Learning to Generate 3D Training Data through Hybrid Gradient Supplementary Material

Dawei Yang[1,2]
[1]University of Michigan
ydawei@umich.edu

Jia Deng[2]
[2]Princeton University
jiadeng@cs.princeton.edu

A. Metrics

Here we detail the metrics that we used in the paper. Assume $\mathbf{n}_i$ and $\mathbf{n}_i^*$ are the unit normal vector at i-th pixel (of N total) in the prediction and ground truth normal maps, respectively. d_i and d_i^* are depth values of the i-th pixel in the prediction and ground truth depth maps, respectively.

- Mean Angle Error (MAE): $\frac{1}{N}\sum_i \arccos(\mathbf{n}_i \cdot \mathbf{n}_i^*)$
- Median Angle Error (MAE): $\underset{i}{\text{median}}[\arccos(\mathbf{n}_i \cdot \mathbf{n}_i^*)]$
- Threshold δ: Percentage of $\mathbf{n}_i$ such that $\arccos(\mathbf{n}_i \cdot \mathbf{n}_i^*) \leq \delta$
- Mean Squared Error (MSE): $\frac{1}{N}\sum_i [\arccos(\mathbf{n}_i \cdot \mathbf{n}_i^*)]^2$
- Absolute Relative Difference: $\frac{1}{N}\sum_i |d_i - d_i^*|/d_i^*$
- Squared Relative Difference: $\frac{1}{N}\sum_i (d_i - d_i^*)^2/d_i^*$
- RMSE (linear): $\sqrt{\frac{1}{N}\sum_i (d_i - d_i^*)^2}$
- RMSE (log): $\sqrt{\frac{1}{N}\sum_i (\log d_i - \log d_i^*)^2}$
- RMSE (log, scale-invariant): $\sqrt{\frac{1}{N}\sum_i (\log d_i - \log d_i^* \cdot [\frac{1}{N}\sum_i (\log d_i - \log d_i^*)])^2}$

B. MIT-Berkeley Intrinsic Image Dataset

Our decision vector β for PCFG is a 29-d vector, with 4 dimensions representing the probabilities of sampling different primitives, 2 for sampling union or difference, 1 for whether to expand the tree node or replace it with a terminal, 6 for translation mean/variance, 6 for scaling log mean/variance, 2 for sphere radius log mean/variance, 2 for box length mean and variance, 4 for cylinder radius and height log mean/variance, and 2 for tetrahedron length log mean/variance.

For optimizing β, we use the mean angle error loss on the validation set as the generalization loss. Note that some dimensions of β are constrained (such as probability needs to be non-negative), so we simply clip the value of β to valid ranges when sampling near β for finite difference computation and updating β. We present the qualitative results in Fig. S.1.

C. NYU Depth V2

The decision vector β is 108-d. It includes the parameters for mixtures of Gaussians/Von Mises for 6 degree-of-freedom (vertical, horizontal and fordinal displacement, yaw, pitch, roll rotation) for shapes in the scene and the camera. Each mixture contains 9 parameters (3 probabilities, 3 means and 3 variances). Examples of perturbed scenes and the original scenes are shown in Fig. S.2. The distributions for translation perturbation of shapes are shown in Fig. S.5.

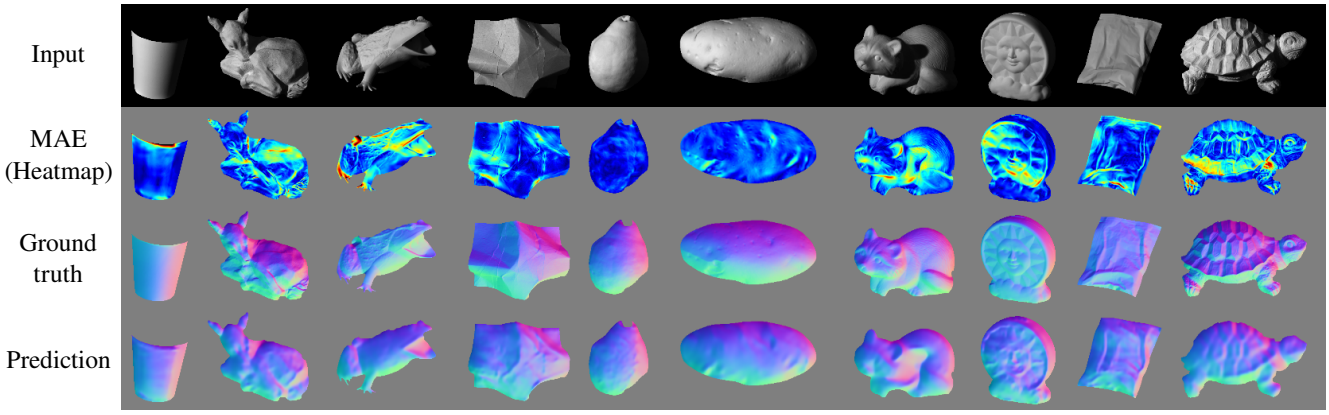


Figure S.1: The test set of the MIT-Berkeley Intrinsic Images dataset.

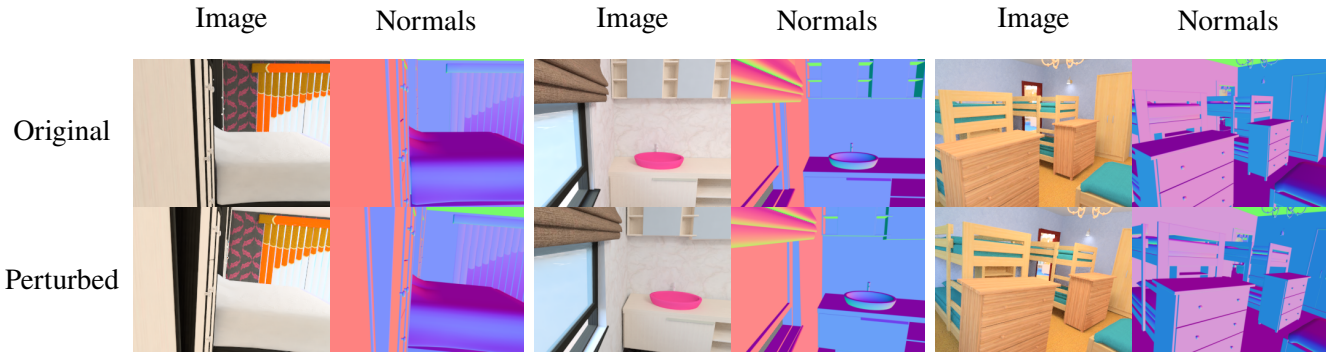


Figure S.2: The original scenes in the SUNCG dataset, and our scenes with camera and objects perturbed using our PCFG.

D. Basel Face Model

The decision vector β has 204 dimensions. We use an off-the-shelf 3DMM implementation[1] to generate face meshes and textures for training. The 3DMM has 199 parameters for face identity, 29 for expression and 199 for texture. In implementation, we use 10 principal dimension for face/expression/texture respectively, and randomly sample from a mixture of 3 multivariate Gaussians. Note that the dimensions are independent, so we have a total of 183 parameters for generating the face mesh. For the 3-dof face pose angle, we also use mixtures of 3 von Mises, which have 21 parameters in total.

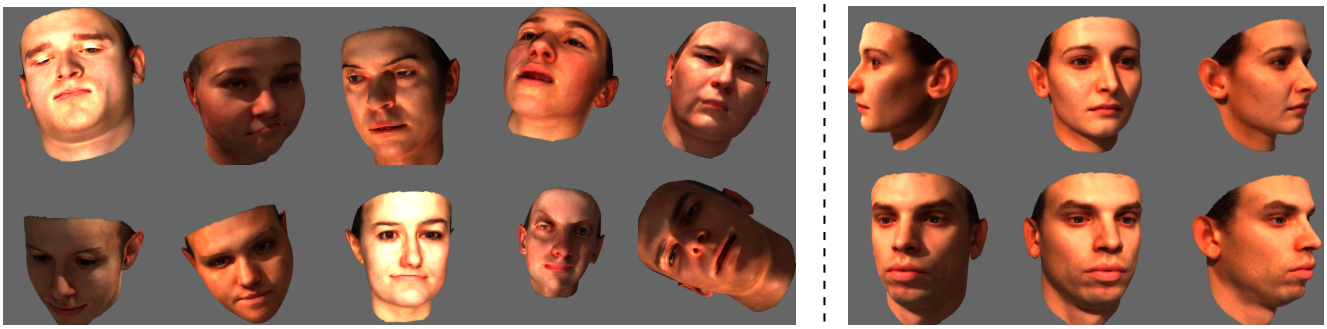
For rendering the training set, we apply a human skin subsurface model using Blender [1], with a random white directional light uniformly distributed on $-z$ hemisphere. For rendering the test set (the scanned faces in the Basel Face Model), we render with the same 3 lighting angles and 9 pose angles, and the same camera intrinsics as in the original dataset. Fig. S.3 shows the training images randomly generated by the PCFG (left) and example test images(right).

E. Synthetic Texture Generation for Intrinsic Image Decomposition

The decision vector β has 36 dimensions. To sample a texture, we first sample the number of polygons using a zero-truncated Poisson distribution. For each polygon, we then sample the number of vertices (from 3 to 6) according to the probabilities specified in β. The vertex coordinates of the polygon follow mixed truncated Gaussians. The polygons are then perturbed using Perlin noise: we first build the signed distance map to the boundary of the polygons, and then perturb the distance using Perlin noise. Finally, we re-compute the boundary according to the distance map to produce the perturbed polygons. These perturbed polygons are then painted onto a canvas to form a texture. This procedure is shown in Fig. S.4.

For rendering, we assume the shading is greyscale, and set up a random white directional light. We use Blender [1] to render the albedo image and the shading image, then multiple the two images together as the final rendered image. We render SUNCG [3] shapes using our synthetic textures for training, and render ShapeNet [2] with original textures for validation and

[1] https://github.com/YadiraF/face3d



Synthetic faces generated from the PCFG | Re-renderings of real scans

Figure S.3: Training images generated using PCFG with 3DMM face model, and 6 example images from the test set.

test. The examples are also shown in Fig. S.4.

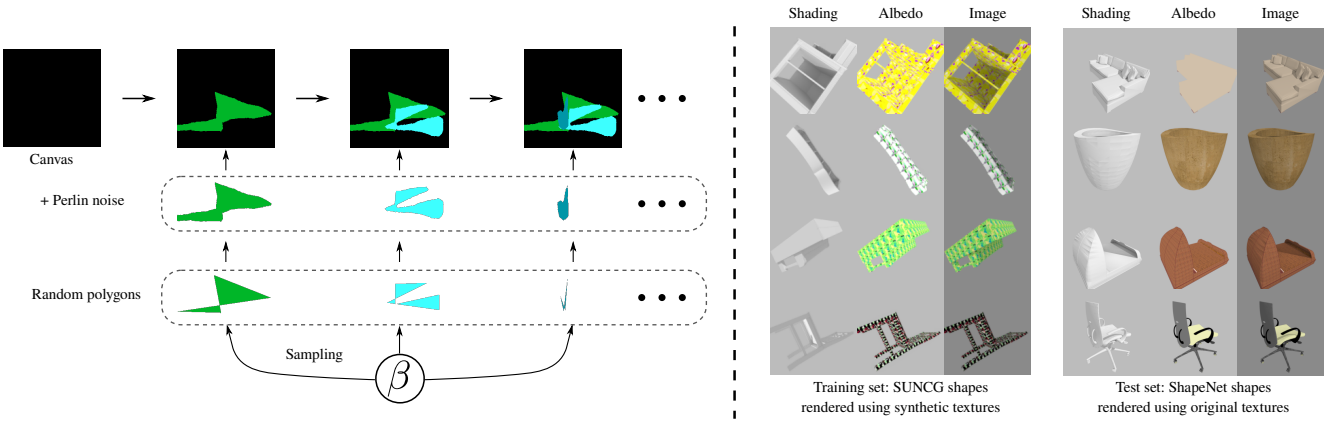


Figure S.4: Our texture generation pipeline and example images of the training and test set.

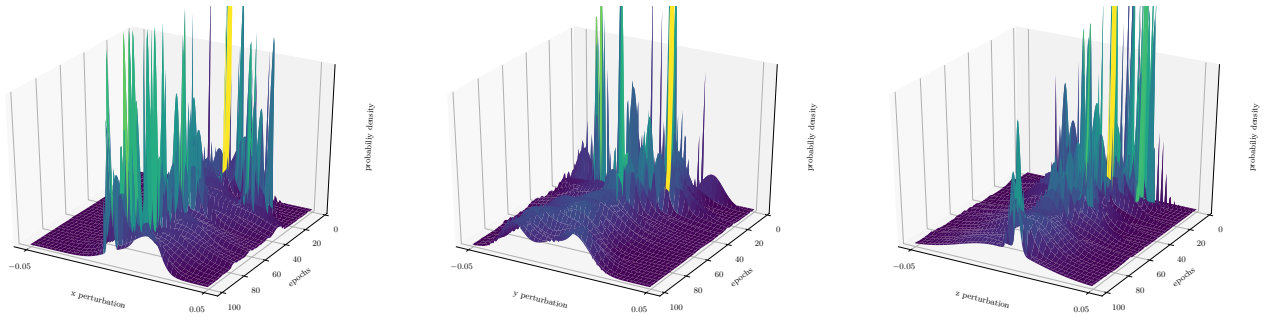


Figure S.5: How probability distributions change over time for SUNCG perturbation parameters. The three images plot the probability densitity of shape displacement along x, y, z axes respectively.

Acknowledgments This work is partially supported by the National Science Foundation under Grant No. 1617767.

References

[1] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam, 2019. 2

[2] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015). 2

[3] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. *Proceedings of 29th IEEE Conference on Computer Vision and Pattern Recognition*, 2017). 2