# 3D Reconstruction from Monocular Video with Spatially-Varying Lighting

**Aniket Gupta**

Northeastern University
Boston, MA
gupta.anik@northeastern.edu

**Dhruv Bansal**

Northeastern University
Boston, MA
bansal.d@northeastern.edu

## 1 Abstract

3D scene reconstruction is an essential yet challenging problem in Computer Vision which offers applications in the domains of planning, navigation and Augmented Reality. Existing methods for this problem estimate the 3D point cloud from multiple depth images or by directly reconstruct local surfaces represented as sparse TSDF volumes for each video fragment. While these methods give great results, one missing component in these approaches is the lighting condition estimation. For realistic Augmented Reality effects, we want to have consistent rendering with the environmental lighting conditions. For example, casting shadows for a newly introduced object in this generated 3D world consistent with environmental lighting. The goal of this project is to augment current state of the art approaches in 3D reconstruction to estimate lighting condition. We also aim to keep the approach computationally inexpensive to make it useful for real world applications.

## 2 Relevant work

In 3D computer vision, 3D scene reconstruction is one of the central ideas. It has been shows to greatly benefit image understanding tasks like semantic segmentation and human action recognition. Moreover, for applications in Augmented Reality tasks, accurate 3D scene is a primary requirement to produce a realistic and immersive experience. Most of the current state of the art approaches for 3D scene reconstruction adopt the depth fusion approach. In (1) Depth maps from single view key frames are generated first using depth estimation approaches. These depth maps are later converted into point clouds and fused into a Truncated Signed Distance Function (TSDF). The reconstructed mesh is than extracted from the TSDF volume. This pipeline has two drawbacks. First being that the overlapping view between different key frames is computed many times causing redundant computation which makes the model slow. Second, since for each key frame a depth map is individually estimated, the scale factor may vary which can cause a scattered output when fusing them together into the TSDF volume. (2) proposes a novel approach to directly reconstruct local surfaces represented as sparse TSDF volumes for each video fragment by a neural network. In this approach, a learning based TSDF fusion module is used to guide the network to fuse features from the previous fragments.

While these methods produce great results in 3D scene reconstruction, they do not estimate the lighting conditions of the environment. Estimating lighting conditions is yet another challenging problem. Given only the observed pixel values, the problem of estimating the geometry of the objects and their interaction with lighting is difficult. Some of the existing learning-based methods (3) use 2D CNNs and formulate this problem as image to image translation where lighting is represented as spherical Gaussian. But these approaches lack one degree of freedom (depth). To improve on this, (4) proposes a Volumetric Spherical Gaussian 3D representation for lighting, which is a voxel representation for the scene surfaces. Spherical Gaussian parameters in each voxel control the emission direction and sharpness of the light source, which captures view-dependent effects and can handle strong directional lighting.

With (4) the lighting estimation problem was solved with end to end training. Another approach (5) also solves the light estimation problem but rather than estimation spherical gaussian they estimated spherical harmonics which indeed ignores spatially varying effect as mentioned in (4). The approach implements point cloud generation from the input then from which Spherical Harmonics coefficients are estimated. The main advantage of PointAR is that it can be used on mobile devices but the data required as an input is not mobile cameras friendly. The input required for (5) is comprised of RGB-D images, mobile camera observation position and a panaromic view which is very much inconvinient when compared to (4) which only requires a normal RGB image.

Although, (6) proposed the estimation of SVBRDF, texture, shapes alongside the lighting estimation with just RGB image which is very unlikely to (5) as RGB images are easily accesible similar to (4), to estimate the Material and Geometry predictions alongside the lighting. With per pixel Spatially Varying Spherical Gaussian (SVSG) they are able to capture some high frequency effects. Also the rendering layer used is differentiable which makes generalization better which is required in the real world. With lighting estimation (6) allows us with object insertion with editing object material in the images and can be useful in augment reality applications. The pipeline proposed using different network architectures for different tasks such as use of MGNet for prediction of Albedo, Normal, Rough and Depth and LightNet for light estimation with a custom rendering layer ending using BSNet as a bilateral solver for refinement. With the help of (6) we can do 3D object insertion and scene editing in images.

## 3 Method

Our complete model consists of two separate branches for lighting estimation and 3D scene reconstruction. The scene reconstruction branch makes use of the model architecture used in (2) and the lighting estimation branch estimate the spherical gaussian parameters for light representation which can be used to generate a light map. Modern renderers can directly take Spherical gaussians as input for light and thus we can render the complete 3D reconstructed model with estimated light.

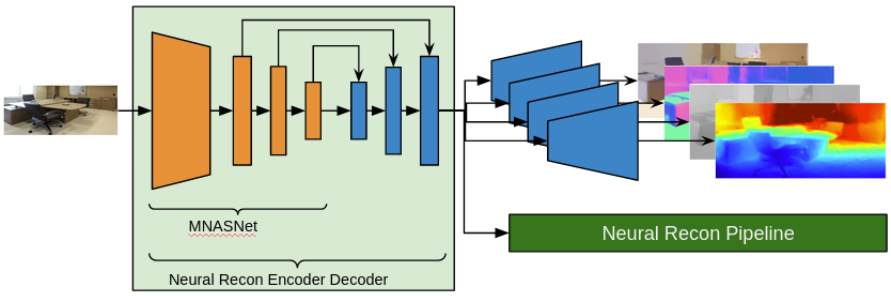


Figure 1: Model Architecture

### 3.1 Light Estimation

We use the output of encoder used in (2) as a input to the lighting estimation branch which has 4 decoders to produce albedo, normal, roughness and depth maps. The output of these four maps will be concatenated and passed through the LightNet model as in (6) which will use the pretrained weights from the same paper and will be fine tuned later on complete dataset.

## 4 Experiments

For the method we are trying to implement we are required with RGB image, depth maps, camera poses: intrensic and extrensic for the training. Once our proposed final model is trained it can be used with just a single RGB image.

Table 1: Quantitative comparison of Mean error values

| | Li. et al (6) | Ours |
|---|---|---|
| Albedo | 1.16 | 3.98 |
| Normal | 4.51 | 8.52 |
| Depth | 7.20 | 9.25 |
| Roughness | 1.70 | 4.33 |

### 4.1 Datasets

The dataset which covers all are requirements is the InteriorNet: Mega-scale Multi-sensor Photo-realistic Indoor Scenes Dataset. We don't have access to that yet for which we have already mailed the dataset owners to provide us with the dataset. Any other dataset other than InteriorNet have one or the other things missing from the above mentions. Some of the datasets which are helpful to bring us close to our final goal are mentioned below:

#### 4.1.1 ScanNet(V2) Dataset

ScanNet is an RGB-D video dataset containing 2.5 million views in a total of 1613 scenes, annotated with 3D camera poses, surface reconstructions, and instance-level semantic segmentation labels. If we are unable to get access to the InteriorNet dataset in time, the plan is to run the Scannet dataset through some state of the art model to generate the required outputs and use that as ground truth for a proof of concept.

#### 4.1.2 OpenRooms Dataset

The dataset is divided on the basis of different combinations which are different sets of camera views, same lighting but different materials, same material different lighting and a combination of different materials and lightings. This dataset is the closest we have to InteriorNet, it consists of images, materials, geometry, masks, SVLighting, SVSG, Shading, Environment maps with Direct illumination, Direct Shadings, Sementic Segmentation Labels, Light Source information and friction coeffecient.

### 4.2 Metrics

(2) considers F-score as the most suitable metric to measure 3D reconstruction quality since both the accuracy and completeness of the reconstruction are considered in the final result. In the scope of this report, we have not made any changes to the 3D reconstruction branch and hence we have assumed their evaluation to be correct. For the lighting estimation branch, we were only able to train our model on a very small batch of the complete dataset (about 10GB out of the total 110 GB dataset) so we have not run it through a test set yet as the model is bound to underfit. But for future work, we plan to use scale invariant L2 error for albedo (A), scale invariant log2 error for depth (D) and L2 error for normal (N) and roughness (R). The mean error values for all four decoders as of now are listed in Table 1. The values are compared against the results obtained by (6) for a clear picture of how our model performed in the initial stages.

## References


[1] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "Demon: Depth and motion network for learning monocular stereo," 12 2016.

[2] J. Sun, Y. Xie, L. Chen, X. Zhou, and H. Bao, "Neuralrecon: Real-time coherent 3D reconstruction from monocular video," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 15 593–15 602, 2021.

[3] M. Garon, K. Sunkavalli, S. Hadap, N. Carr, and J.-F. Lalonde, "Fast spatially-varying indoor lighting estimation," 2019.

[4] Z. Wang, J. Philion, S. Fidler, and J. Kautz, "Learning Indoor Inverse Rendering with 3D Spatially-Varying Lighting," 2021. [Online]. Available: http://arxiv.org/abs/2109.06061

[5] Y. Zhao and T. Guo, "Pointar: Efficient lighting estimation for mobile augmented reality."

[6] Z. Li, M. Shafiei, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker, "Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image."