Supplementary material for "DRWR: A Differentiable Renderer without Rendering for Unsupervised 3D Structure Learning from Silhouette Images"

Zhizhong Han [1 2] **Chao Chen** [1] **Yu-Shen Liu** [1] **Matthias Zwicker** [2]

1. Implementation details

1.1. Network architecture

The network we leverage for point cloud reconstruction from single images includes a 2D encoder and a 3D point cloud decoder. The 2D encoder is a CNN with 7 layers. The first layer has a 5×5 kernel with 16 channels and a stride of 2. Each of the remaining layers has 3 kernels and comes in pairs, where one layer in the pair has a stride of 2 and the other has a stride of 1. The number of channels grows by a factor of 2 after each strided layer. These convolutional layers are followed by two fully connected layers with a dimension of 1024. The 3D point cloud decoder has one fully connected layer with a dimension of 1024 and then predicts a point cloud. The point cloud of N points is predicted as a vector with a dimension of $3N$ (point coordinates).

1.2. Evaluation Metric

We evaluate the accuracy of predicted point clouds by comparing the predicted point clouds with the ground truth point clouds in terms of Chamfer distance (CD) defined below,

$$\begin{aligned} CD(S_1, S_2) =& \frac{1}{|S_1|} \sum_{p \in S_1} \underset{q \in S_2}{min} \|p - q\|_2 \\ &+ \frac{1}{|S_2|} \sum_{q \in S_2} \underset{p \in S_1}{min} \|q - p\|_2. \end{aligned} \tag{1}$$

where S_1 is the predicted point cloud and S_2 is the ground truth point cloud, p is a point on S_1 and q is a point on S_2, $|S_1|$ and $|S_2|$ are the numbers of points on S_1 and S_2, respectively.

To compare with voxel-based results in terms of Intersection over Union (IoU), we discretize the 3D space holding the predicted point clouds or the ground truth point clouds into a 3D voxel grid with a resolution of 32^3, where a voxel is set to 1 if it contains a point. Our results of IoU are computed by comparing the 3D grid voxelized from the predicted point cloud with the 3D grid voxelized from the ground truth point cloud.

1.3. Interpolated shapes

Our interpolated shapes are demonstrated in Fig.12 in our draft. We leverage interpolated features to generate interpolated shapes using a trained 3D point cloud decoder.

Specifically, we first randomly select two generated 3D point clouds, and use the two features which are the input to the 3D point cloud decoder in the generation of these two 3D point clouds to represent a start location and an end location in the feature space, respectively. Then, we compute the 7 interpolated features along the line connecting the start location and the end location. Finally, we generate 7 interpolated shapes using the trained 3D point cloud decoder with the corresponding interpolated features as input.

[1]School of Software, BNRist, Tsinghua University, Beijing 100084, P. R. China [2]Department of Computer Science, University of Maryland, College Park, USA. Correspondence to: Yu-Shen Liu <liuyushen@tsinghua.edu.cn>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.