UFORecon: Generalizable Sparse-View Surface Reconstruction from Arbitrary and UnFavOrable Sets

Supplementary Material

A. More Implementation Details

In this section, we provide more details on our model architectures not fully covered in the main paper.

Network details. Given multi-view images, we first use the Feature Pyramid Network [27] to extract image features. We set the number of channels of each level to 32, 16, and 8, respectively. Similar to TransMVSNet [9], we utilize a sequence of attention blocks to extract cross-view features. To build correlation frustums from each view, every source image is considered as a reference image, while the other images are considered as the source images. We use 4 attention blocks and we use multi-head attention with the number of heads of 8.

Building Correlation Frustums. We build a cascaded correlation frustum for each source view as Sec 4.3. We clarify and elaborate more on the aggregation function $\langle\cdot\rangle$ of the correlation frustums here. For each level $l = 1, ..., L$, we obtain pairwise feature correlation as follows:

$$c_{ij}^{(d)}(\mathbf{p}) = F_i \cdot \hat{F}_{j\to i}^{(d)}, \tag{10}$$

where $\cdot$ denotes a dot product and $\hat{F}_{j\to i}^{(d)}$ denotes a warped j-th source feature map to i-th source feature at depth hypothesis d. Then we add all $N-1$ possible pairs to assign a pixel-wise weight map with its maximum correlation along the depth dimension as follows:

$$C_i^{(d)}(\mathbf{p}) = \sum_{\substack{j=1 \\ j\neq i}}^{N} \max_d \left\{ c_{ij}^{(d)}(\mathbf{p}) \right\} \cdot c_{ij}^{(d)}(\mathbf{p}). \tag{11}$$

Finally, we concatenate all depth hypotheses and add correlation frustums of all source viewpoints:

$$C'(\mathbf{p}) = \sum_{i=1}^{N} C_i(\mathbf{p}), \tag{12}$$

where $C'(\mathbf{p}) \in \mathcal{R}^{1\times d\times h\times w}$ denotes the combined correlation frustums from all source views. This combined correlation frustum is regularized with 3D CNNs and output $V_i \in \mathcal{R}^{c\times d\times h\times w}$, which is used for estimating intermediate depths. We further regularize V_i to represent global feature volume V_i' and we obtain the global feature vector via trilinear interpolation for each level l.

B. Detailed Description of Baseline Methods

In our research, we compared our method with four different types of surface reconstruction approaches across all experiments. (1) The most directly comparable category is generalizable implicit surface reconstruction [26, 28, 36, 48], where our method is included. These approaches assume a very limited number of views and unseen scenes during inference. (2) Generalizable neural rendering methods [4, 44, 52] also use a very limited number of views to perform novel view synthesis from unseen scenes. These methods learn geometry through volume density rather than SDF, leading to ambiguity as they lack an explicit surface definition. By applying thresholding on volume density and with Marching Cubes [29], geometry can be obtained. (3) Neural implicit surface reconstruction [43, 51] techniques extract accurate geometry from densely captured multi-view images of a single scene via optimization. Since these methods cannot generalize across multiple scenes, per-scene optimization is required to estimate geometry, unlike generalizable methods. (4) We compare with a conventional matching-based method, COLMAP [38]. It utilizes feature matching and triangulation from multiple images, providing a baseline for comparing the effectiveness of neural network-based methods. (5) Additionally, we include learning-based MVS methods [9, 49].

We only use generalizable implicit surface reconstruction methods for comparing view-combination generalizability in Table 2 as they outperform all the other approaches in terms of generalizable surface reconstruction.

C. More Information on the VC Score

Derivation from view selection score. In this section, we provide more details on the View-Combination (*VC*) score which is derived from the view selection score [49] as described in Sec. 3. View selection score $s(i, j)$ indicates the informativeness of a view pair between image i and image j.

$$s(i,j) = \sum_{\mathbf{p}} \mathcal{G}(\theta_{ij}(\mathbf{p})), \tag{13}$$

$$\theta_{ij}(\mathbf{p}) = \left(\frac{180}{\pi}\right) \arccos\left((c_i - \mathbf{p}) \cdot (c_j - \mathbf{p})\right), \tag{14}$$

where $\mathbf{p}$ is a common track in both view i and j, while c_i and c_j denote the camera center of view i and view j, respectively. To obtain $\mathbf{p}$, we use off-the-shelf reconstruction software, COLMAP [38]. In addition, $\mathcal{G}$ is a piecewise Gaussian function that peaks at certain angle θ_0 as follows:

Set	Method	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Mean (CD) ↓
Favorable (VC: 1788)	VolRecon [36]	1.20	2.59	1.56	1.08	1.43	1.92	1.11	1.48	1.42	1.05	1.19	1.38	0.74	1.23	1.27	1.38
	ReTR [26]	1.05	2.31	1.44	0.98	1.18	1.52	0.88	1.35	1.30	0.87	1.07	0.77	0.59	1.05	1.12	1.17
	Ours	**0.76**	**2.05**	**1.31**	**0.82**	**1.12**	1.18	0.74	**1.17**	**1.11**	**0.71**	**0.88**	0.58	0.54	**0.86**	0.99	**0.99**
	Ours*	0.77	2.10	1.34	0.87	1.15	**1.16**	**0.71**	1.25	1.17	0.81	0.90	**0.57**	**0.51**	**0.86**	**0.97**	1.01
Normal (VC: 192)	VolRecon [36]	2.63	4.22	2.89	2.49	2.93	2.50	1.68	1.84	2.02	1.76	2.35	2.64	1.16	2.17	1.76	2.34
	ReTR [26]	2.06	3.72	2.54	2.51	1.75	2.11	1.49	1.57	1.74	1.35	1.88	2.05	1.00	1.74	1.48	1.93
	Ours	1.30	2.59	1.51	1.39	**1.04**	1.28	0.80	1.37	1.16	0.95	0.98	0.90	0.54	1.06	1.08	1.20
	Ours*	**1.02**	**2.21**	**1.42**	**1.00**	1.23	**1.24**	**0.72**	**1.36**	**1.03**	**0.77**	**0.86**	**0.84**	**0.47**	**0.91**	**0.96**	**1.07**
Unfavorable (VC: 57)	VolRecon [36]	3.43	3.64	4.26	4.63	2.43	3.40	2.81	2.41	2.36	2.49	3.79	3.55	1.44	3.60	3.38	3.18
	ReTR [26]	3.00	3.98	3.78	4.22	2.22	2.93	3.00	2.51	2.24	2.36	2.36	3.92	1.63	2.83	3.07	2.94
	Ours	1.39	2.25	1.65	1.96	1.53	1.61	1.22	1.92	1.36	1.66	1.75	1.29	0.73	1.70	1.39	1.56
	Ours*	**1.31**	**2.00**	**1.41**	**1.36**	**1.24**	**1.58**	**1.06**	**1.44**	**1.37**	**0.99**	**1.45**	**0.96**	**0.58**	**1.34**	**1.09**	**1.28**

Table 5. **Quantitative results on different View-Combination (VC) score levels.** Each method's performance is indicated across all scans, with the mean Chamfer Distance (CD). An asterisk (*) indicates methods trained with the random set training strategy (Sec.4.6). We use camera indices (23,24,33), (16,26,42), and (1,16,36) for *Favorable*, *Normal*, *Unfavorable* sets respectively. The corresponding average *VC* scores are given in parentheses.

$$\mathcal{G}(\theta) = \begin{cases} \exp\left(-\frac{(\theta-\theta_0)^2}{2\sigma_1^2}\right), & \theta \leq \theta_0, \\ \exp\left(-\frac{(\theta-\theta_0)^2}{2\sigma_2^2}\right), & \theta > \theta_0, \end{cases} \tag{15}$$

where we set $\theta_0 = 5$, $\sigma_1 = 1$, and $\sigma_2 = 10$ following the original definition [49]. While the view selection score originally calculates between two image pair, we define *VC* score to represent a collective informativeness of a combination for the reconstruction by averaging the values for all possible pairs as follows:

$$VC = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} s(i,j), \tag{16}$$

where n denotes the number of source images.

VC score statistics. For more statistical information about the *VC* score, we report the distribution of the *VC* score on the DTU test dataset that we used for the experiments. We average the *VC* scores of all test scenes for each combination. As shown in Figure 7, we evenly split the entire scores into three groups, *Favorable, Normal* and *Unfavorable*. The distribution is bell-shaped where the *Normal* set generally has the highest frequencies. To reproduce the Table 1 and Table 2, we use the combination with the camera indices $(23, 24, 33)$ as a favorable set following the previous generalizable surface reconstruction protocols [26, 28, 36, 48], and $(1, 16, 36)$ for the unfavorable set.

D. Results on Various VC Score Levels

We have reported the reconstruction performance on both *Unfavorable* (Table 2) and *Favorable* (Table 1) sets. Additionally, we provide a comprehensive report on performance across all *VC* score levels, which includes results for the *Normal* set. It is important to note that, unless specified otherwise, we adhered to the same training protocol as in previous studies [26, 36], utilizing only fixed sets of the best view combinations.

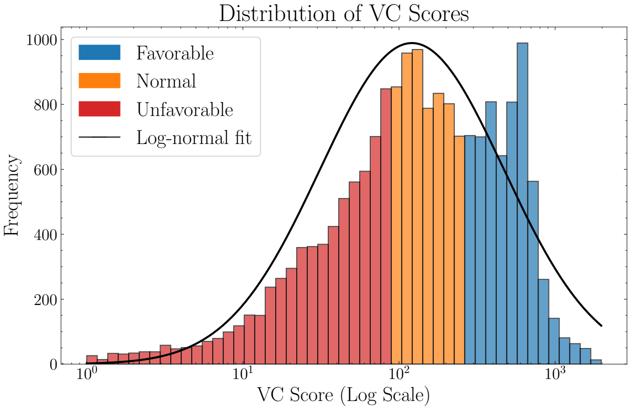

Figure 7. **Distribution of the View-Combination (*VC*) scores on the test scenes**. We visualize the statistical distribution of all scenes to give a sense of how *VC* score is distributed. We average all the scores of the test scenes given a camera view combination. A curved line denotes a fitted Gaussian distribution to VC scores in the log scale.

As demonstrated in Table 5, our method consistently shows superior performance across all VC levels and scenes. Furthermore, when employing the random set strategy for training (denoted as Ours* in Table 5), we observed an enhanced performance, particularly in the *Normal* and *Unfavorable* sets. This finding confirms the effectiveness of the random set training strategy in improving generalizability across different view combinations.

E. Random Set Training on Baseline Methods

The random set training strategy, intuitively, should be beneficial in reconstructing from *Unfavorable* sets, as it exposes the network to a variety of view combinations during training. Meanwhile, this approach might slightly de-

Method	Favorable	Unfavorable
VolRecon [36]	1.42	3.18
VolRecon* [36]	2.74 (+1.32, 92%)	3.88 (+0.7, 22%)
ReTR [26]	1.17	2.94
ReTR* [26]	1.62 (+0.45, 38%)	2.88 (-0.06, 2%)
Ours	0.99	1.56
Ours*	1.01 (+0.02, 2%)	1.28 (-0.28, 18%)

Table 6. **Effect of employing the random set training.** This table shows the Chamfer Distance (CD) metrics across different methods, comparing the performance on both *Favorable* and *Unfavorable* sets. The lower CD represents the better results. (*) denote the use of a random set training strategy (Sec. 4.6). The values in parentheses represent the change in performance relative to the standard training approach, with percentages indicating the rate of degradation (+) or improvement (-).

grade the performance in *Favorable* sets compared to methods trained solely with fixed *Favorable* sets.

To assess the impact of random set training, we applied this strategy to baseline methods [26, 36] and reported the results in Table 6. In the case of VolRecon [36], random set training resulted in suboptimal performance in both *Favorable* and *Unfavorable* test scenarios. We conjecture that the introduction of arbitrary view combinations as inputs may disrupt the training dynamics in VolRecon [36].

In contrast, for both ReTR [26] and our method, the adoption of random set training improved performance in *Unfavorable* scenarios, albeit at the expense of slightly diminished results in *Favorable* sets. Notably, our approach with random set training (Ours*) exhibited a minor performance degradation in *Favorable* sets (-2%), while achieving a significant enhancement in *Unfavorable* sets (+18%). This improvement can be attributed to our framework's ability to learn correlations among source images.

F. Impact of the Number of Views

We explored the effect of viewpoint density on our method by varying the number of source views. As detailed in Table 7, we observed a gradual improvement in performance with increasing viewpoint density. Using more views alleviates the challenges associated with reconstructing difficult regions, such as those that are occluded or not commonly shared across views.

G. Comparison on Depth Estimation.

MVS methods and other baseline methods experience significant performance degradation in unfavorable settings (Table 8), which are more susceptible to self-occlusion. The correlation frustum is one of our key contributions (not introduced in C2F2NeuS) to model global correlation among input views, and the reconstruction transformer estimates SDFs in a correlation-aware manner. We achieve robust performance in handling arbitrary view combinations.

Number of Views	VolRecon [36]	Ours
2	1.72	1.15
3	1.38	1.00
4	1.35	0.97
5	**1.33**	**0.96**

Table 7. **Effect of the number of views on reconstruction.** The table compares the Chamfer Distance for VolRecon [36] and our method across varying number of viewpoints. As the number of views increases, the reconstruction quality improves for both methods, with our method showing superior performance in all tested cases.

	Favorable			Unfavorable		
Method	<1mm ↑	<4mm ↑	Abs. ↓	<1mm ↑	<4mm ↑	Abs. ↓
MVSNet [43]	29.95	72.33	13.62	5.82	19.72	72.74
TransMVSNet[8]	38.95	82.91	13.53	6.21	15.60	161.71
VolRecon	43.60	82.53	7.12	3.93	11.43	185.95
ReTR [22]	51.95	85.06	**5.41**	10.83	19.26	58.02
Ours	**54.80**	**86.69**	6.09	**51.30**	**65.14**	**17.02**

Table 8. 3-Views Depth Map Comparison on DTU Datasets.

H. Comparative Ablation Study on Depth Supervision

As shown in Table 9, our method works well even without depth supervision. Especially in unfavorable sets, our method without depth supervision outperforms the baselines even with depth supervision. Depth supervision in SparseNeuS tends to result in over-smooth surfaces, which is also reported in [36].

	Favorable (CD↓)		Unfavorable (CD↓)	
Method	w/o $\mathcal{L}_d$	w/ $\mathcal{L}_d$	w/o $\mathcal{L}_d$	w/ $\mathcal{L}_d$
SparseNeuS [24]	1.64	4.22	4.16	5.55
VolRecon [30]	2.06	1.38	8.65	3.18
ReTR [22]	1.45	1.17	4.41	2.94
Ours	**1.26**	**0.99**	**2.26**	**1.56**

Table 9. Ablation study of Depth Supervision ($\mathcal{L}_d$).

I. Regarding Pracitcal Relevance.

Apparently, obtaining accurate camera poses from a sparse set of images presents challenges. In response, a parallel line of research has focused on estimating relative camera poses from sparse views [6, 39], and it is anticipated that we can apply these methods to obtain more accurate camera poses from sparse, unfavorable set of views. Another research direction explores reconstructing 3D in a pose-free manner only using source images [10, 17], in which case our proposed modeling of correlation across images becomes even more meaningful. These approaches potentially address concerns on obtaining camera poses. We believe that obtaining accurate geometry under pose-free settings remains underdeveloped and seems to be a promising future work.

J. More Qualitative Results

Lastly, we provide additional visual examples that highlight the effectiveness of our method. We include all test scenes except the scenes reported in the main paper. Similar to Figure 4, we include the results of the comparison with baseline methods across *Favorable*, *Normal*, and *Unfavorable* sets.

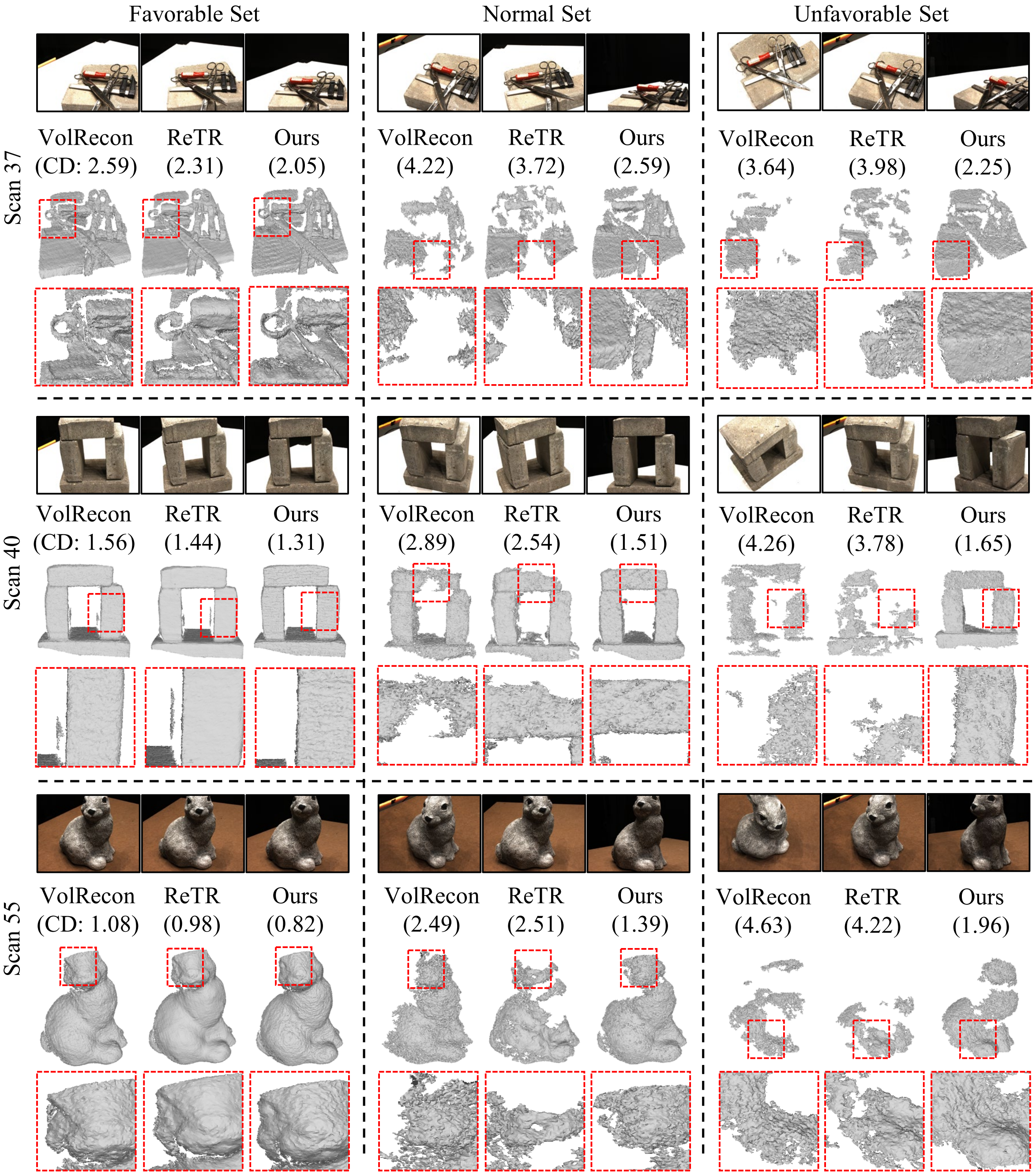

Figure 8. **A qualitative results of reconstruction across various VC Levels.** The numbers in parentheses denote the Chamfer Distance.

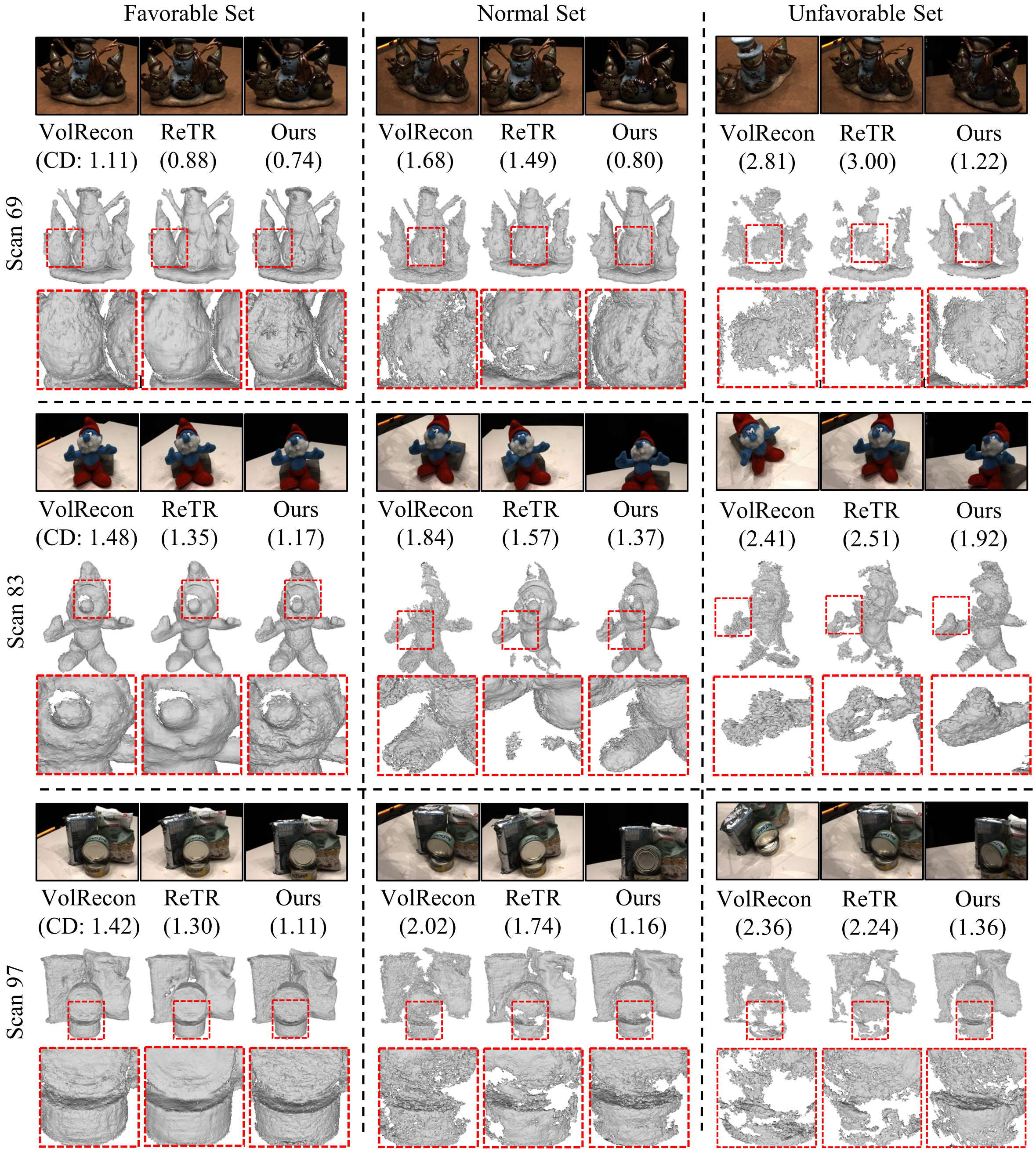

Figure 9. **A qualitative results of reconstruction across various VC Levels.** The numbers in parentheses denote the Chamfer Distance.

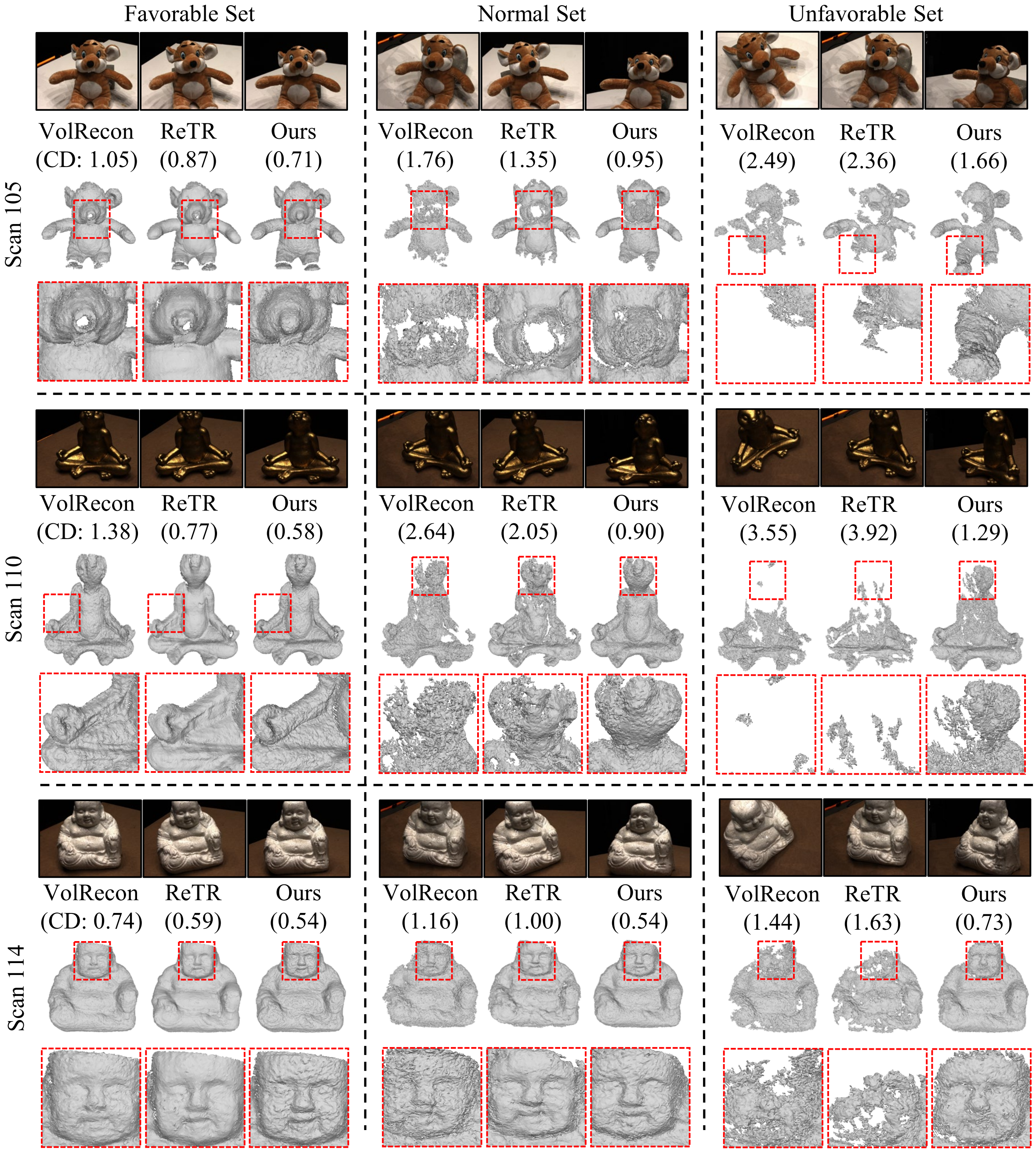

Figure 10. **A qualitative results of reconstruction across various VC Levels.** The numbers in parentheses denote the Chamfer Distance.

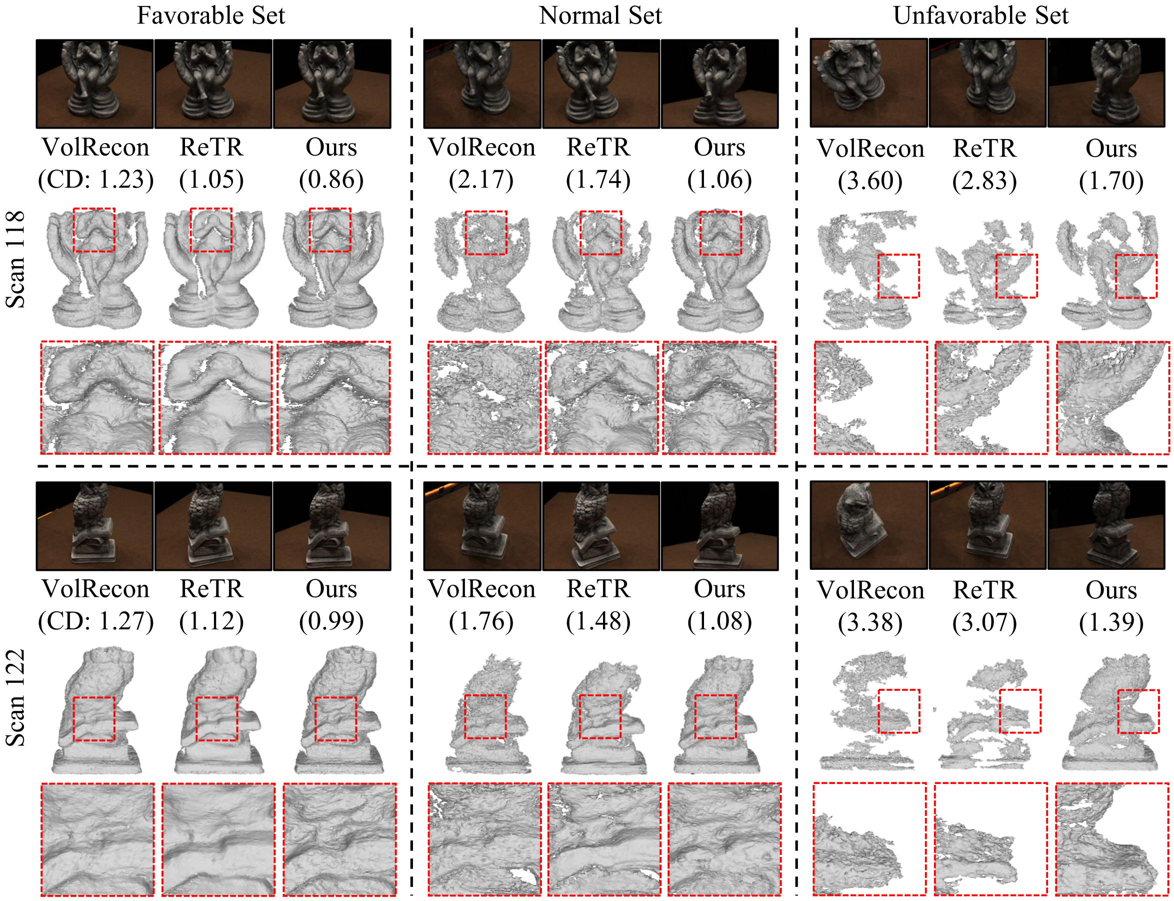

Figure 11. **A qualitative results of reconstruction across various VC Levels.** The numbers in parentheses denote the Chamfer Distance.