3D-LMNet: Latent Embedding Matching for Accurate and Diverse 3D Point Cloud Reconstruction from a Single Image : Supplementary Material

Priyanka Mandikal*
priyanka.mandikal@gmail.com
Navaneet Murthy*
navaneetl@iisc.ac.in
Mayank Agarwal*
mayankgrwl97@gmail.com
R Venkatesh Babu
venky@iisc.ac.in

Video Analytics Lab,
Department of Computational and Data Sciences,
Indian Institute of Science,
Bangalore, India

1 Training Dataset Details

We train all our networks on synthetic models from the ShapeNet [1] dataset. We use the same 80%-20% train/test split provided by [2] consisting of models from 13 different categories, so as to be comparable with the previous works. We use the input images provided by [2], where each model is pre-rendered from 24 different azimuth angles. We crop the images to 128 × 128 resolution before passing it through our network. For generating the ground truth point cloud, we uniformly sample 16384 points on the mesh surface using farthest point sampling.

2 Network Architectures

We provide network architecture details for the point cloud and image encoders and the common decoder in Tables 1,2 and 3. It should be noted that 3D-LMNet has a total of 22.7M parameters, while PSGN [3] has nearly double the number with 42.9M parameters.

S.No.	Layer	Filter Size	Output Size	Params
1	conv	1x1	2048x64	0.4K
2	conv	1x1	2048x128	8.6K
3	conv	1x1	2048x128	16.8K
4	conv	1x1	2048x256	33.5K
5	conv	1x1	2048x512	132.6K
6	maxpool	-	512	0

Table 1: Point Cloud Encoder Architecture

* equal contribution

S.No.	Layer	Filter Size/ Stride	Output Size
1	conv	3x3/1	64x64x32
2	conv	3x3/1	64x64x32
3	conv	3x3/2	32x32x64
4	conv	3x3/1	32x32x64
5	conv	3x3/1	32x32x64
6	conv	3x3/2	16x16x128
7	conv	3x3/1	16x16x128
8	conv	3x3/1	16x16x128
9	conv	3x3/2	8x8x256
10	conv	3x3/1	8x8x256
11	conv	3x3/1	8x8x256
16	conv	5x5/2	4x4x512
17	linear	-	128

Table 2: Image Encoder Architecture

S.No.	Layer	Output Size
1	linear	256
2	linear	256
3	linear	1024*3

Table 3: Decoder Architecture

3 Quantitative Comparison of 3D-LMNet Variants on ShapeNet

We report the category-wise Chamfer and EMD error metrics for all our latent matching variants on the validation split provided by [2] for the ShapeNet dataset [1] in Table 4. Our latent matching approaches (3D-LMNet-$\mathcal{L}_1$ and $\mathcal{L}_2$) significantly outperform the network trained directly with Chamfer loss (3D-LMNet-Chamfer). 3D-LMNet-$\mathcal{L}_1$ is better in all categories in Chamfer scores, and all but one category in terms of EMD scores.

4 Reconstructions on ShapeNet

Qualitative comparison with state-of-art and baseline for single-view reconstruction on ShapeNet validation set are provided in Figs. 1 and 2. Note that the samples are randomly selected.

5 Reconstructions on Pix3D

Qualitative comparison with state-of-art and baseline for single-view reconstruction on the real-world Pix3D dataset are shown in Fig. 3. Note that the samples are randomly selected.

6 Generating Multiple Plausible Outputs

We provide more examples for the probabilistic latent matching scheme explained in the paper in Fig. 4. We notice variations in legs, handles and back of the chair models.

Category	Chamfer			EMD		
	3D-LMNet Chamfer	3D-LMNet $\mathcal{L}_2$	3D-LMNet $\mathcal{L}_1$	3D-LMNet Chamfer	3D-LMNet $\mathcal{L}_2$	3D-LMNet $\mathcal{L}_1$
airplane	4.47	3.39	**3.34**	7.35	4.81	**4.77**
bench	5.03	4.74	**4.55**	5.38	5.17	**4.99**
cabinet	6.76	6.26	**6.09**	7.03	6.73	**6.35**
car	4.70	4.61	**4.55**	4.31	4.20	**4.10**
chair	6.72	6.54	**6.41**	8.16	8.11	**8.02**
lamp	8.31	7.28	**7.10**	17.21	16.03	**15.80**
monitor	6.96	6.65	**6.40**	7.66	7.53	**7.13**
rifle	3.03	2.79	**2.75**	6.67	**6.06**	6.08
sofa	6.20	6.00	**5.85**	5.97	5.80	**5.65**
speaker	8.77	8.33	**8.10**	9.20	9.61	**9.15**
table	6.59	6.16	**6.05**	8.34	7.95	**7.82**
telephone	5.62	4.87	**4.63**	7.50	5.79	**5.43**
vessel	4.76	4.45	**4.37**	6.92	5.84	**5.68**
mean	5.99	5.54	**5.40**	7.82	7.20	**7.00**

Table 4: Category-wise 3D reconstruction metrics for different latent matching variants of 3D-LMNet on the ShapeNet dataset [1]. All metrics are scaled by 100.

7 Auto-Encoder Results

7.1 Reconstructions

3D point cloud reconstruction results are shown in Fig. 5. The reconstructions are very similar to the ground truth point clouds in appearance and spread.

7.2 Latent Space Interpolations

We analyze the quality of the learnt latent space of the auto-encoder by manipulating the latent vector z, and visually observing the generated reconstructions. Fig.6 shows the resulting reconstructions as we linearly interpolate between two different models in the test set. We find that the interpolations are smooth and the intermediate reconstructions form valid models even in the cross-category setting.

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.

[2] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3D-r2n2: A unified approach for single and multi-view 3D object reconstruction. In *European Conference on Computer Vision*, pages 628–644. Springer, 2016.

[3] Haoqiang Fan, Hao Su, and Leonidas Guibas. A point set generation network for 3D object reconstruction from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 38, 2017.

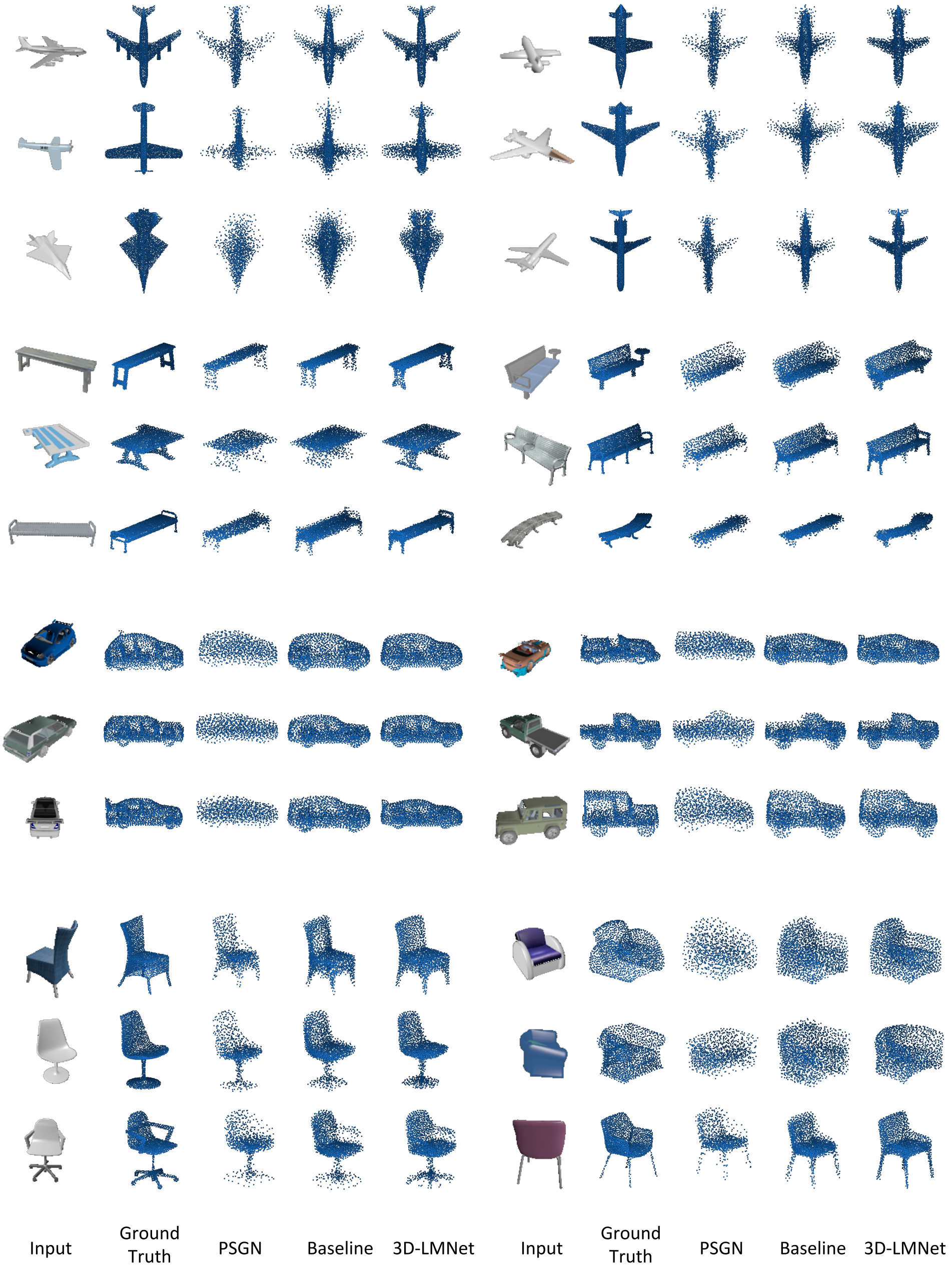

Figure 1: Reconstructions on ShapeNet. 3D reconstructions on randomly sampled input images from the validation set of ShapeNet. Note that although the baseline reconstructions for cars obtain a good shape, the points are unevenly distributed which results in high EMD error metrics (main text Table 4). On the other hand, 3D-LMNet reconstructions are well distributed and obtain lower EMD error metrics. Results best viewed zoomed.

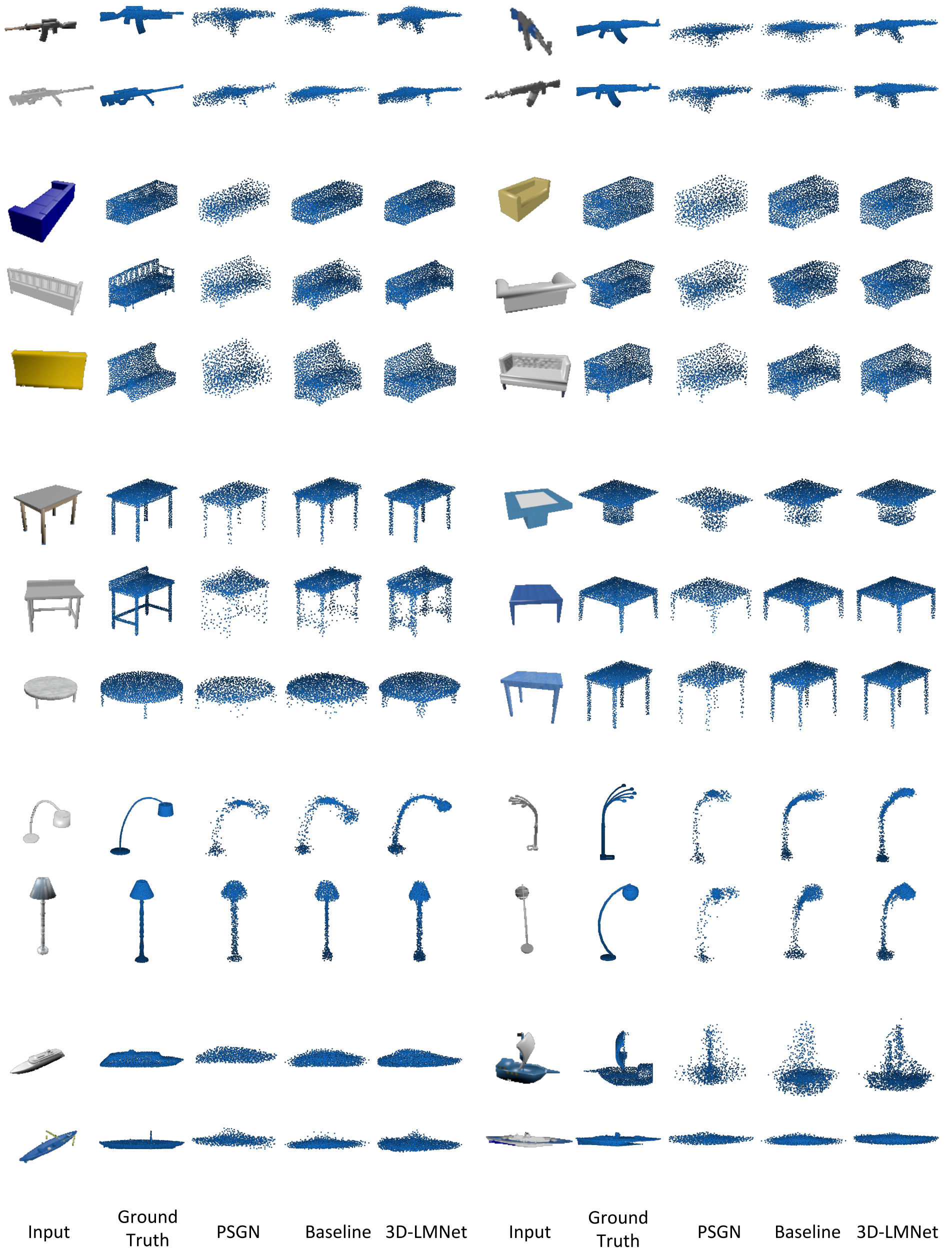

Figure 2: Reconstructions on ShapeNet. 3D reconstructions on randomly sampled input images from the validation set of ShapeNet. Results best viewed zoomed.

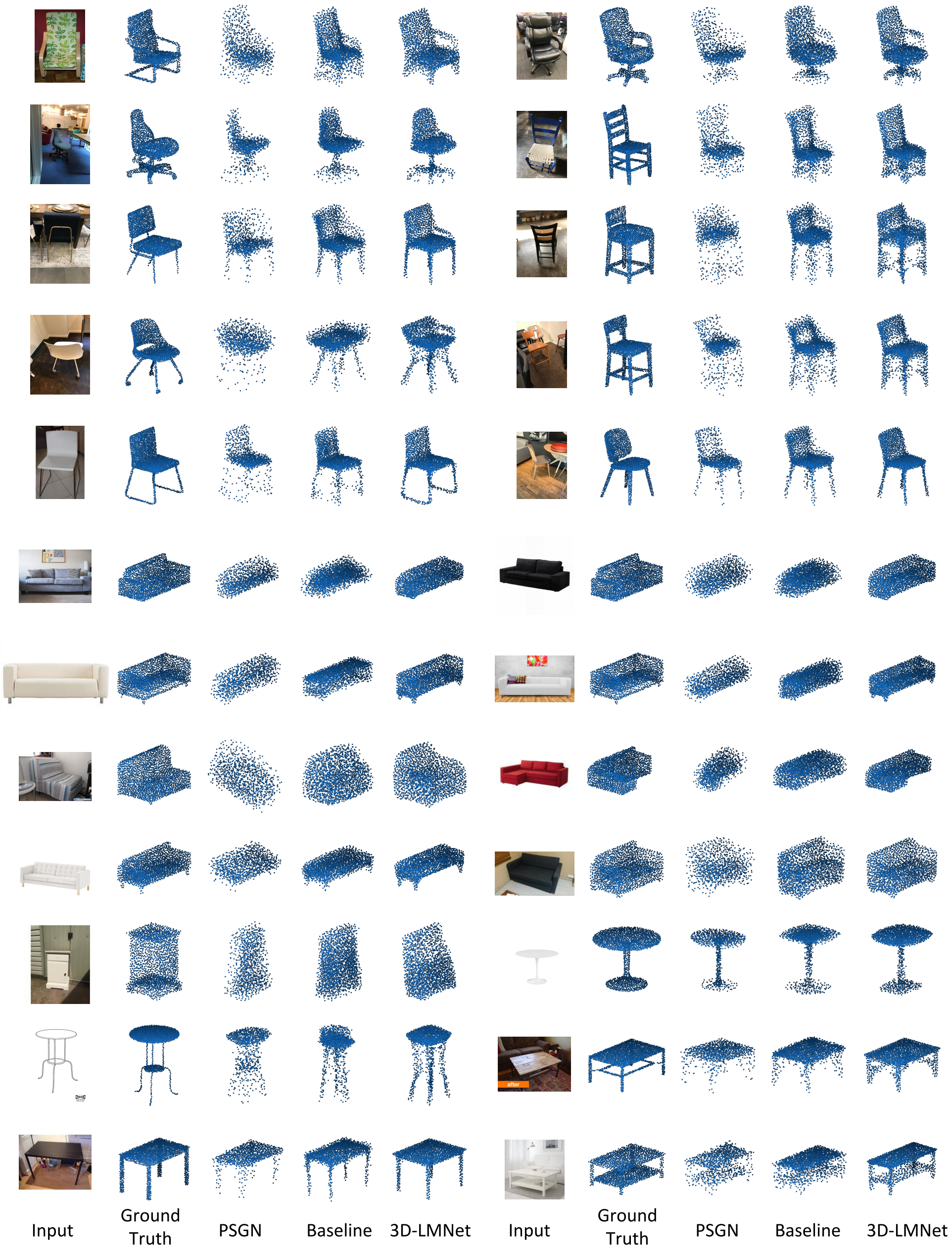

Figure 3: Reconstructions on Pix3d. 3D reconstructions on randomly sampled input images from Pix3D. Results best viewed zoomed.

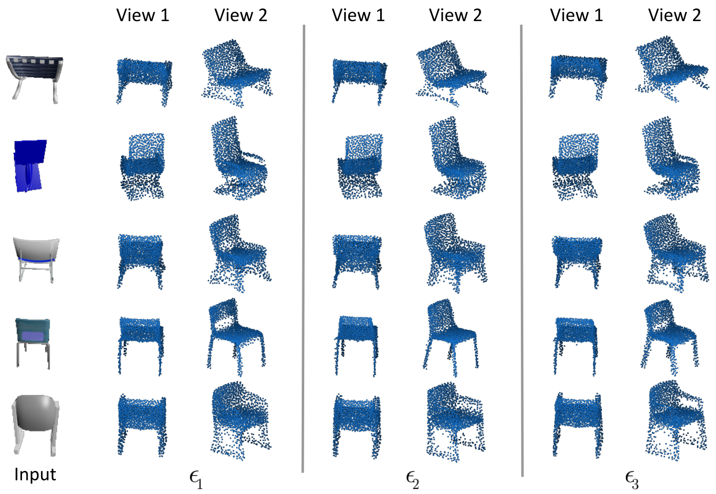

Figure 4: Qualitative results for probabilistic latent matching. Multiple reconstructions for ambiguous input views are obtained by sampling ε. Reconstruction results are shown from two different viewing angles for each ε so as to highlight the correspondence with the input image.

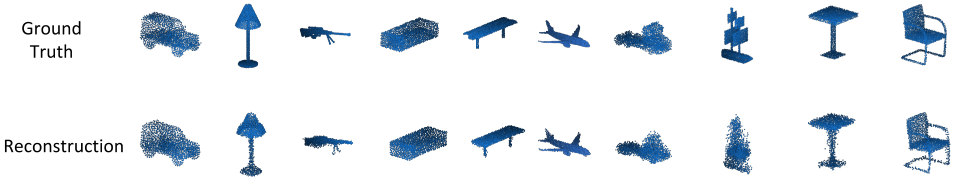

Figure 5: Auto-Encoder Reconstructions

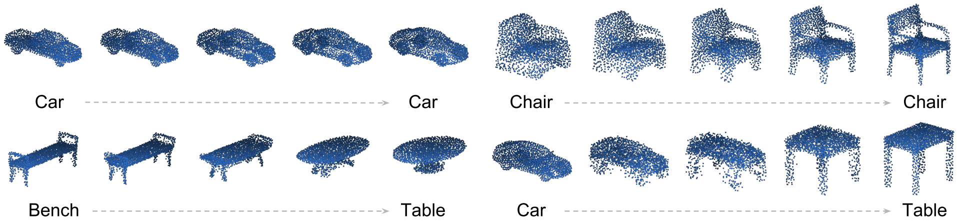

Figure 6: Auto-Encoder Interpolations