Supplemental Material for Multi-View Point Regression Network for Single View Reconstruction

Anomymous AAAI submision
Paper ID 2760

In this supplemental material, we provide additional technical details, extra analysis experiments including more qualitative results and representation analysis to the main paper.

Proof of Quasi-volume Term

The volume discrepancy for continuous surface is defined in Equation 3. To discretize the volume discrepancy, we start from local parameterization, illustrated in Fig. A. Let $\mathbf{p}$ denote the parameterization of a triangle $\Delta = (\mathbf{x}_a, \mathbf{x}_b, \mathbf{x}_c)$. The parameterization of point $\mathbf{x}$ on Δ is given by $\mathbf{x}(\mathbf{p}) = \mathbf{x}_a + p_x\overrightarrow{\mathbf{x}_a\mathbf{x}_b} + p_y\overrightarrow{\mathbf{x}_a\mathbf{x}_c}$, which is a bilinear mapping $[0,1]^2 \mapsto \mathbb{R}^3$. By simple derivation, we get $\mathrm{d}\mathbf{x} = 2|\Delta|\mathrm{d}\mathbf{p}$. Here $|\Delta| = \frac{1}{2}||\overrightarrow{\mathbf{x}_a\mathbf{x}_b} \times \overrightarrow{\mathbf{x}_a\mathbf{x}_c}||_2$ is the area of Δ and $\mathbf{n}_\Delta = \frac{\overrightarrow{\mathbf{x}_a\mathbf{x}_b} \times \overrightarrow{\mathbf{x}_a\mathbf{x}_c}}{||\overrightarrow{\mathbf{x}_a\mathbf{x}_b} \times \overrightarrow{\mathbf{x}_a\mathbf{x}_c}||_2} = \frac{\overrightarrow{\mathbf{x}_a\mathbf{x}_b} \times \overrightarrow{\mathbf{x}_a\mathbf{x}_c}}{2|\Delta|}$ is the outward normal of triangle Δ. We consider a generic functional such that, $\mathcal{L}(\mathcal{S}) = \int_{\mathcal{S}} h(\mathbf{x}, \mathbf{n})\mathrm{d}\mathbf{x} = \sum_{t\in\mathcal{M}} |\Delta| \int_\rho h(\mathbf{x}, \mathbf{n}_\Delta)\mathrm{d}\mathbf{p}$, where $h(\mathbf{x}, \mathbf{n})$ is a function associated with point $\mathbf{x}$ and its normal $\mathbf{n}$, $\rho = \{\mathbf{p}|p_x \in [0,1] \text{ and } p_y \in [0, 1-p_x]\}$ denotes the local parametric domain. Then, the volume discrepancy in Equation 3 is discretized as:

$$\begin{aligned}\mathcal{L}_{vol}(\mathcal{M}, \widetilde{\mathcal{M}}) &= \sum_{\Delta\in\mathcal{M}} |\Delta| \int_\rho (\widetilde{\mathbf{x}} - \mathbf{x})\mathbf{n}_\Delta \mathrm{d}\mathbf{p} \\ &= \sum_{\mathbf{x}_k\in\mathcal{X}} ||\widetilde{\mathbf{x}}_k - \mathbf{x}_k||_2 \mathbf{n}_k \qquad \text{(A)}\end{aligned}$$

where $\mathbf{n}_k = \sum_{\Delta\in\Omega_k} |\Delta|\mathbf{n}_\Delta$ is the area-weighted normal at vertex $\mathbf{x}_k$, Ω_k denotes the 1-ring triangles of $\mathbf{x}_k$.

As stated in Section 3, $\mathcal{S}$ is partitioned into several visible surfaces $\mathcal{S}_i$ from multiple viewpoints $\mathbf{c}_i$, $i = 1, ..., N$. Equation A can be rewritten as an integral over the surface by counting only visible points.

More Reconstruction Result

More reconstruction results compared to Su et al.(Fan, Su, and Guibas 2017) and 3D-R2N2 (Choy et al. 2016) are shown in Fig. F. More qualitative results are shown in Fig. H and Fig. G.

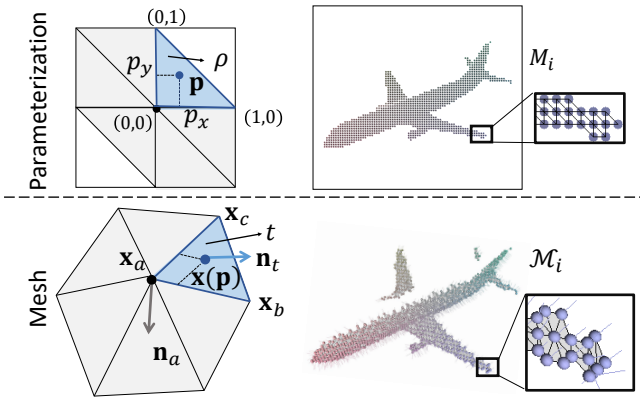

Figure A: Parameterization of discretized surface.

Representation Analysis

Discriminative representation. We attempt to show the discriminative ability of the learned latent space by classification. A linear SVM (Cortes and Vapnik 1995) is adopted to classify the learned 2048-dimension features z on the ShapeNet-13 testing dataset. The confusion matrix of the classification result is illustrated in Fig. B and the mean accuracy is about 92.7%, which demonstrates that the leaned features are discriminative with semantics. Also, we use t-SNE (Van Der Maaten 2014) to embed the learned features into a 2D space. Fig. E shows the embedding space. Models of similar shapes are grouped together, indicating that the proposed network encodes geometric information well.

Shape arithmetic. To demonstrate the learned feature space to be representative and discriminative on shape parts, we conduct an arithmetic experiment on the learned features. Three features are randomly selected and performed with the $A+B-C$ operation as the authors of TL-embeding (Girdhar et al. 2016) do. The resulting feature is then used to generate a new point cloud. Some typical results are shown in Fig. C.

Limitation

We also discovered some limitations of our method. The network tends to mimic the unseen models as the seen ones. We test the data that do not belong to the categories in ShapeNet-13, on which the network is trained. Fig. D shows some

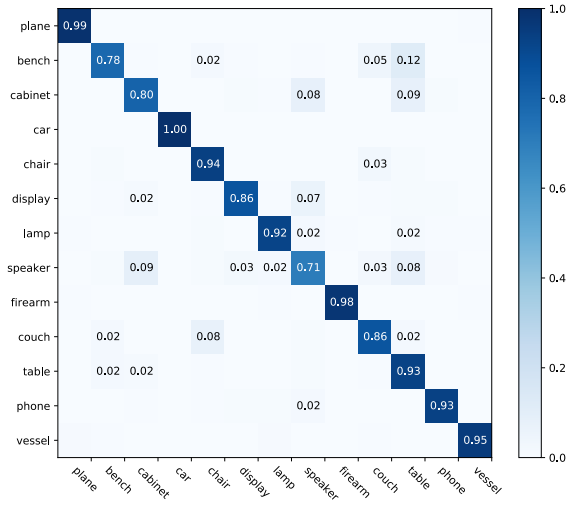

Figure B: Confusion matrix of classification on testing latent features.

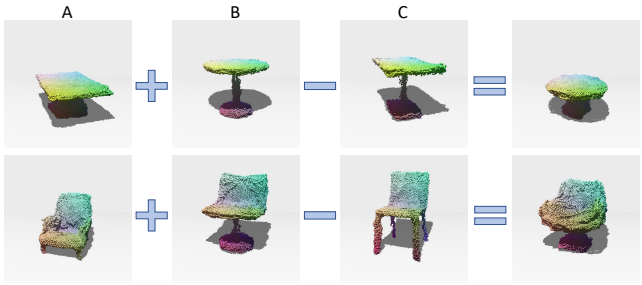

Figure C: Shape arithmetic results. In the first row, adding a round table with a thin leg to a rectangular table with a big leg and removing a rectangular table with a thin leg results in a round table with a big leg. In the second row, we obtain an easy chair with one leg by adding and removing a similar-looking chair with one and four legs respectively to an easy chair with four legs.

failure cases. For example, in (a) and (b), the bathtub is reconstructed as a couch and the mail box is recovered as a cabinet. Our network may focus on local details more than global shapes when the network has no idea of the shapes. The generated results shown in Fig. D (c) and (d) have some distorted surfaces because few ellipsoid structures exist in ShapeNet-13.

References

Choy, C. B.; Xu, D.; Gwak, J.; Chen, K.; and Savarese, S. 2016. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision (ECCV)*.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine learning* 20(3):273–297.

Fan, H.; Su, H.; and Guibas, L. 2017. A point set generation network for 3d object reconstruction from a single image. *IEEE International Conference on Computer Vision (ICCV)*.

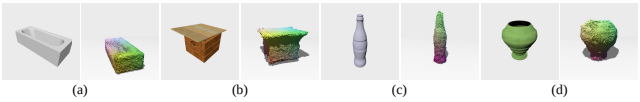

Figure D: Examples of failure cases. (a) The bathtub is reconstructed as a couch. (b) The mail box is recovered as a cabinet with four small legs. (c) and (d) are of distorted surfaces because few ellipsoid structures exist in ShapeNet-13.

Figure E: 2D embedding of learned latent features with t-SNE. Models of similar shapes are grouped together.

Girdhar, R.; Fouhey, D. F.; Rodriguez, M.; and Gupta, A. 2016. Learning a predictable and generative vector representation for objects. In *European Conference on Computer Vision (ECCV)*.

Van Der Maaten, L. 2014. Accelerating t-sne using tree-based algorithms. *Journal of machine learning research*.

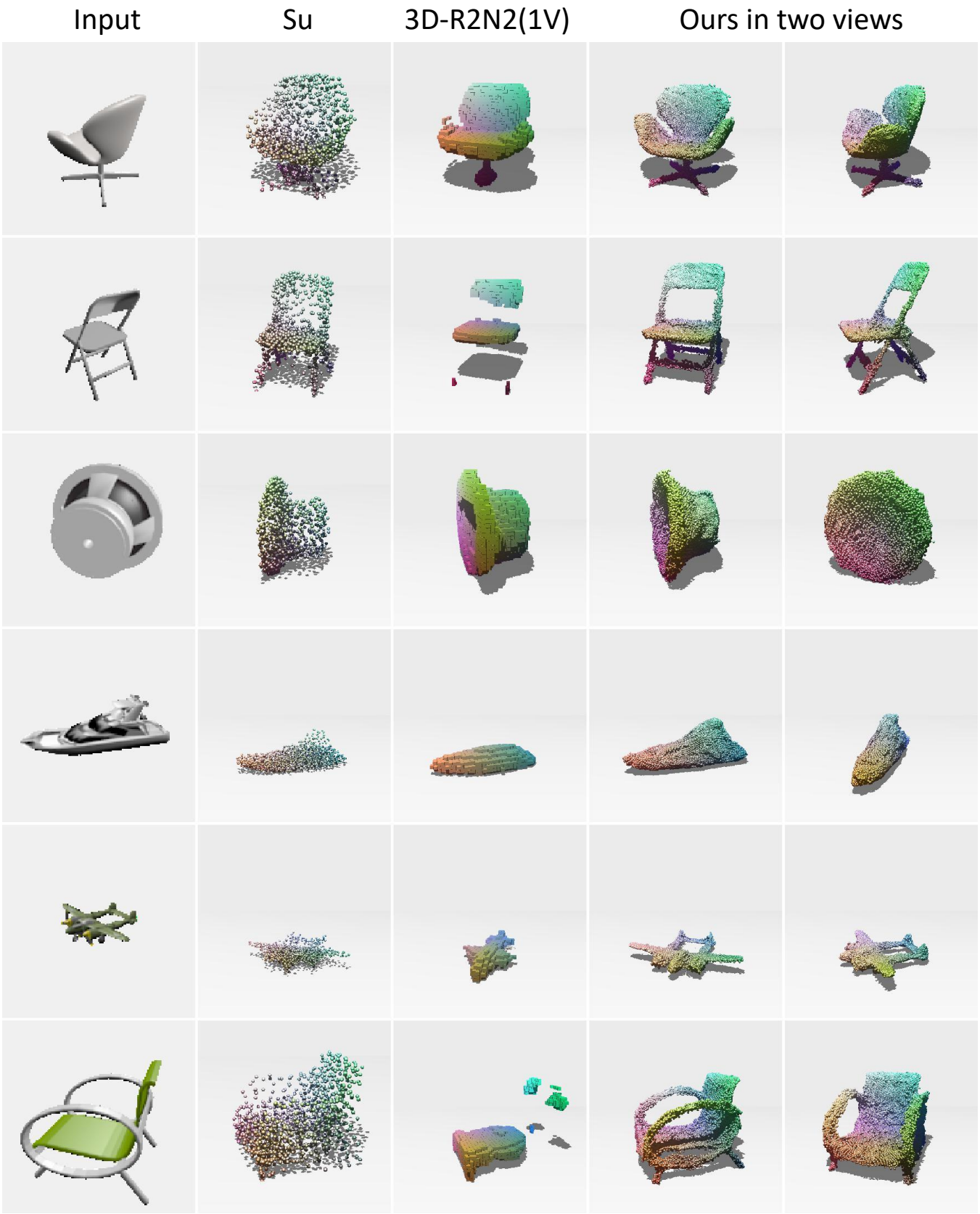

Figure F: More qualitative reconstruction results compared to Su(Fan, Su, and Guibas 2017) and 3D-R2N2(Choy et al. 2016). Note that for 3D-R2N2, we take a single input image for fair comparison. The rightmost two columns show our generated models in two views. Our results preserve more details and recover concave structures well.

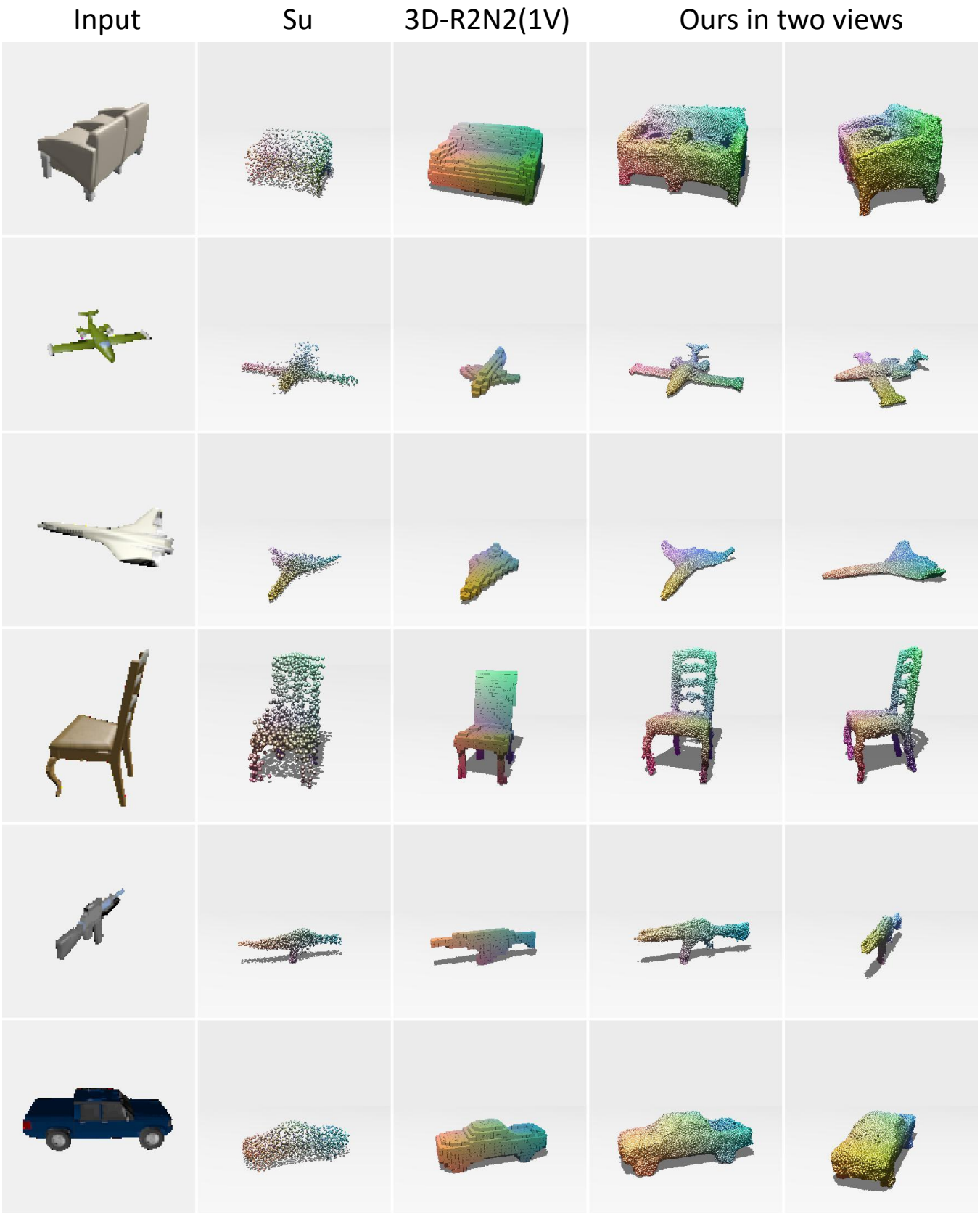

Figure F: More qualitative reconstruction results compared to Su(Fan, Su, and Guibas 2017) and 3D-R2N2(Choy et al. 2016). Note that for 3D-R2N2, we take a single input image for fair comparison. The rightmost two columns show our generated models in two views. Our results preserve more details and recover concave structures well.

Figure G: More reconstruction results from the testing dataset.

Figure H: More qualitative reconstruction results.