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.

$$EMD(PC, PC_T) = \min_{\phi:PC\rightarrow PC_T} \sum_{p\in PC} \|p-\phi(p)\|_2$$

$$CD(PC, PC_T) = \sum_{p_1\in PC} \min_{p_2\in PC_T} \|p_1-p_2\|_2^2 + \sum_{p_2\in PC_T} \min_{p_1\in PC} \|p_1-p_2\|_2^2$$

In the equations, PC and PC_T are the sampled point clouds from predicted and ground truth mesh.

Solutions

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

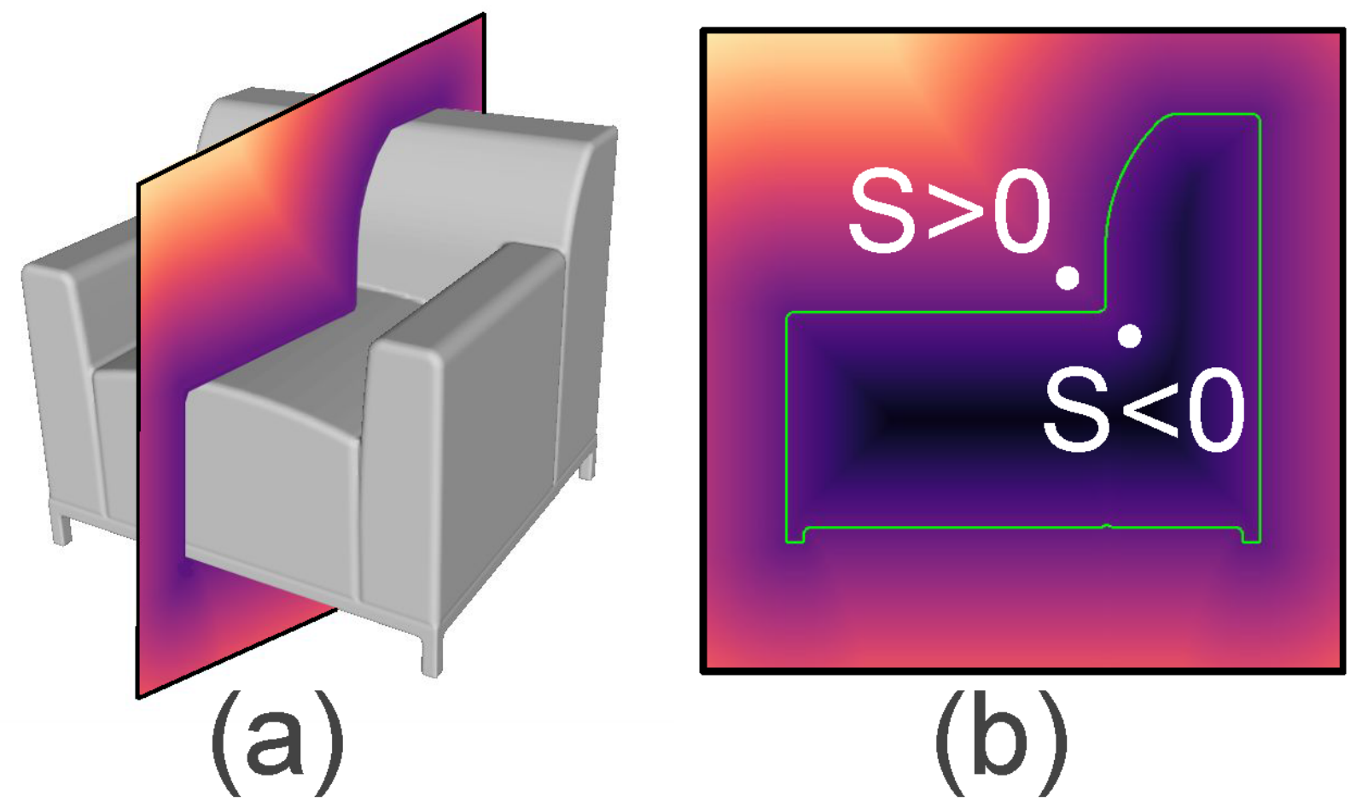
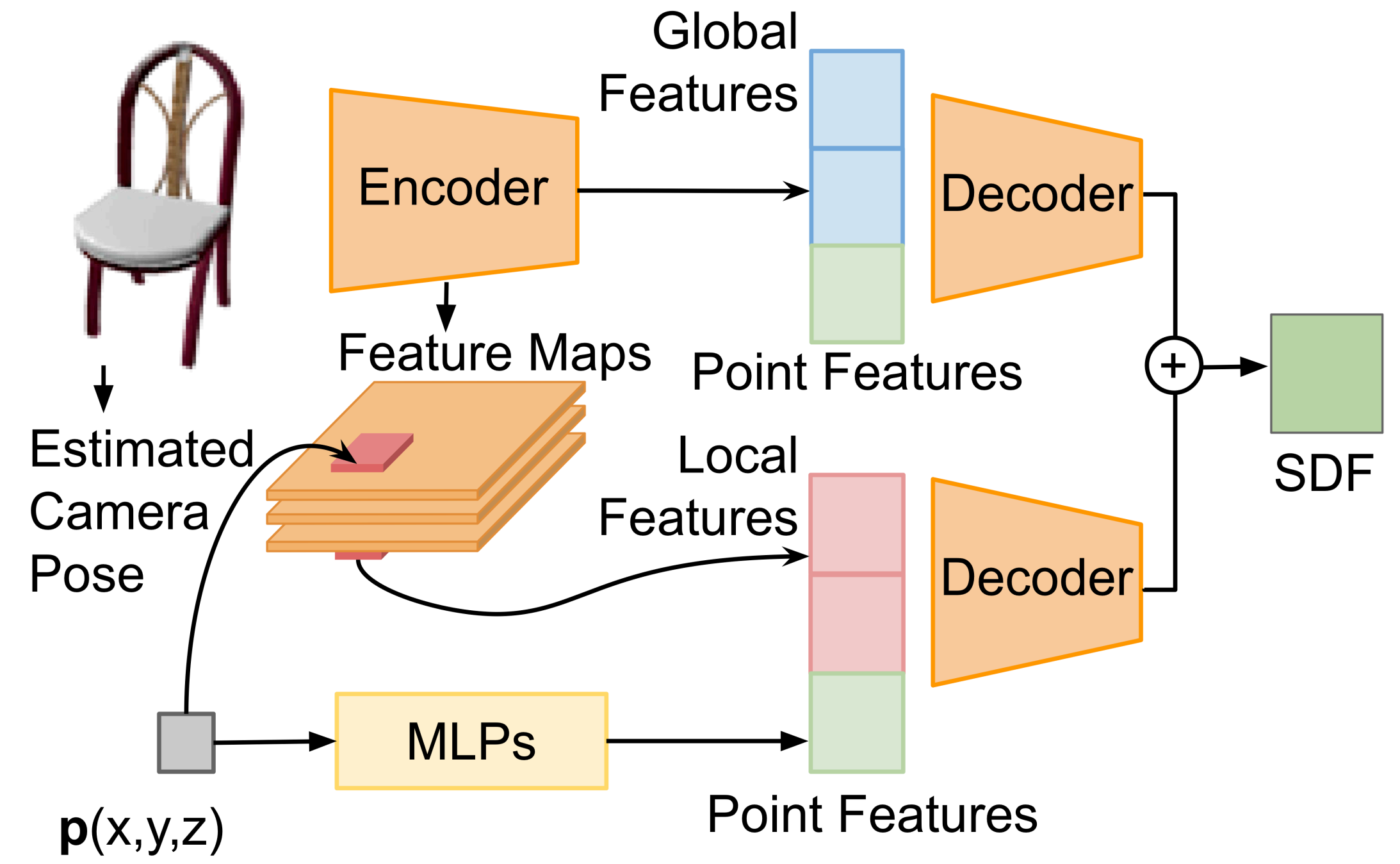


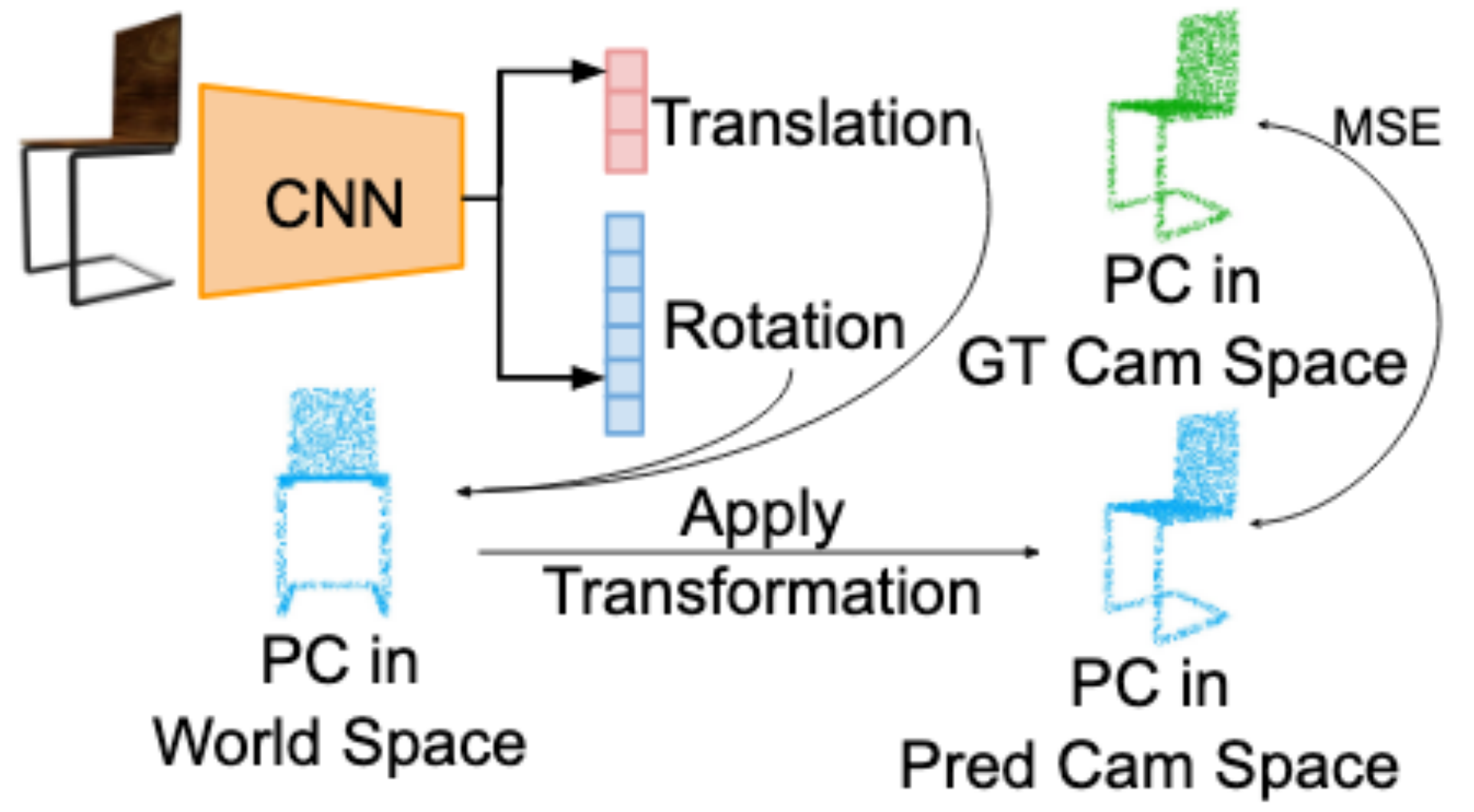
Illustration of SDF:
(a) Rendered 3D surface with S = 0.
(b) Cross-section of the SDF. A point is outside the surface if s > 0, inside if s < 0, and on the surface if s= 0.



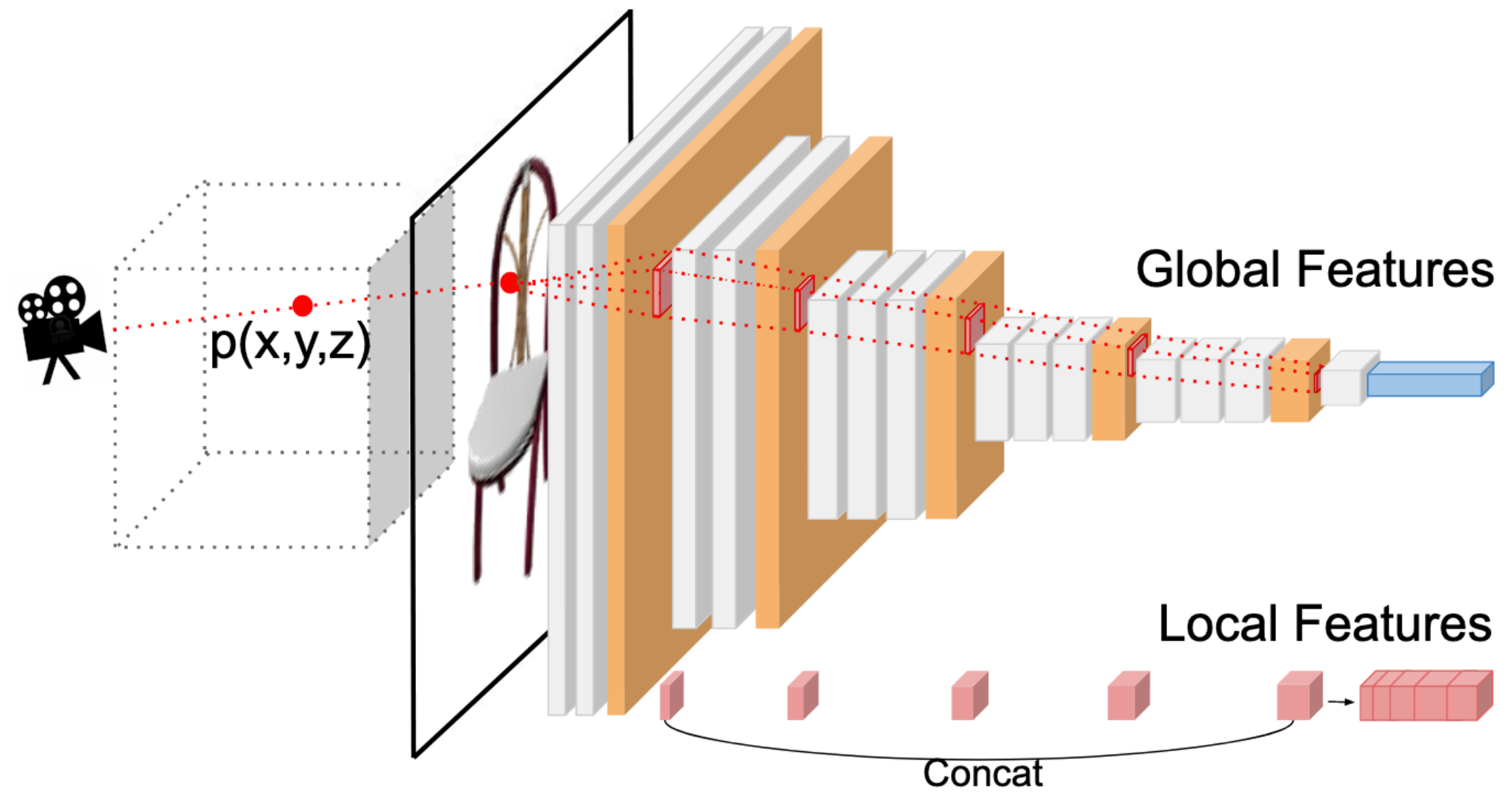
Model Overview: we use global, local and point features to predict SDF for many locations in the space.

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 Estimation Network. 'PC' denotes point cloud. 'GT Cam' and 'Pred Cam' denote the ground truth and predicted cameras.



Local feature extraction. We use the estimated camera parameters to project 3D point p onto the image plane. Then we identify the projected location on each feature map layer of the encoder.

Objective Functions

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

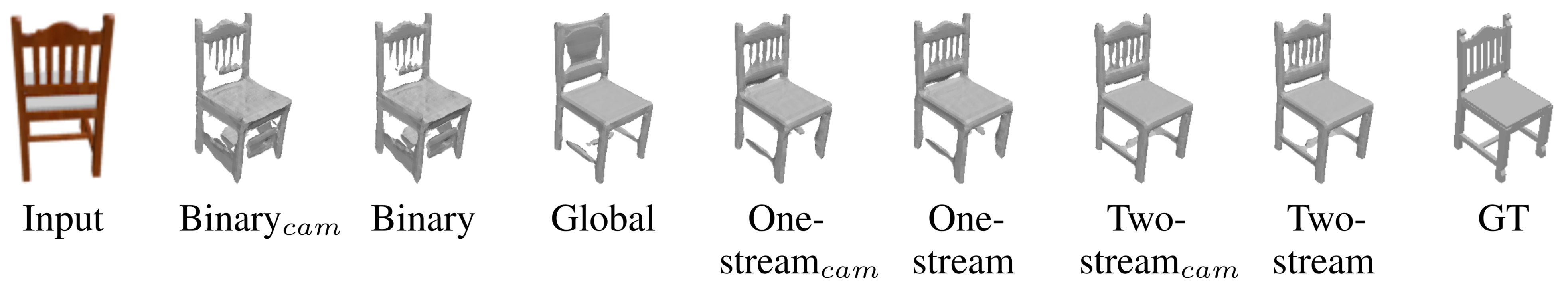
Camera Pose Estimation: $PC_w \in R^{N\times 3}$ is the point cloud in the world space, N is number of points in PC_w. For each $P_w \in PC_w$, P_G represents the corresponding ground truth point location in the camera space and $||\cdot||^2_2$ is the squared L 2 distance.

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

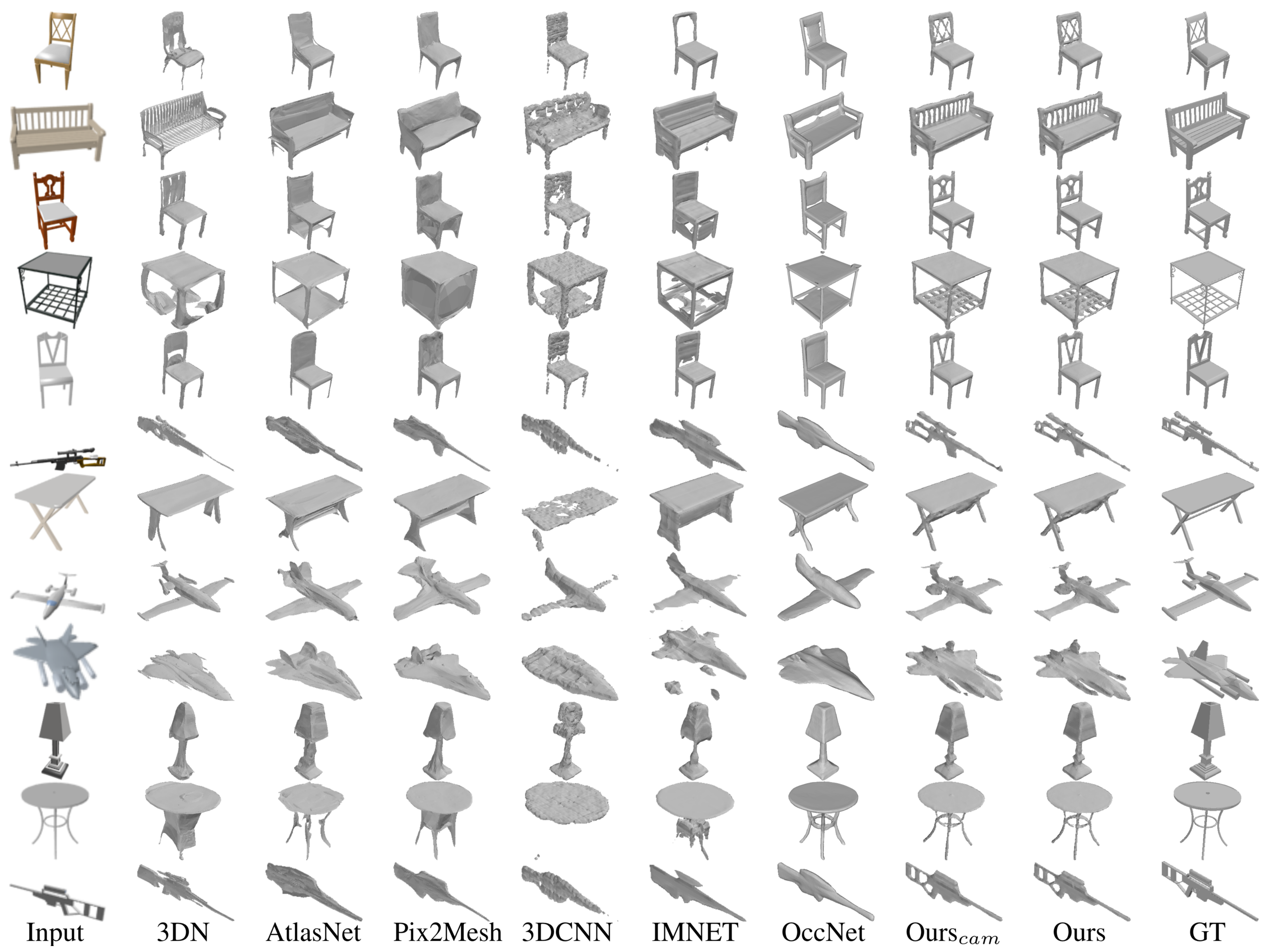
$$m = \begin{cases} m_1, & \text{if } SDF^I(\mathbf{p}) < \delta \\ m_2, & \text{otherwise,} \end{cases}$$

SDF Generation: |·| is the L1-norm. m_1, m_2 are different weights, and for points whose signed distance is below a certain threshold δ, we use a higher weight of m_1.

Results and Evaluations

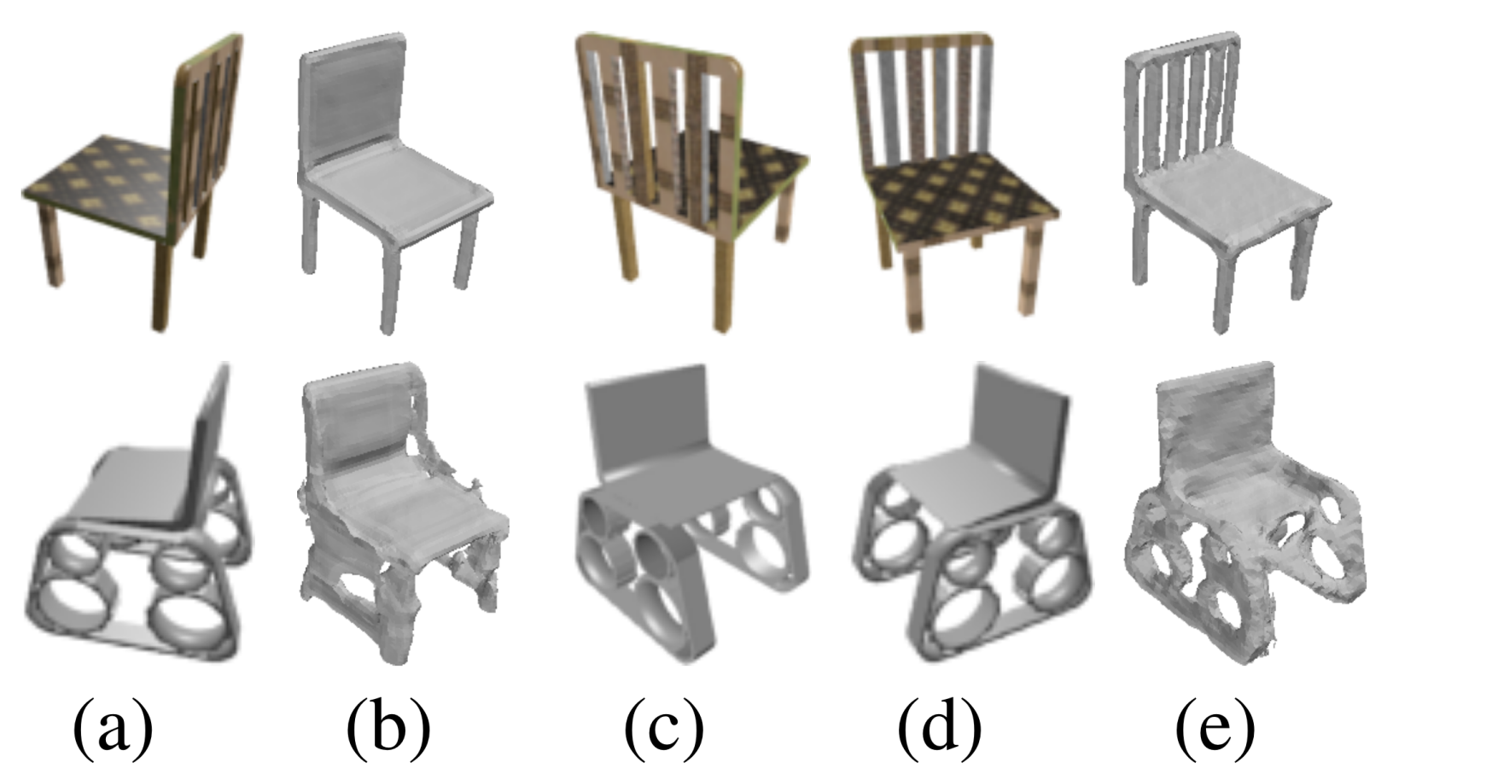


Ablation studies: 'GT' denotes ground truth shapes, 'cam' denotes models with estimated camera pose. 'Binary' denotes prediction of inside or outside the surface instead of a distance value. "One-stream" denotes concatenate global and local features instead of adding them in the network. DISN uses two-stream and predict real signed distance value to the object surface.

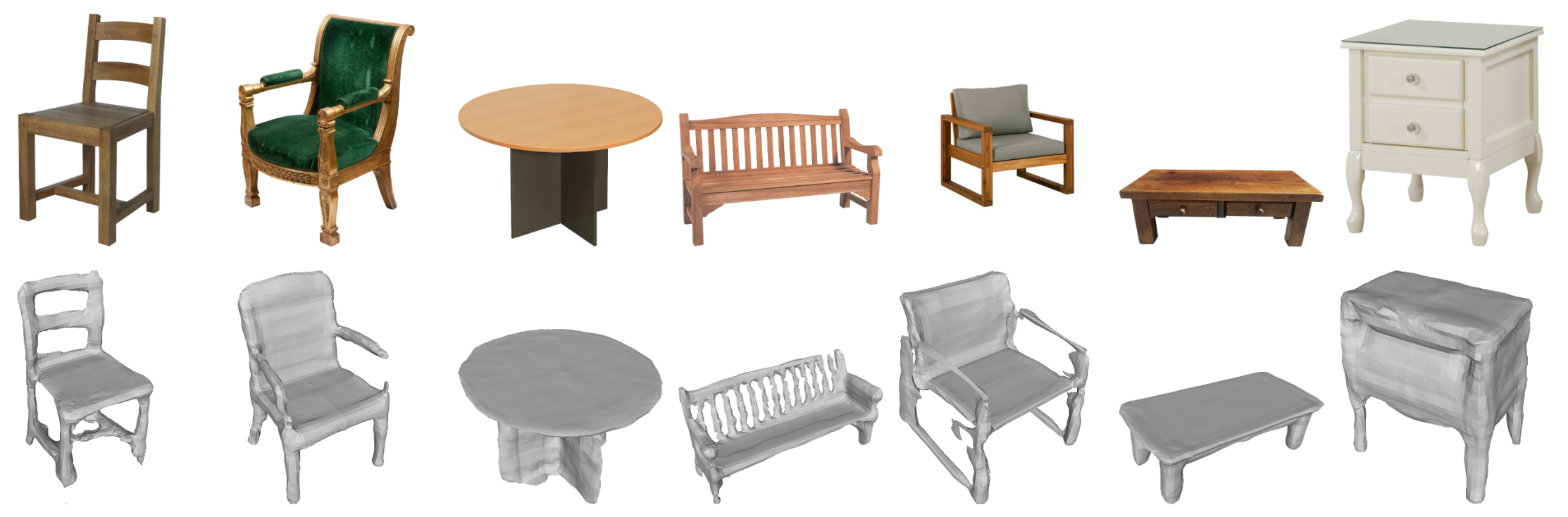


		plane	bench	box	car	chair	display	lamp	speaker	rifle	sofa	table	phone	boat	Mean
IoU	AtlasNet	39.2	34.2	20.7	22.0	25.7	36.4	21.3	23.2	45.3	27.9	23.3	42.5	28.1	30.0
	Pxl2mesh	51.5	40.7	43.4	50.1	40.2	55.9	29.1	52.3	50.9	60.0	31.2	69.4	40.1	47.3
	3DN	54.3	39.8	49.4	59.4	34.4	47.2	35.4	45.3	57.6	60.7	31.3	71.4	46.4	48.7
	IMNET	55.4	49.5	51.5	74.5	52.2	56.2	29.6	52.6	52.3	64.1	45.0	70.9	56.6	54.6
	3D CNN	50.6	44.3	52.3	**76.9**	52.6	51.5	36.2	58.0	50.5	**67.2**	50.3	70.9	**57.4**	55.3
	OccNet	54.7	45.2	**73.2**	73.1	50.2	47.9	**37.0**	**65.3**	45.8	67.1	**50.6**	70.9	52.1	56.4
	DISN$_{cam}$	**57.5**	**52.9**	52.3	74.3	**54.3**	**56.4**	34.7	54.9	**59.2**	65.9	47.9	**72.9**	55.9	**57.0**
	DISN	61.7	54.2	53.1	77.0	54.9	57.7	39.7	55.9	68.0	67.1	48.9	73.6	60.2	59.4

Quantitative and qualitative results on ShapeNet single-view 3D reconstruction benchmark



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



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