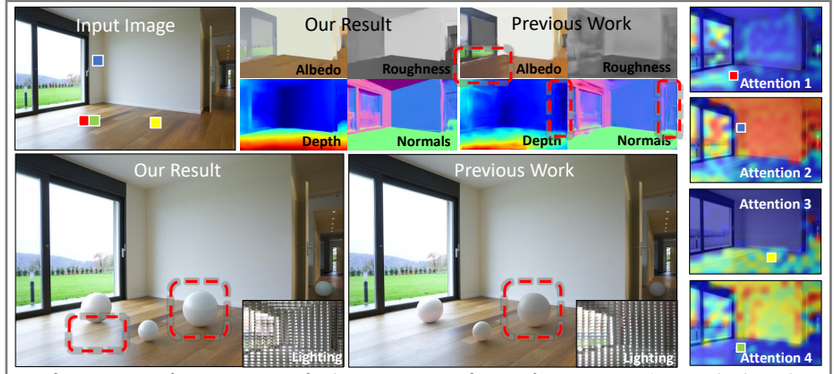
IRISformer: Dense Vision Transformers for Single-Image Inverse Rendering in Indoor Scenes

Rui Zhu[1] Zhengqin Li[1] Janarbek Matai[2] Fatih Porikli[2] Manmohan Chandraker[1]
[1]UC San Diego [2]Qualcomm AI Research

UCSDCSE Computer Science and Engineering

Qualcomm

Overview



IRISformer, a specific instantiation of a dense vision transformer, for inverse rendering in both single-task and multi-task settings.

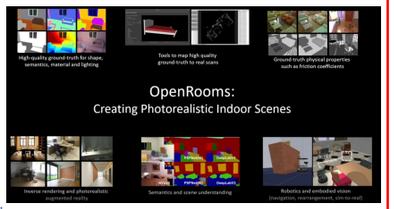
Given a single real-world image, (upper-left) IRISformer simultaneously infers material (albedo and roughness), geometry (depth and normals), and spatially-varying lighting of the scene. The estimation enables virtual object insertion where we demonstrate high-quality photorealistic renderings in challenging lighting conditions compared to previous works [1,2] (lower-left).

Motivated by the intuition that the long-range attention learned by transformers is ideally suited to reason long-range interactions to account for shadows, highlights and interreflections→, we propose to use Transformers in place of CNNs in inverse rendering pipeline improve estimations of all modalities.

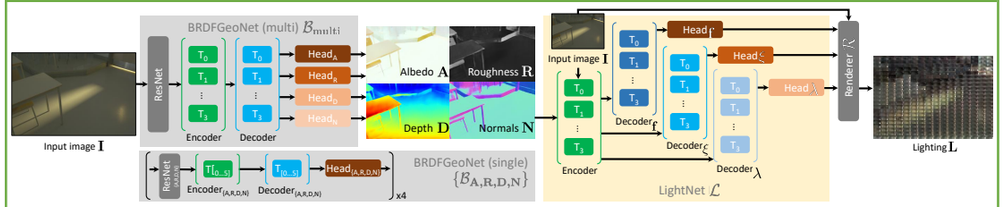
↗ The learned attention is visualized for selected patches, indicating global context and implicitly learned semantic notions.

Datasets

- **Training**: our in-house developed OpenRooms dataset [2] for large-scale photorealistic renderings of indoor scenes, with ground truth full 3D geometry, material, lighting and semantics. OpenRooms is used to train the entire pipeline from scratch, with full supervision on all tasks →.
- **Evaluation**:
  - albedo estimation after finetuning: IIW dataset
  - geometry estimation (depth, normals) after finetuning: NYUv2 dataset
  - object insertion/material editing: natural images dataset from Garon et. al [3].



Method



↑Two-stage inverse rendering pipeline with Transformers as backbone:

1. albedo, roughness, depth, normals;
2. per-pixel lighting (spherical Gaussian (SG) mixture).

Full-supervision using ground truth is imposed on all estimations in Stage 1. In Stage 2, taking estimation of SG, a lighting renderer renders a per-pixel lighting map, on which we may impose a fully-supervised lighting reconstruction error, and a self-supervised image re-rendering error to jointly constrain all estimations.

We also explore single-task and multi-task versions to account for trade-off between model capacity and model size↗.

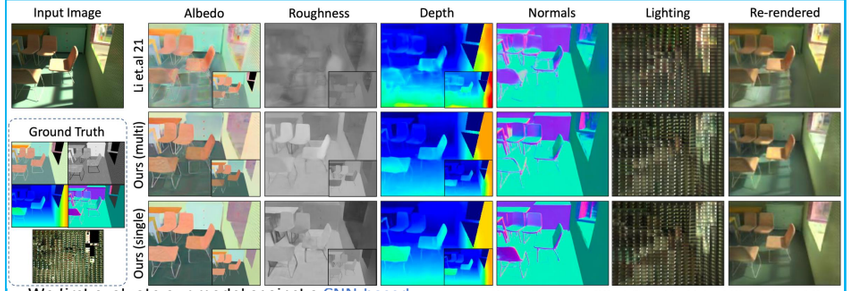
	single-6	single-4	multi	Li'20 [23]
Model Size (MB)	7,305	6,256	1,539	795
Inference (ms)	141.9	125.9	91.9	45.2
A+R+D+N	6.00	6.08	6.44	7.65
L+I	12.14	12.85	12.54	18.72

Table 5. Analysis on multiple design choices: IRISformer (single-task with 6 or 4 layers in **BRDFGeoNet**, multi-task with 4 layers), and CNN-based architecture from Li *et al.* [23] on OR [23].

[1] Li et al., 2020, Inverse Rendering for Complex Indoor Scenes
[2] Li et al., 2021, OpenRooms
[3] Garon et al., 2019, Fast spatially-varying indoor lighting
[4] Barron et al., 2013, Intrinsic scene properties
[5] Gardner et al., 2017, Learning to predict indoor illumination
[6] Li et al., 2018, CGIntrinsics

References

Evaluation: synthetic images



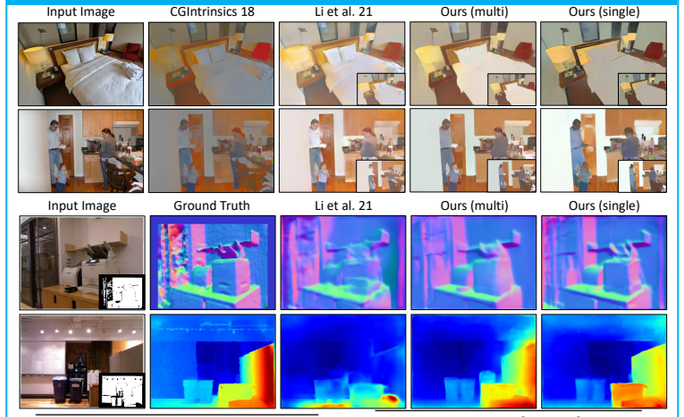
We first evaluate our model against a CNN-based baseline from Li et. al→.

In the above sample ↑ with strong highlights and complex scene geometry, we achieve much better estimations in all modalities, for example on the ground area and the chairs, our results are much more spatially consistent, with less artifacts and more details. We also achieve better decoupling of albedo from geometry and lighting.

Method	A↓	R↓	D↓	N↓	L↓	I↓	L+I↓
IRISformer (multi)	0.51	5.52	1.72	2.05	12.50	1.15	12.54
IRISformer (multi+BS)	0.51	5.50	1.71	2.05	12.47	1.15	12.58
IRISformer (single)	**0.43**	5.50	**1.42**	**1.89**	**12.04**	0.99	**12.14**
IRISformer (single+BS)	**0.43**	**5.48**	1.44	**1.89**	12.08	0.97	12.17
Ours (direct)	-	-	-	-	12.29	1.29	12.42
Li'21 [23]	0.52	6.31	2.20	2.61	18.63	**0.88**	18.72
Li'21+BS [23]	0.48	6.30	1.91	2.61	18.61	**0.88**	18.70

Table 1. Errors of BRDF, geometry and lighting with a base of 10^{-2} on OpenRooms [23]. Lower is better. For lighting estimation, **L** is the lighting reconstruction error, **I** is the rendering error and **L+I** is the combined lighting loss for which **LightNet** is trained.

Evaluation: real-world images



Method	Finetuning Datasets	WHDR↓
IRISformer (multi)	OR+IIW	13.1
IRISformer (single)	OR+IIW	**12.0**
Li'21 [23]	OR+IIW	16.4
Li'20 [19]	CGM+IIW	15.9
NIR'19 [33]	CGP+IIW	16.8
CGIntrinsics'18 [20]	CGI+IIW	17.5

Table 2. Intrinsic decomposition on IIW [4]. Lower is better.

Method	Mean(°)↓	Med.(°)↓	Depth↓
IRISformer (multi)	23.5	16.3	0.162
IRISformer (single)	**20.2**	**13.4**	**0.132**
Li'21 [23]	25.3	18.0	0.171
Li'20 [19]	24.1	17.3	0.184
NIR'19 [33]	21.1	16.9	-
Zhang'17 [47]	21.7	14.8	-

Table 3. Normal (mean and median) and depth (mean on inverse depth) prediction results on NYUv2 [34]. Lower is better.

Evaluation: object insertion/material editing

Jointly evaluate geometry and lighting via rendering virtual objects into the scene.

Sample 1: better highlights by ours, on the center object ↗. **Sample 2&3:** more globally spatially consistent across bunnies in multiple locations, in both the lighting intensities and directions→.

We also carry out an A/B study using the insertion results↓, and material editing ↘.

Gardner'17 [11]	Garon'19 [12]	Li'21 [23]	Ground Truth
0.24	0.30	0.47	0.58

Table 4. A user study on object insertion, where we compare IRISformer with each of the previous work or ground truth and report the percentage of feedbacks where other method is considered to be more photorealistic than ours.