Multiview Aggregation for Learning Category-Specific Shape Reconstruction

We thank the reviewers for their valuable comments, and are happy to see feedback such as "concepts presented in the paper are solid" (**R1**), "well demonstrated ... proposed approach outperforms other 3D shape reconstruction methods" (**R2**), and "makes sense on an intuitive level" (**R3**). The reviewers agree that NOX maps are "definitely novel ... predicting object shape beyond whats visible in the image" (**R1**), "Being able to predict such a representation from a single image/images is of great significance" (**R2**). In addition to the positives identified by the reviewers, NOX maps provide strong 2D–3D correspondences, can support articulating object categories, encode camera pose implicitly (see supplementary), and allow multiview aggregation both at the NOCS and feature-level (using permutation equivariant layers). We answer questions, address factual errors, and present more details to improve our manuscript.

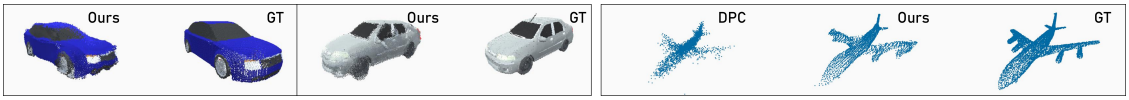
Experiments (R1, R2, R3): We will improve the clarity of the experiments section in the final version and add additional details from below.

First, we address **R1**'s questions and concerns. In Table 2, why does the quality worsen for multiview setting for cars compared to single view? We believe that this is the result of cars having a simpler convex shape that does not benefit from feature-level multiview aggregation. We observe significant improvements for more complex shapes like chairs and airplanes for the same setting. While probabilistically sampling an isosurface in 3D-R2N2 makes sense, we chose to sample the center of each voxel for easier comparison following prior work which does the same [13]. In Table 3, different categories having different variants that perform well is likely due to specific shape distribution for each category. Similar to Table 2, we observe that more complex shapes like airplanes or chairs benefit from more views and information aggregation at both the feature and NOCS level. While for simpler shapes like cars, feature-space aggregation benefits less. The differences in chairs between Table 4 and 5 are due to different experimental setting. In particular, one is a single view model and the other is trained on up to 5 views but evaluated on 1 view. In Table 3, the "Fixed Multi" models were trained with 2, 3, or 5 views respectively.

R2 questions the claims about feature-level aggregation in Tables 2 and 3. Our main claim is that adding more views improves test-time reconstruction (numbers get smaller from left to right in Table 3), not that training on more views improves performance (numbers do not always get smaller from variable 5 to 10). We will add this detail in the paper.

We disagree with **R3** that our reference methods do not allow fair comparison. The methods we compare against are the few methods that cover a subset of the problems we solve: variable multiview input at train/test time (3D-R2N2), and reconstruct complete 3D shape as a point cloud (3D-R2N2 and DPC). Our method further supports articulating categories and can encode camera pose. "Dated" methods are still valid prior art to compare against especially when they solve similar problems.

Qualitative Results (R1, R2, R3): Due to strict page limits, we did not include more qualitative results and comparisons but will include more like the images below to the final version.



Related Work (R1, R2, R3): Thanks for pointers to make our related work more exhaustive (we will add missing references and correct existing ones). Specifically, we will add a discussion about relationship and differences to geometry images, position maps, and back surface prediction. In short, the differences are direct 2D–3D correspondences, the ability to use existing CNN machinery, and implicit encoding of camera pose.

We thank **R3** for the reference to the review paper and to Matryoshka Networks which we will add. As mentioned in Figure 3, our approach is fully capable of representing the interior parts of an object far more compactly than Matryosha networks with only K layers. We can also trivially merge these layers using set union as opposed to recursive composition. We choose to only use the first and last intersections due to computational efficiency. Thus, we disagree with **R3**'s subjective characterization of our work as a "minor variation" of well-known concepts. Our independent discovery of a more compact encoding of shape suggests otherwise.

Other Questions: We do not employ any thresholding for the set union operation nor do we filter the estimated NOX maps. Our method is robust even without postprocessing but median and bilateral filter do improve (will add to supplementary document). In all experiments, we trained one network for each category as is normal practice but we also trained joint networks for all categories and are happy to include these results. We expect the results of "What Do Single-view 3D Reconstruction Networks Learn?" (**R1**, **R2**) to hold to our approach as we share similar encoder-decoder to other discussed work. However, we believe that we are learning an implicit distribution of shapes (albeit as NOCS maps) for each category and can model intra-categroy topology variation. The ground truth point set is the set union of the ground truth NOX maps which are sampled during the rendering process. Our multiview NOX nets are trained from scratch similar to the single-view nets.