We thank the reviewers for the helpful feedback. All reviewers noted the novelty and superior performance of DISN, and the clarity of the exposition. We introduce a detail-preserving implicit surface network that generates high-quality 3D shapes by extracting local features. Although our quantitative results don't exceed existing methods by a large margin, we believe *the qualitative results (including 30 pages of figures in Supplementary)* are adequate to show that DISN achieves state-of-the-art performance on single-view reconstruction. We address major concerns raised by the reviewers.

R1.Q1: F-Score Comparison Please refer to Table 1 for F-Score results. We will add this experiment in our revision.

Threshold(%)	0.5%	1%	2%	5%	10%	20%
3DCNN	0.064	0.295	0.691	0.935	0.984	0.997
IMNet	0.063	0.286	0.673	0.922	0.977	0.995
DISN gt cam	0.079	0.327	0.718	0.943	0.984	0.996
DISN est cam	0.070	0.307	0.700	0.940	0.986	0.998

Table 1: F-Score for varying thresholds (% of reconstruction volume side length, same as "What Do Single-view 3D Reconstruction Networks Learn?"(Tatarchenko et al, CVPR 2019)) on all categories.

R1.Q2: Comparison with a baseline replacing local feature extraction with background subtraction Please refer to Figure 1. Compare to results from global features, the baseline can generate some holes. However, it suffers from inaccurate reprojection and noncontinuous SDF prediction.

R1.Q3: What global and local stream's sdf look like As shown in Figure 1, the global branch produces the overall shape and the local branch adjusts the sdf based on local details.

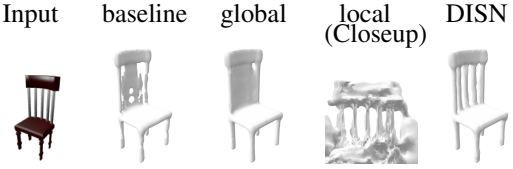

Figure 1: Baseline is as suggested by R1. We also show the reconstruction of DISN's global branch and the chair back details generated by DISN's local branch.

R1.Q4: Relations to Pix2Mesh Local Feature Module Pixel2mesh reconstructs 3D model by deforming an ellipsoid, which makes them impossible to produce different topology and generate different shape details. However, by taking advantage of predicting implicit surface, DISN is able to generate shape details according to local features.

R2.Q1: Statement of contributions is misleading Thanks for pointing this out. We will further emphasize local feature extraction as our main contribution in our revision.

R2.Q2: Add citations for the related voxel and point cloud methods We will conduct a more comprehensive literature review and also add the citations.

R2.Q3: Relations to "PIFu" Thanks for suggesting this related work which was released at the same time as NeurIPS 2019 submission deadline. We will discuss the relations of our paper and this paper in our final version.

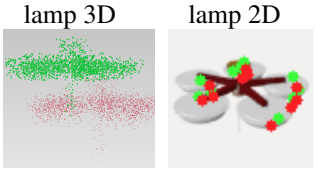

Figure 2: Green dots are sampled ground truth points. Red dots are projected points using estimated camera parameters. The lamp has an offset in 3d space.

R3.Q1: Quantitative results are not superior to previous methods by a large margin. 1) Although DISN doesn't outperform 'global' by a large margin quantitatively, the qualitative results illustrates that DISN is able to reconstruct various shape details that all methods without local feature extraction fail to generate as acknowledged by other reviewers. To the best of our knowledge, DISN is the first work that is able to generate fine-grained details in 3D shapes (such as holes in the rifle in our teaser). Moreover, we also provide a wide range of qualitative results in Supplementary, which sufficiently show our superior performance compared to state-of-the-art methods. 2) "What Do Single-view 3D Reconstruction Networks Learn?"(Tatarchenko et al, CVPR 2019) shows that the qualitative results are not necessarily related to the quality of generated shape details. Please also refer to R1.Q1 for F-score comparison where DISN constantly outperforms the state-of-the-art methods. In Figure 1 we also show the importance of our local feature extraction to the detail reconstruction.

R3.Q2a: Camera prediction for symmetric objects We show some examples of camera prediction of symmetric objects in Figure 2 that have large projection errors in 3D but small errors on 2D. 1) In most cases, incorrect camera prediction due to symmetric ambiguity has no impact on 2d reprojection, therefore no impact on local features query. 2) Even the 2d reprojected location has a shift of 2.95 pixels on average (Table 2 in our paper), this shift will be decreased when we query local features on higher-level feature maps with smaller dimensions.

R3.Q2b: Small Camera variance of Choy's Rendering Dataset Camera prediction is not our main contribution. To make a fair comparison, we use the same dataset as previous methods. In Choy's dataset, the distance and field of view are fixed instead of elevation and cyclorotation. The degree of freedom is 3 instead of 1. As suggested by the reviewer, we enlarge the camera variation by making the camera not to point towards the origin. We simply change C_x, C_y in intrinsic matrix from fixed to random numbers in $(-40, 40)$ and trained a new camera prediction network by predicting both extrinsic and intrinsic parameters. The reprojection error for this new setting is 3.54, while the reprojection error for "Choy" dataset is 2.95 (Table 2 in our paper). Therefore, our camera prediction method is robust to larger variation and higher DOF. We are also preparing a new dataset for 3D reconstruction by rendering with greater camera variation. We will train a new camera prediction network and add this experiment to our final version.

R3.Q3a: Confusing notations. We apologize for the confusion and will revise in the final version.

R3.Q3b: One stream with gt camera is worse than estimated. The ground truth and the estimated results should be swapped. We will correct the typo in the final version.