# 3D Object Styled Generation and Multi-Modal Image Retrieval

**Muhammad Yusuf Hassan***
IIT Gandhinagar, India
https://md-hassan.github.io/
md.hassan@iitgn.ac.in

**Shivasankaran V P***
IIT Gandhinagar, India
https://shiva-sankaran.github.io/
vp.shivasan@iitgn.ac.in

**Prajwal Singh**
IIT Gandhinagar, India
https://scholar.google.com/citations?user=w5GbFHIAAAAJ&hl=en
singh_prajwal@iitgn.ac.in

**Shanmuganathan Raman**
IIT Gandhinagar, India
https://people.iitgn.ac.in/~shanmuga/index.html
shanmuga@iitgn.ac.in

## Abstract

In this work, we explore the ramifications of the recent advancements in visual models in conjunction with natural language supervision. Specifically, we discuss two downstream tasks, namely 3D Object Styled Generation and Multi-Modal Image Retrieval. For 3D object styled generation, we provide proof of concept for creating styled 3D objects from textual descriptions. For multi-modal image retrieval, we prove the hypothesis on which the current SOTA model works and release our replicated code, which matches the SOTA performance.

**2012 ACM Subject Classification** Computing methodologies → Computer vision; Computing methodologies → Computer graphics

**Keywords and phrases** Computer Vision, Multimodality, Image Retrieval, Generation, Style Transfer

...

## 1 Introduction

Intelligent systems capable of processing Vision-Language data by taking the best of both worlds open myriad new applications of Artificial Intelligence like object generation/retrieval from textual descriptions. The popularity DALL-E [12] gained commercially in the past year, and openAI[1] planning DALL-E's commercial launch[2] attests to the utility of vision tools conditioned on language. Additionally, with more and more visual-language datasets [1] [15] [5] [16] being made public, there is an influx of research works exploiting these datasets to build high utility tools.

Along with DALL-E, OpenAI released CLIP [11], a deep learning model which can learn visual concepts from natural language supervision. The authors of CLIP showcased the transferable power of the learned image features by beating previous SOTA zero-shot image classification models by a considerable margin [11]. CLIP's ability to learn a shared embedding space for vision and language data can mold various previous vision models into vision-language models. This work explores two

*Equal contribution

[1] https://openai.com/

[2] https://openai.com/blog/dall-e-now-available-in-beta/

vision-language tasks: "3D-object styled generation" and "image retrieval through sketch and textual descriptions." The primary contributions of this work are as follows

- We provide proof of concept for 3D-object-styled generation from textual description.
- We release our implementation for SOTA multi-modal image retrieval. The official code has yet to be released, and there exists no other public implementation).

The rest of the report is divided as follows: section 2 discusses the previous works on both the problems, section 3 revisits CLIP, section 4and 5 delineate the explored methodologies, section 6 describe the experiment settings, section 7 summarizes our findings, section 8 comprises of the proposed future work and finally section 9 concludes this report.

# 2 Related Work

## 2.1 3D Object Styled Generation

After the success of 2D image generative models, the focus has shifted to the 3D domain. There have been attempts at 3D object generation in various formats like point cloud [20], voxel [19], mesh [8], etc. The novelty of CLIP-Forge is its zero-shot learning capability, which it derives by leveraging the CLIP model. The zero-shot learning paradigm was popularized by the works [9], and [6] on image classification.

## 2.2 Multi-Modal Image Retrieval

Several works have explored sketch-based image retrieval (SBIR), such as [10], [21], while [2], [3] have worked on it in a zero-shot context. The cross-modal task of text-based image retrieval (TBIR) has also seen recent success using cross-attention between images and text as in [22]. Multimodal sketch retrieval has also been researched extensively, as in [18] and [17]. The work in [4] shows that simply adding the representations of the two modalities is quite effective for retrieval.

# 3 CLIP

CLIP (Contrastive Language–Image Pre-training) [11] is a zero-shot multimodal learning model proposed by OpenAI through a simple pre-training task. The model aims to find a joint embedding space for both images and texts by contrastively forcing the image encoder and text encoder to find embedding vectors corresponding to the relation between the image and the text.

In other words, CLIP tries to bring the embeddings of an image and text close together if they both belong to the same class and correspondingly otherwise. The result of this learning paradigm, in essence, permits the interchangeability of text and image embeddings of the same data in different modalities and creates new opportunities by leveraging this learning paradigm in downstream tasks.

# 4 3D Object Styled Generation

## 4.1 CLIP-Forge

We first explore 3D model generation using text prompts. The CLIP-Forge [13] model generates plausible 3D voxel models of common objects, given a textual input. The work seems to be proposed mostly as a proof-of-concept, owing to the small number of classes of objects that the model can generate.

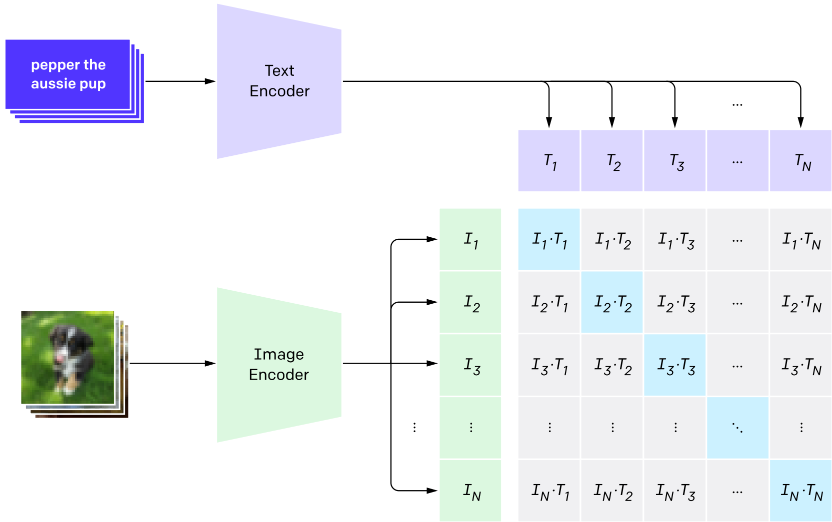


**Figure 1** Contrastrive pre-training in CLIP. Image taken from here[3]

CLIP-Forge leverages the powerful text-image understanding of CLIP to encode embeddings that contain information from both text and multi-view 2D projections. An implicit decoder is trained to generate 3D voxel models from these embeddings.

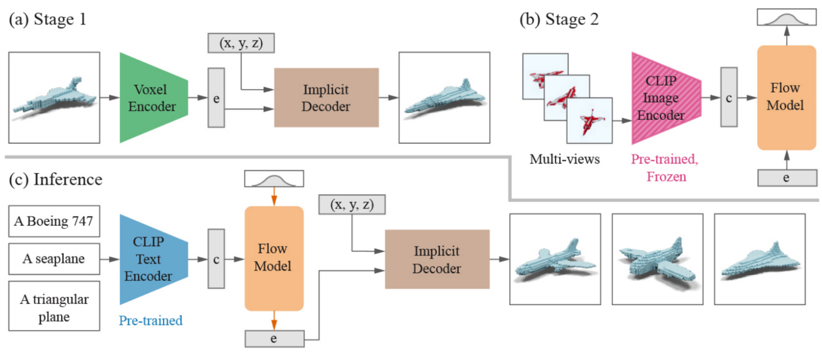


**Figure 2** The CLIP-Forge pipeline. Image taken from the paper [13]

## 4.2 Text2Mesh

The Text2Mesh [7] model stylizes a given 3D mesh by predicting color and local textures based on a given text prompt. The model iteratively tries to minimize a CLIP-based semantic loss between the text and intermediate renderings to generate a plausible output.

# 5 Multi-Modal Image Retrieval

In this section, we discuss the model "TASK-former" proposed P. Sangkloy et al. [14] for image retrieval using text description and sketch as input. The authors of TASK-former argue that "both the modalities complement each other in a manner that cannot be achieved easily by either alone."

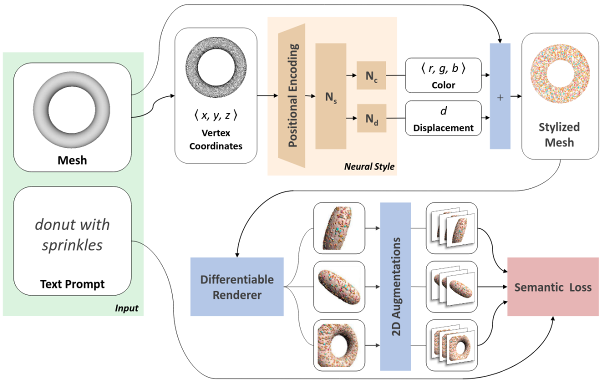


Figure 3 The Text2Mesh pipeline. Image taken from the paper [7]

To empirically prove their argument, they propose TASK-former as an extension of CLIP's learning paradigm for multi-modal image retrieval.

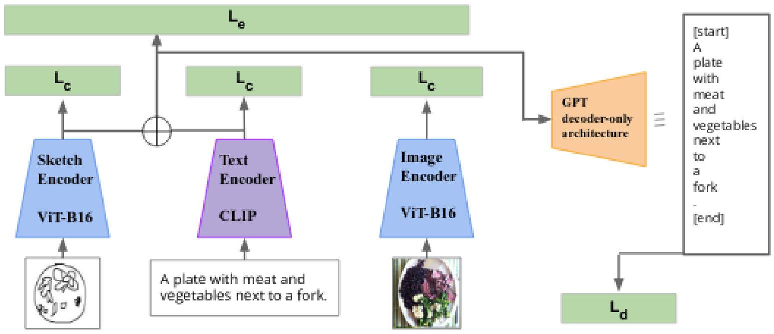


Figure 4 TASK-former. Image taken from here the paper [14]

TASK-former accepts an optional sketch as input in addition to the text query. The sketch will supplement the text query by carrying information that is difficult to express as text, like relative positioning and sizes.

## 5.1 Training

TASK-former extensively uses CLIP's image and text encoders to find both modalities' embeddings. They propose loss terms in addition to CLIP's symmetric cross-entropy to use both text and sketch. The final loss is a weighted sum of embedding loss (CLIP), classification loss, and caption generation loss with a weights ratio of 100,10,1, respectively.

### 5.1.1 Embedding Loss($L_e$)

They use CLIP's contrastive learning loss to find a shared embedding space. Contrastive loss is applied between image embeddings and sketch + text embeddings. They experiment with three combinations between sketch and text embeddings: addition, element-wise max, and concatenation. Experimentally addition operation gave the best results.

### 5.1.2 Classification Loss($L_c$)

They employ classification loss to retain object-related features for all three embeddings.

### 5.1.3 Caption Generation Loss($L_d$)

Caption generation loss ensures the combined embedding has enough information to reconstruct the original text caption.

# 6 Experiments

## 6.1 3D Object Styled Generation

We note that CLIP-Forge does not produce models with colors and texture, nor does Text2Mesh generate models using text. In order to produce a complete pipeline that produces 3D models using only text, we sought to combine the above two methods.

Our proposed experimental pipeline is illustrated in Figure 5. It takes a text input and generates a 3D voxel model using the CLIP-Forge model. The voxel model is converted into a mesh as required by the next part of the pipeline. Finally, the generated mesh is sent through Text2Mesh, conditioned by a textual input describing the required style and texture.

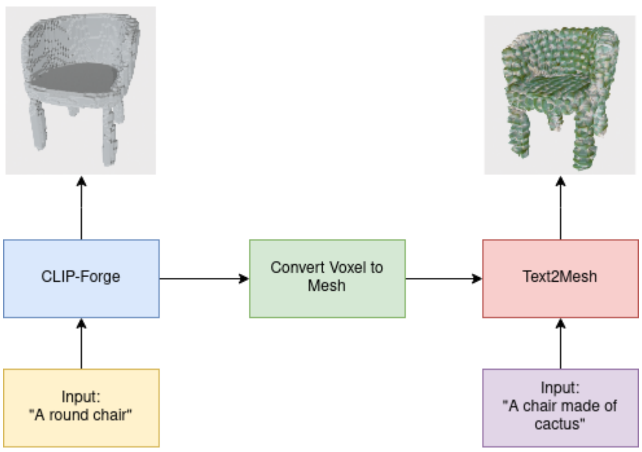


**Figure 5** An overview of our proposed generation pipeline.

## 6.2 Image Retrieval

We implement the architecture described in Fig 4 taken from the original paper. We train TASK-former with the three losses as explained in section 5.1. We use a simple 2-layer MLP network as the classification head on top of all three encoders. Our GPT model for caption generation has six decoder layers and eight attention heads with 512 hidden dimensions. We train the network for 50 epochs on 1 GeForce GTX TITAN X.

# 7 Results

## 7.1 3D Object Styled Generation

The results obtained from our proposed pipeline are shown in 6. We note that the colors generated on these models are satisfactory. However, the geometric texture is not as good. We speculate that this is partly due to the low-resolution voxel models that CLIP-Forge generates, which have to be later converted to mesh for input to Text2Mesh. Our experiments on higher-quality mesh inputs to Text2Mesh show that the model's output is highly dependent on the quality of the input mesh.

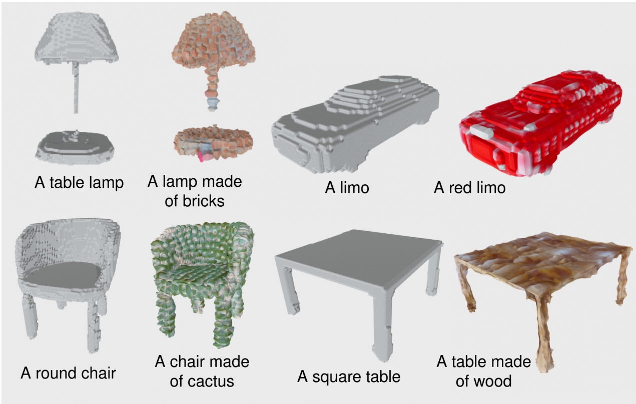


**Figure 6** Sample outputs from our proposed generation pipeline. Models on the left are models generated using CLIP-Forge; on the right are models stylized using Text2Mesh.

## 7.2 Image Retrieval

We replicate the results proposed in the original paper [14]. Our replicated code is made publicly available on `https://github.com/md-hassan/Sketch-Text-Image-Retrieval`. Moreover, we do slightly better for R@5 and R@10 compared to the metrics reported in their paper. This improvement could be attributed to the potential difference in the classification of MLP and GPT variants used in the original paper and our implementation, as they have not mentioned any architecture descriptions of the MLP or GPT used in their paper.

| Method | R@1 | R@5 | R@10 |
|---|---|---|---|
| CLIP (Zero shot) | 0.378 | 0.624 | 0.722 |
| TASK-former: Feature Max | 0.443 | 0.704 | 0.804 |
| TASK-former: Feature concat | 0.357 | 0.650 | 0.768 |
| TASK-former: Feature add | 0.609 | 0.847 | 0.917 |
| TASK-former(Ours)(Sketch) | 0.473 | 0.791 | 0.866 |
| TASK-former(Ours)(Text) | 0.577 | 0.845 | 0.891 |
| TASK-former(Ours)(Sketch + Text) | 0.603 | 0.873 | 0.941 |

**Table 1** TASK-former image retrieval metrics as given by the authors in the original paper. TASK-former is trained with $L_e + L_c + L_d$ for this table. Refer to the original paper for complete metrics. Our implementation uses feature addition

Further, we additionally experiment by retrieving using only sketch embedding and text embedding and prove the TASK-former author's hypothesis that sketch and text embeddings combinedly give better results than either independently. The results can be seen in Table 1.

## 8 Future Work

Future work on 3D model styled generation could be on generating higher resolution and more detailed voxel model outputs. Since this phase is highly influential on the style generation phase, it is vital to make improvements here. Further, we can work on inputting a single text phrase rather than using different text inputs for model generation and style generation. We could do this by encoding the text input and sending the same encoding as input to both phases.

One another possible way of leading this project with our findings is to translate the discussed image-retrieval methods for 3D-object retrieval.

## 9 Conclusion

In this work, we provide proof of concept for a 3D object styled generation pipeline. Further, we successfully replicate the work of [14] and prove their hypothesis that adding sketch and text embeddings gives better results than simply using either for image retrieval.

## 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 Sounak Dey, Pau Riba, Anjan Dutta, Josep Lladós, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2174–2183, 2019.

3 Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5084–5093, 2019.

4 Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021.

5 Huaizu Jiang, Xiaojian Ma, Weili Nie, Zhiding Yu, Yuke Zhu, and Anima Anandkumar. Bongard-hoi: Benchmarking few-shot visual reasoning for human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19056–19065, 2022.

6 Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958, 2009.

7 Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13492–13502, June 2022.

8 Charlie Nash, Yaroslav Ganin, S. M. Ali Eslami, and Peter W. Battaglia. Polygen: An autoregressive generative model of 3d meshes. *ArXiv*, abs/2002.10880, 2020.

9 Mark Palatucci, Dean A. Pomerleau, Geoffrey E. Hinton, and Tom Michael Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009.

10 Anubha Pandey, Ashish Mishra, Vinay Kumar Verma, Anurag Mittal, and Hema A. Murthy. Stacked adversarial network for zero-shot sketch based image retrieval. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2529–2538, 2020.

11 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from

natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

**12** Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

**13** Aditya Sanghi, Hang Chu, J. Lambourne, Ye Wang, Chin-Yi Cheng, and Marco Fumero. Clip-forge: Towards zero-shot text-to-shape generation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18582–18592, 2022.

**14** Patsorn Sangkloy, Wittawat Jitkrittum, Diyi Yang, and James Hays. A sketch is worth a thousand words: Image retrieval with text and sketch. In *European Conference on Computer Vision*, pages 251–267. Springer, 2022.

**15** Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2017.

**16** Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

**17** Ivona Tautkute, Tomasz Trzciński, Aleksander P. Skorupa, Łukasz Brocki, and Krzysztof Marasek. Deepstyle: Multimodal search engine for fashion and interior design. *IEEE Access*, 7:84613–84628, 2019.

**18** Nam S. Vo, Lu Jiang, Chen Sun, Kevin P. Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval - an empirical odyssey. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6432–6441, 2019.

**19** Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Joshua B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NIPS*, 2016.

**20** Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge J. Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4540–4549, 2019.

**21** Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen Change Loy. Sketch me that shoe. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 799–807, 2016.

**22** Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and S. Li. Context-aware attention network for image-text retrieval. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3533–3542, 2020.