Supplemental Material
Mask2CAD: 3D Shape Prediction by Learning to Segment and Retrieve

Weicheng Kuo[1,2], Anelia Angelova[1,2], Tsung-Yi Lin[1,2], and Angela Dai[3]

[1] Google AI
[2] Robotics at Google
[3] Technical University of Munich
{weicheng,anelia,tsungyi}@google.com, angela.dai@tum.de

1 Additional Results on Pix3D

In Figure 1, we show more qualitative results of our approach on Pix3D [2]. Furthermore, we conduct ablation studies to shed light on the roles of each component in the system. Our analysis shows that shape, pose, and translation are all important for estimating the viewer-centered geometry, with shape retrieval having the most room for improvement, and box detection having the least. The analysis was done by replacing each predicted component with its ground truth counterpart. In terms of Mesh AP, groundtruth shapes help by +14.6, rotation by +10.2, and translation by +7.5. Surprisingly, groundtruth 2D boxes offer no improvement because the detections on Pix3D are very good (90 Box AP, similar to Mesh R-CNN) and the small advantage is offset by the distribution shift between train (jittered boxes) and test (perfect boxes) time. This agrees with what Mesh R-CNN reports, i.e. they also observed a loss when using ground truth boxes (6 point loss on Mesh AP50).

2 Network Architecture Details

The Mask2CAD image-stream network architecture comprises 2D detection as bounding box, class label, and instance mask prediction, as well as our 3D shape retrieval and pose estimation. For the 2D detection, our architecture borrows from that of ShapeMask [1]. For the 3D inference with shape embedding, pose classification, and pose regression, and object center prediction, these branches all use the same architecture as the coarse mask prediction branch of [1] (with the exception of the output layers). The inputs of these branches are the features from the region of interest (ROI) of detection backbone feature pyramid network. We detail each branch in Table 1, 2, and 3.

3 t-SNE visualizations for image-CAD embeddings

Figures 2, 3, 4 show the t-SNE visualizations of the image-shape embedding spaces for the bed, wardrobe, desk, table, tool, misc, and chair classes.

Fig. 1. Additional qualitative results of Mask2CAD on Pix3D [2].

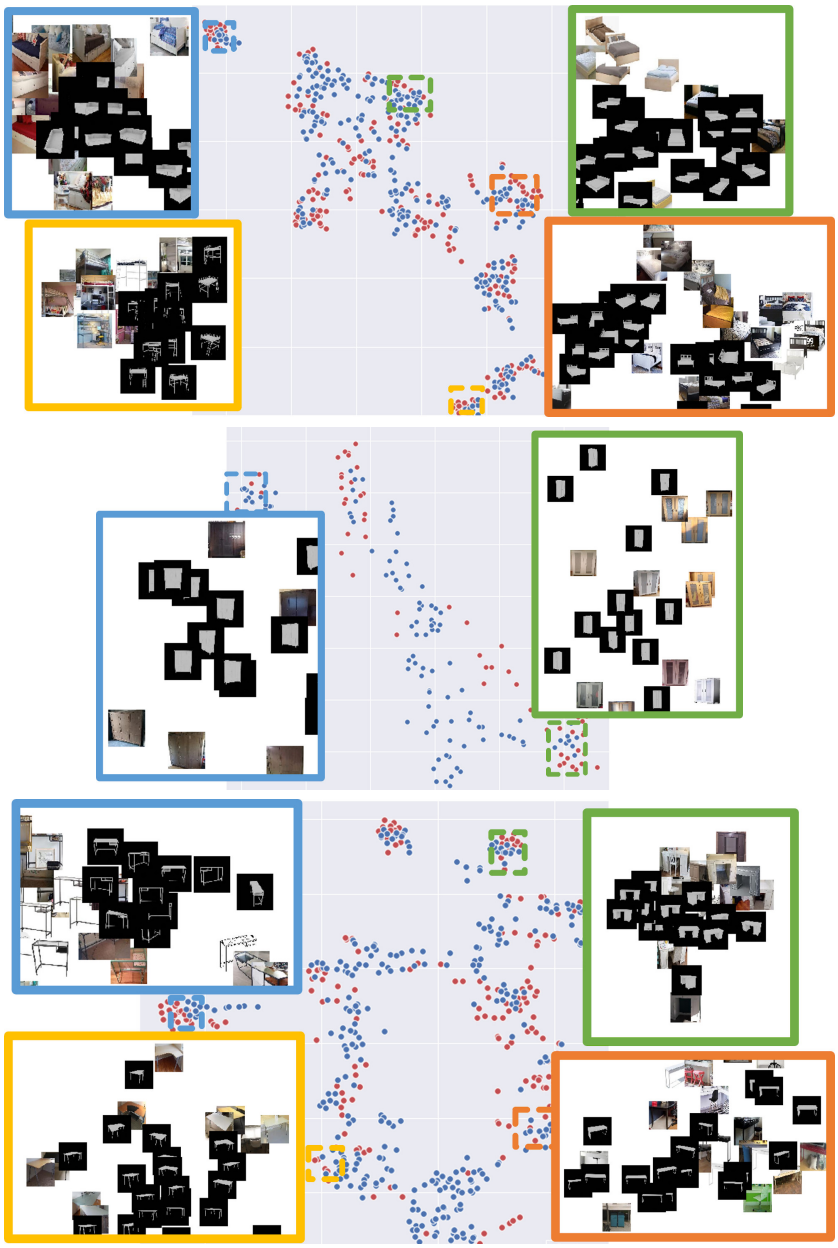

Fig. 2. t-SNE embeddings of Mask2CAD for the bed (top), wardrobe (middle) and desk (bottom) classes. Red points correspond to images, and blue to shapes.

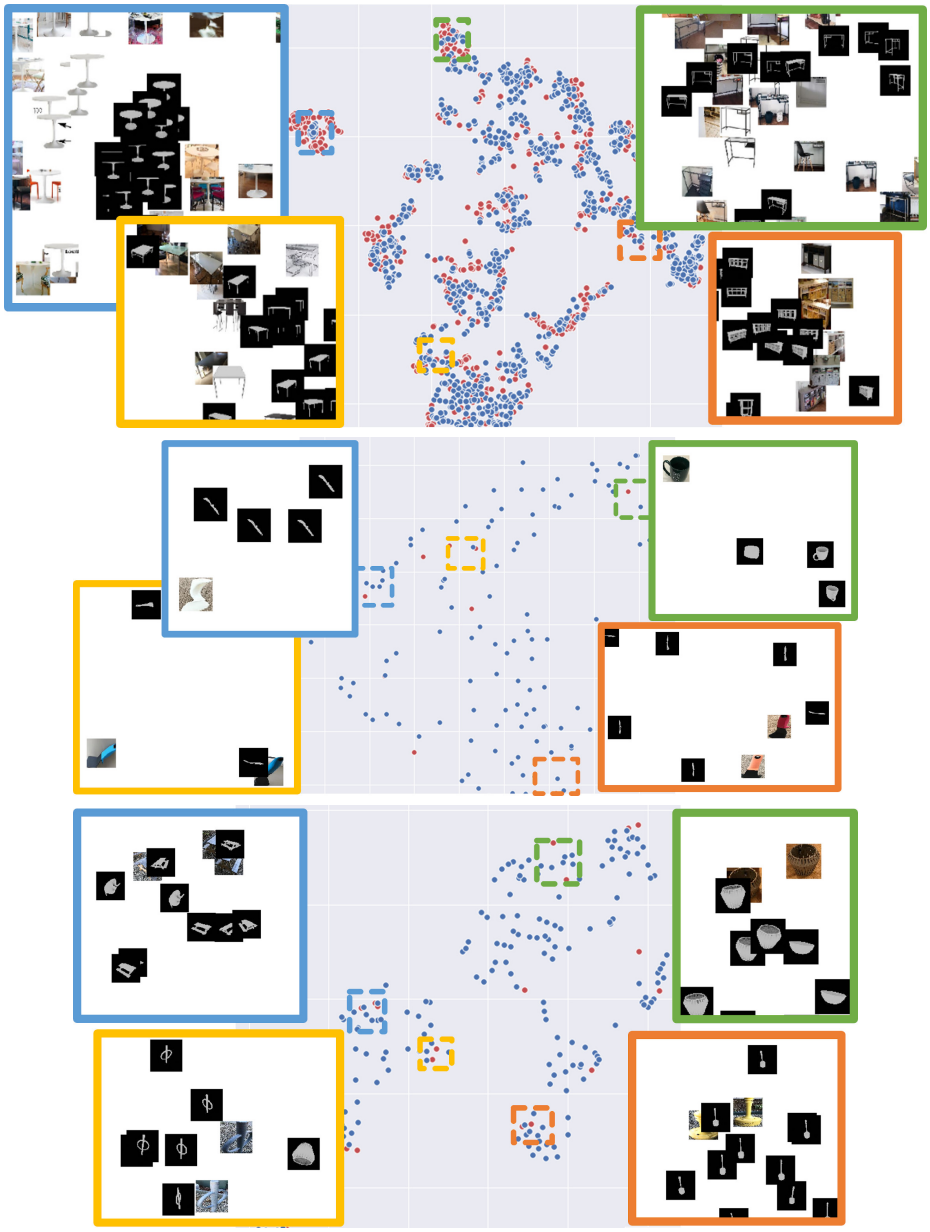

Fig. 3. t-SNE embeddings of Mask2CAD for the table (top), tool (middle) and misc (bottom) classes. Red points correspond to images, and blue to shapes.

Index	Inputs	Operation	Output shape
(1)	Input	Region of Interest (ROI) features	$32 \times 32 \times 256$
(2)	(1)	$3\times$ (Conv($256 \rightarrow 256, 3 \times 3$) + BatchNorm + ReLU)	$32 \times 32 \times 256$
(3)	(2)	(Conv($256 \rightarrow 256, 3 \times 3$) + BatchNorm + ReLU)	$32 \times 32 \times 128$
(4)	(3)	AveragePool(axes=(0, 1))	128

Table 1. Network architecture of the shape embedding branch. The last convolution layer downsamples the number of channels from 256 to 128.

Index	Inputs	Operation	Output shape
(1)	Input	Region of Interest (ROI) features	$32 \times 32 \times 256$
(2)	(1)	$4\times$ (Conv($256 \rightarrow 256, 3 \times 3$) + BatchNorm + ReLU)	$32 \times 32 \times 256$
(3)	(2)	AveragePool(axes=(0, 1))	256
(4)	(3)	Linear($256 \rightarrow N_{pose} \times N_{class}$)	160

Table 2. Network architecture of the pose prediction branch. For pose classification, the output is $N_{pose} = 16$ for each class $N_{class} = 10$. For the following pose regression after this classification, the architecture is identical except for using $N_{pose} = 4$ for predicting the regression quaternion instead of the 16 medoid bins.

Index	Inputs	Operation	Output shape
(1)	Input	Region of Interest (ROI) features	$32 \times 32 \times 256$
(2)	(1)	$4\times$ (Conv($256 \rightarrow 256, 3 \times 3$) + BatchNorm + ReLU)	$32 \times 32 \times 256$
(3)	(2)	AveragePool(axes=(0, 1))	256
(4)	(3)	Linear($256 \rightarrow N_{center} \times N_{class}$)	20

Table 3. Network architecture of the object center regression branch. The output is $N_{center} = 2$ for each class $N_{class} = 10$, where N_{center} equals 2 for (δ_x, δ_y).

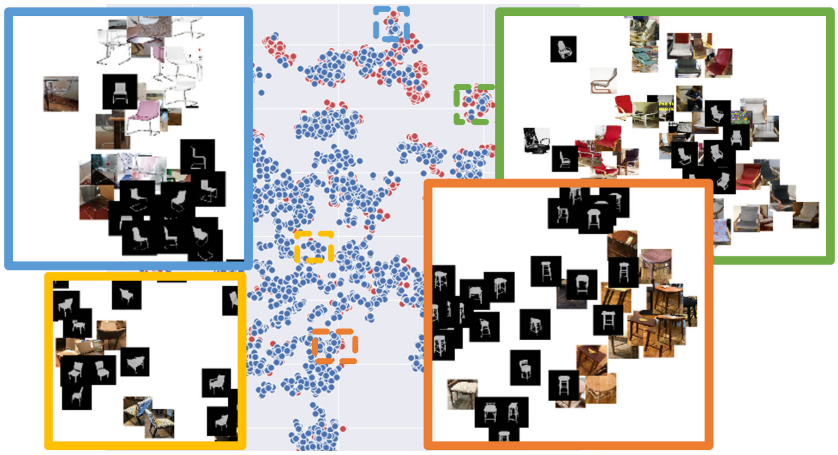

Fig. 4. t-SNE embedding of Mask2CAD for the chair class. Red points correspond to images, and blue to shapes.

References

1. Kuo, W., Angelova, A., Malik, J., Lin, T.Y.: Shapemask: Learning to segment novel objects by refining shape priors. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9207–9216 (2019)
2. Sun, X., Wu, J., Zhang, X., Zhang, Z., Zhang, C., Xue, T., Tenenbaum, J.B., Freeman, W.T.: Pix3D: Dataset and methods for single-image 3D shape modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2974–2983 (2018)