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

Rui Zhu1, Zhengqin Li1, Janarbek Matai1, Fatih Porikli2, Manmohan Chandraker1
1UC San Diego 2Qualcomm AI Research

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 photo-realistic renderings in challenging lighting conditions compared to previous work [1,2] (lower left).

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

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

Datasets

- Training: our in-house developed OpenRooms dataset [2] for large-scale photo-realistic 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: IW dataset
  - geometry estimation (depth, normals) after finetuning: NYUV2 dataset
  - object insertion/material editing: natural images dataset from Garon et. al. [3]

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 better decoupling of albedos from geometry and lighting.

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.⇒

Evaluation: synthetic images

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&B: 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-L, and material editing⇒

Table 1: Error on BRDF rendering between ours and the ground truth.

References

[1] Gardner et al., 2017, Learning to predict indoor illumination
[2] Li et al., 2021, Fast spatially-varying indoor lighting
[3] Garon et al., 2021, Learning to predict indoor illumination
[4] Li et al., 2018, Commixtes

Table 5. Analysis on multiple design choices: IRISformer (single) vs. BRDFGeoNet, multi-task vs. single-task, and CNN-based architectures from Li et al. [17] or CGIntrinsics [18].