Supplementary Document

TEXTCRAFT: ZERO-SHOT GENERATION OF HIGH-FIDELITY AND DIVERSE SHAPES FROM TEXT

Anonymous authors
Paper under double-blind review

1 LIMITATIONS AND FUTURE WORK

Despite providing higher quality over baseline methods, TextCraft still suffers from limitations in capturing smaller details (e.g., chair with a hole on its back) and generating shapes from text prompts that involve counting (e.g., chair with *four* slats). We believe that the quality of shapes can be improved by using implicit representations. Furthermore, it also fails on many text queries which are not present in the prior knowledge of CLIP or where there is less alignment between text and image embeddings. The method also lacks texture details which are a critical component of 3D pipelines. In future work, we will address these limitations by exploring better shape representations that capture even finer details and investigate neural networks that can count.

2 COMPONENTS AFFECTING RECONSTRUCTION QUALITY

We compare different design decisions for the VQ-VAE at 32^3 resolution . The results are shown in Table 1. In the first 4 rows, we investigate if VQ-Autoencoder is sensitive to the codebook loss hyperparameter β and find that the quality of reconstruction is not affected much when we vary it between 0.1 to 1. Next, we evaluate in the next 2 rows if the size of codebook embedding has an effect on reconstruction quality and also find minimal change. Finally, we add residual connections to the decoder and encoder, and find that this significantly improves the reconstruction quality. For all the experiments with VQ-VAE, we use the last row of Table 1 as hyperparameters. Moreover, we use the exact settings for 64^3 VQ-VAE, with the addition of another ResNet block to the encoder and the decoder.

3 TRAINING AND EXPERIMENT DETAILS

We use the Adam Optimizer (Kingma & Ba, 2014) with a learning rate of 1e-4 for all stages of training. For both the 32^3 and 64^3 VQ-VAE, we apply the ResNet architecture on a convolutional encoder and decoder with a codebook size of 512 and embedding dimensions of 64. We choose a grid size of 4^3 for the 32^3 VQ-VAE and a grid size of 8^3 for the 64^3 VQ-VAE . For the Stage 2 and Stage 3 transformers, we use a bidirectional transformer with 8 attention blocks, 8 attention heads, and a token size of 256. For Stage 2, we train the network for 250 epochs with a batch size of 32 whereas for Stage 3 we train for 300 epochs. We do not use any dropout in the transformers. For all the experiments, we use 24 renderings Choy et al. (2016) of ShapeNet13. We run both the coarse and fine transformer for 12 steps during inference.

For results in Table 1, all annealing schemes use the starting scale parameters as 4.5 with 12 sampling steps. Note for these experiments we use the best seed similar to the protocol used in CLIP-Forge. In the rest of the paper, we average over 3 seeds. We also calculate all the results on the same classifier and resolution as CLIP-Forge.

4 QUALITATIVE RESULTS

Figure. 1 and Figure. 2 provide more qualitative text-conditioned shape generation results of TextCraft for both ShapeNet13 and ShapeNet55.

beta	*emb dims*	*encoder*	*decoder*	**IOU↑**	**MSE↓**
0.1	64	VoxEnc	VoxEnc	0.8876	0.005740
0.25	64	VoxEnc	VoxEnc	0.8872	0.005808
0.50	64	VoxEnc	VoxEnc	0.8856	0.005918
1.0	64	VoxEnc	VoxEnc	0.8831	0.006081
0.25	32	VoxEnc	VoxEnc	0.8856	0.005898
0.25	128	VoxEnc	VoxEnc	0.8820	0.006092
0.1	64	Res-VoxEnc	Res-VoxEnc	0.9148	0.004333

Table 1: Different hyperparameters for stage 1 VQ-VAE at 32^3 resolution

References

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*, pp. 628–644. Springer, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

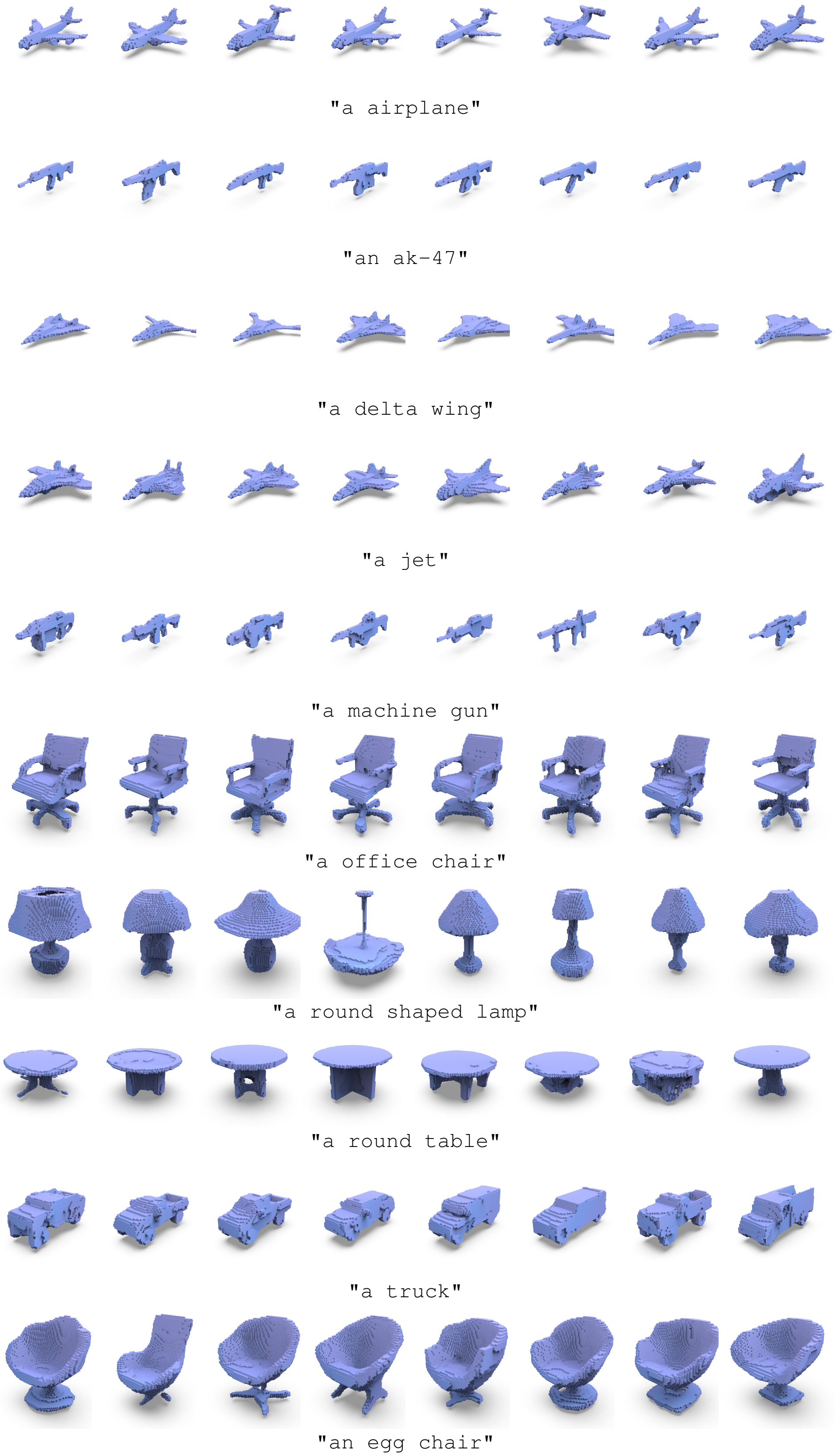
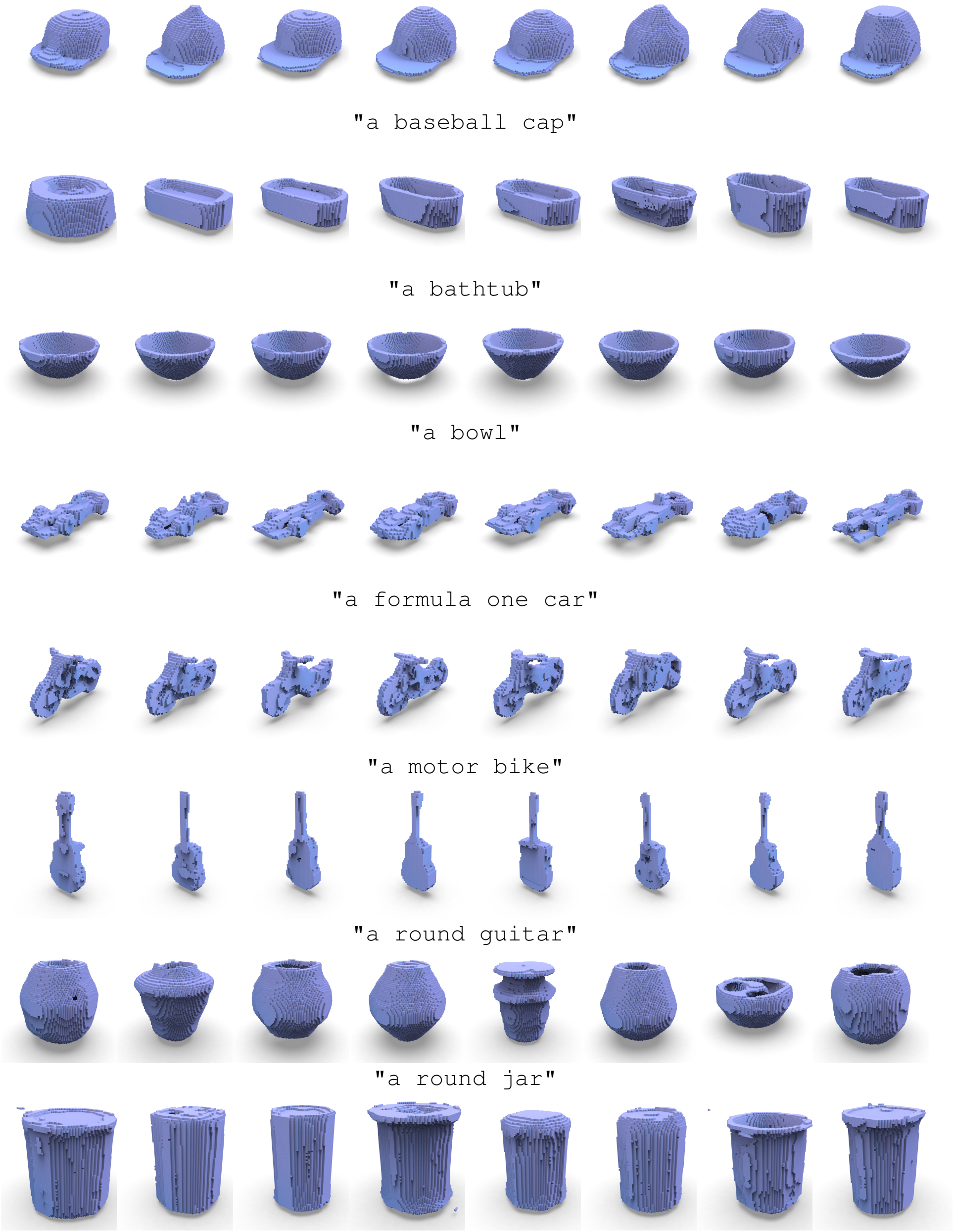


Figure 1: Multiple generated 3D shapes by TextCraft with different text input. The text inputs are (sub-)category names of ShapeNet13, and phases with semantic attributes.



"a baseball cap"

"a bathtub"

"a bowl"

"a formula one car"

"a motor bike"

"a round guitar"

"a round jar"

"a trash can"

Figure 2: Multiple generated 3D shapes by TextCraft with different text input. The text inputs are (sub-)category names of ShapeNet55, and phases with semantic attributes.