optimize for the BRDF and CRF parameters by re-rendering with the cached shading maps. This stage takes 20 minutes using 11 GB of VRAM on the kitchen scene. Here, we use SGD optimizer for 3 epoch and batch size 32768 rays with initial learning rate of $lr = 1e + 1$ and decay after every epoch by a factor of 0.2. Following IRIS [18], we regularize the roughness and metallic channels to stay close to diffuse materials ($w_{rough} = 1e - 3$, $w_{metal} = 5e - 3$).

B. Additional Results

Relighting. Our supplementary video provides comparisons and relightings with rendered trajectories. We provide additional results for all the synthetic scenes in Figure 12, for the real ScanNet++ (2a1b555966, 651dc6b4f1, a003a6585e) scenes in Figure 13. We compare also on the real scenes of FIPT [25] in Figure 14, which uses photometric stereo for the mesh reconstruction. We provide more results on syntetic and ScanNet++ [28] scenes in Figure 15.

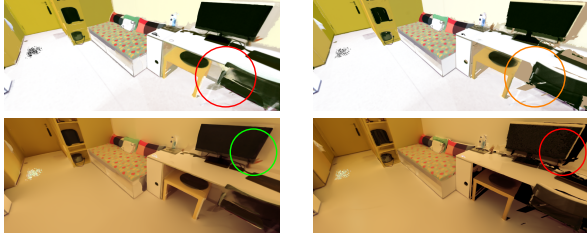


Figure 10. **Geometry Ablation.** The estimated albedo is similar with laser scan (left) and photometric reconstruction (right [18]) reconstruction.



Figure 11. **Failure Cases.** Even though our parametric formulation makes our method robust, geometric reconstruction errors propagate into our final reconstruction, such as missing thin structures, transparent objects or floating artifacts.

Cross-View Aggregation Ablation. We provide additional ablation results on our cross-view aggregation strategies in Figure 16.

Temperature Ablation. We ablate the effect of the temperature annealing factor in Figure 9 to evaluate the effectiveness of the assignments. The prediction assignments $\alpha_{i,k}$ converge to binary (red=select, blue=drop). Their entropy (right) drops from 0.5 (naive averaging) to 0 (mode selection); $\tau_{\text{anneal}}=0.85$ balances convergence and exploration.

Geometry Ablation. In Figure 10 we show qualitative comparison between input geometries obtained with laser scan or with photometric reconstruction on a ScanNet++ scene (7e09430da7), showing that our method gives comparable results even with lower quality geometry.

Failure cases. Since our method depends on reconstructed geometry, we inherit their limitations. Thin structures and semi-transparent object are hard to reconstruct, often leading to missing or incorrect geometry. We show such failures in Figure 11 Furthermore, consistently wrong predictions can leak into the reconstruction. The book in Figure 2 shows white and grey caused by incorrect predictions.

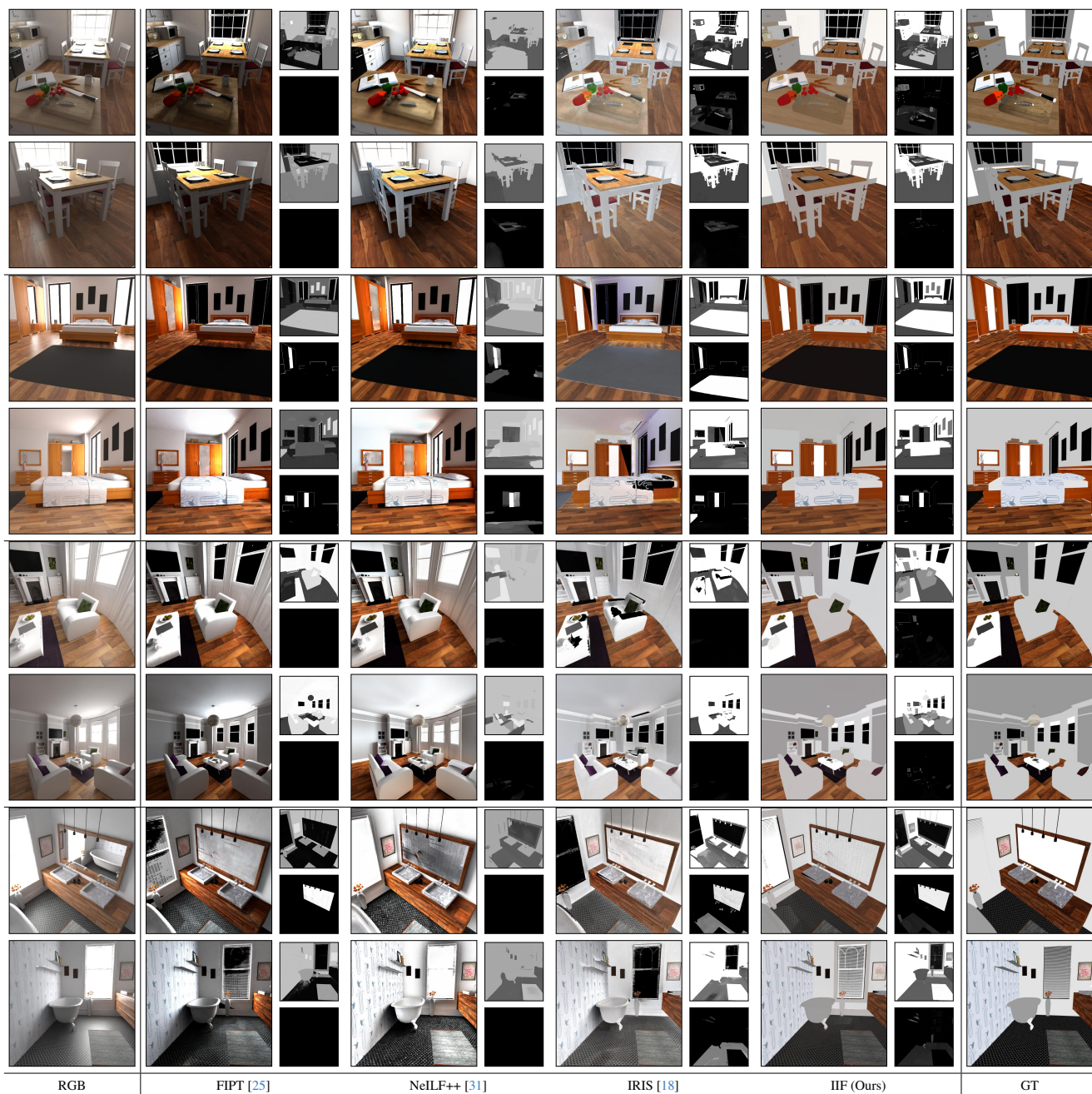


Figure 12. **Synthetic comparisons.** Additional samples on the synthetic scenes.

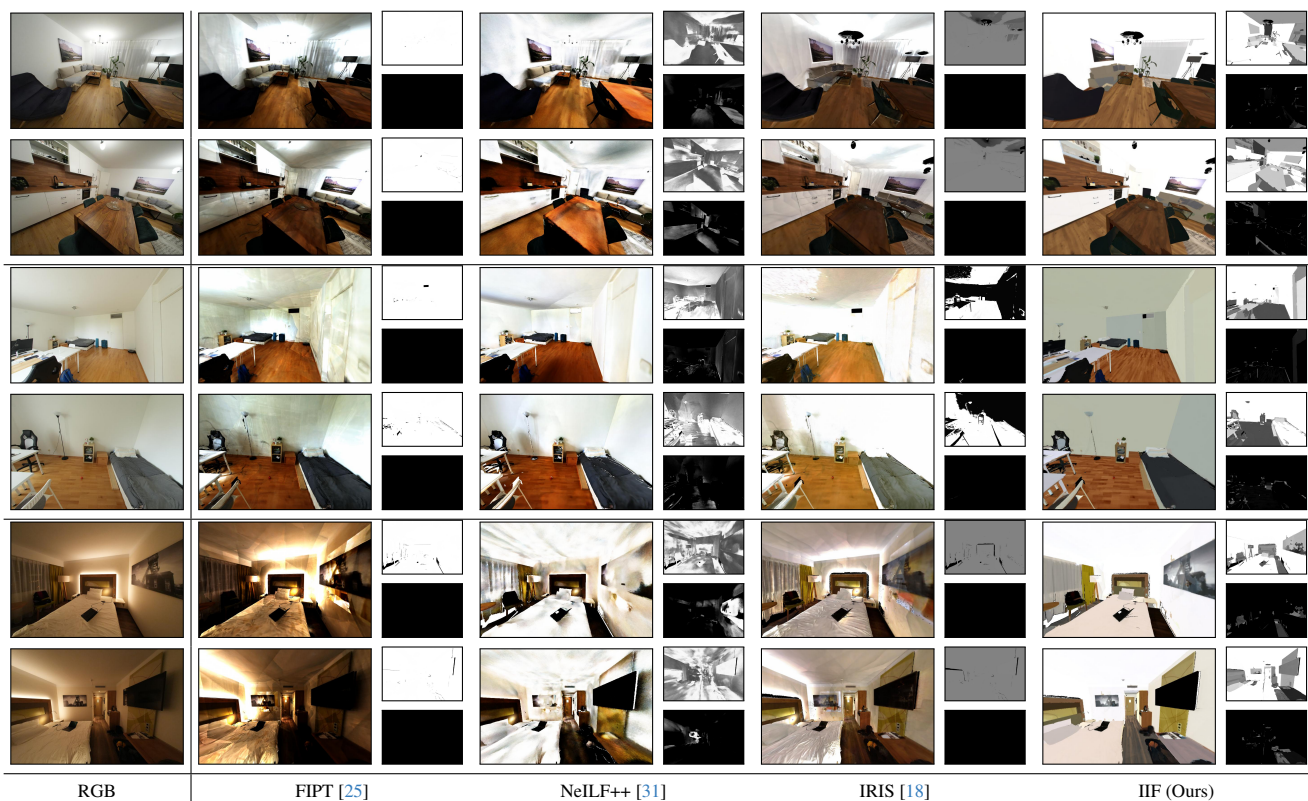


Figure 13. **Scannet++ [28] comparisons.** Additional samples on ScanNet++ [28] scenes.

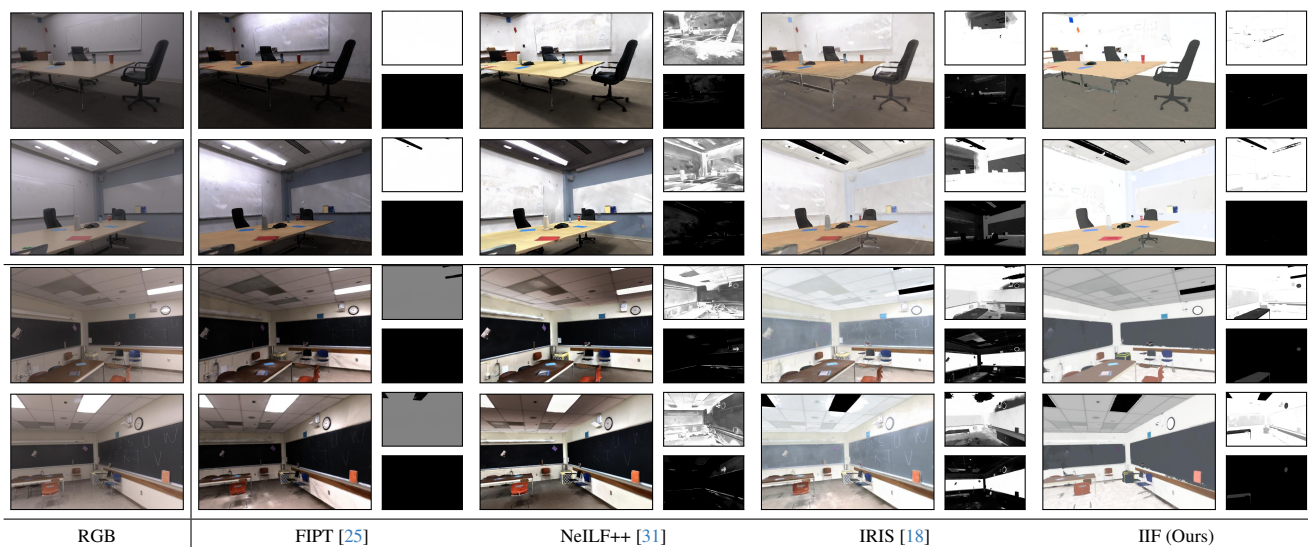


Figure 14. **Real [25] comparisons.** Additional samples on the real scenes of FIPT [25].



Figure 15. **Additional relightings.** We show additional relighting results on synthetic and ScanNet++ scenes over a smooth trajectory of an emissive sphere. For additional interpolations, please refer to our video.



Figure 16. **Cross-view aggregation (additional results).** Single-view material estimation can yield detailed but inconsistent predictions (Fig. 2). IRIS [18] uses per-object aggregation, losing patterns. Per-texel aggregation maintains patterns but introduces seams. Our parametric modeling (§ 3.1) provides a low-dimensional space of consistent 3D aggregations. Distribution matching (§ 3.2) selects the best predictions per view to preserve fine details.