3D Shape Completion using Multi-resolution Spectral Encoding

Supplementary Material

Pallabjyoti Deka, Saumik Bhattacharya, Debashis Sen and Prabir Kumar Biswas
Indian Institute of Technology, Kharagpur
pallabjyotidk@gmail.com, {saumik, dsen, pkb}@ece.iitkgp.ac.in

The supplementary material is organized as follows - In Sec. 1, we discuss the network architecture of the encoders (Sec. 1.1) and that of the decoders in the complete encoder-decoder model (Sec. 1.2). Details regarding the datasets are discussed in Sec. 2 for both ShapeNet [1] (Sec. 2.1) and ScanNet [5] (Sec. 2.2). In the Sec. 3, we provide results on various experiments including evaluation of our model on seen categories of ScanNet dataset (Sec. 3.1), empirical time analysis of our Spectral Module (Sec. 3.2), effect of different components in terms of F1 score (Sec. 3.3), effect of different train/test category split (Sec. 3.4) and ablation studies on the performance considering individual resolutions and multi-resolution (Sec. 3.5).

1. Model Architecture Details

1.1. Architecture of the Model Encoders

In Sec. 3.1 of the *main paper*, we elaborate our model's encoding methods and their pretraining. Here, we provide comprehensive details about the network architecture employed in the blocks of the encoders. Fig. 1 shows the detailed architecture of the encoder blocks. In Fig. 1(a), we show the overview of the encoders from Fig. 2 of the *main paper*. Both the encoders for the partial scan input and the shape prior input have the same architecture. Figs. 1(b), 1(c), and 1(e) show the inside architecture of each of the ResNet blocks. Fig. 1(e) also shows the *kernel* and *stride* used for the third ResNet block to accommodate the three different resolutions. The architecture of the Spectral Module (SM) within the encoders is detailed in Fig. 1(d). The details of the Attention Refinement (AR) block used in this encoding process are given in Fig. 1(f). We use *padding=1* for every convolution operation except in the SM where we keep *padding=0*.

In the SM, we use an efficient implementation of the FFT and call it Real FFT [8], where only half of the spectrum is computed. We could do so as the input to the FFT is a real signal, which lends certain symmetry properties to the spectrum [4]. The real and imaginary components resulting from the FFT operation are stacked together and a sequence of Conv-BN-ReLU operations is performed as shown in Fig. 1(d). We then split this result into the real and imaginary parts to perform an inverse Fast Fourier Transform (FFT) operation to get a real-valued output as explained in [2].

1.2. Architecture of the Model Decoders

In Sec 3.2 of the *main paper*, we discuss our encoder-decoder network model in detail. Here, in Fig. 2(a), we show the overview of that network from Fig. 3 of the *main paper*. Figs. 2(b), (c) and (d) give the architecture details of the three decoders in Fig. 2(a). We use $\mathcal{D}^{32}$, $\mathcal{D}^{8}$ and $\mathcal{D}^{4}$ in a hierarchical manner to get the fully reconstructed shape S.

2. Datasets

2.1. ShapeNet data

In our method, the synthetic dataset ShapeNet [1] is used for both training and testing purposes. Truncated-Signed Distance Fields (SDF) with a truncation value of 2.5 voxel units is used as the data format throughout. For a particular ground truth, four different views of a partial input are taken into account. The training set contains 18 categories - *table, chair, sofa, cabinet, clock, bookshelf, piano, microwave, stove, file cabinet, trash bin, bowl, monitor, keyboard, dishwasher, washing machine, pots, faucet, and guitar*. We test the model on eight novel (unseen) categories - *bathtub, lamp, bed, bag, printer, laptop, bench, and basket*. The evaluation of our model on this dataset is in Tab. 1 of the *main paper*, where our model outperforms the others.

2.2. ScanNet data

To evaluate our model on ScanNet [5], we fine-tune our ShapeNet model on some categories available in the training set of ScanNet. The real-world data is represented using truncated-SDFs as well, but with a truncation value of 3 voxel units. In our fine-tuning set, the categories are - *chair, table, sofa, trash bin, cabinet, bookshelf, file cabinet, and monitor*. We test our model in both seen and unseen categories. The six unseen categories are - *bathtub, lamp,*

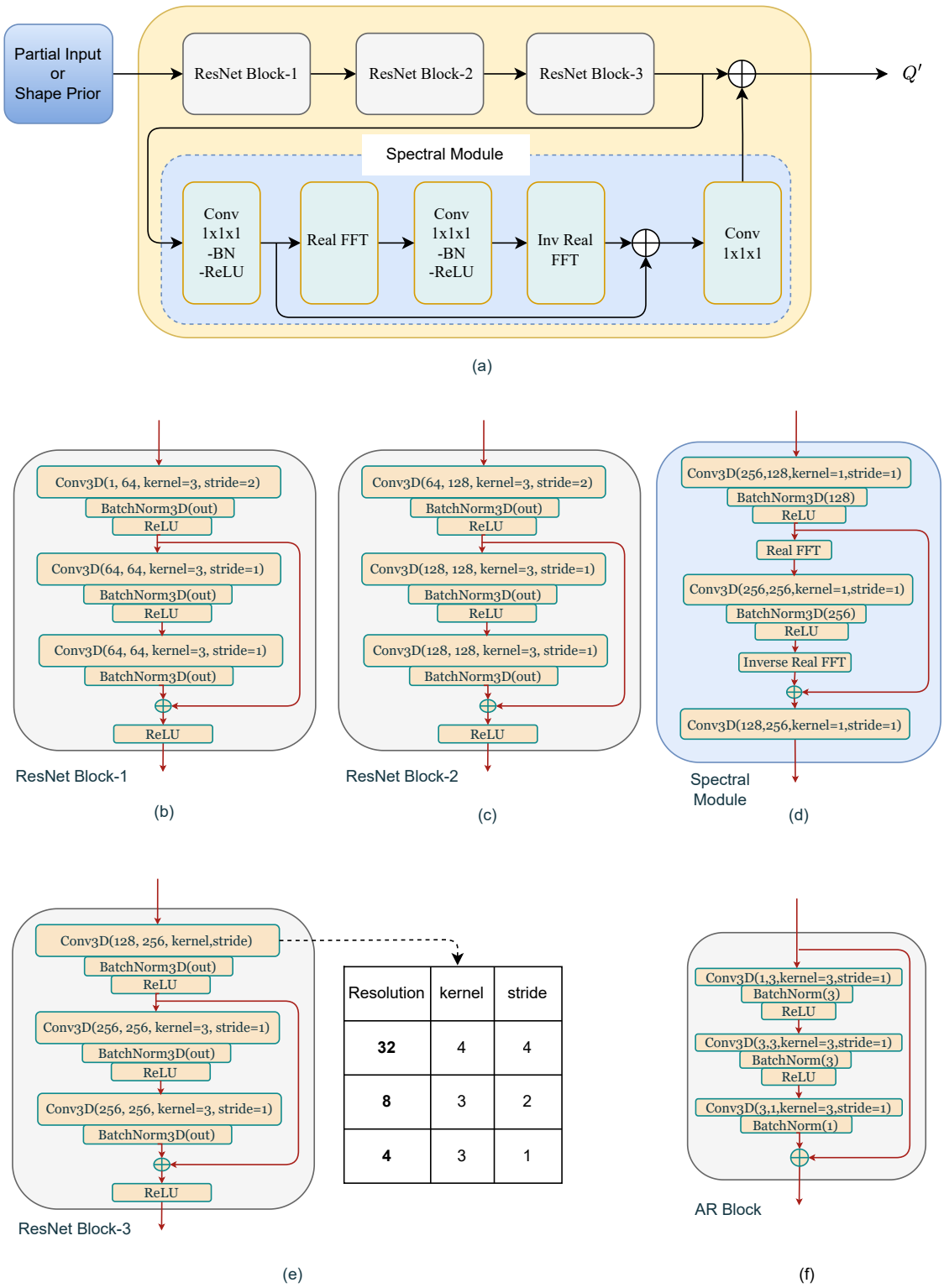

Figure 1. Detailed overview of various components of the encoders shown earlier in Fig. 2 of the *main paper*: (a) shows the encoders acting on the partial scan input and the shape prior input. They are separate encoders with the same architecture, and hence, one of them is shown; (b), (c), and (e) show the details of each ResNet block; (d) shows the details of the spectral module (SM); (f) shows the details of the AR block. $\bigoplus$ denotes element-wise addition.

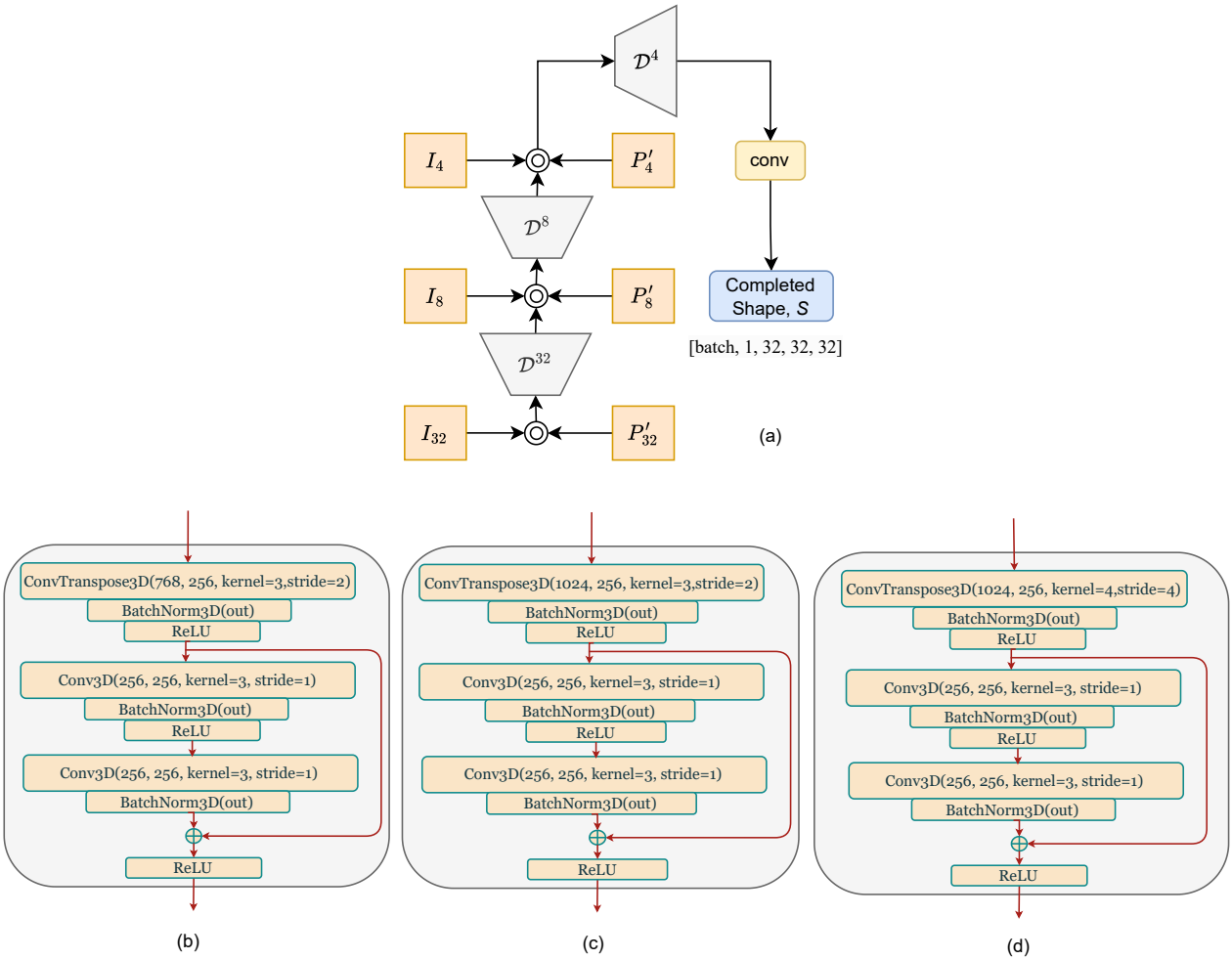

Figure 2. (a) shows the overview of our decoder module as explained earlier in Fig. 3 of the *main paper*; (b), (c), and (d) detail the $\mathcal{D}^{32}$, $\mathcal{D}^8$ and $\mathcal{D}^4$ decoders, respectively. $\odot$ denotes concatenation and $\bigoplus$ denotes element-wise addition.

bed, bag, basket, and printer and the seven seen categories are - *chair, table, sofa, trash bin, cabinet, bookshelf, and monitor*.

3. Experiments

3.1. Evaluation on seen categories of ScanNet dataset

In Tab. 1, we compare our results with the current state-of-the-art methods IF-Nets [3], 3D-EPN [5], Few-Shot [9], PatchComplete [7] and AutoSDF [6] on seen categories of the ScanNet dataset. It is evident that our method outperforms the recent methods on average for the seen categories. The evaluation of our model on the unseen categories of the same dataset is in Tab. 2 of the *main paper*, where our model outperforms the others.

3.2. Efficiency of the Spectral Module

In Tab. 2, we present an analysis of the time required to train encoders for each resolution with and without the Spectral Module (SM). As shown in the table, the use of SM results in only a marginal increase in training time, as mentioned in Sec. 4.4 of our *main paper*.

3.3. Effect of different components in terms of F1 score

An ablation study on the different components of our method was conducted using Chamfer Distance (CD) and Intersection-over-Union (IoU) as evaluation metrics, as shown in Tables 3 and 4 of the main paper. The improvements due to the use of $\mathcal{L}_{grad}$, SM, and AR block can be further corroborated using $F1$ measure as shown in Tab. 3.

	Chamfer Distance ↓ ($\times 10^2$)						IoU↑					
	IFN	**3DEPN**	**FS**	**PC**	**ASDF**	**Ours**	**IFN**	**3DEPN**	**FS**	**PC**	**ASDF**	**Ours**
Table	10.15	8.74	7.13	6.60	6.72	**5.99**	0.46	0.47	0.50	0.54	0.49	**0.55**
Sofa	7.87	4.94	**4.28**	4.53	4.58	4.39	0.67	0.69	**0.75**	0.73	0.72	0.73
Trash Bin	5.23	5.03	5.65	4.44	4.48	**4.09**	0.62	0.61	0.70	0.68	0.66	**0.72**
Bookshelf	5.17	4.87	4.33	3.80	4.12	**3.63**	0.58	0.53	**0.65**	0.61	0.61	0.63
Chair	7.93	9.99	6.88	7.14	**6.00**	6.15	0.43	0.40	0.46	0.45	**0.49**	**0.49**
Monitor	6.39	5.75	4.98	4.74	5.92	**4.36**	0.53	0.52	0.59	0.56	0.49	**0.60**
Cabinet	5.64	4.60	4.36	**4.17**	4.53	4.26	0.74	0.76	**0.80**	0.79	0.78	0.78
Inst. Avg.	7.65	7.94	6.18	6.02	5.68	**5.40**	0.51	0.50	0.56	0.55	0.55	**0.58**
Cat. Avg.	6.91	6.27	5.37	5.06	5.20	**4.70**	0.58	0.57	0.63	0.62	0.61	**0.64**

Table 1. Evaluation of the different approaches on seen categories of ScanNet dataset (real data).

Resolution	**Base w/o SM**	**Base with SM**
32	1.132	1.136
8	1.749	1.785
4	2.784	2.914

Table 2. The table shows the time taken to train the encoders at each resolution of our model (in hours). All the experiments in this table are performed for 80 epochs keeping batch size and learning rate to be 32 and 0.001, respectively. **Base** indicates performing the training of the model using only $\mathcal{L}_{one}$ loss as described in the *main paper (Sec 3.1)* and SM refers to Spectral Module.

	Base	**SM**	$\mathcal{L}_{total}$	**AR and** $\mathcal{L}_{total}$	**Ours**
Inst. F1 ↑	0.7584	<u>0.7769</u>	0.7706	0.7754	**0.7821**
Cat. F1 ↑	0.7649	<u>0.7836</u>	0.7770	0.7812	**0.7879**

Table 3. Ablation study on the components of our method using F1 score as the evaluation metric considering the unseen categories of the ShapeNet dataset. The second-best scores are underlined.

3.4. Effect of train/test category split

We shuffle the overall categories between train/test data to show the robustness of our method. We randomly selected one category from the train set and another category from test set, and swapped them ('basket' & 'bookshelf' in Split 1 and 'lamp' & 'chair' in Split 2, respectively). From the Tab. 4, it is evident that our method performs equally across different splits.

3.5. Ablation on different resolutions

Our 3D shape completion approach comprises a multi-resolution encoder-decoder network. In Tab. 5 and Tab. 6, we analyze the performance of our proposed approach when the multiple resolutions are considered in comparison to when only the individual resolutions are considered separately. As seen in the table, our model performs better with the multi-resolution approach rather than the single-resolution ones.

	CD ↓ ($\times 10^2$)		**IoU** ↑	
	Inst. Avg.	**Cat. avg.**	**Inst. Avg.**	**Cat. Avg.**
Split 1	4.04	4.06	0.655	0.665
Split 2	4.08	4.14	0.667	0.676
Original Split	4.08	4.13	0.664	0.673

Table 4. Performance of our model on different train/test splits on the unseen categories of ShapeNet data

	Chamfer Distance ↓ ($\times 10^2$)			
	Ours(32^3)	**Ours**(8^3)	**Ours**(4^3)	**Ours**
Bag	6.55	4.55	4.42	**3.94**
Lamp	32.39	7.44	5.73	**4.68**
Bathtub	6.55	4.01	3.73	**3.52**
Bed	8.51	4.90	4.50	**4.35**
Basket	20.62	7.33	5.45	**5.03**
Printer	7.98	4.78	4.58	**4.47**
Laptop	15.84	4.13	3.85	**3.51**
Bench	10.18	4.43	3.77	**3.58**
Inst. Avg.	13.82	5.12	4.44	**4.08**
Cat. Avg.	13.58	6.03	4.50	**4.13**

Table 5. Performance analysis of the proposed approach based on Chamfer distance for single and multi resolutions on the ShapeNet unseen data, where 'Ours' indicates multi-resolution.

	IoU ↑			
	Ours(32^3)	**Ours**(8^3)	**Ours**(4^3)	**Ours**
Bag	0.680	0.755	0.750	**0.780**
Lamp	0.285	0.498	0.555	**0.587**
Bathtub	0.447	0.620	0.665	**0.695**
Bed	0.491	0.641	0.661	**0.678**
Basket	0.390	0.573	0.607	**0.635**
Printer	0.638	0.762	0.770	**0.780**
Laptop	0.210	0.590	0.636	**0.668**
Bench	0.310	0.483	0.538	**0.558**
Inst. Avg.	0.408	0.603	0.639	**0.664**
Cat. Avg.	0.431	0.615	0.648	**0.673**

Table 6. Performance analysis of the proposed approach based on Intersection-over-Union (IoU) for single and multi resolutions on the ShapeNet unseen data

3.6. Ablation on the effect of shape priors

In Tab. 7, we demonstrate that reducing the number of shape priors utilized in the proposed method by 50% leads to a decrease in the overall performance of the model. Conversely, increasing the number of shape priors by 10% does not improve the average performance, but it demands a higher storage space requirement. This shows that the number of priors employed by our approach is sufficient for the shape completion task.

	CD ↓ ($\times 10^2$)		IoU ↑	
	Inst. Avg.	**Cat. Avg.**	**Inst. Avg.**	**Cat. Avg.**
Ours	4.08	4.13	0.664	0.673
Ours (50% priors)	4.22	4.28	0.657	0.666
Ours (110% priors)	4.16	4.24	0.659	0.667

Table 7. Performance analysis of the proposed method based on Chamfer Distance (CD) and Intersection-over-Union (IoU) for different shape priors on the ShapeNet unseen data.

3.7. Additional visual shape completion results

Fig. 4 and Fig. 5 of the *main paper* provided a few visual shape completion results of our proposed approach in comparison to the state-of-the-art. Here, in Fig. 3 and Fig. 4, we provide additional comparisons obtained using our model on objects from the ShapeNet [1] and ScanNet [5] datasets, respectively, which demonstrate the effectiveness of our approach.

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1, 5, 6

[2] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4479–4488. Curran Associates, Inc., 2020. 1

[3] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6970–6981, 2020. 3

[4] James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19:297–301, 1965. 1

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1, 3, 5, 6

[6] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 306–315, 2022. 3, 6

[7] Yuchen Rao, Yinyu Nie, and Angela Dai. Patchcomplete: Learning multi-resolution patch priors for 3d shape completion on unseen categories. *Advances in Neural Information Processing Systems*, 35:34436–34450, 2022. 3, 6

[8] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 1

[9] Bram Wallace and Bharath Hariharan. Few-shot generalization for single-image 3d reconstruction via priors. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3818–3827, 2019. 3

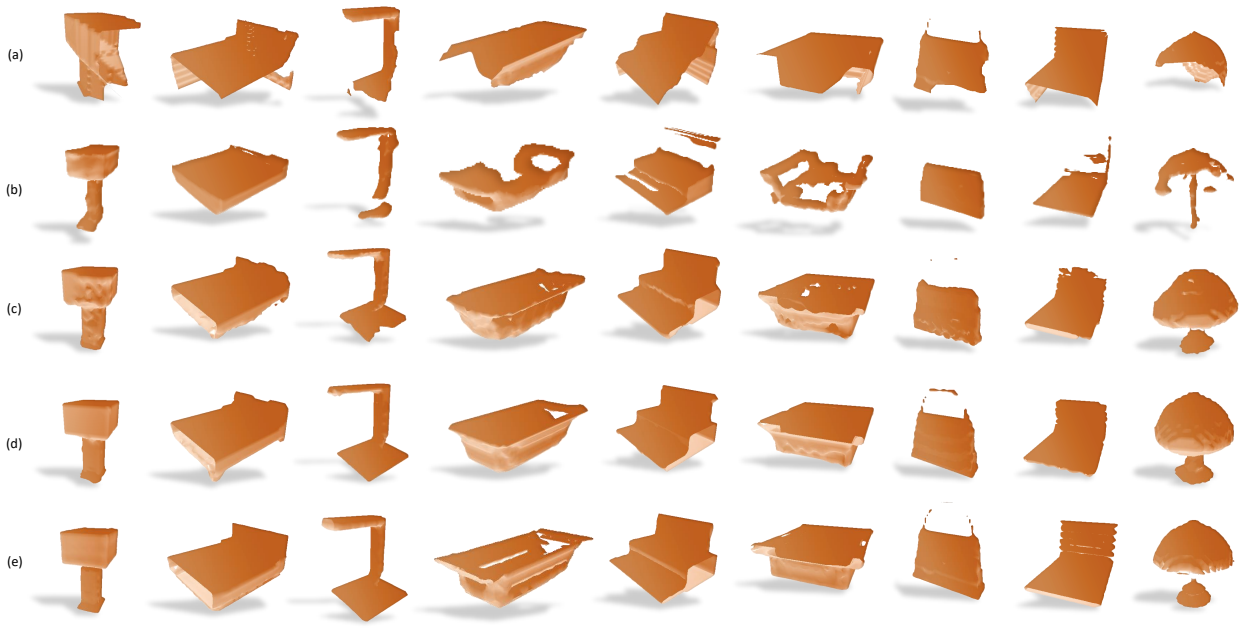

Figure 3. Qualitative comparison with state-of-the-art methods on ShapeNet [1] data. In the figure above, (a)-input shape, (b)-AutoSDF [6], (c)-PatchComplete [7], (d)-Ours and (e)-ground truth objects.

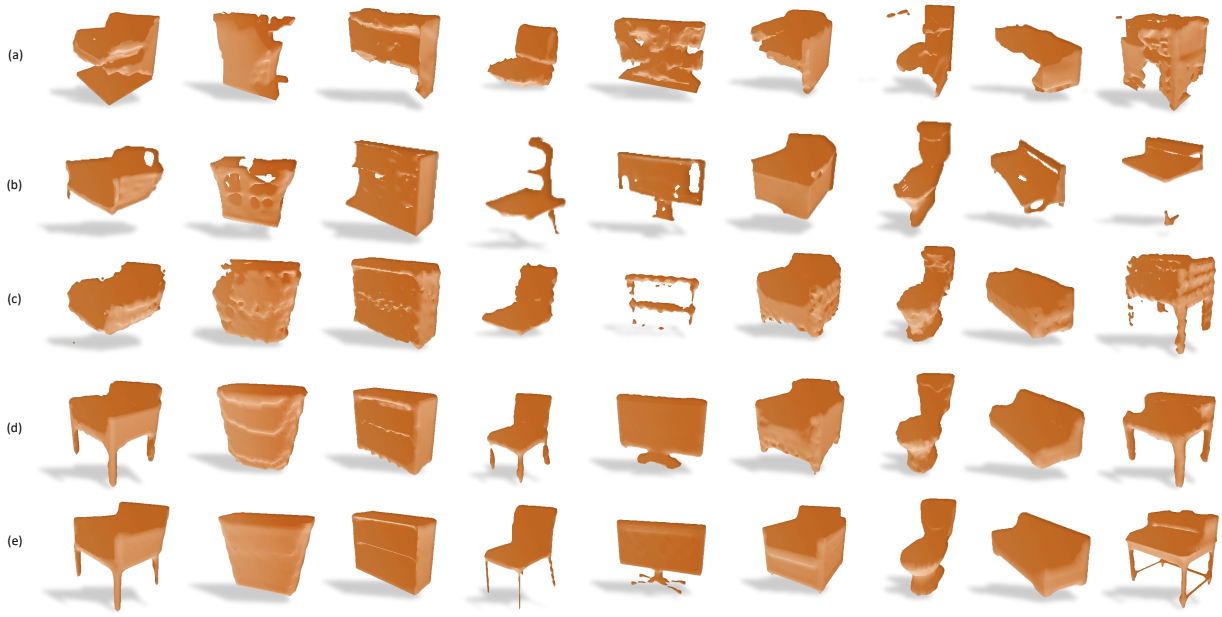

Figure 4. Qualitative comparison with state-of-the-art methods on ScanNet [5] data. In the figure above, (a)-input shape, (b)-AutoSDF [6], (c)-PatchComplete [7], (d)-Ours and (e)-ground truth objects.