DISN: Deep Implicit Surface Network for High-quality Single-view 3D Reconstruction

Weiyue Wang*, Qiangeng Xu*, Duygu Ceylan, Radomir Mech, Ulrich Neumann

Task and Motivations

Task: Generate a high-quality detail-rich 3D mesh from a single 2D image

Input Image — Reconstructed 3D Model

Motivations:

1. Mesh surface is the most desired 3D representation for 3D object reconstruction.

2. Voxels and point clouds have limited resolution. Modeling explicit surfaces (e.g. mesh) is constrained by fixed topology.

3. Generation loss such as Chamfer Distance (CD) and Earth-mover Distance (EMD) penalize mostly on overall shape. The details are usually missing.

Solutions

1. DISN predicts Signed Distance Function (SDF) for locations in the space. SDF has no topology restriction, no limitations on resolution. Then uses Marching cubes to generate meshes.

2. DISN uses the local features at the projected location, the global features, and the point features to predict the SDF. Global features preserve overall shape, local features preserve fine-grained details.

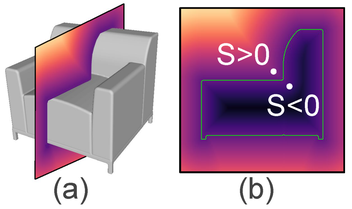
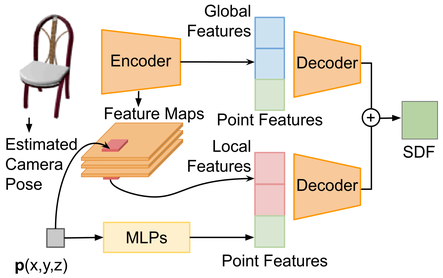


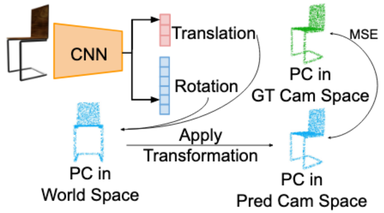
Illustration of SDF:
A point is outside the surface if s > 0, inside if s < 0, and on the surface if s= 0.



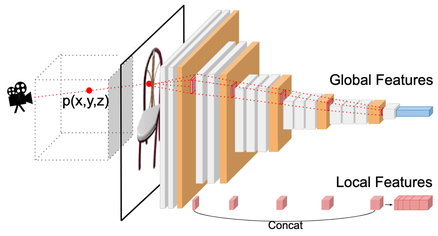
Model Overview

Camera Pose Estimation and SDF Generation

1. Estimate the camera pose, the 6 DoF transformation from the camera coordinate to world coordinate.
2. Extract local features: We use the camera pose to find a 3D point's 2D location on the image and extract local feature patches from multiple network layers.



Camera Pose Network. 'PC' means point cloud. 'GT and Pred Cam' denote the ground truth and predicted cameras.



Local Feature Extraction. Project 3D point p onto the image plane and extract features on the location from multiple layers.

Objective Functions

$$L_{cam} = \frac{\sum_{\mathbf{p}_w \in PC_w} ||\mathbf{p}_G - (\mathbf{R}\mathbf{p}_w + \mathbf{t}))||_2^2}{\sum_{\mathbf{p}_w \in PC_w} 1}$$

Camera Pose Estimation: $PC_w \in R^{N\times 3}$ is the point cloud in the world space. Each P_w, P_G is the ground truth location in the camera space.

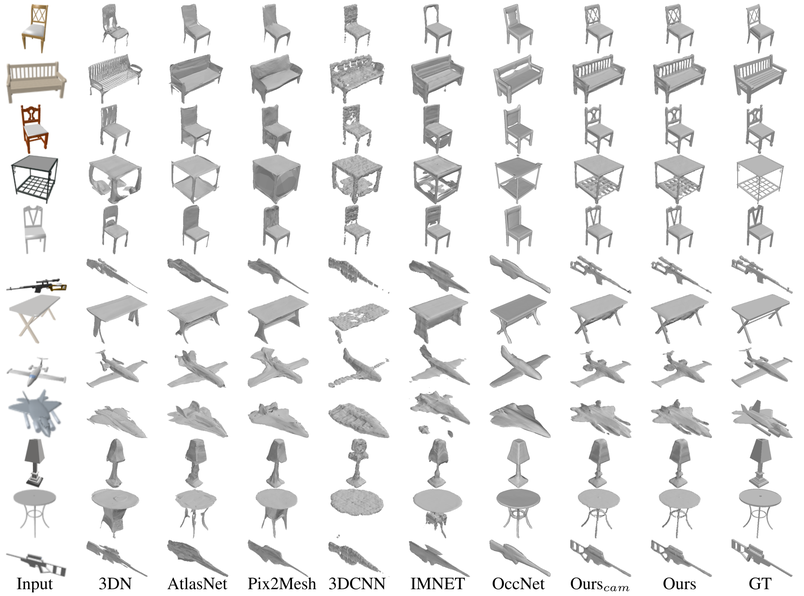
$$L_{SDF} = \sum_{\mathbf{p}} m|f(I,\mathbf{p}) - SDF^I(\mathbf{p})|$$

$$m = \begin{cases} m_1, & \text{if } SDF^I(\mathbf{p}) < \delta \\ m_2, & \text{otherwise,} \end{cases}$$

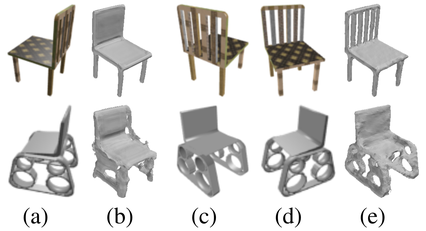
SDF Generation: $|\cdot|$ is the L1-norm. m_1, m_2 are weights for points whose signed distance is below or above a certain threshold δ. $m_1 > m_2$

Results and Evaluations

Ablation studies: 'GT' denotes ground truth, 'cam' denotes using estimated camera pose. 'Binary' denotes prediction of inside or outside the surface instead of a distance value. "One-stream" denotes concatenate global and local features instead of adding them in the network.



Qualitative results on ShapeNet single-view 3D reconstruction benchmark. DISN is the only one that can preserve both the overall shape and the fine-grained details.



DISN can be extended for Multi-view reconstruction:
(a): Single-view input.
(b): Reconstruction result from (a).
(c)&(d): Two other views input.
(e): Multi-view reconstruction results from (a), (c) and (d).

We train DISN on rendered images of ShapeNet and test it on real online images on the first row.