A. Overview

In this document we provide technical details and additional discussions and results to the main paper.

In Section B, we describe the details about our network. Section C-D are some discussions about our work. Section E provides some additional qualitative results on real-scanned data. Section F-I show ablation studies about some choices when we design the network. Section J shows some failure cases of our work.

B. Details of the network architecture

In multi-level feature extraction. we use the implementation of PointNet++. Following the notations in PointNet++, we use $(K, r, [l_1, ..., l_d])$ to represent a level with K local regions of ball radius r, and $[l_1, ..., l_d]$ the d fully connected layers with width l_i $(i = 1, ..., d)$. The parameters we use are shown in Table 1.

In feature expansion and reconstruction, the separated MLPs applied for each level feature consist of fully connected layers with width [256, 128], and the shared MLP used for coordinates reconstruction is with width [64, 3]. Note that the MLP used in RFA (Residual feature aggregation) is involved in the separated MLPs which means we use the same architecture for both RFA and GLFA, but the functionality of the first fully connected layer in RFA is to transform the computed difference to the feature space.

The MLP used in the attention module is with width [16, 8, 1], which outputs a scalar as the score for each point. In local folding unit, the MLP used for generating the concatenated feature to the final coordinates is with width [512, 512, 3]. Please see our code with the link in the attached text file for more details about the implementation.

	Parameters	Output	Interpolated
Level 1	$K = N, r = 0.1, mlp = [32, 32, 64]$	$N \times$ 64	$N \times$ 64
Level 2	$K = N/2, r = 0.2, mlp = [64, 64, 128]$	$N/2 \times$ 128	$N \times$ 64
Level 3	$K = N/4, r = 0.3, mlp = [128, 128, 256]$	$N/4 \times$ 256	$N \times$ 64
Level 4	$K = N/8, r = 0.4, mlp = [256, 256, 512]$	$N/8 \times$ 512	$N \times$ 64
Level 5	$K = N/16, r = 0.5, mlp = [512, 512, 1024]$	$N/16 \times$ 1024	$N \times$ 64
Level 6	$K = N/32, r = 0.6, mlp = [512, 512, 1024]$	$N/32 \times$ 1024	$N \times$ 64
Global	$N = 1, mlp = [512, 512, 1024]$	1×1024	$N \times$ 1024

Table 1. Parameters used in multi-level feature extraction process.

C. Symmetrical characteristic during completion

During the completion process, we find our network try to learn the symmetrical characteristic of the object to complete the model. As shown in Figure

1, it can be seen that the $Y_{missing}$ is close to the partial input after a proper transformation. This indicates that the details can be preserved not only in the partial input but also in the predicted symmetrical part taking advantages of the symmetrical characteristic.

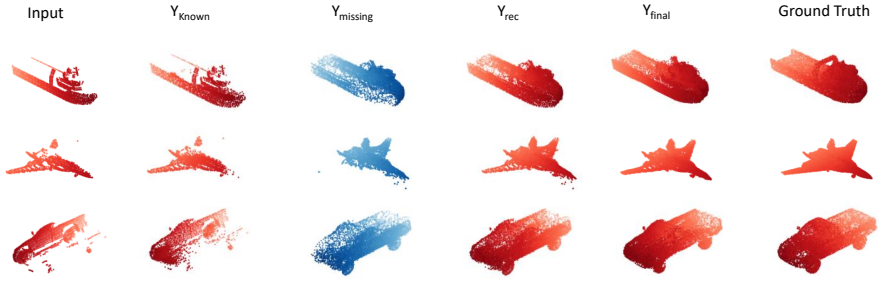

Fig. 1. Symmetrical characteristic during the completion.

D. Is our network just doing classification?

A recent work "What Do Single-view 3D Reconstruction Networks Learn?" [1] claims that some state-of-the-art single-view 3D reconstruction methods do not actually perform reconstruction but classification. We find there are several evidences to prove our network is not just doing classification. An intuitive evidence is that our network can predict different details for every different model (see Figure 4 in our paper). If our network is just doing classification, it will predict a certain shape from a certain category. Another evidence is that instead of just using global feature to generate the complete shape, we leverage local features with proposed separated feature aggregation. It prevents the situation that the network just uses the global feature to produce a approximate shape from a certain category.

"What Do Single-view 3D Reconstruction Networks Learn?" [1] proposed the cluster and the retrieval baselines to see if the performance of the learning based single-view 3D reconstruction methods are close to just using the cluster or retrieval baseline. We intend to compare our method with these two baselines, but they are all based on single-view color image reconstruction. Following these baselines, we design another retrieval baseline based on point cloud: we retrieval a complete ground truth point cloud from the training set which is closest to the input point cloud in terms of Chamfer Distance. We evaluate our work using F-score metric which is proposed in [1] as shown in Table 2. We can see that our method significantly outperform the retrieval baseline. Note that other baseline works also outperform the retrieval baseline. Qualitative results are also shown in Figure 2, where our method refers to NSFA-RFA. It can be observed that the

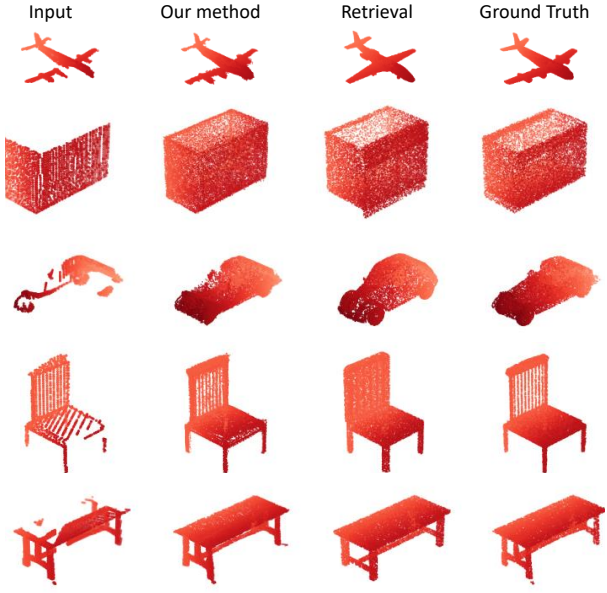

Fig. 2. Comparisons between our method and retrieval baseline.

Method	Avg	airplane	cabinet	car	chair	lamp	sofa	table	vessel
FC	0.65	0.87	0.57	0.70	0.57	0.59	0.50	0.69	0.67
Folding	0.62	0.84	0.59	0.66	0.55	0.51	0.51	0.69	0.63
TopNet	0.67	0.86	0.61	0.70	0.59	0.60	0.56	0.72	0.71
PCN	0.69	0.88	0.65	**0.72**	0.62	0.63	0.58	0.76	0.69
Retrieval	0.44	0.71	0.37	0.38	0.39	0.42	0.29	0.49	0.47
GLFA	0.74	**0.91**	0.64	0.68	0.70	0.80	0.62	0.81	0.79
RFA	**0.76**	**0.91**	**0.66**	**0.72**	**0.74**	**0.82**	**0.63**	**0.83**	**0.79**

Table 2. Evaluation of all the baselines, retrieval baseline and our methods on F-score metric.

retrieval method can get complete shapes without any outliers, but our method can generate more accurate results.

E. More results on real-scanned data.

Beside real-scanned data from Kitti, we also scanned some models using the Structure Sensor to see our network performance. We show some results in Figure 3. The first column shows the color image and the second row shows the scanned partial model. We manually extract the partial object from the scanned model which is shown in the third column. The final column shows the completion results. We can see our network can recover the complete shape on these models.

Note that in Figure 3(c), as our training set seems do not contains a biplane type, the produced result of the biplane might involve more noises than others.

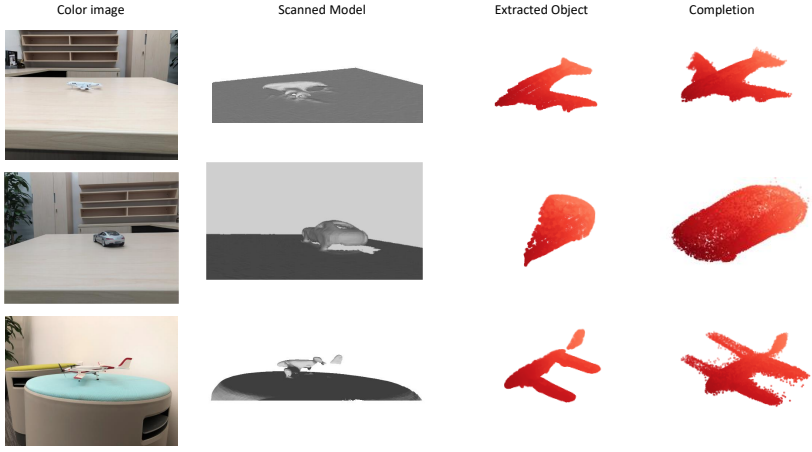

Fig. 3. Results on real-scanned data which is scanned by Structure Sensor.

F. How we decide m in GLFA

As mentioned in the GLFA (Global & local feature aggregation) session, we aggregate first m level features to form f_{known} and last m level features to express $f_{missing}$. We test the effectiveness of m with networks of different feature extraction levels. The results are shown in Figure 4. Specifically, when we set $m = \lfloor \frac{n}{2} \rfloor + 1$, there are overlapped layers between f_{known} and $f_{missing}$. For example, with network of 5 levels, m equals 3, which means f_{known} and $f_{missing}$ both involve the third-level feature. It can be seen that all the networks achieve best performance when $m = \lfloor \frac{n}{2} \rfloor + 1$. In contrast, when $m = \lfloor \frac{n}{2} \rfloor$ which indicates there are no overlapped layers between f_{known} and $f_{missing}$, the performance of all networks drops dramatically. We consider the reason is that it needs at least one overlapped layer to create the correlation between f_{known} and $f_{missing}$. Meanwhile, with the number of overlapped layers increasing, the performance of the networks also drops. This maybe because providing much more overlapped layers would obscure the boundary of the local feature and global feature.

G. The choices of $C_{missing}$

As mentioned in the Separated Feature Aggregation part, we concatenate the missing part features with the coordinates $C_{missing}$. At first, we just assign

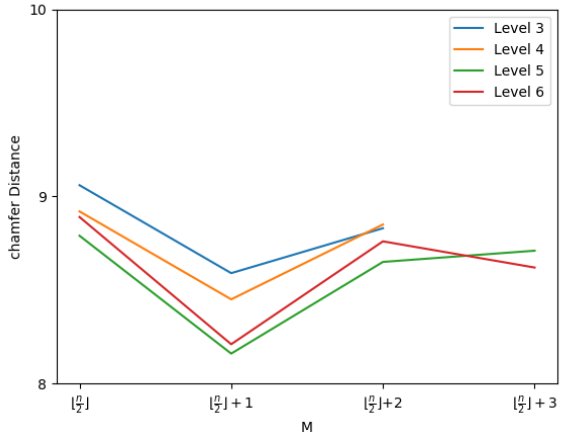

Fig. 4. Effect of m in GLFA.

C_{origin} to $C_{missing}$, but we found the network can converge faster if we set $C_{missing}$ with more proper initial coordinates. Thus we involves T-Net to represent $C_{missing}$. We compare the performance of using C_{origin} and T-Net respectively, with results shown in Table . Both networks are trained in 25 epochs. We can see involves T-Net can improve the network performance on most of categories. But for NSFA-RFA on novel category, it seems that it is hard to generate proper $C_{missing}$ for residual feature aggregation. In this case, $C_{missing}$ might degrade to C_{origin}, where T-Net just does any transformation.

	NSFA-RFA		NSFA-GLFA	
	Known	Novel	Known	Novel
C_{origin}	8.56	**10.06**	9.48	11.70
T-Net	**8.06**	10.08	**8.14**	**9.98**

Table 3. Quantitative evaluation of $C_{missing}$ when using C_{origin} and T-Net on Chamfer Distance multiplied by 10^4.

H. Evaluation of the components in loss function.

In the loss function, the repulsion term $\mathcal{L}_{rep}(Y_{coarse})$ is used for making the results to be uniformly distributed. The other $\mathcal{L}_{CD}$ components are used for intermediate and final supervision to guarantee each step results. We do evaluation by removing each term with results shown in Table 4.

By removing	Known categories	Novel categories
$\mathcal{L}_{rep}(Y_{coarse})$	8.21	11.25
$\mathcal{L}_{CD}(Y_{rec}, Y_{gt})$	8.18	11.05
$\mathcal{L}_{CD}(Y_{coarse}, Y_{gt})$	8.13	11.40
Full loss	**8.06**	**10.80**

Table 4. Quantitative evaluation of loss components with NSFA-RFA on Chamfer Distance multiplied by 10^4.

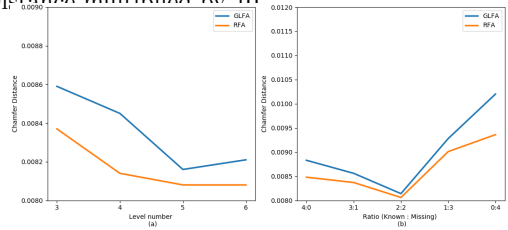

Fig. 5. Effects of the level number (a) and the combination ratio (b) to the evaluation metric.

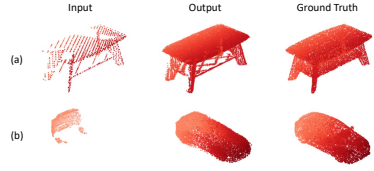

Fig. 6. Failure cases.

I. Ablation Studies

Design choices involved in our network include choosing the number of level in multi-level features extraction and the combination ratio when we aggregate the known and missing part features. These parameters can influence performance so we analyze the performance of our method as a function of these parameters.

Level number. We test the performance of our network with different feature extraction level number for both NSFA-GLFA and NSFA-RFA on the testing set of known categories. Figure 5 (a) shows the Chamfer distance of our network with different level number. On the whole, the Chamfer distance reduces as the level number increases for both strategies and NSFA-RFA performs a little better than NSFA-GLFA. From level 5 to 6, the performance of NSFA-GLFA decreases. This maybe because we add much more level features to both the known and missing parts, which blurs the boundaries of the global and local features.

Combination ratio. Another important factor is the mixture ratio. We test different mixture ratio with the baseline network. The results are shown in Figure 5 (b). When the combination ratio between known part and missing part is close to 1:1, the network achieves best performance. With ratio 1:3 and 0:4, both the performance of NSFA-RFA and NSFA-GLFA drops largely. We consider the reason is that providing few known part features makes it hard for the network to keep the details of the original model. With ratio 4:0, NSFA-RFA and NSFA-GLFA degrade to the networks directly concatenating multi-level features during the feature aggregation.

J. Failure Cases

We find some failure cases during the experiments which are shown in Figure 6. They can be categorized into two cases. The first one is that, as illustrated in

Figure 6(a), the partial model contains discontinuous parts (the strips under the desktop) which are caused by the view point, but the complete model has continuous shape in that area. Our network seems to regard the discontinuous parts as the details of the model and try to keep them during the completion. The other case is that the provided partial model does not have enough cues for the network to predict the details of the model as shown in Figure 6(b). Our network can recognize the model is a car but can not predict its specific details

References

1. Tatarchenko, M., Richter, S.R., Ranftl, R., Li, Z., Koltun, V., Brox, T.: What do single-view 3d reconstruction networks learn? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3405–3414 (2019)