

CLIP-Forge: Towards Zero-Shot Text-to-Shape Generation

Aditya Sanghi¹

Ye Wang²

¹Autodesk AI Lab

Hang Chu¹

Chin-Yi Cheng¹

²Autodesk Research

Joseph G. Lambourne¹

Marco Fumero¹

{aditya.sanghi,hang.chu,joseph.lambourne,ye.wang,chin-yi.cheng,marco.fumero}@autodesk.com

Abstract

While recent progress has been made in text-to-image generation, text-to-shape generation remains a challenging problem due to the unavailability of paired text and shape data at a large scale. We present a simple yet effective method for zero-shot text-to-shape generation based on a two-stage training process, which only depends on an unlabelled shape dataset and a pre-trained image-text network such as CLIP. Our method not only demonstrates promising zero-shot generalization, but also avoids expensive inference time optimization and can generate multiple shapes for a given text.



“a cuboid sofa” “a round sofa” “an airplane” “a space shuttle” “an suv” “a pickup truck”

Figure 1: CLIP-Forge generates meaningful shapes without using any shape-text pairing labels.

1 Introduction

Generating 3D shapes from text input has been a challenging and interesting research problem with both significant scientific and applied value [14, 21, 16, 17]. Artificial intelligence and cognitive science researchers have long sought to bridge the two modalities of natural language and geometric shape [35, 3]. Such models are a key enabling component to new smart tools in creative design and manufacture as well as animation and games [6].

Significant progress has been made on text-to-image generation. DALL-E [32] and its associated pre-trained visual-textual embedding model CLIP [31] has shown promising results on text-to-image generation. Notably, they have demonstrated strong zero-shot generalization while evaluated on tasks the model has not been specifically trained on. Shape generation is a more fundamental problem as images are projections of the inherently 3D physical world. Therefore, one may wonder if the success in 2D can be transferred to the 3D setting. This turns out to be a non-trivial problem. Unlike the text-to-image case, where paired data is abundant, it is impractical to acquire huge paired datasets of texts and shapes.

Leveraging the progress of text-to-image generation, we present CLIP-Forge, a two-stage training method for zero-shot text-to-shape generation. As shown in Fig. 2, we obtain a latent space for shapes via training an autoencoding occupancy network [23], and train a normalizing flow model [9] to connect the CLIP encoding of 2D renderings of 3D shapes and the latent shape encoding. CLIP-Forge requires no text label for shapes, and avoids the expensive optimizations as employed in existing approaches [13, 36].

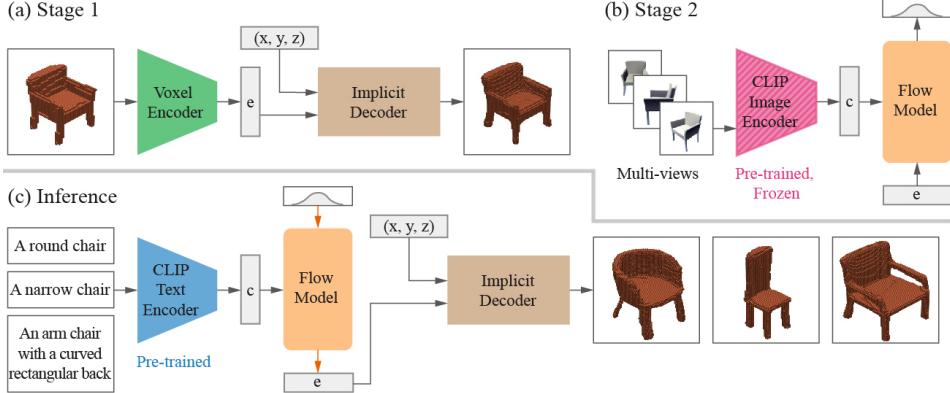


Figure 2: An overview of our method: top shows the two training stages and bottom shows inference.

The main contributions of this paper are as follows:

- We present a method, CLIP-Forge that generates 3D shapes directly from text, with an efficient generation process requiring no inference time optimization, and can generate multiple shapes for a given text.
- Our method requires no text labels as training data, offering an opportunity to leverage wider shape datasets.
- We show results for zero-shot generalization, such as interpreting descriptive text (e.g. “a round sofa”), and interpreting unseen vocabulary (e.g. “a space shuttle”), shown in Fig. 1.

2 Related Work

CLIP and Zero-Shot Transfer: A major building block of this work is CLIP [31], which shows groundbreaking zero-shot capability by a mechanism to connect text and image by bringing them closer in the latent space. Works such as ALIGN [18], have used a similar framework on noisy datasets. Recently, pre-trained CLIP has been used for several zero-shot downstream applications [18, 12, 34, 30, 13, 27]. Research most similar to ours are zero-shot image and drawing synthesis [30, 13], which involve iteratively optimizing a random image to increase certain CLIP activations. There is still no clear way to apply previous approaches to 3D. Our approach conditions a shape prior network with CLIP features, which has the advantages of significant speed-up and the capability to generate multiple shapes from a single text.

3D Shape Generation and Language: Recently, there has been tremendous progress in generating different 3D data such as point clouds [2, 20, 38], voxels [37], implicit representations [23, 29, 7] and meshes [25]. We adopt implicit representations in this work due to their superior quality. More recently, works such as [5, 1] have used text to localize objects in 3D scene. Unlike [6], which also generates shapes using text descriptions, our approach is zero-shot and requires no labels.

Multi-Stage Training: In this work we follow a multi-stage training paradigm, where we first learn the embeddings of data and then learn a probabilistic model of the encoding. Such a framework has been explored in image generation [26, 22, 11] and 3D shape generation [2, 7]. In this work, we first train a 3D shape autoencoder and then model the embeddings using normalizing flow [33, 9, 28].

3 Method

Our method requires a collection of 3D meshes without any associated text labels, which takes the format of $\mathcal{S} = \{(\mathbf{I}_n, \mathbf{V}_n, \mathbf{P}_n, \mathbf{O}_n)\}_{n=1}^N$ comprising of rendered image \mathbf{I}_n , voxel grid \mathbf{V}_n , query points \mathbf{P}_n and occupancies \mathbf{O}_n . The training phase comprises of two stages.

1. We train an autoencoder with a voxel encoder and an implicit decoder. Once the training of the autoencoder is complete we obtain shape embedding e_n for each 3D shape in \mathcal{S} .

2. We train a conditioned normalizing flow network to model and generate \mathbf{e}_n . It is conditioned with image features obtained from the CLIP image encoder using \mathbf{I}_n .

During inference, we first convert the text to the same latent space as image features using the CLIP text encoder. We then condition the normalizing flow network with the given text features and a random vector sampled from the uniform Gaussian distribution to obtain a shape embedding. Finally, this shape embedding is converted to a 3D shape using the implicit decoder.

Training Stage 1: We train an autoencoder to extract shape embeddings $\{\mathbf{e}_n\}_{n=1}^N$ for the training dataset. The input is a batch of \mathbf{V}_n of resolution 32^3 . We feed this input to the voxel encoder to obtain a batch of \mathbf{e}_n . The voxel encoder comprises of batch-normalized 3D convolution layers followed by linear layers. $\{\mathbf{e}_n\}_{n=1}^N$ are augmented with Gaussian noise, which empirically improve the generation quality. We use an implicit decoder inspired by the Occupancy Networks [23], which takes concatenated \mathbf{e}_n and \mathbf{P}_n as input. The implicit decoder consists of linear layers with residual connections and predicts \mathbf{O}_n .

Training Stage 2: We train a normalizing flow network using $\{\mathbf{e}_n\}_{n=1}^N$ and their corresponding rendered images $\{\mathbf{I}_n\}_{n=1}^N$. Note that each \mathbf{I}_n can include multiple images of the same shape from different viewpoints. We model the conditional distribution of \mathbf{e}_n using RealNVP [9] with 5 layers, which transforms the distribution of \mathbf{e}_n into a normal distribution. We obtain the condition vectors $\{\mathbf{c}_n\}_{n=1}^N$ by passing $\{\mathbf{I}_n\}_{n=1}^N$ through the ViT [10] based CLIP image encoder, whose weights are frozen. \mathbf{c}_n is concatenated with the transformed feature vector at each scale and translation coupling layers of RealNVP.

Inference: First, we convert the text into the text embedding using the CLIP text encoder. As the CLIP image and text encoders are trained to bring the image and text embeddings in the same latent space, we can simply use the text embedding as the condition vector to the normalizing flow model in stage 2. Moreover, the normal distribution allows us to sample multiple times to obtain multiple shape embeddings of a single condition vector. We obtain the mean shape embedding by using the mean of the normal distribution. These shape embeddings are then converted to 3D shapes using the implicit decoder in stage 1.

4 Experiments

We use the ShapeNet [4] dataset which consists of 13 rigid object classes. We use the processed version of the data which consists of rendered images, voxel grids, query points and their occupancies from shapes as provided in [8, 23]. For both training stages, we use the ADAM optimizer [19] with a learning rate of 1e-4 and a batch size of 32.

Qualitative Evaluation: We qualitatively evaluate generative capabilities of our method. Results are shown in Fig. 3. First, we show that our network can generate multiple shapes using a single text query. Next, we show that our network can generate shapes based on sub-category, common semantic words, and common shape attributes. We also show failure cases of our method where it has limited generalization abilities (e.g. “a cat” that is distinct from any categories in the dataset). Finally, we show generations by interpolation between two text inputs in Fig. 4.

Perceptual Evaluation: To evaluate CLIP-Forge’s ability to make use of attributes and sub-category information in text, we conducted a perceptual evaluation using Amazon SageMaker Ground Truth and crowd workers from Mechanical Turk [24]. The crowd workers were presented with a pair of images as shown in Figure 5. One image was generated from text containing just the generic category name (e.g. “a chair”), while the other was generated using text containing an attribute (e.g. “a round chair”) or a sub-category (e.g. “a bar stool”). The crowd workers were asked to identify which image best matched the text containing the attribute or sub-category. We tested 97 pairs of images, which were each rated by 27 crowd workers. For 67% of the image pairs, the crowd workers identified the images generated by the text containing the attribute or sub-category as best matching the description.

Effect of using Multiple Views: We evaluate if using more views helps the generation quality using the FID [15] metric. We first take 110 text queries and generate a mean shape embedding for each text query. We then generate 32^3 resolution 3D objects for all the text queries. We compare the generated 3D shapes with the test dataset of ShapeNet. FID depends on a pre-trained network, for which we train a voxel classifier on the 13 ShapeNet classes and use the feature vector from the fourth layer. We report the results in Fig. 5. It can be seen that using more views help improve the generation quality.

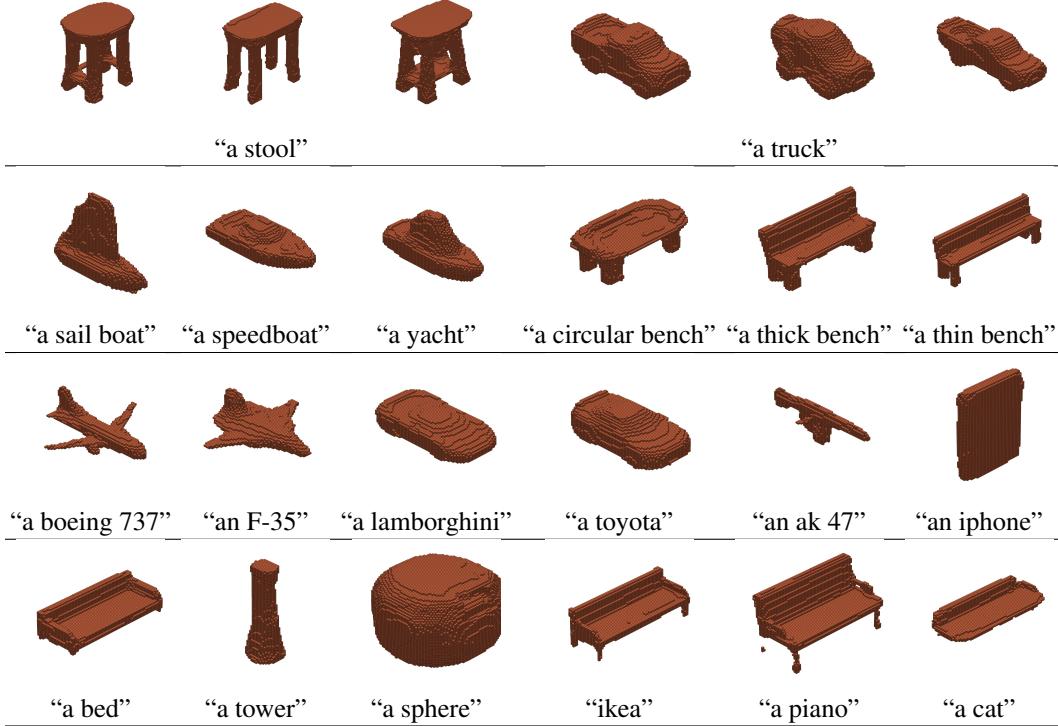


Figure 3: CLIP-Forge results. Rows from top to bottom: multiple generations with a single text, more qualitative results, common name text query results, and generalization to unseen categories.



Figure 4: CLIP-Forge generations by interpolating from “a round table” to “a square table”.

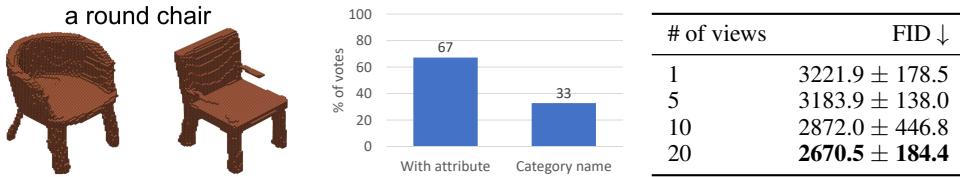


Figure 5: Left: an example question for crowd workers. Middle: perceptual evaluation result. Right: effect of using multiple views.

5 Conclusions and Future Work

We proposed CLIP-Forge, a new method to generate 3D shape from text. The method requires no labels and has potential for leveraging larger shape datasets. Comparing to previous work, our method does not require expensive optimization at inference time, making it suitable for more interactive tools for design and manufacturing. In future work, we aim to extend the training scale, conduct more comprehensive quantitative and qualitative evaluations, and further explore the effect of data augmentation techniques.

References

- [1] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European Conference on Computer Vision*, pages 422–440. Springer, 2020.
- [2] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018.
- [3] P. Achlioptas, J. Fan, R. X. D. Hawkins, N. D. Goodman, and L. J. Guibas. Learning to refer to 3d objects with natural language. 2018.
- [4] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. Shapenet: An information-rich 3d model repository, 2015. cite arxiv:1512.03012.
- [5] D. Z. Chen, A. X. Chang, and M. Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 202–221. Springer, 2020.
- [6] K. Chen, C. B. Choy, M. Savva, A. X. Chang, T. Funkhouser, and S. Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Asian Conference on Computer Vision*, pages 100–116. Springer, 2018.
- [7] Z. Chen and H. Zhang. Learning implicit fields for generative shape modeling. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [8] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. 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.
- [9] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.
- [10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [11] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021.
- [12] H. Fang, P. Xiong, L. Xu, and Y. Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021.
- [13] K. Frans, L. B. Soros, and O. Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders, 2021.
- [14] Z. Han, M. Shang, X. Wang, Y.-S. Liu, and M. Zwicker. Y2seq2seq: Cross-modal representation learning for 3d shape and text by joint reconstruction and prediction of view and word sequences. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):126–133, Jul. 2019.
- [15] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [16] P.-H. Huang, H.-H. Lee, H.-T. Chen, and T.-L. Liu. Text-guided graph neural networks for referring 3d instance segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1610–1618, May 2021.
- [17] F. Huq, A. Iqbal, and N. Ahmed. Holistic static and animated 3d scene generation from diverse text descriptions. *CoRR*, abs/2010.01549, 2020.
- [18] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.
- [19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

- [20] C.-L. Li, M. Zaheer, Y. Zhang, B. Poczos, and R. Salakhutdinov. Point cloud gan. *arXiv preprint arXiv:1810.05795*, 2018.
- [21] A. S. Lin, L. Wu, R. Corona, K. W. H. Tai, Q. Huang, and R. J. Mooney. Generating animated videos of human activities from natural language descriptions. 2018.
- [22] J. Liu, Y. Yao, and J. Ren. An acceleration framework for high resolution image synthesis. *arXiv preprint arXiv:1909.03611*, 2019.
- [23] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [24] A. Mishra. Machine learning in the aws cloud: Add intelligence to applications with amazon sagemaker and amazon rekognition, 2019.
- [25] C. Nash, Y. Ganin, S. M. A. Eslami, and P. W. Battaglia. Polygen: An autoregressive generative model of 3d meshes. *ICML*, 2020.
- [26] A. v. d. Oord, O. Vinyals, and K. Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017.
- [27] D. Pachomov, S. Hira, N. Wagle, K. E. Green, and N. Navab. Segmentation in style: Unsupervised semantic image segmentation with stylegan and clip. *arXiv preprint arXiv:2107.12518*, 2021.
- [28] G. Papamakarios, T. Pavlakou, and I. Murray. Masked autoregressive flow for density estimation. *arXiv preprint arXiv:1705.07057*, 2017.
- [29] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [30] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021.
- [31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.
- [32] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation, 2021.
- [33] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.
- [34] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, and K. Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.
- [35] C. Tang, X. Yang, B. Wu, Z. Han, and Y. Chang. Part2word: Learning joint embedding of point clouds and text by matching parts to words, 2021.
- [36] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 2018.
- [37] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS’16, page 82–90, Red Hook, NY, USA, 2016. Curran Associates Inc.
- [38] G. Yang, X. Huang, Z. Hao, M.-Y. Liu, S. Belongie, and B. Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4541–4550, 2019.