Towards Neural Representations of Heterogeneous Translucent Voxelised Media

Tom Gillooly[1]
thomas.b.gillooly@ntnu.no
Jon Y. Hardeberg[1], Abhijeet Ghosh[2], G. Claudio Guarnera[3]

[1] Norwegian University of Science and Technology
[2] Imperial College London
[3] University of York

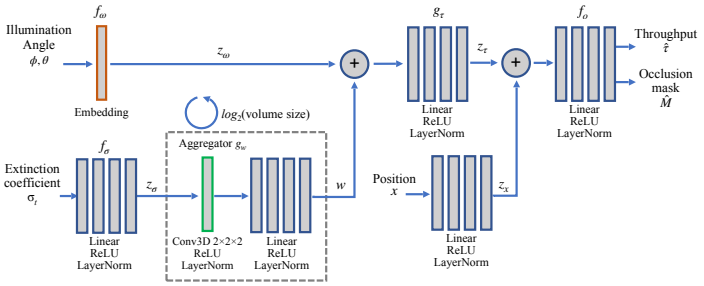
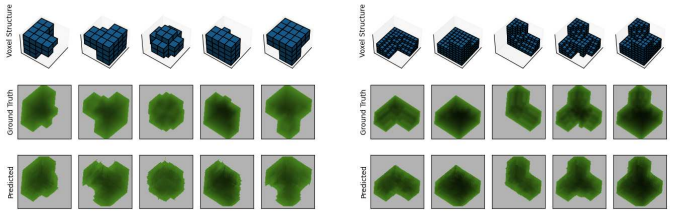


Figure 1: Full neural model. Vector concatenation is indicated by "+".



(a) Cube size 4, ρ_c=0.961, wMAPE=0.049 (b) Cube size 8, ρ_c=0.920, wMAPE=0.093

Figure 2: Neural renders of voxel structures at various scales.

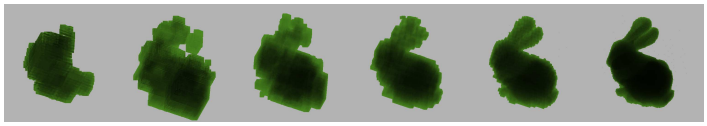


Figure 3: Neural rendering of a voxelized Stanford bunny. From left to right: full volume rendering with one feature vector, then progressively splitting the volume in half and using a feature vector per sub-volume.

When rendering homogeneous media, the log transmittance is the product of the extinction coefficient and the path length. Therefore, a reduction in path length can be offset by an increase in the extinction coefficient to maintain the same transmittance. We refer to this increased extinction coefficient as the equivalent optical parameter. However in heterogeneous structures, depending on the incident angle and position, the path may encounter varying optical parameters. Therefore, an equivalent optical parameter can no longer be expressed as a single scalar value; instead it must be represented as a function of material composition, incident angle, and incident position. Existing work on heterogeneous voxelised structures retains the full voxel grid in feature space [3], while other research on neural network-based rendering of translucent media relies on handcrafted volume representations [1]. In contrast, we learn a single vector representation by progressively aggregating latent optical parameter representations of a voxelised structure and train a neural rendering pipeline to convert these representations into throughput values.

Specifically, we individually encode raw optical parameters in a 3D structure $z_\sigma = f_\sigma(\sigma_t) \in \mathbb{R}^{N\times N\times N\times D}$, which are then aggregated with a learned function g_w, implemented as a 3D convolutional kernel:

$$w_n = g_w(w_{n-1}) \quad w_0 := z_\sigma \tag{1}$$

Once the volume has been fully aggregated, the resulting latent vector can be conditioned on an encoded position value to produce both a throughput prediction $\hat{\tau}$ and an occlusion prediction $\hat{M}$:

$$\hat{\tau} = f_\tau(z_\tau, z_x)_0 \quad \hat{M} = f_\tau(z_\tau, z_x)_1 \tag{2}$$

where $z_\tau = g_\tau(z_\omega, w_n)$ is the latent volume representation conditioned on the angle of illumination $z_\omega = f_\omega(\phi, \theta)$, and $z_x = f_x(x)$ is a vector representing the query location on the surface of the volume. The occlusion prediction indicates whether the incident illumination has passed cleanly through the volume without intersecting any voxels. Model accuracy is evaluated with the Concordance Correlation Coefficient ρ_c [2] and Weighted Mean Absolute Percentage Error (wMAPE), the latter given by the ratio $\sum_{i=1}^{n}|A_i - F_i| / \sum_{i=1}^{n}|A_i|$.

Our dataset consists of a set of voxel grids with varying occupancy, where an increase in voxel count is matched with a decrease in extinction coefficient to give overall identical throughput. Our motivation is to ensure that structures with equivalent properties are present in the dataset so that this regularity can be learned. Morphological operations are then applied, thus leading to structures with greater complexity.

To render the different configurations, we ray trace a single voxel and query the throughput by hit location on voxel layouts of increasing scale (see Fig 2 for some examples). The model performs well in predicting occlusion and transparency relative to the ground truth for cube sizes of 1, 2, and 4, each of which are present in the training data. On unseen larger volumes the wMAPE score begins to deteriorate, and the model appears to produce output for a lower frequency version of the input volume, resulting in correct overall shape but a loss of detail.

To render larger and more complex volumes, we divide the volume into sub-volumes, render them separately, and then composite them into a final volume, as shown in Fig 3. In the rightmost image, we split the full volume into $2 \times 2 \times 2$ sub-volumes, resulting in only 104 unique latent vectors that can be queried in parallel for throughput values, rather than processing the full 64^3 individual voxels.

Learning equivalence across different structures. We generate a set of optically equivalent structures and visualise the latent vector z_τ using t-SNE [4]. The output for an example configuration is shown in Fig 4. As the structures are equivalent per-angle, we expect that they cluster by illumination angle in high-dimensional space. As the figure shows (top row), this is indeed the case for the three volume sizes in the training dataset. While larger volumes are mapped to a different cluster (bottom row), each equivalent structure for these larger volumes does show similar clustering. Therefore, the latent space exhibits learned periodicity.

[1] Simon Kallweit, Thomas Müller, Brian Mcwilliams, Markus Gross, and Jan Novák. Deep scattering: Rendering atmospheric clouds with radiance-predicting neural networks. *ACM Transactions on Graphics (TOG)*, 36(6):1–11, 2017.

[2] I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

[3] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2019.

[4] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

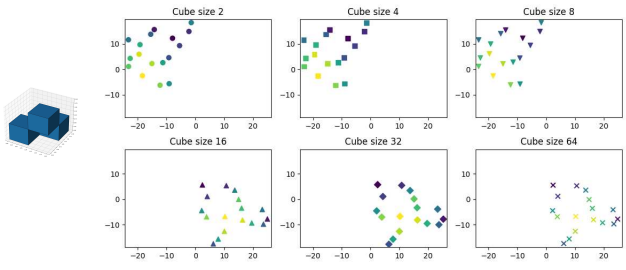


Figure 4: t-SNE plots of z_τ for optically equivalent structures. Different colours correspond to different conditioning illumination angles.