THE UNIVERSITY OF CHICAGO

TEL AVIV UNIVERSITY

Text2Mesh

Text-Driven Stylization for Meshes


Oscar Michel*


Richard Liu*

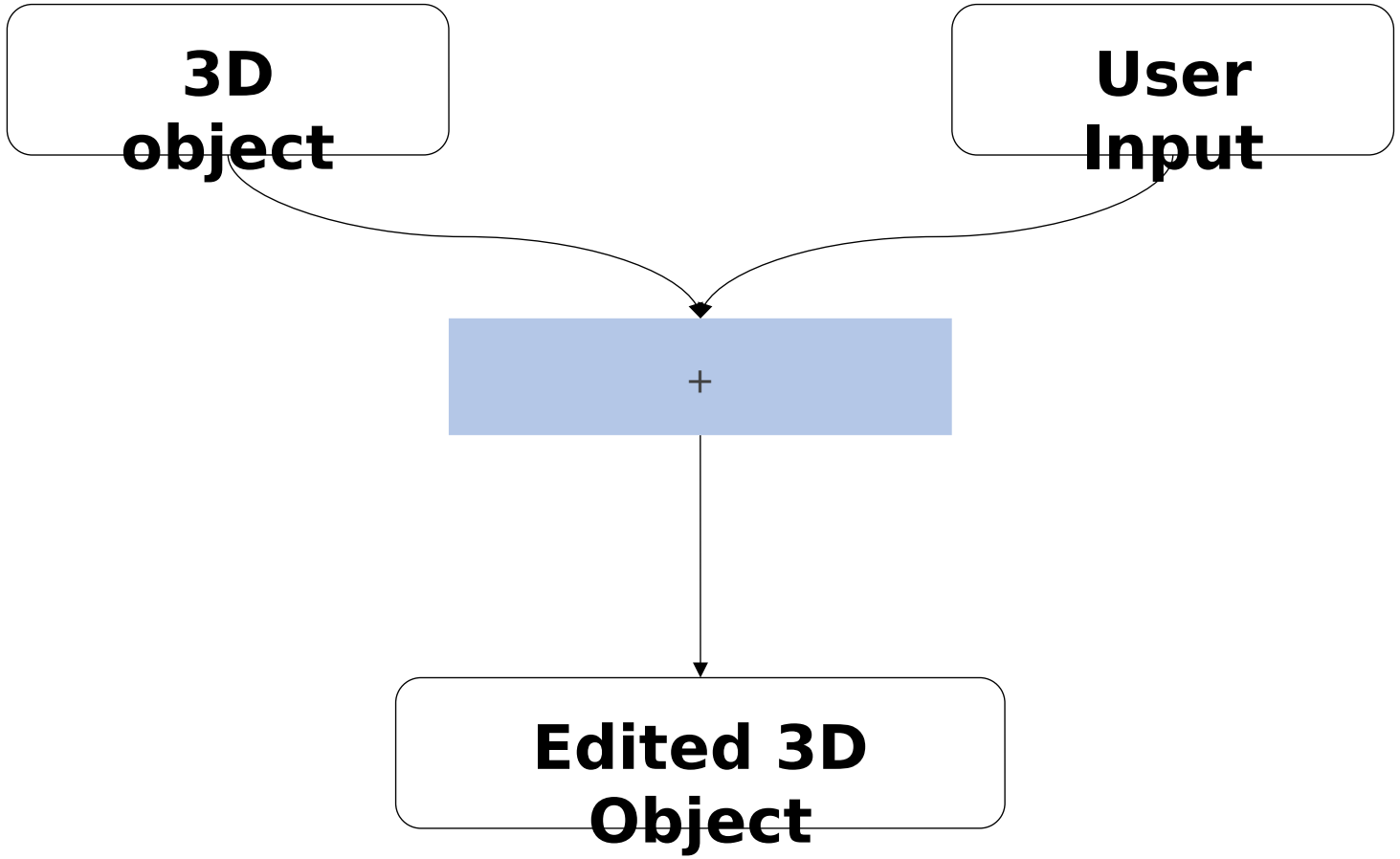

Roi Bar-On*


Sagie Benaim

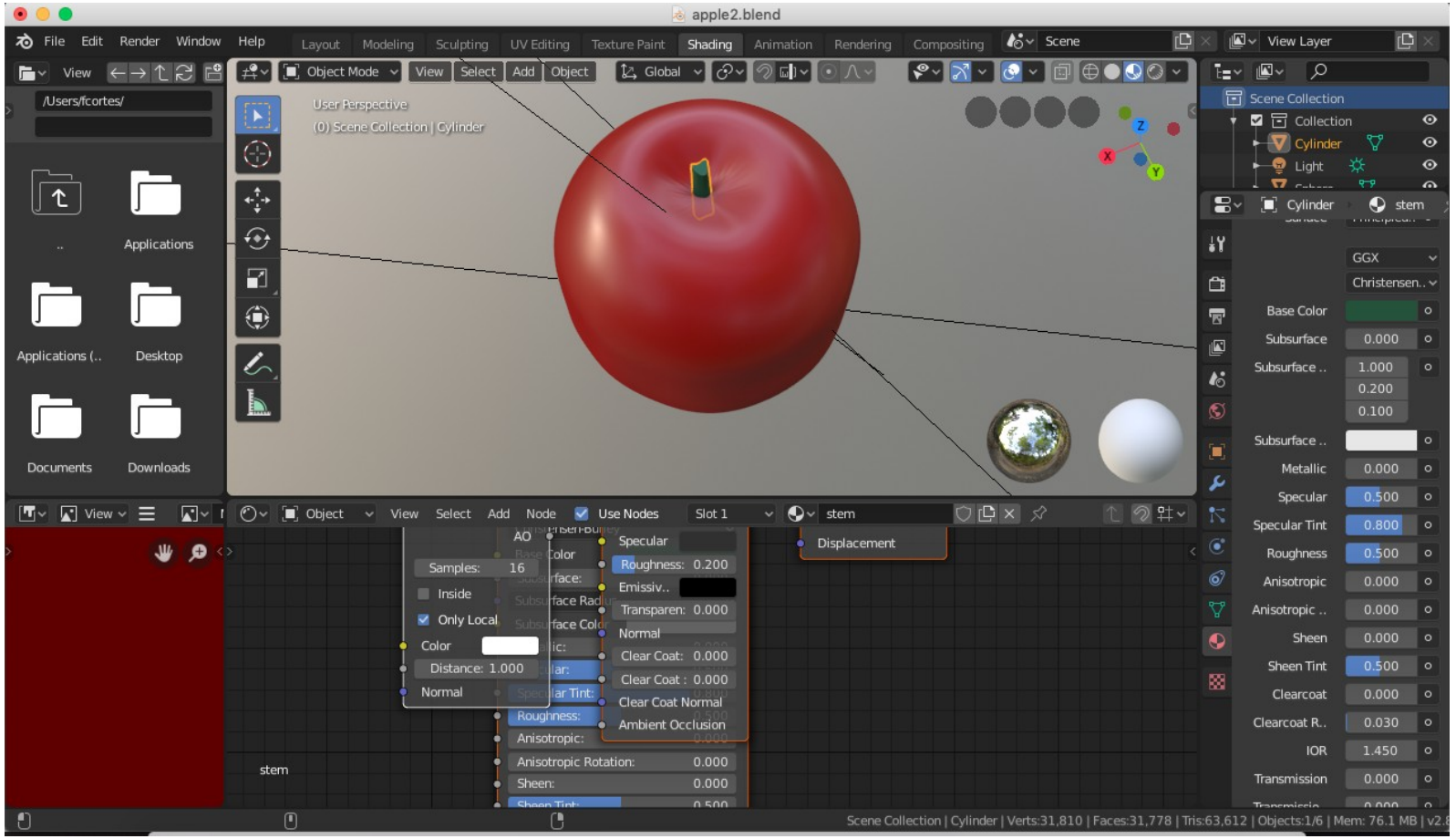

Rana Hanocka

* Equal Contribution

3D Object Editing



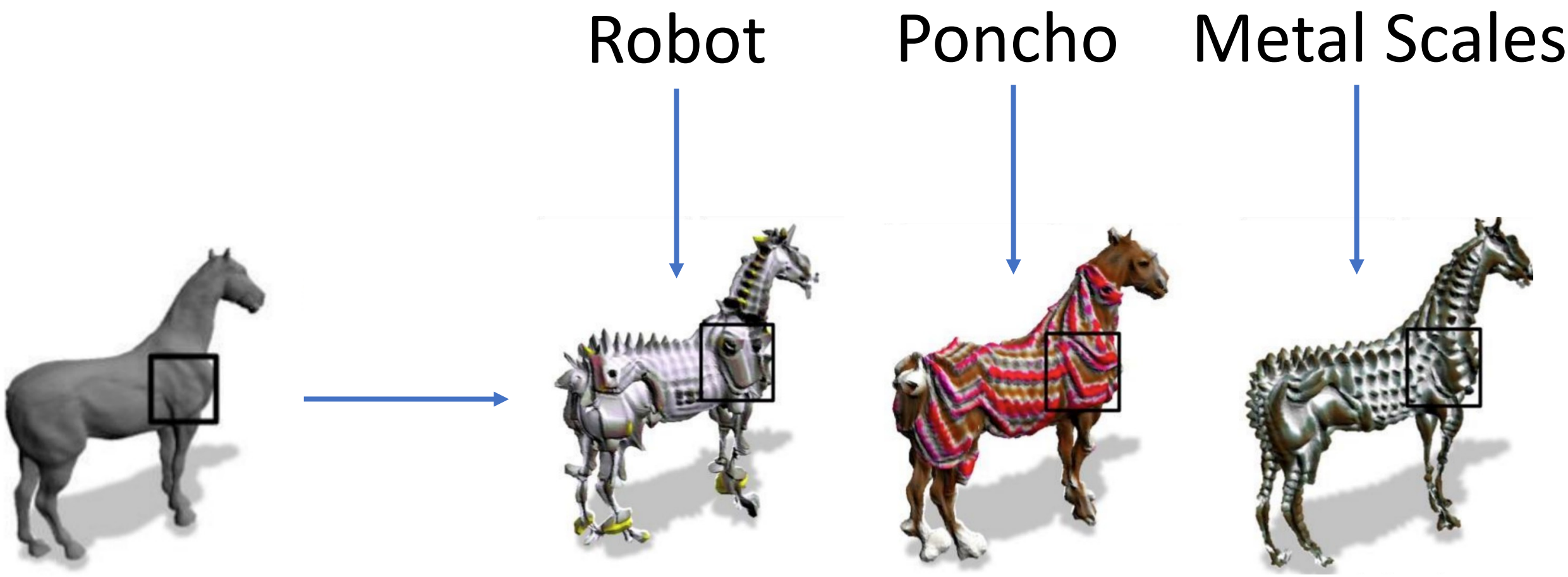
3D Object Editing: Unintuitive and Complicated



3D Object Editing: Lack of Data

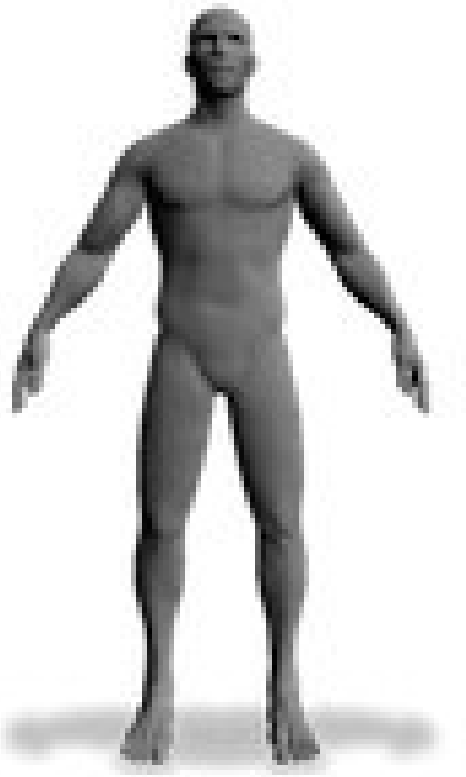
ShapeNet (3D)	ImageNet (2D)	Dataset
220k	14M	Objects
3k	20k	Categories

Goal: Zero-Shot Intuitive Editing of 3D Objects



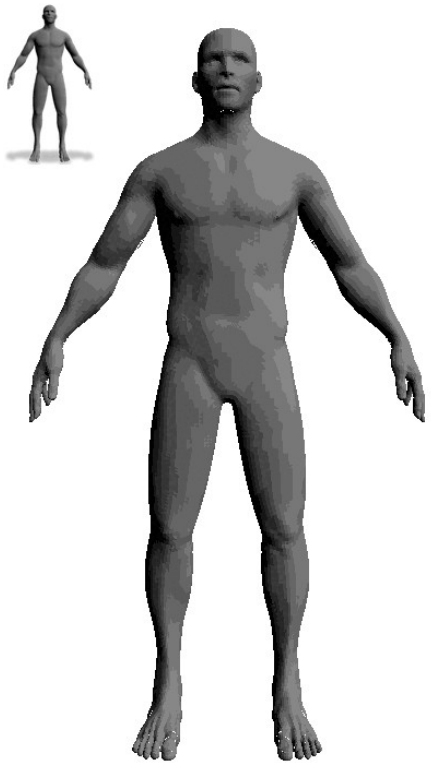
Part Aware Global Semantics

Iron
Man



Variety of Textures and Materials

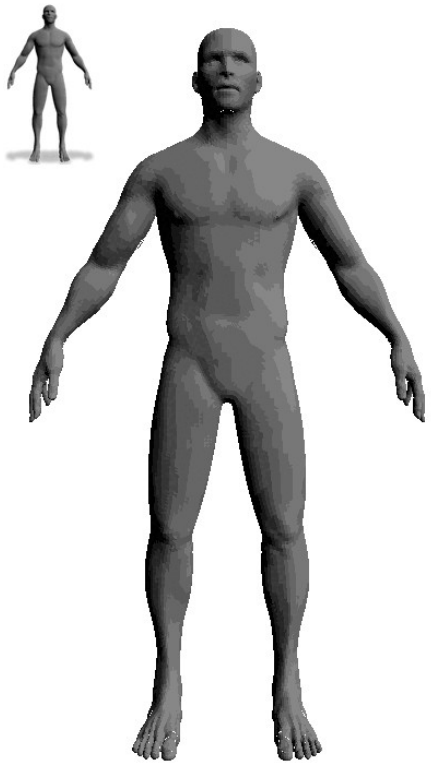
Iron Man



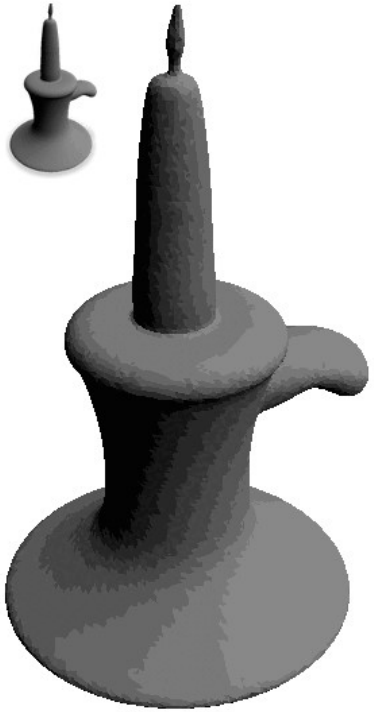
Colorful Crochet Candle

Structured Textures with Lighting

Iron Man



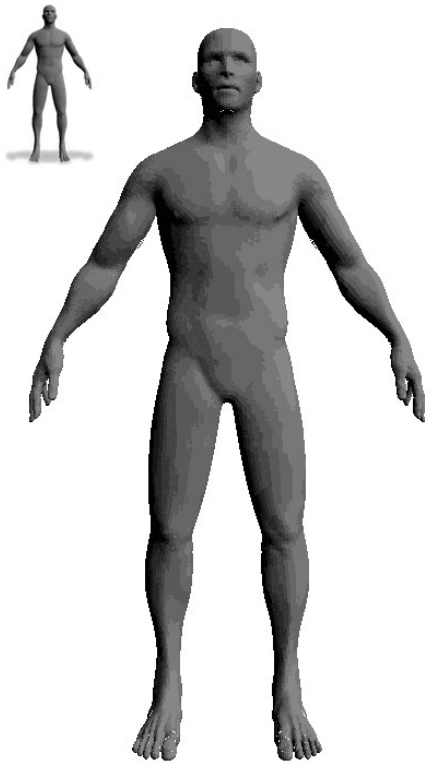
Colorful Crochet Candle



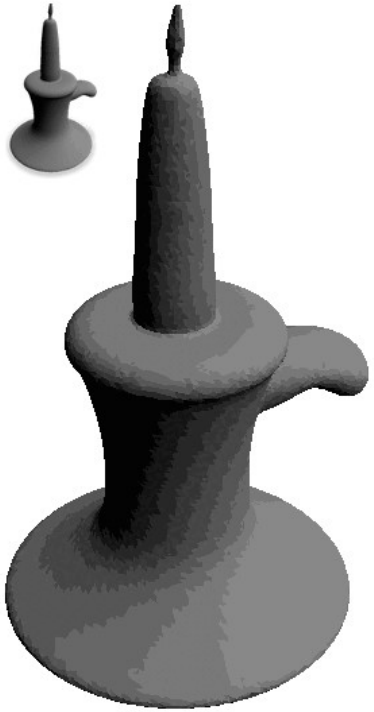
Brick Lamp

Out of Domain Generations

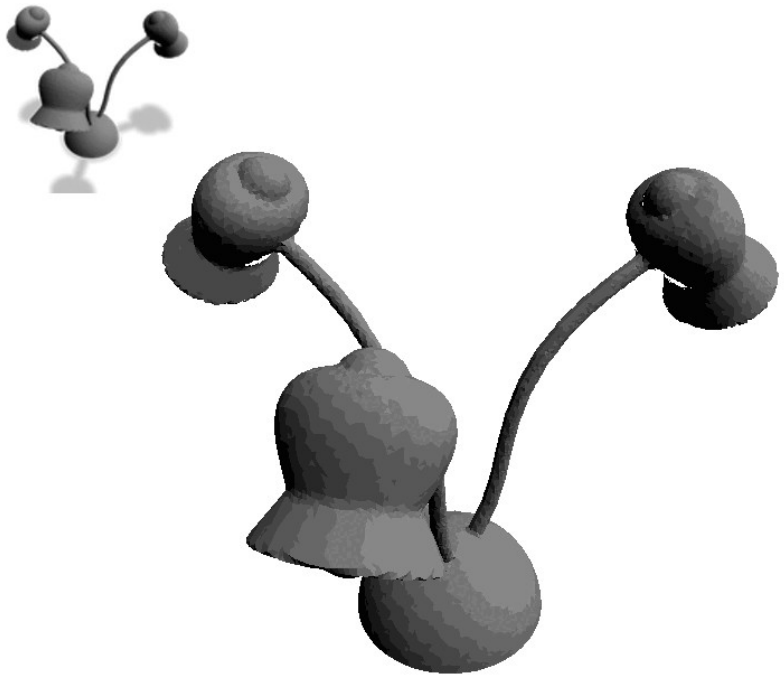
Iron Man



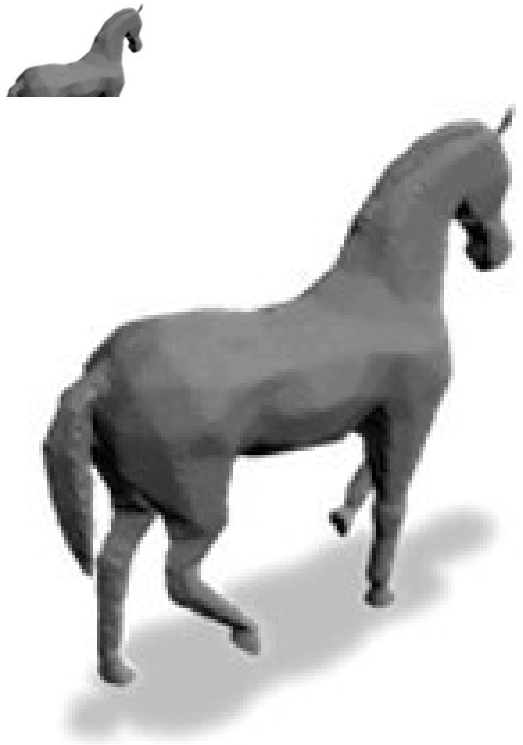
Colorful Crochet Candle



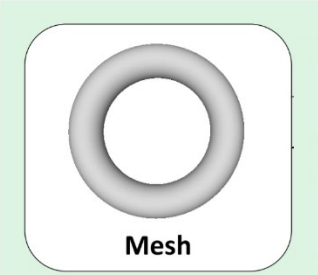
Brick Lamp



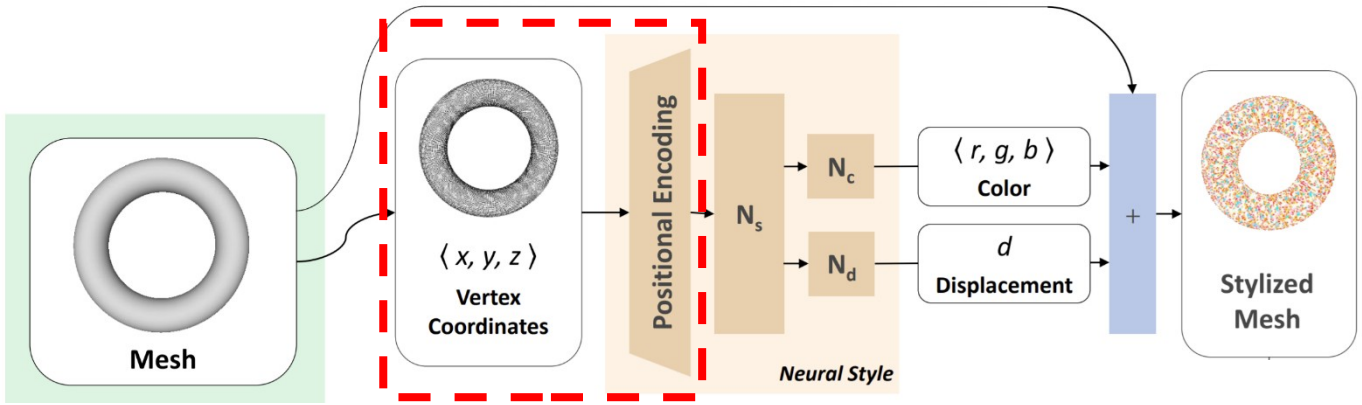
Astronaut Horse



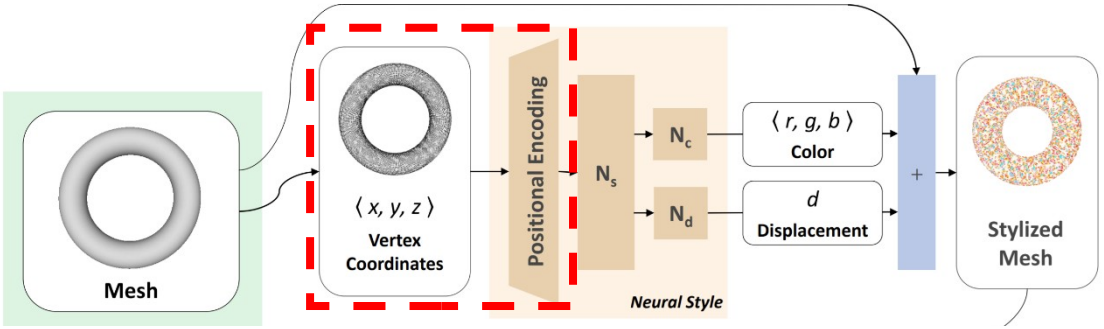
Input



Neural Style Field



Positional Encoding



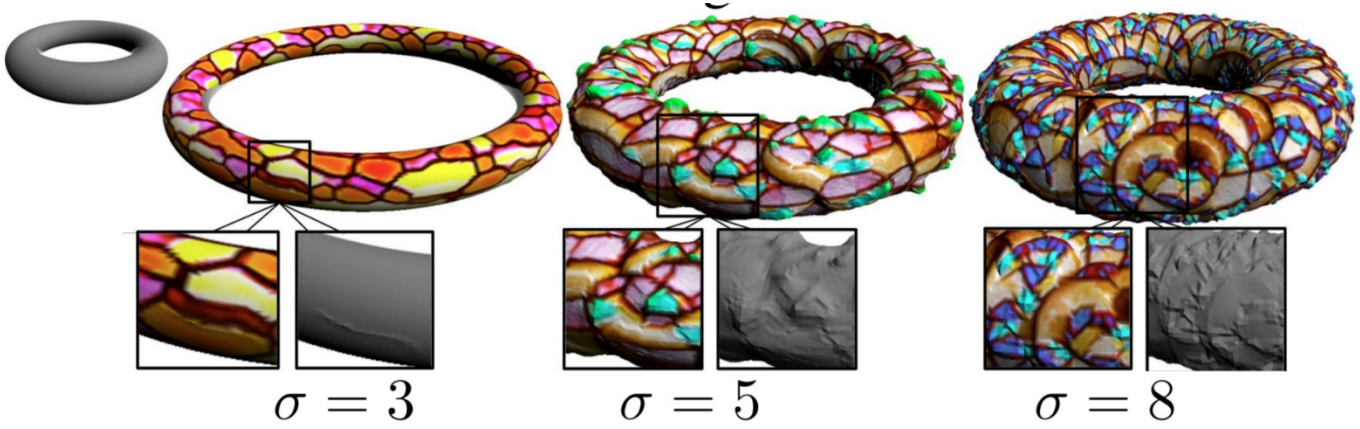
- Frequency based encoding:

$$\gamma(p) = \left[\cos(2\pi \mathbf{B} p), \sin(2\pi \mathbf{B} p)\right]^{\mathrm{T}}$$

randomly drawn from

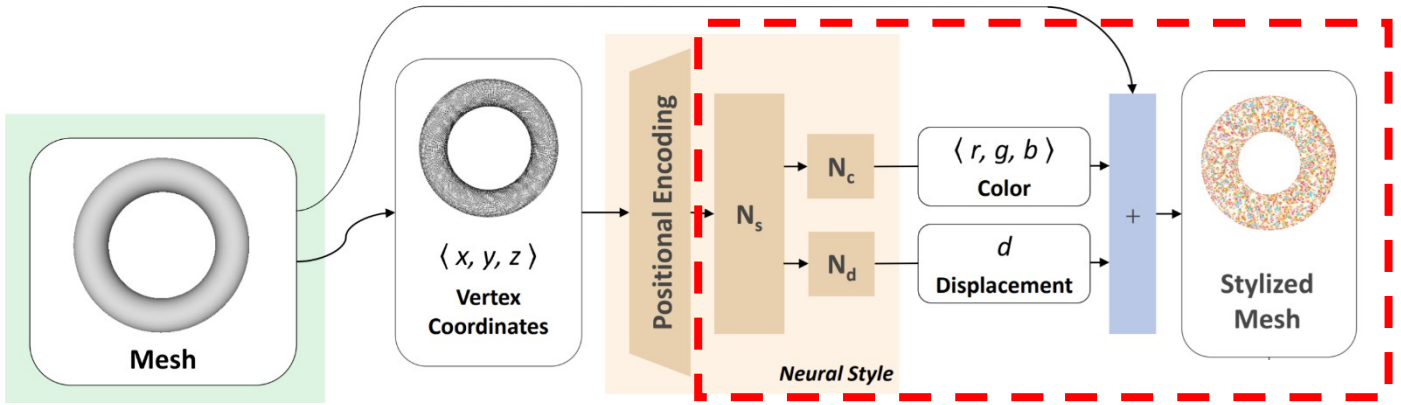
- is a hyperparameter which controls the output frequency:

'Stained glass donught'

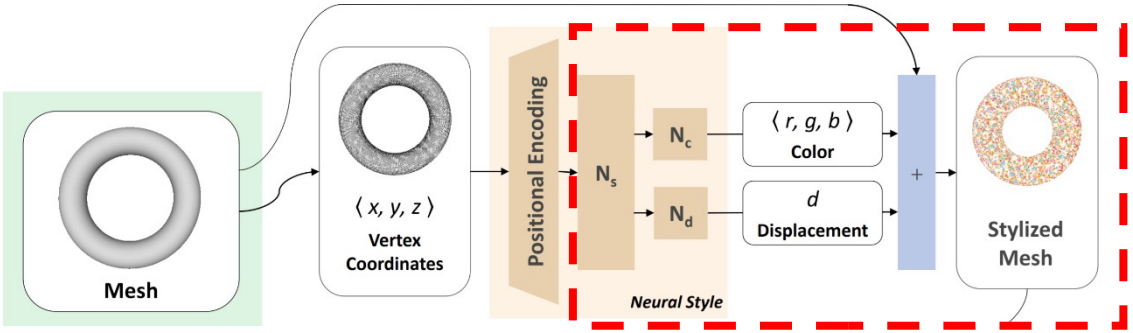


M. Tancik et al., “Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains”. NeurIPS 2020.

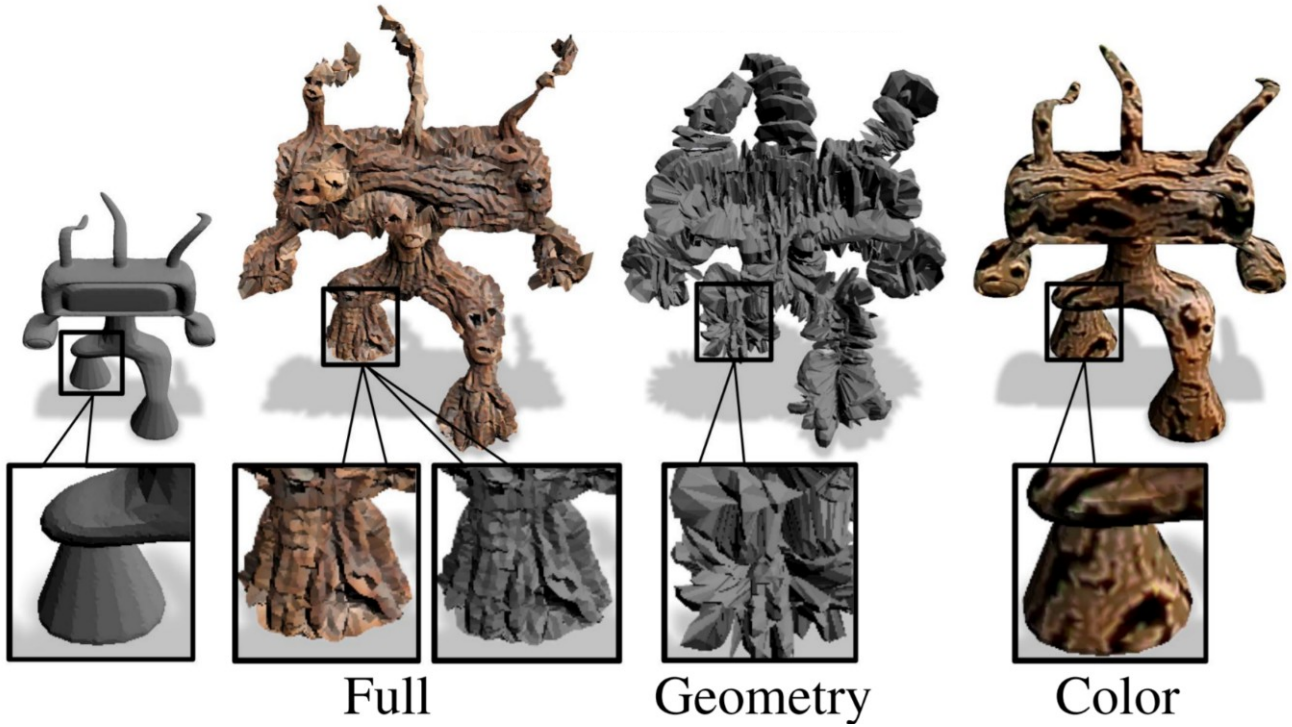
Neural Style Field



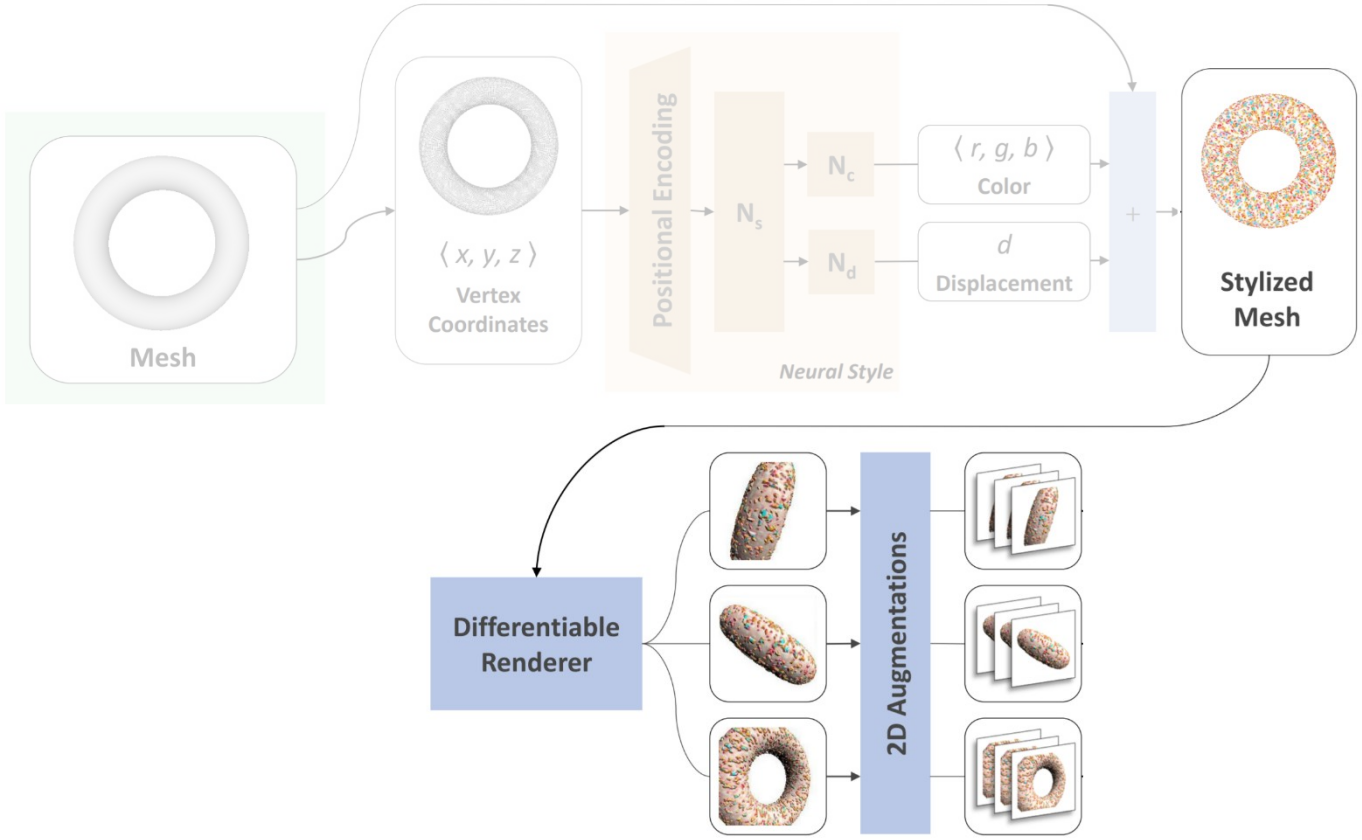
Geometry and Color



'Alien made of bark'

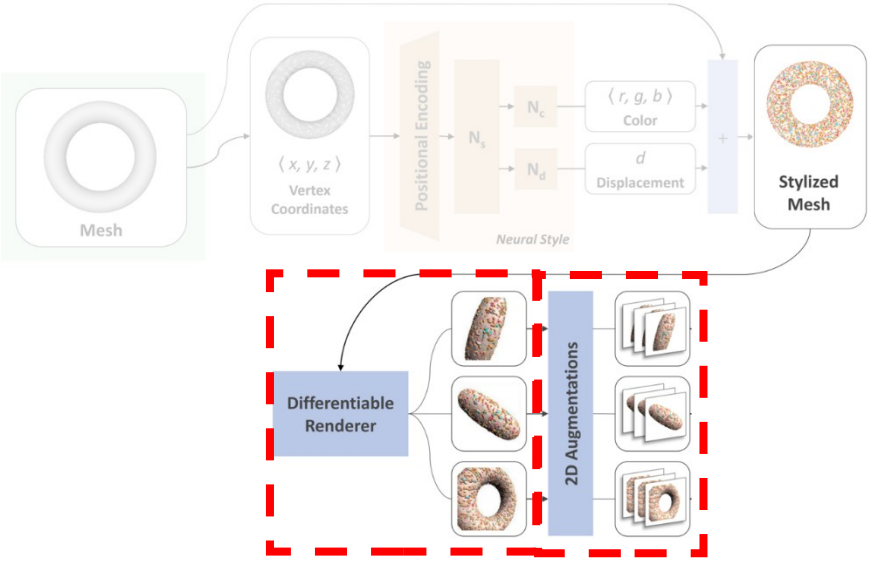


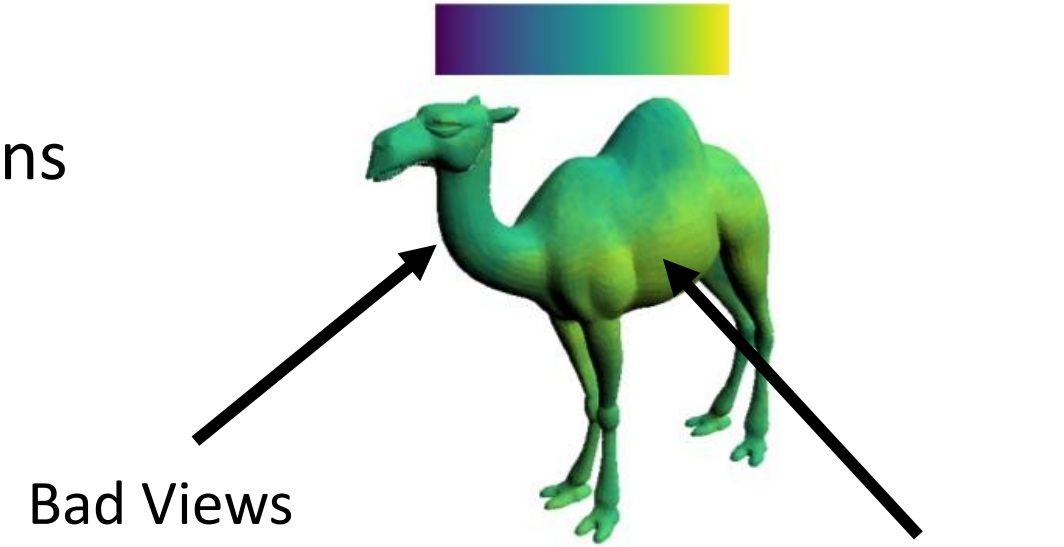
Neural Rendering and Augmentations



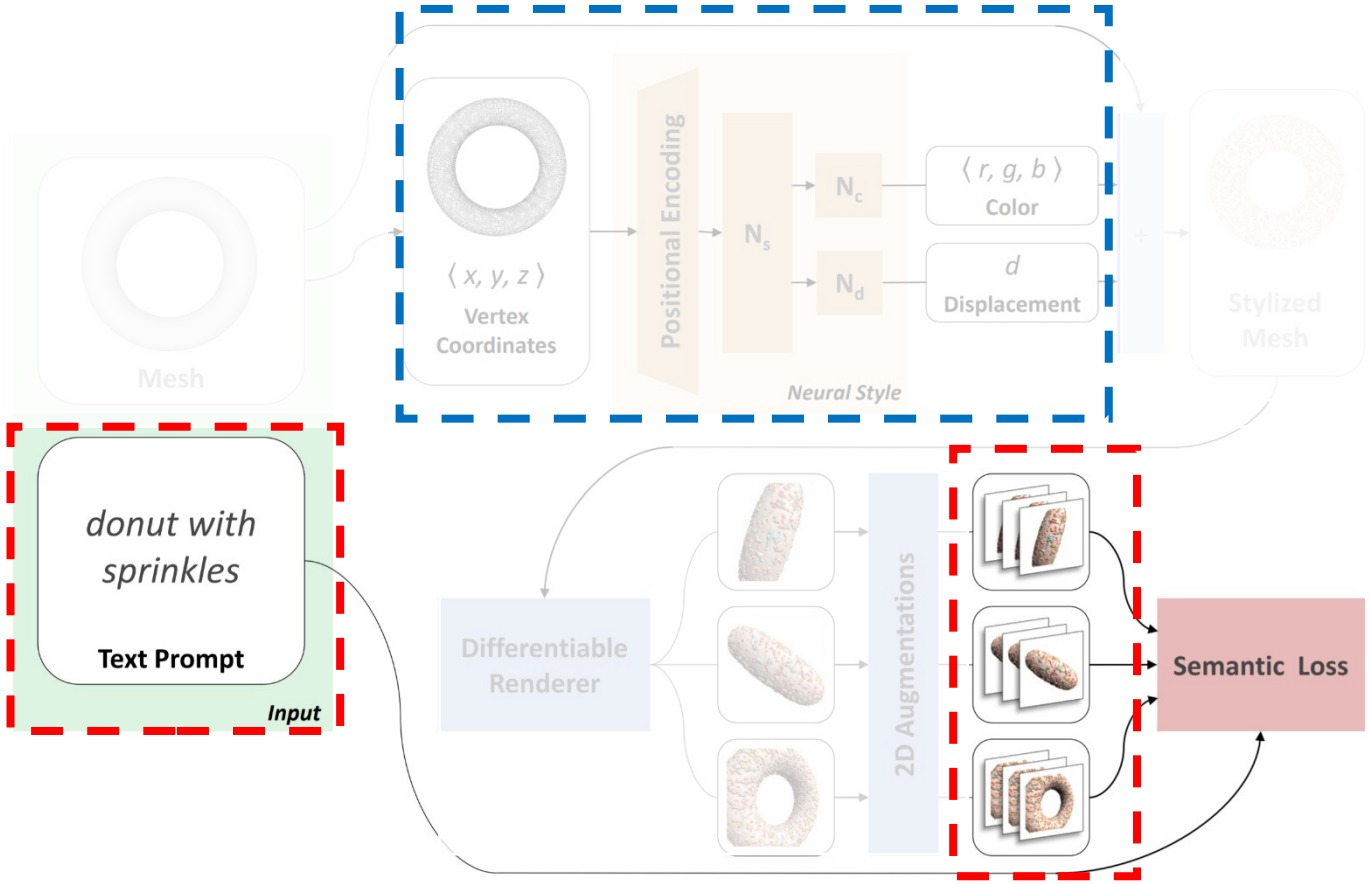
How are views selected?

- *Anchor view v:* view with high similarity to target in CLIP space
- Many such views exist!
- Sample random views from a where .
- 5 views are sufficient.
- Global and local set of augmentations

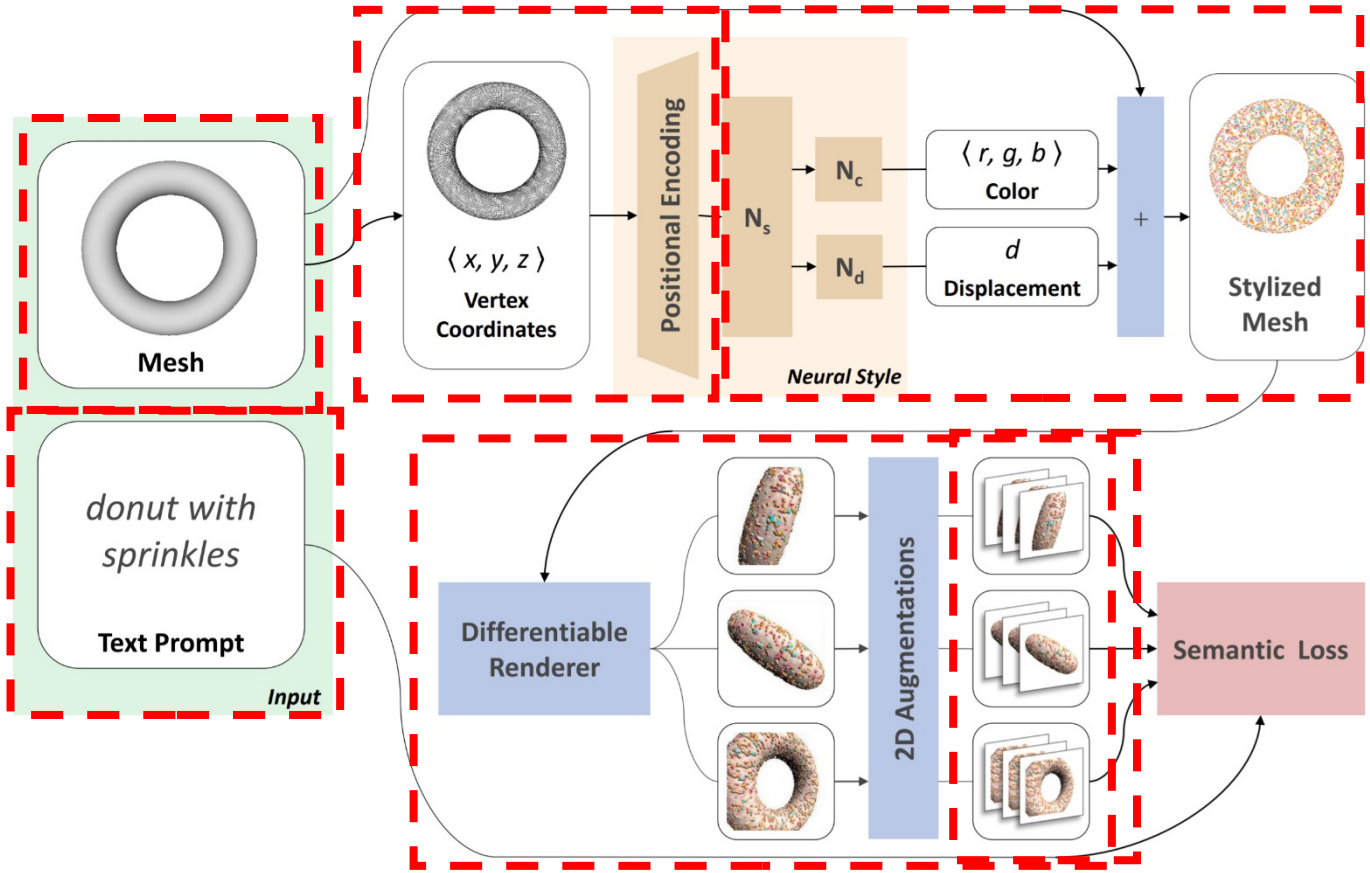




CLIP Based Semantic Loss

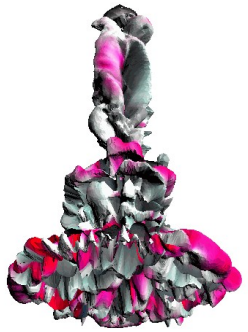


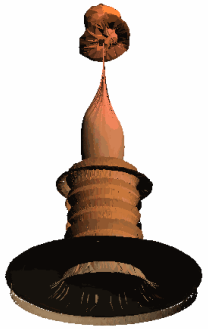
Full Method


Mesh
donut with sprinkles
Text Prompt
Input
⟨ x, y, z ⟩
Vertex Coordinates
Positional Encoding
N_s
N_c
N_d
Neural Style
⟨ r, g, b ⟩
Color
d
Displacement
+
Stylized Mesh
Differentiable Renderer
2D Augmentations
Semantic Loss

Components Introduce Prior

‘Candle made of bark’





-Net -Aug -FFN -Crops -Displ

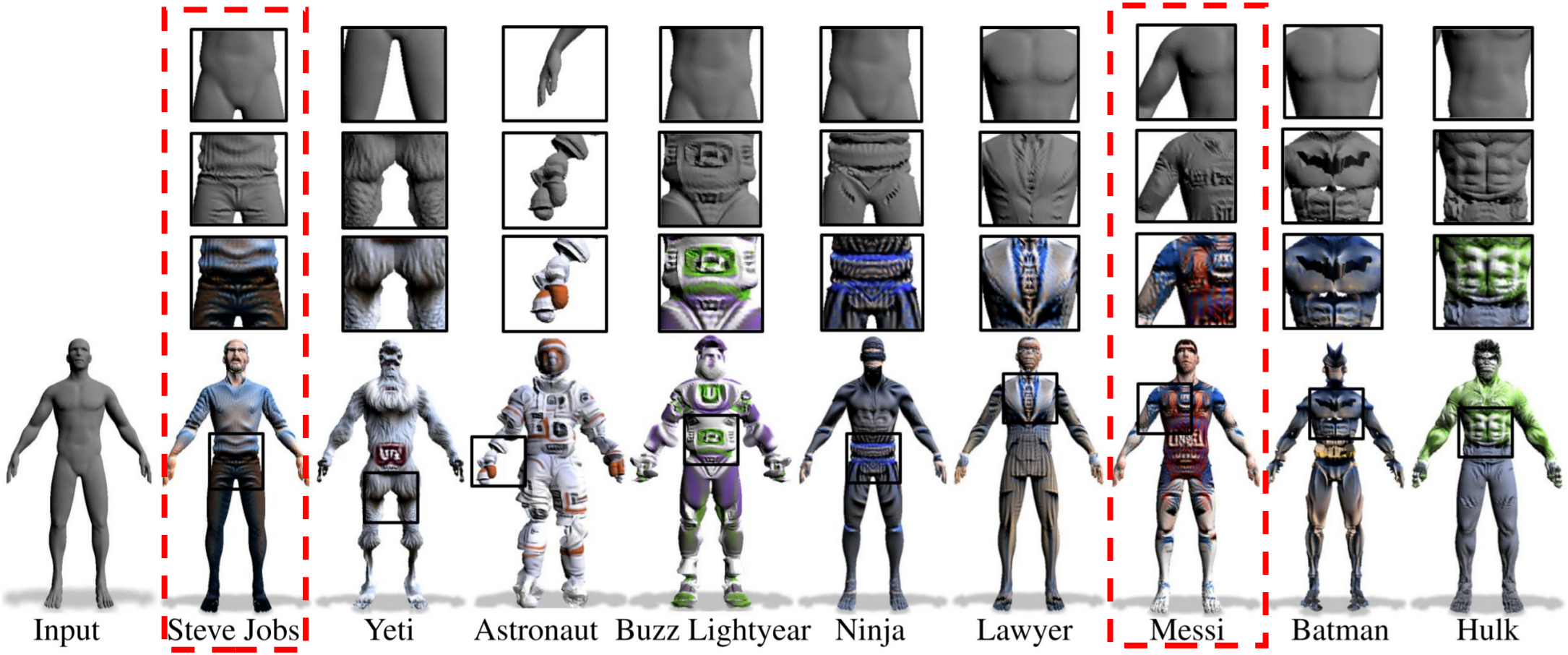
Results

Results

Morphing

a camel made of brown crochet

Humans


Input
Steve Jobs
Yeti
Astronaut
Buzz Lightyear
Ninja
Lawyer
Messi
Batman
Hulk

Increasing Granularity of Text

“Lamp”

Increasing Granularity of Text

“Luxo lamp”

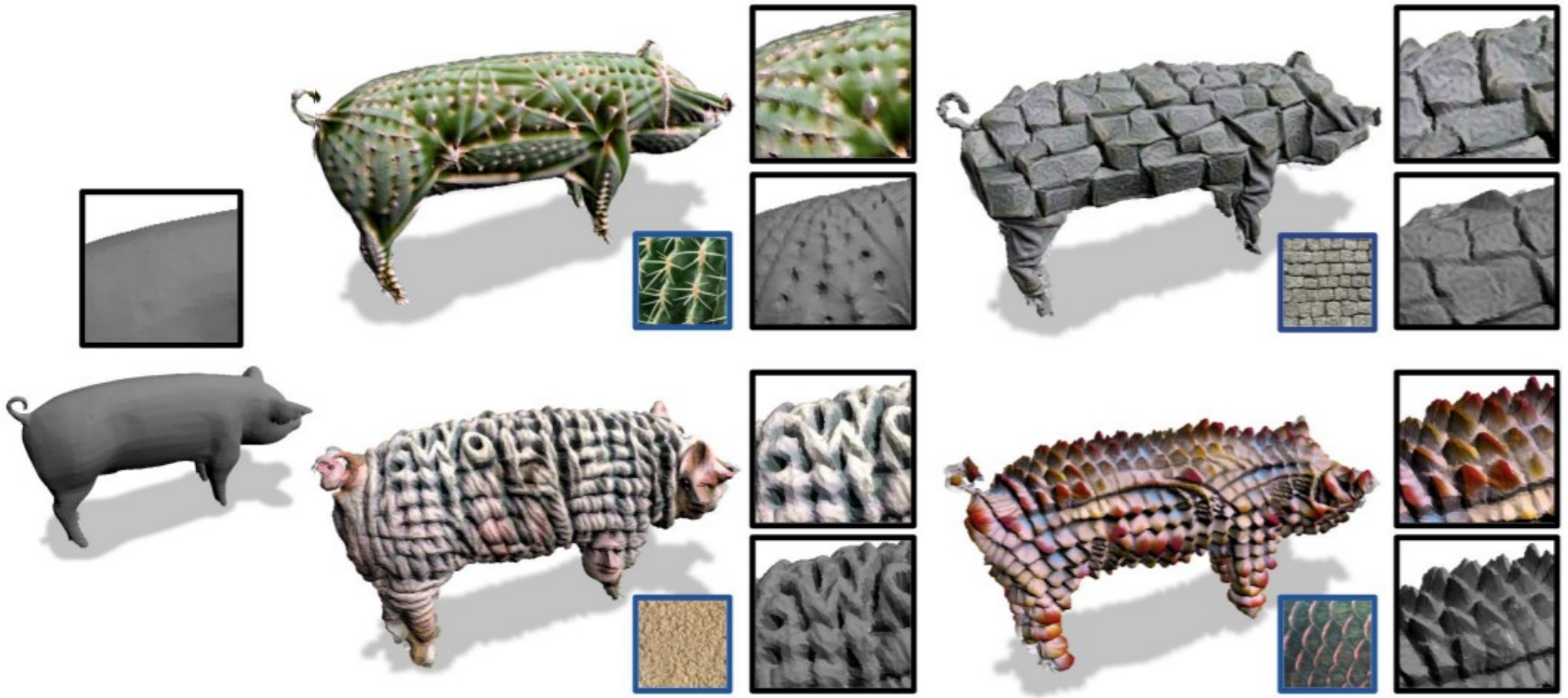
Increasing Granularity of Text

“Blue steel luxo lamp”

Increasing Granularity of Text

“Blue steel luxo lamp with corrugated metal”

Different Target Modality: **Image Target**



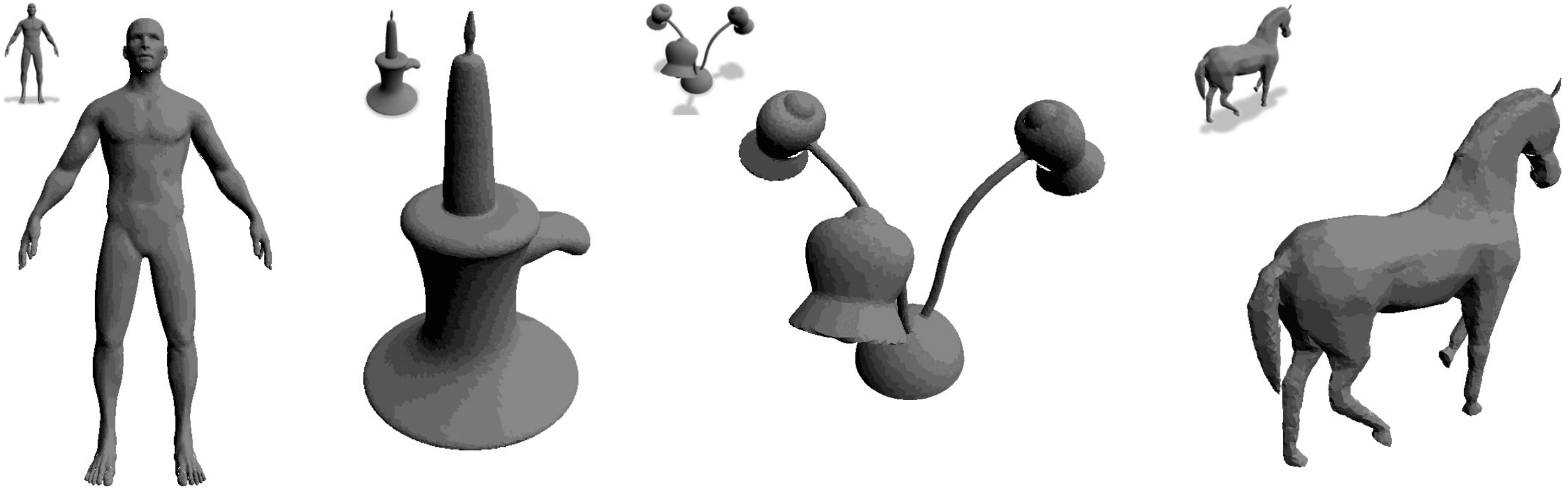
Different Target Modality: **Target Mesh**

Conclusion

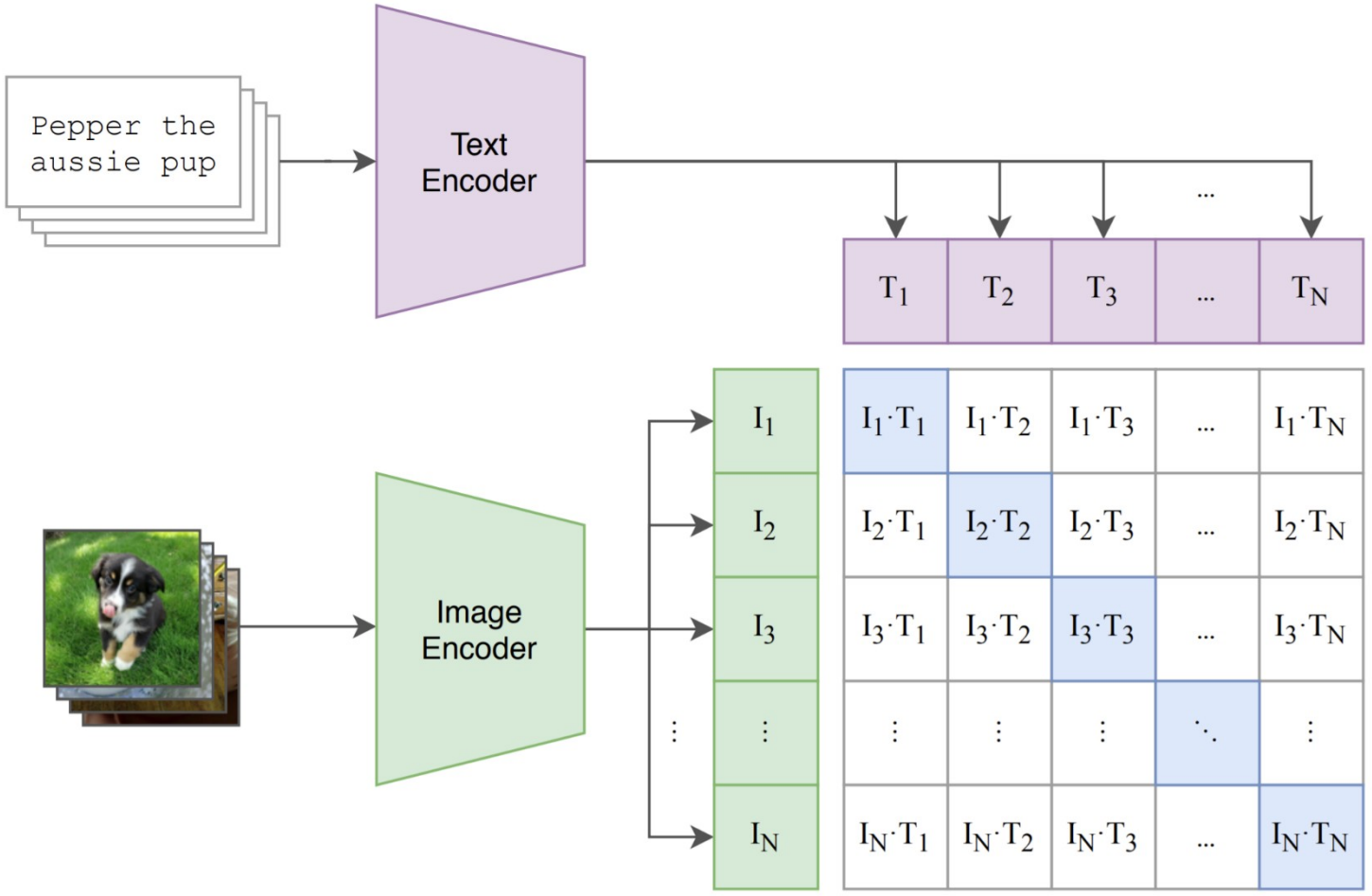
- Intuitive user control using text.
- Zero Shot! No GAN or 3D dataset required.
- Arbitrarily high resolutions can be rendered.
- Disentanglement into an explicit mesh *content* and an *implicit* neural style field.
- Fine grained control both in terms of text and 3D shape.
- In-the-wild meshes, arbitrary styles. Out-of-domain stylizations.

Thank You

Visit https://threedle.github.io/text2mesh/ for more



What is CLIP?



A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision”. ICML 2021.