3D Object Generation and Multi-modal Retrieval

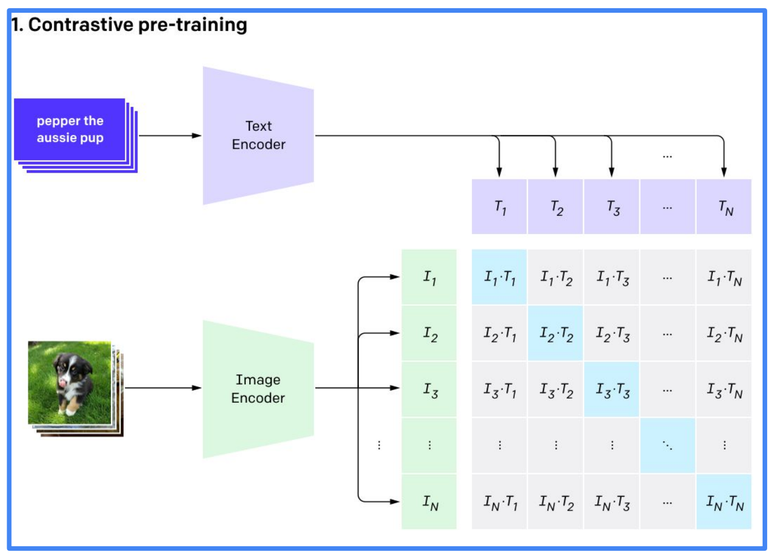
Muhammad Yusuf Hassan*; Shivasankaran V P*; Prajwal Singh, Shanmuganathan Raman

IIT Gandhinagar

md.hassan, vp.shivasan, singh_prajwal, shanmuga@iitgn.ac.in

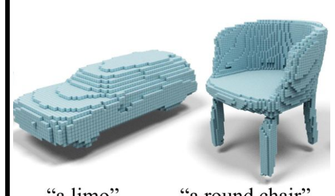
1. Motivation

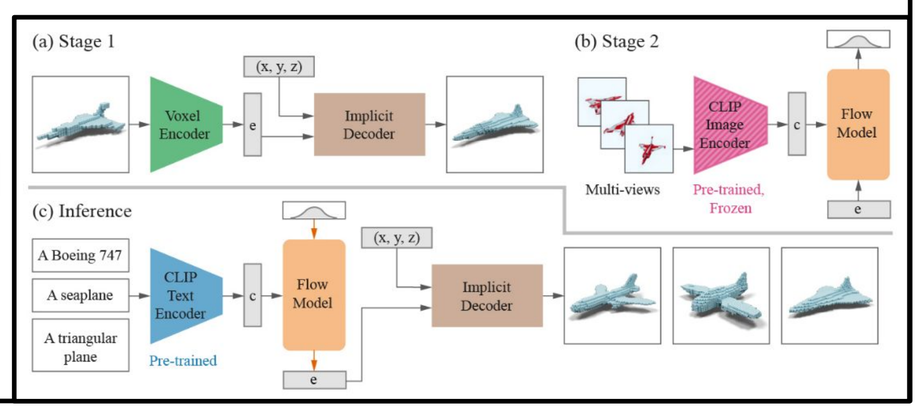
The introduction of the CLIP (Radford et al.) model has effectively found a common information subspace for both visual and textual data, which opens multiple doors to cross-modal tasks. In this work, we explore two of these tasks 3D object generation and multi-modal object retrieval.



2. CLIP-Forge

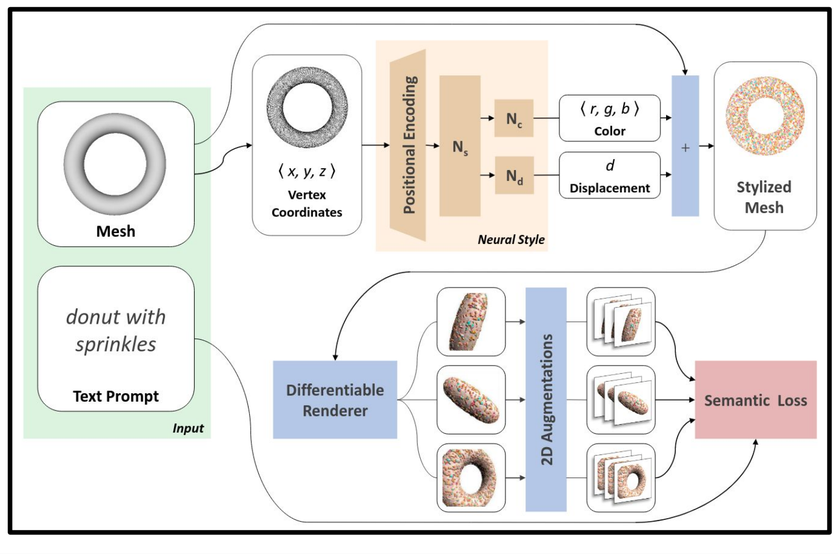
CLIP-Forge (Sanghi et al.) takes a text prompt as input and generates a plausible 3D voxel model. It leverages the powerful text-image understanding of CLIP and adapts it to the 3D domain using an implicit decoder for generation.

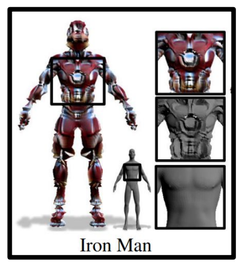




3. Text2Mesh

Text2Mesh (Michel et al.) modifies an input 3D mesh to approximate the color and texture as described by an input text prompt. It tries to minimize a CLIP-based semantic loss between the text and intermediate renderings to generate a plausible output.

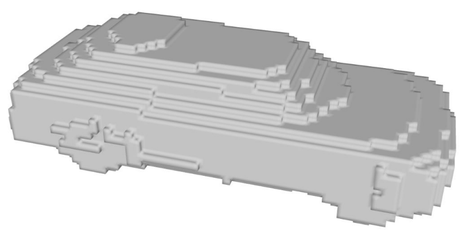




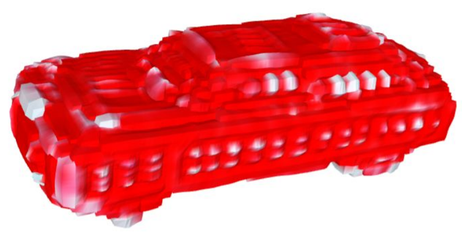
4. Combining CLIP-Forge and Text2Mesh

We ran experiments on a combined pipeline of the CLIP-Forge and Text2Mesh models in an attempt to generate high-quality textured 3D models from only text inputs.

CLIP-Forge's output
Input text: "a limo"

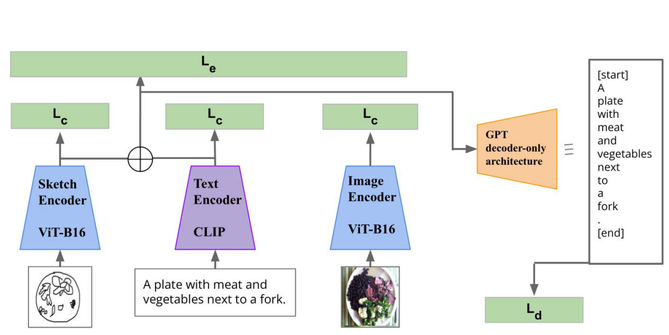


Text2Mesh's output.
Input text: "a red limo"



4. TaskFormer

TaskFormer (P. Sangkloy et al.) works on the hypothesis that both text and sketch modalities complement each other for the image retrieval task. It finds a common subspace for all three modalities of data (image, sketch, text).

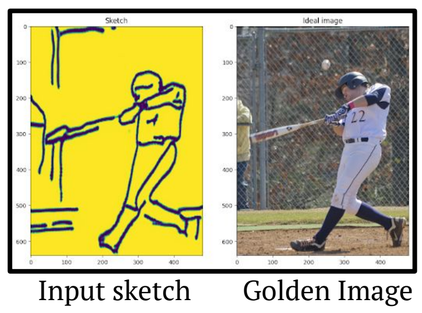


5. Implementation and Evaluation

We replicate the results obtained by the authors. However, we also note that using pure textual descriptions performs better than sketch+text, which casts doubt on the method of combining sketch and textual embeddings.

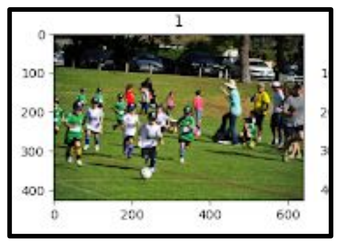
Modalities	R@1	R@5	R@10
Sketch	23.30	56.60	68.12
Text	44.96	65.98	73.46
Sketch + Text	26.47	56.02	67.47

Input Text:
"Image of a person"

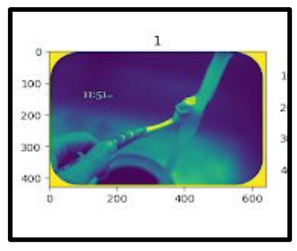


Input sketch Golden Image

Retrieval Text:



Retrieval Shape:



Retrieval Combined: