StructureNet: Hierarchical Graph Networks for 3D Shape Generation

Supplementary Material

KAICHUN MO*, Stanford University
PAUL GUERRERO*, University College London
LI YI, Stanford University
HAO SU, University of California, San Diego
PETER WONKA, KAUST
NILOY J. MITRA, University College London and Adobe Research
LEONIDAS J. GUIBAS, Stanford University and Facebook AI Research

ACM Reference Format:
Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy J. Mitra, and Leonidas J. Guibas. 2019. StructureNet: Hierarchical Graph Networks for 3D Shape Generation: Supplementary Material. *ACM Trans. Graph.* 38, 6, Article 242 (November 2019), 12 pages. https://doi.org/10.1145/3355089.3356527

1 JOINT EMBEDDING OF SHAPES AND IMAGES

The joint embedding of multiple modalities described in Section 6.4 of our paper can also be used for retrieval. For example, instead of looking up the encoded shape that is closest to an encoded image, we can look up the images that are closest to an encoded shape, and thereby get the top-k image matches for a given shape. In Figure 1 we show the top-3 images and point clouds for a given query shape. Qualitatively, most of the retrieved results are a good match to the query shape.

Joint embedding. A two-dimensional t-SNE embedding [Maaten and Hinton 2008] of the joint multi-modal latent space is shown in Figure 2. We show representative samples on a grid, choosing at each location randomly one of the modalities: shapes, images or point clouds. We can see that the distributions of the different modalities align well; nearby samples tend to represent similar shapes. Sofa chairs, for example, are clustered on the left side of the diagram for all modalities, and on the right side, we find chairs with backrests that have multiple vertical bars. Furthermore, the learned latent space is 'structurally smooth' that nearby regions tend to be connected by natural transitions between the structures of the chairs, which is also confirmed by the interpolation experiments in Section 6.3. of the paper.

Part-based retrieval. Retrieval based on individual parts, for example, retrieving chairs with backrests similar to a query shape, can be done by training a separate encoder for each part type that we want to retrieve. The bottom three rows of Figure 1 show part-based retrieval results for the base and backrest of chairs, compared to performing a retrieval based on the full shape. To retrieve images with similar bases, for example, we train an encoder similar to Section 6.4 of our paper, but trained to using the latent space of chair bases only instead of full chairs. Unlike the latent space of the full shape, the latent space of parts is not specifically regularized to be smooth. Still, we can see from the successful retrieval results, that

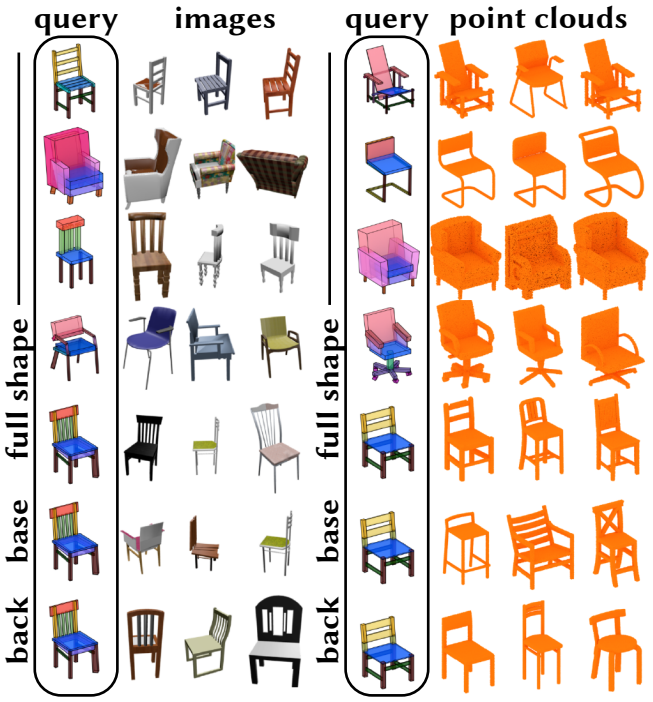

Fig. 1. **Image and point cloud retrieval.** Images and point clouds are retrieved using on a query shape based on the distance from the query in the multi-modal latent space. The bottom three rows compare retrieval with the full shape as query to part-based retrieval using the backrest and chair base only. The retrieved shapes are similar to the query, showing that similar shapes have a small distance in latent space, even across modalities.

*joint first authors

Webpage: https://cs.stanford.edu/~kaichun/structurenet/.
Emails: kaichun@cs.stanford.edu; paul.guerrero@ucl.ac.uk; ericyi@stanford.edu; haosu@eng.ucsd.edu; peter.wonka@kaust.edu.sa; n.mitra@ucl.ac.uk; guibas@cs.stanford.edu.

the latent space of individual parts tends to be meaningful, where feature vectors that have a small distance in latent space correspond to similar parts.

2 ABLATION STUDY

We performed an ablation study to evaluate the contribution of individual components to our method for the shape reconstruction experiments. Results are shown in Table 1. Specifically, we trained 5 variations of our method, removing a combination of components in each. Components we examined are the *message passing* performed in the decoder, where, different from the encoder, it is not strictly necessary to handle relationship edges, the *normal reconstruction loss* $\mathcal{L}_{\text{normal}}$, and the *structure consistency loss* $\mathcal{L}_{\text{sc}}$. The normal reconstruction loss noticeably decreases the geometry reconstruction error $E_{\mathbf{P}}$ and together with the structure consistency loss $\mathcal{L}_{\text{sc}}$, lowers the consistency errors. The normal and structure consistency losses come at the cost of a slightly increased hierarchy error $E_{\mathbf{H}}$, presumably since these losses encourage the network to focus more resources on the part geometry, as opposed to the hierarchy. This cost is reduced by message passing, which significantly lowers the hierarchy error. Finally, we also compare to removing edges altogether, which results in a significant increase in the geometry reconstruction error.

Table 1. **Ablation study.** We compare our full method (bottom row) to a version without combinations of message passing (mp), the normal loss $\mathcal{L}_{\text{normal}}$ (nl), and the structure consistency loss $\mathcal{L}_{\text{sc}}$ (scl). In the top row we show a version that does not use relationship edges. The normal and edge loss both increase consistency significantly, at a small cost in the hierarchy reconstruction. Message passing improves coordination between parts, reducing this cost.

	reconstruction err.			consistency err.	
	$E_{\mathbf{P}}$	$E_{\mathbf{H}}$	$E_{\mathbf{R}}$	E_{rc}	E_{gc}
no edges	0.0662	0.194			0.0288
- (mp, scl, nl)	0.0649	0.192	0.240	0.0323	0.0365
- (mp, scl)	0.0616	0.198	0.243	0.0216	0.0259
- (nl)	0.0631	0.201	0.254	0.0323	0.0380
- (scl)	0.0649	0.201	0.249	0.0194	0.0242
- (mp)	0.0621	0.212	0.250	0.0186	0.0223
StructureNet	0.0620	0.200	0.246	0.0183	0.0226

3 IMPLEMENTATION

We implement StructureNet in PyTorch [Paszke et al. 2017]. All sub-networks of our hierarchical graph networks are implemented as simple Multilayer Perceptrons (MLPs) with ReLU non-linearities, and without batch normalization [Ioffe and Szegedy 2015], except for the specialized encoders for images and unannotated point clouds, and the pre-trained point cloud autoencoder for the part geometry. We use a batch size of 32 shapes. Due to the difficulty of batched training with recursive networks, we compute the loss for each shape separately before summing the per-shape losses up to obtain the loss for the batch. Back-propagation is performed on the batch loss. Typically, our networks for bounding box geometry converge in $1-2$ days, whereas the networks for point cloud geometry require $2-4$ days to train on a single GeForce RTX 2080 Ti and an Intel i9-7940X CPU. Memory consumption is modest, at approximately $1-2$ GB. We will release code and datasets upon acceptance.

4 SEMANTIC HIERARCHIES

We present the PartNet [Mo et al. 2019] semantics hierarchies for chairs (Figure 3), tables (Figure 4) and storage furnitures (Figure 5) that we use in this paper. We assign the semantic labels in the figures with the colors that we use for box-shape and point cloud visualization in the main paper.

5 MORE OBJECT CATEGORIES

In Figures 6 and 7, we show shape generation and interpolation results for two additional object categories in PartNet: vases and trash cans. Additionally, we show a training attempt on a severely under-sampled dataset in Figure 8. See the captions for more detailed descriptions.

6 ADDITIONAL GENERATED SHAPES

We show more StructureNet VAE generation results for box-shapes in Figure 9 and for point cloud shapes in Figure 10.

7 ADDITIONAL SHAPE INTERPOLATIONS

We show more StructureNet VAE interpolation results for box-shapes and point cloud shapes in Figure 11.

8 ADDITIONAL SHAPE ABSTRACTIONS

We show more StructureNet shape abstraction results from 2D images, 3D point clouds or partial scans in Figure 12.

REFERENCES

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

Kaichun Mo, Shilin Zhu, Angel Chang, Li Yi, Subarna Tripathi, Leonidas Guibas, and Hao Su. 2019. PartNet: A Large-scale Benchmark for Fine-grained and Hierarchical Part-level 3D Object Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

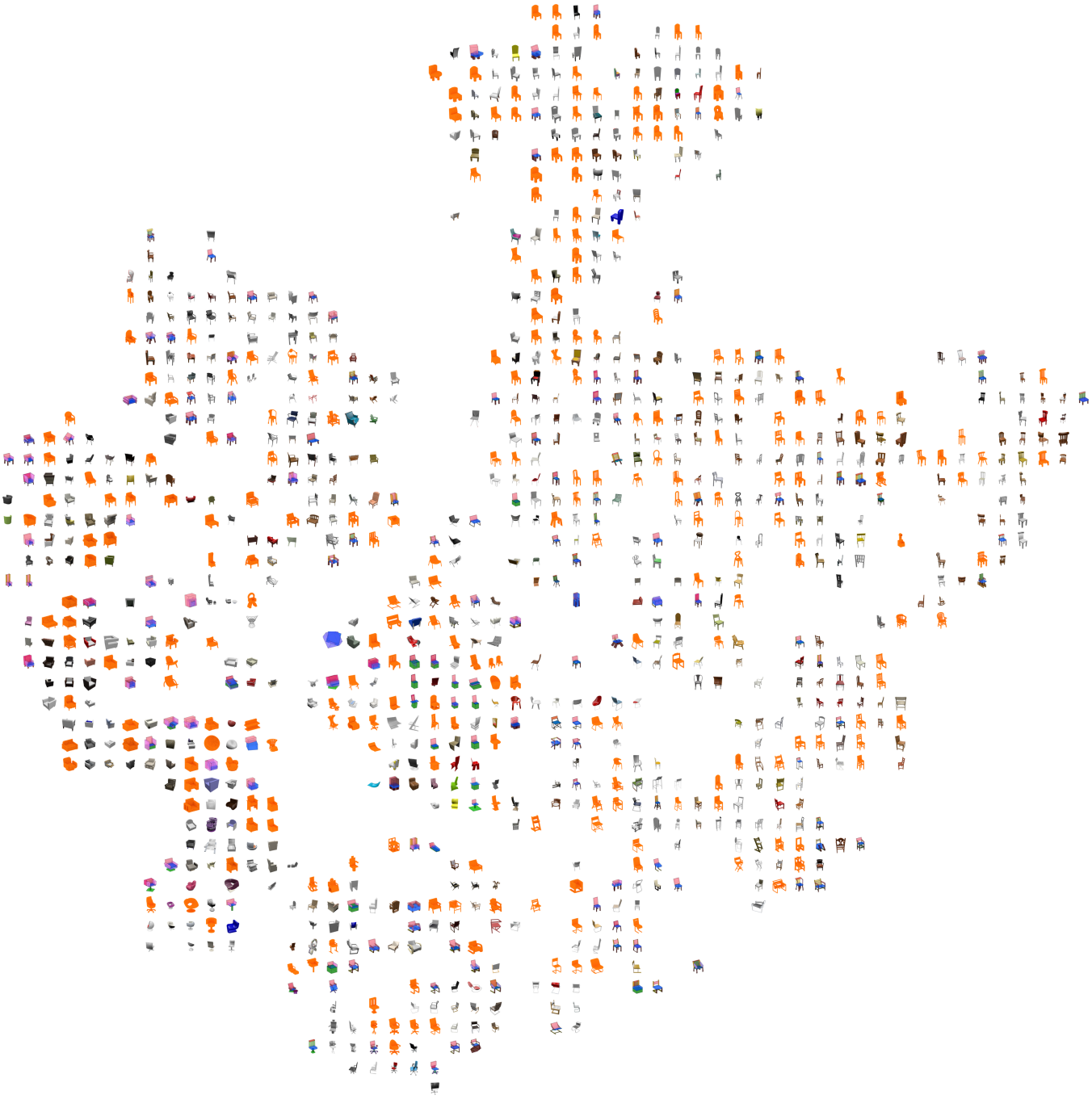

Fig. 2. **Joint embedding of images, point clouds and shapes.** We visualize the multi-modal latent space as a two-dimensional embedding. At each grid point, we randomly show one of the modalities.

Fig. 3. **PartNet semantic hierarchy for chairs.** Dash lines show the OR-nodes and solid lines show the AND-node in PartNet. We assign the semantic labels in the figures with the colors that we use for box-shape and point cloud visualization in the main paper.

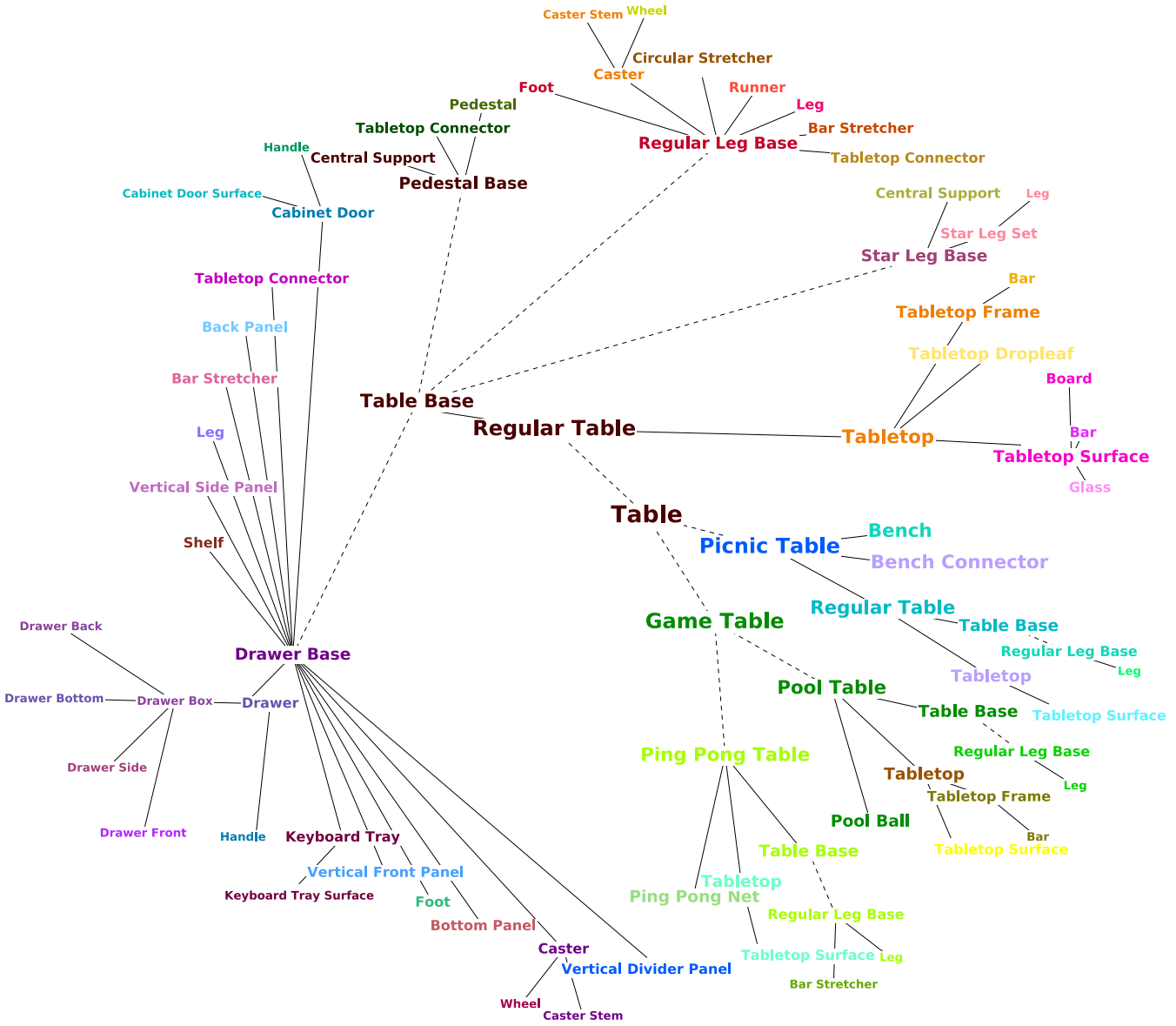

Fig. 4. **PartNet semantic hierarchy for tables.** Dash lines show the OR-nodes and solid lines show the AND-node in PartNet. We assign the semantic labels in the figures with the colors that we use for box-shape and point cloud visualization in the main paper.

Fig. 5. **PartNet semantic hierarchy for storage furnitures.** Dash lines show the OR-nodes and solid lines show the AND-node in PartNet. We assign the semantic labels in the figures with the colors that we use for box-shape and point cloud visualization in the main paper.

Fig. 6. **Shape generation results for vases and trash cans.** The datasets for these categories are smaller than for our main categories: 505 samples for vases and 83 for trashcans. Vases have a less complex structure compared to the other categories, making the quality of the generated geometry more important, while trashcans have a wider range of structures.

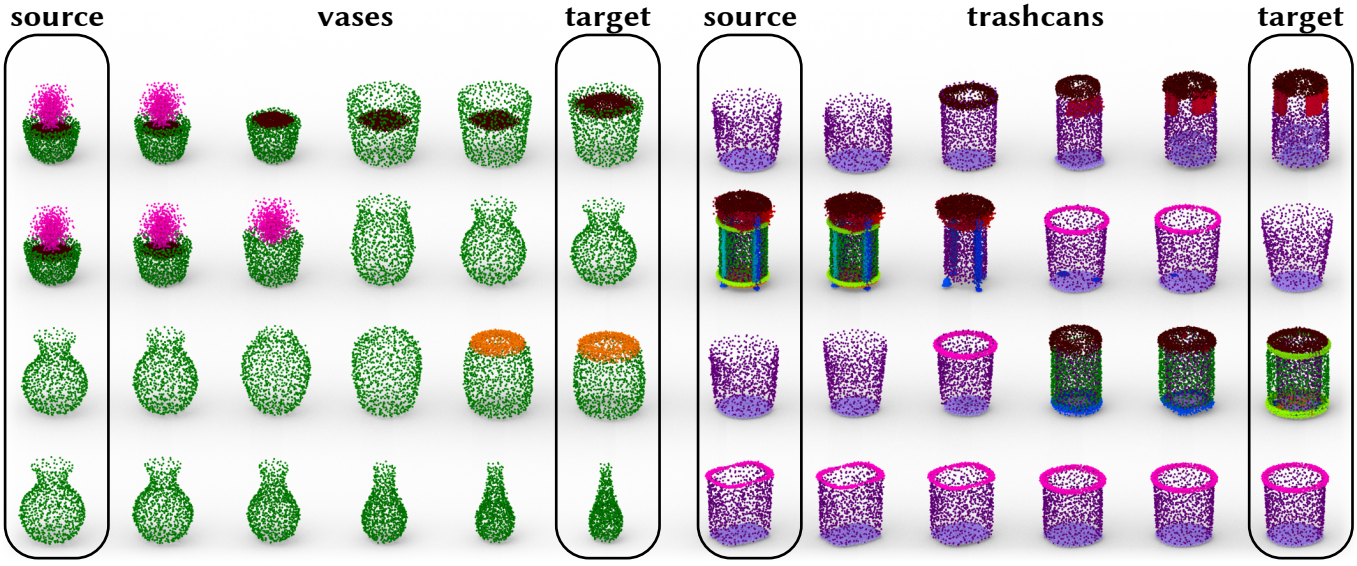

Fig. 7. **Shape interpolation results for vases and trash cans.** Similar to our main categories, structure is interpolated smoothly. The last rows for vases and trash cans show that the part geometry is interpolated smoothly as well.

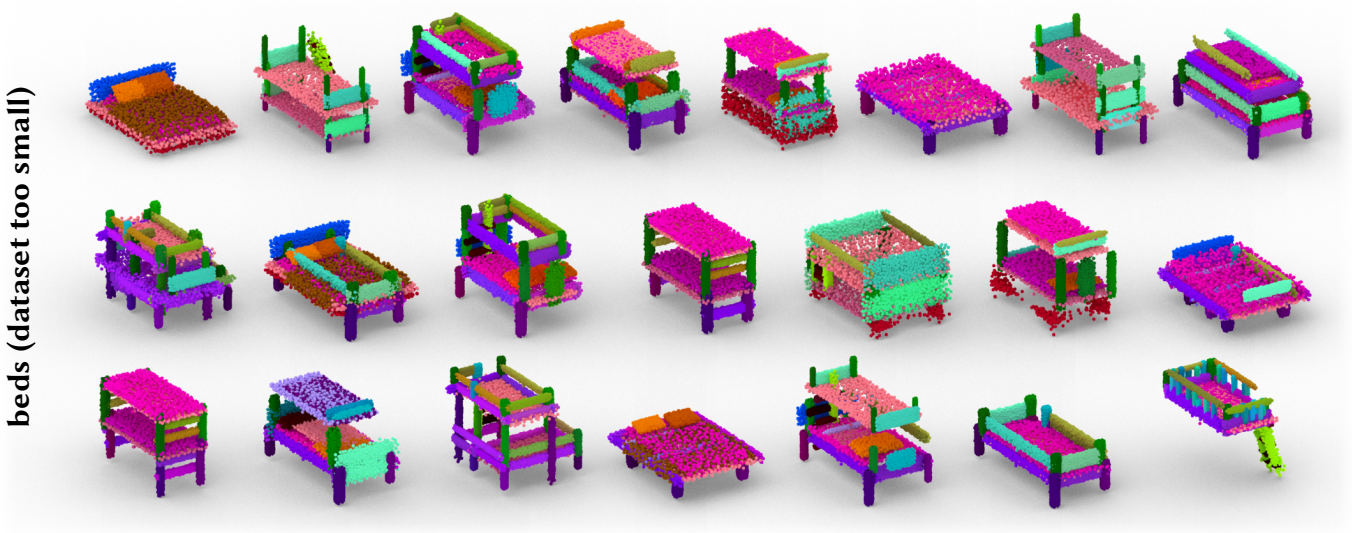

Fig. 8. **Shape generation results for beds.** We also test on this third, severely under-sampled category, with a training set size of 54. As we can see in Figure 8, the network is experimenting with different structures, but the size of our dataset is not large enough for the network to reliably distinguish between realistic and unrealistic beds.

Fig. 9. More Box-shape Generation Results.

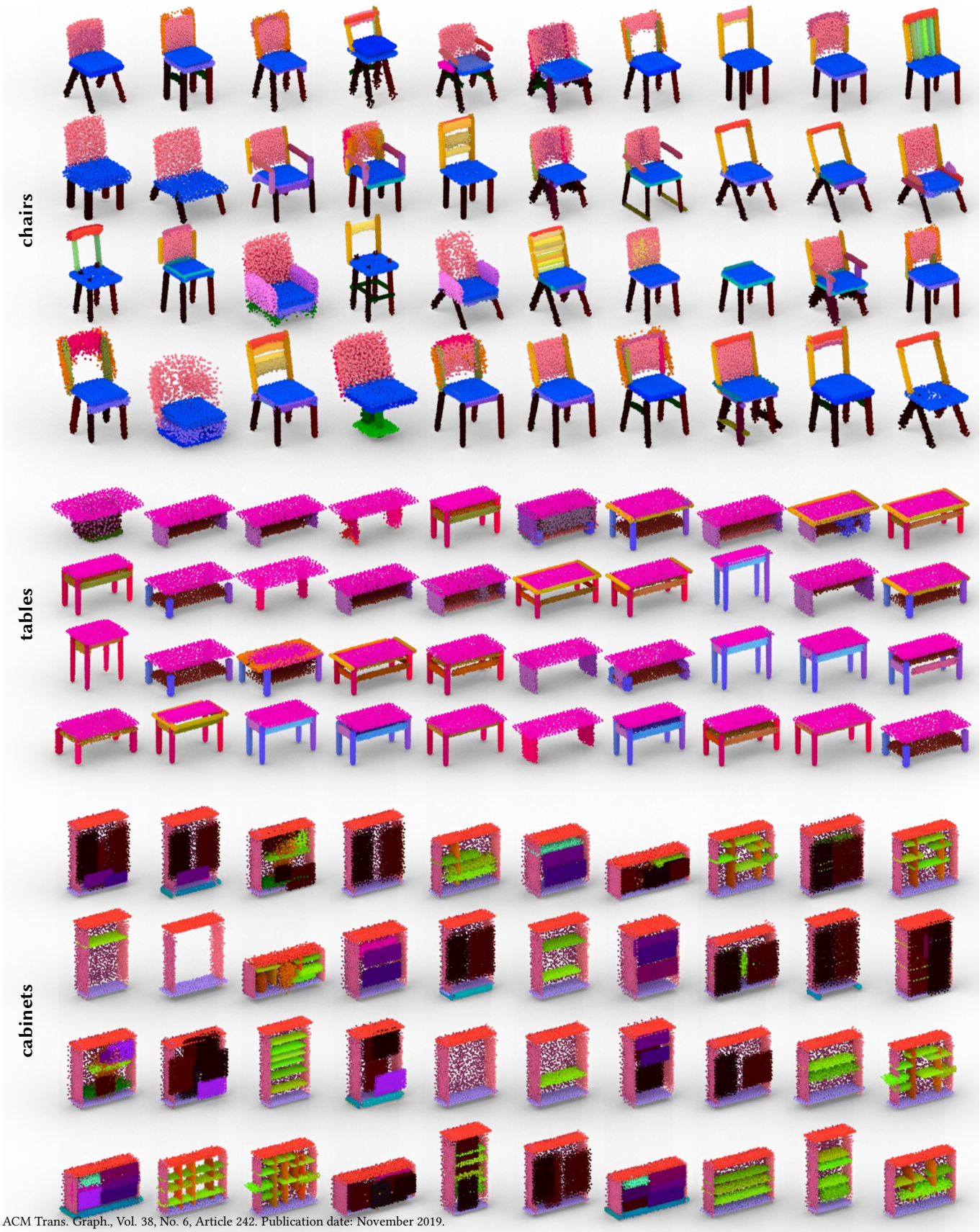

Fig. 10. More Point Cloud Generation Results.

Fig. 11. More Shape Interpolation Results.

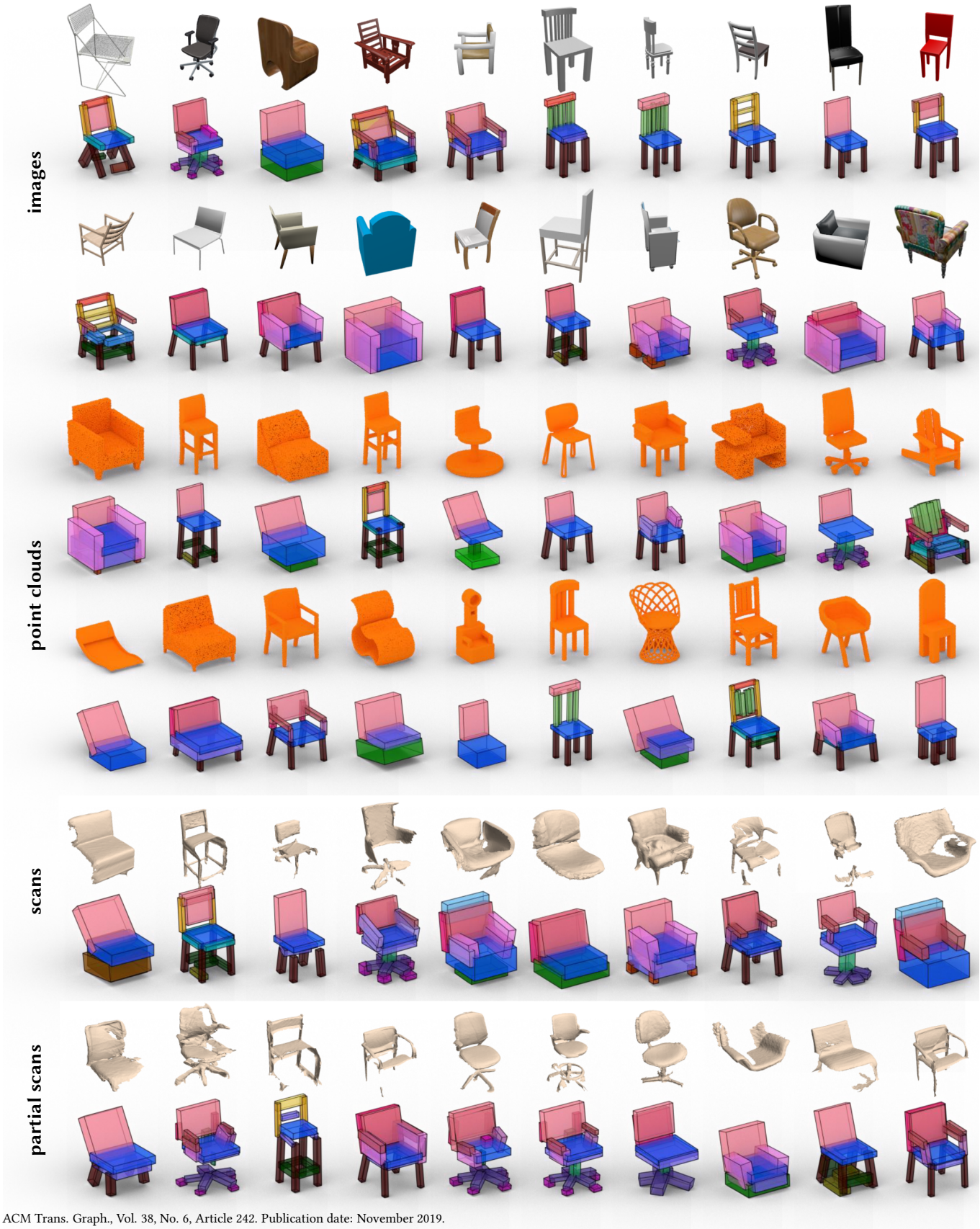

Fig. 12. More Shape Abstraction Results.