TEXTDEFORMER GEOMETRY MANIPULATION USING TEXT GUIDANCE

W Gao, N Aigerman, T Groueix, V G. Kim, R Hanocka

University of Chicago &
Adobe Research

CONTENT

01 INTRODUCTION

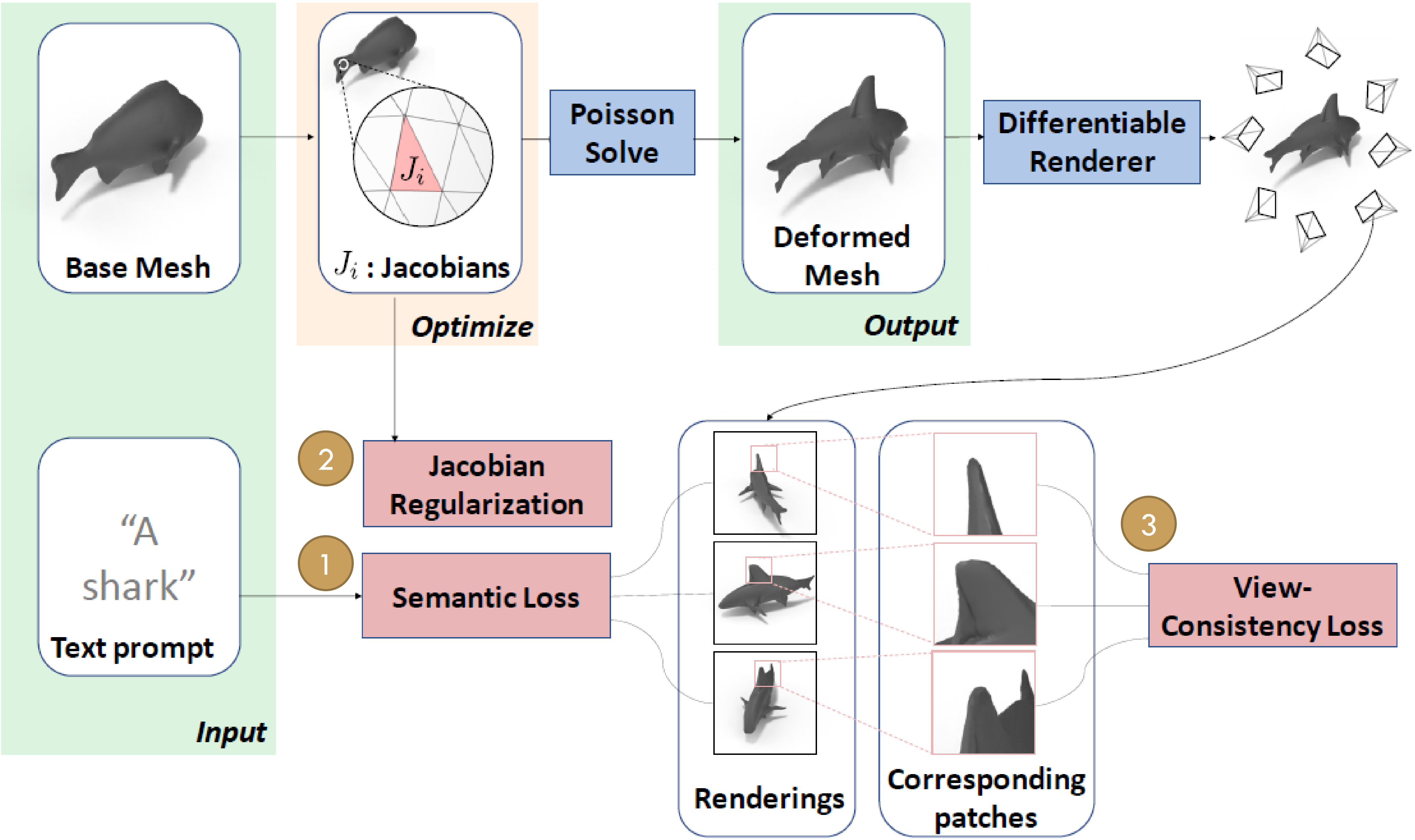
TEXTDEFORMER

- A method to deform 3D meshes into other shapes through text guidance.
- large, low-frequency shape changes, and small high-frequency details.
- not rely on 3D training data
- focus on the shape deformation task.

INTRODUCTION

1) produce high-quality surface geometry, with minimal self-intersections and noisy normal.
   - To deal with the significant artifacts, we optimize matrices representing the deformation's gradients, i.e., the Jacobians of each of the triangles, and compute the deformed vertex positions from them, by solving Poisson's equation.
2) produce plausible results which match the text description.
   - We devise a novel loss, which encourages vertices to achieve similar CLIP features from different viewpoints, thereby leading to global coherency in the deformations.
3) adhere to the input geometry. (e.g., deform the source’s head into the target’s head and not into body).
   - We add a identity-preserving term, which ensures that the deformation optimization step does not stray too far from the initial input mesh, thereby preventing the optimization from ignoring the input geometry.

02 METHOD



Base Mesh
J_i
J_i : Jacobians
Optimize
Poisson Solve
Deformed Mesh
Output
Differentiable Renderer
Input
“A shark”
Text prompt
1
Semantic Loss
2
Jacobian Regularization
Renderings
Corresponding patches
3
View-Consistency Loss

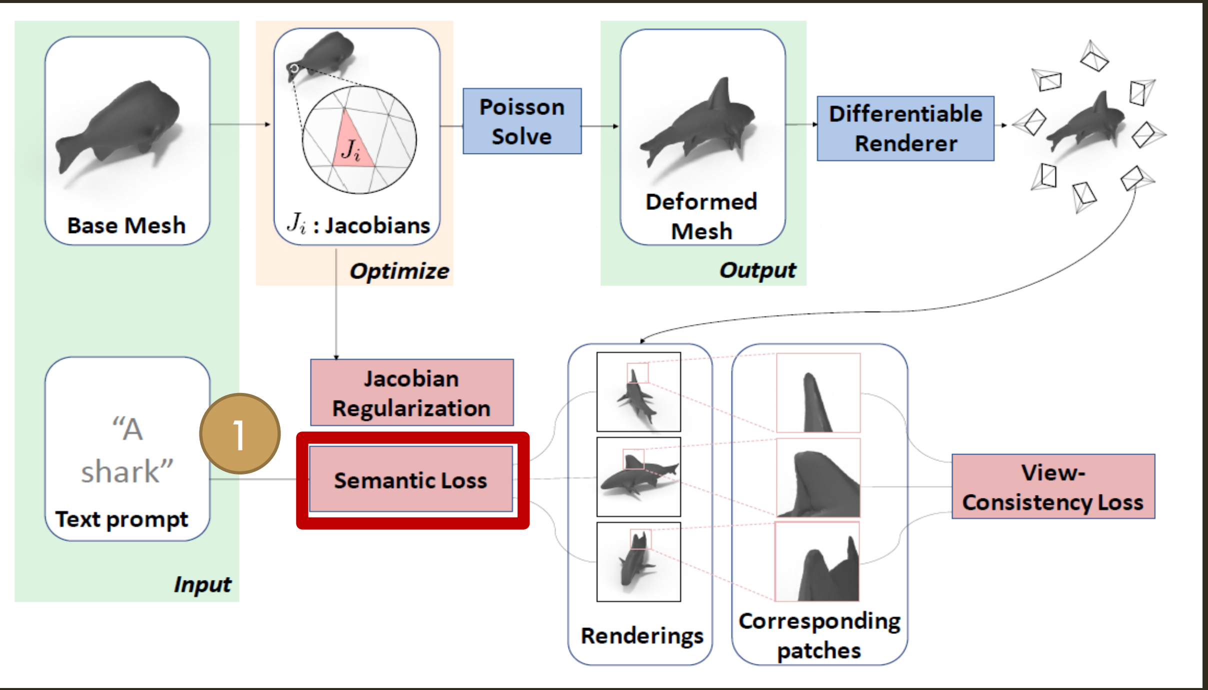
TEXTDEFORMER

TextDeformer deforms a base mesh by optimizing pertriangle Jacobians using natural language as a guide.

We optimize the deformation using three losses:

1) A CLIP-based semantic loss drives the deformation toward the text prompt.
2) our regularization on the Jacobians controls the fidelity to the base mesh.
3) a view-consistency loss matches multiple views of the same surface patch to ensure a coherent deformation.

SEMANTIC LOSS



Deformations through Jacobians

we represent per-triangle Jacobians by matrices $J_i \in \mathbb{R}^{3\times3}$ for every face $f_i \in \mathcal{F}$

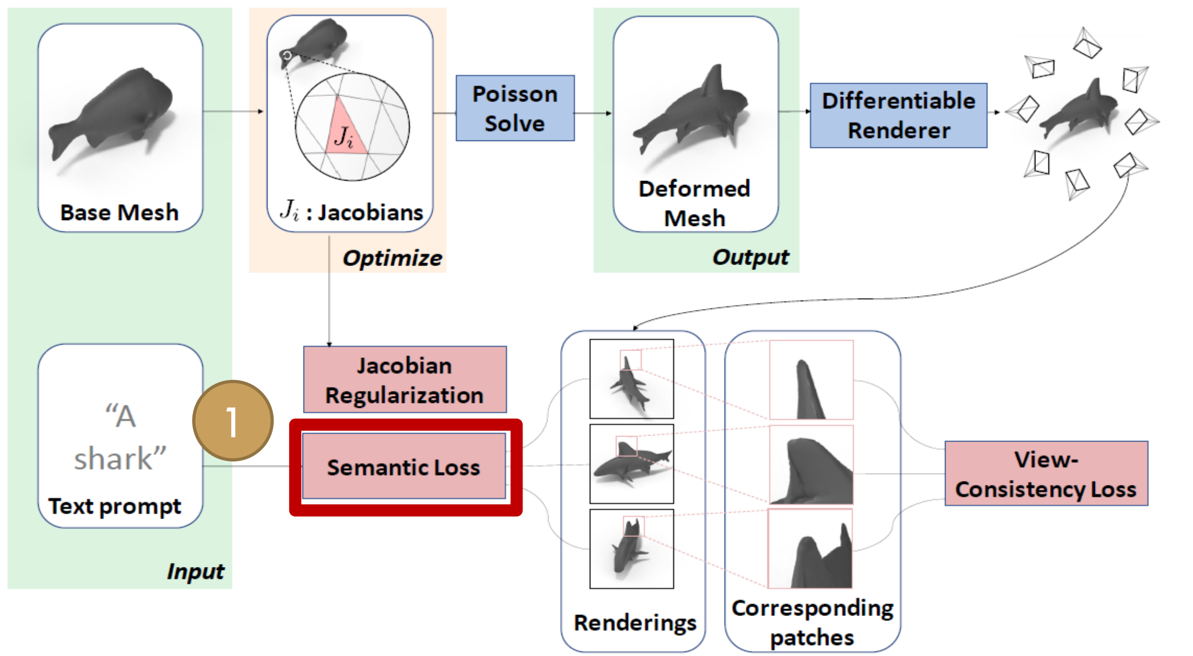
we solve a Poisson problem to compute a deformation map $\Phi*$ as the mapping with Jacobian matrices for each face that are closest to $\{Ji\}$ in the least-squares sense, that is: :

$$\Phi^* = \min_{\Phi} \sum_{f_i \in \mathcal{F}} |f_i| \|\nabla_i(\Phi) - J_i\|_2^2$$

$\nabla i(\Phi)$: the Jacobian of Φ at triangle fi

$|\ fi\ |$: the area of that triangle.

SEMANTIC LOSS



$$\mathcal{L}_{\Delta\mathcal{P}}(\Phi^*,\mathcal{P},\mathcal{P}_0) = \mathrm{sim}\left(\Delta\mathrm{CLIP}(\mathcal{P},\mathcal{P}_0), \Delta\mathrm{CLIP}(\Phi^*(\mathcal{M}),\mathcal{M})\right)$$

$$\Delta\mathrm{CLIP}(\mathcal{P},\mathcal{P}_0) = \mathrm{CLIP}(\mathcal{P}) - \mathrm{CLIP}(\mathcal{P}_0)$$

Language Guidance

pre-trained visionlanguage CLIP

R : differentiable renderer

deformed shape :

$$e_{\mathcal{M}} = \mathrm{CLIP}(\Phi^*(\mathcal{M})) \in \mathbb{R}^{512}$$

language prompt :

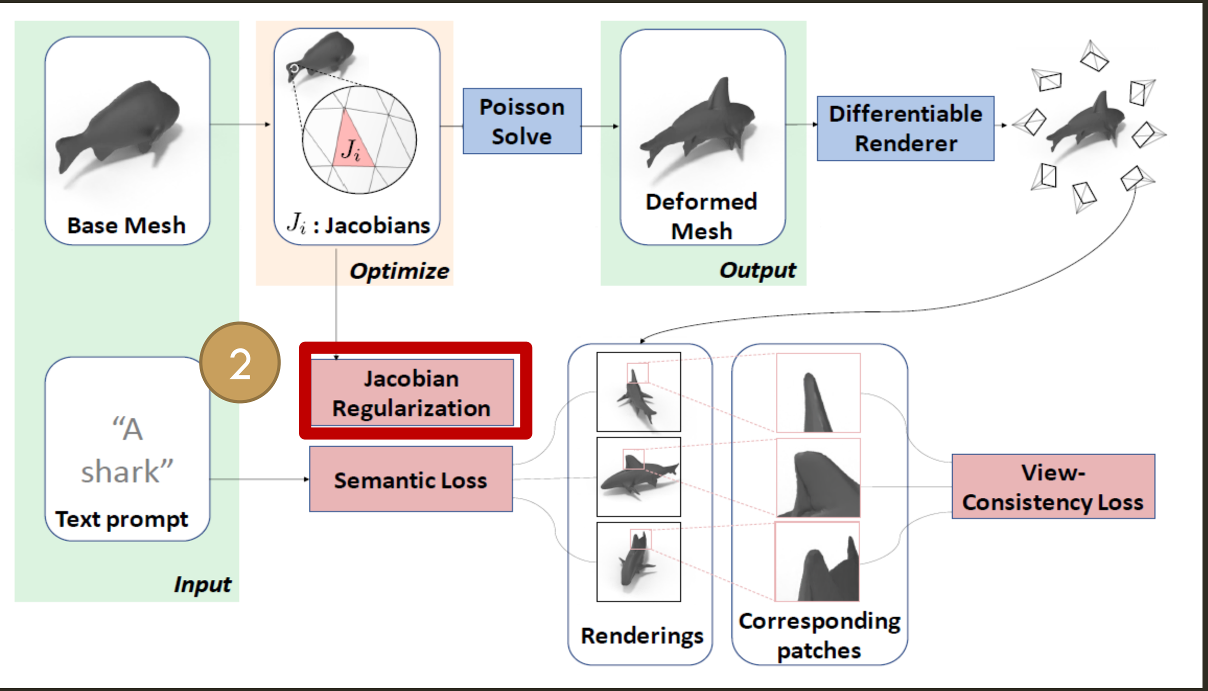
$$e_{\mathcal{P}} = \mathrm{CLIP}(\mathcal{P}) \in \mathbb{R}^{512}$$

we may optimize Φ∗ such that eM and eP agree, by maximizing the cosine similarity between the embeddings:

$$\mathcal{L}_{\mathcal{P}}(\Phi^*,\mathcal{M},\mathcal{P}) = \mathrm{sim}\left(e_{\mathcal{M}}, e_{\mathcal{P}}\right)$$

Where sim(·, ·) stands for cosine similarity.

JACOBIAN REGULARIZATION

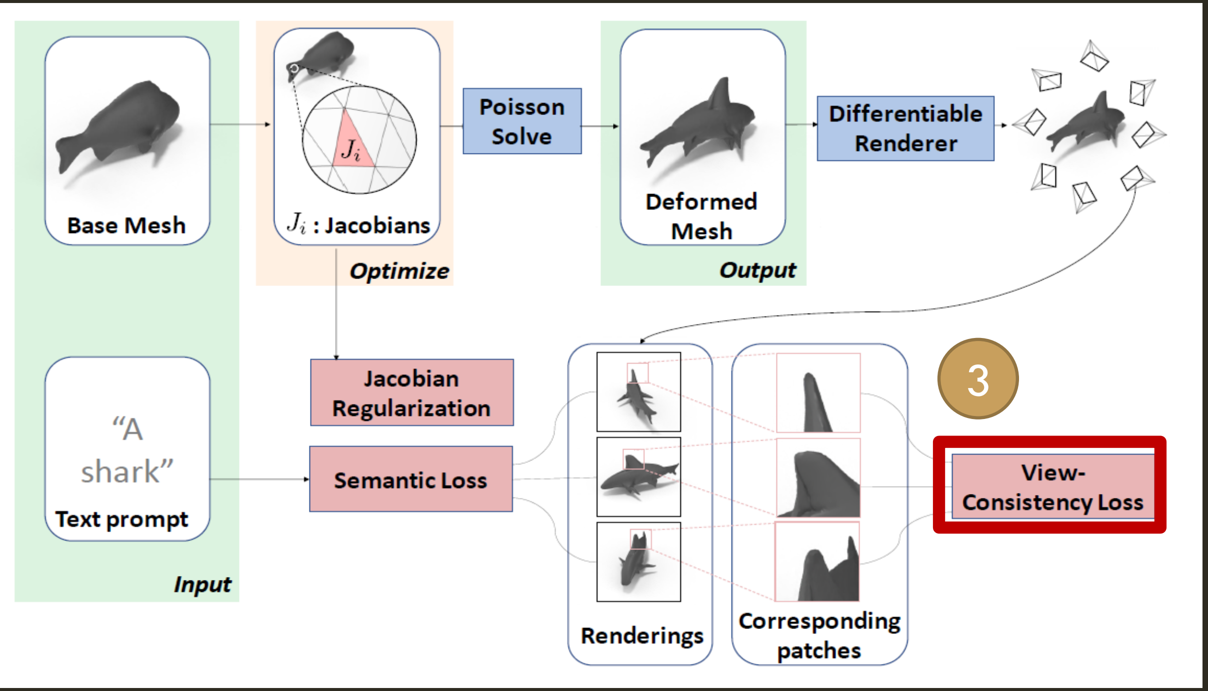


To prevent the deformation from straying too far from the input undeformed geometry, we introduce another regularization term on the predicted Jacobians, which penalizes the difference between the Jacobians $\{Ji\}$ and the identity, i.e., no deformation:

$$\mathcal{L}_I(t_j) = \alpha \sum_{i=1}^{|\mathcal{F}|} \|J_i - I\|_2$$

α : a hyper-parameter which may be tuned to control the strength of the deformations defined by $\{Ji\}$.

VIEW-CONSISTENCY LOSS



the patch-level deep features of CLIP's vision transformer (ViT)

encourage vertices to have similar deep features across renders from different viewpoints:

$$\mathcal{L}_{VC}(v) = \sum_{i=1}^{|\mathcal{R}(\mathcal{M})|} \sum_{\substack{j=1 \\ j \neq i}}^{|\mathcal{R}(\mathcal{M})|} \mathrm{sim}\left(\mathcal{T}_k(P(v, r_i)), \mathcal{T}_k(P(v, r_j))\right)$$

for some chosen layer T_k

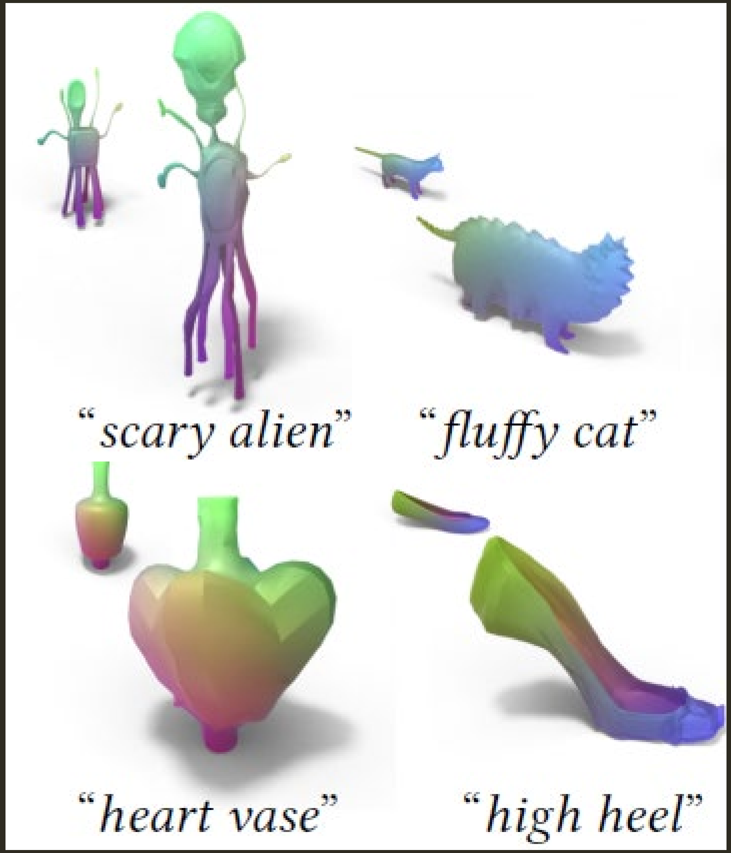
We penalize this loss over all vertices $v \in M$:

$$\mathcal{L}_{VC}(\mathcal{M}) = \beta \sum_{v \in \mathcal{V}} \mathcal{L}_{VC}(v)$$

β : another tunable hyper-parameter.

03 EXPERIMENTS

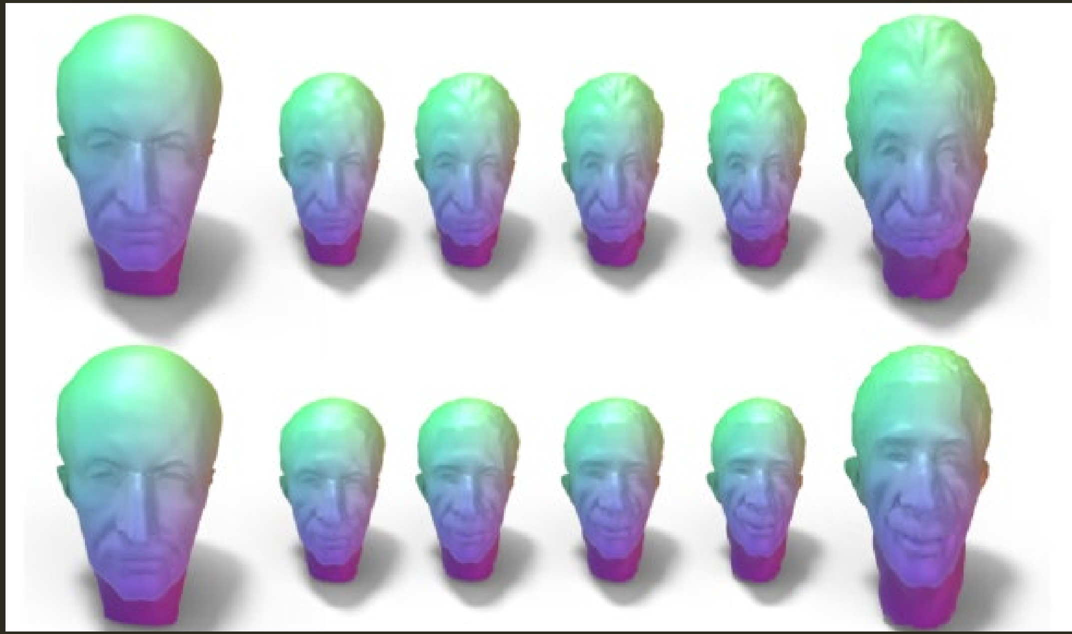
GENERALITY OF TEXTDEFORMER



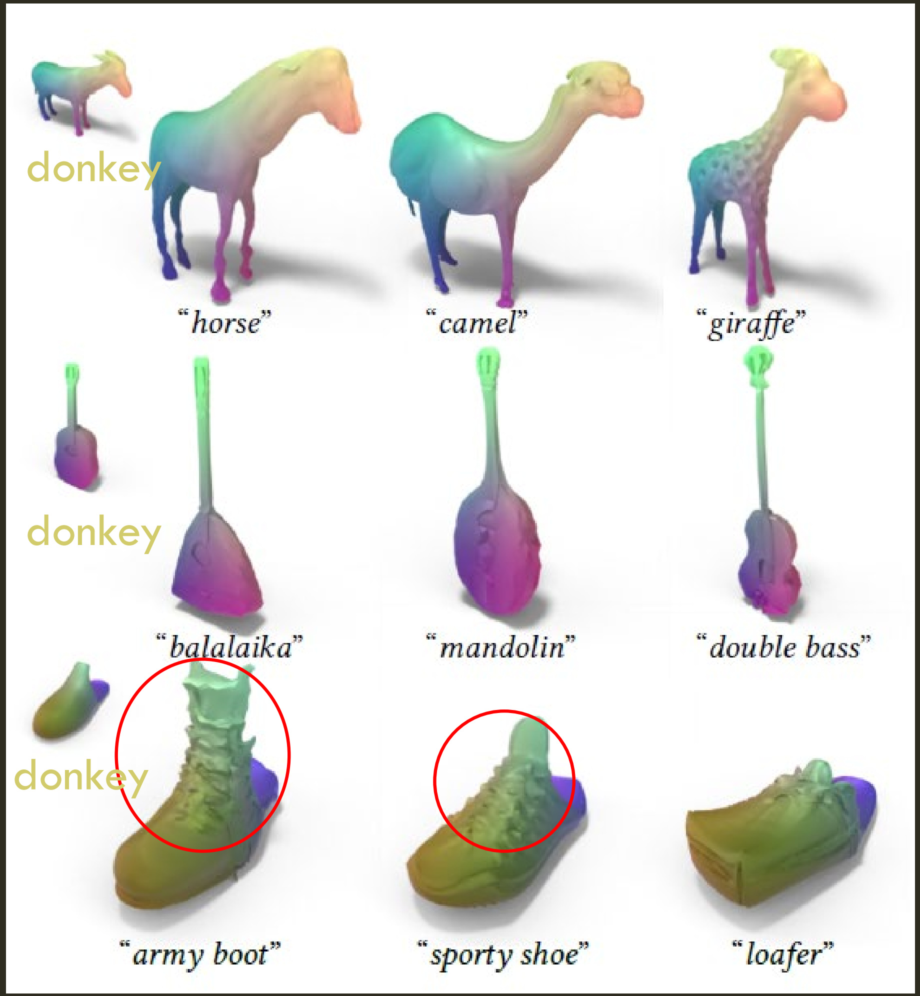
- Adjective Targets
- Related Targets
- Unrelated Targets

EXPRESSIVENESS OF TEXTDEFORMER

1. Frequency
   - ➢ low-frequency deformations : other shapes
   - ➢ high-frequency deformations : fine details
2. Dense Matching
   - ➢ nose to nose, eyes to eyes etc.



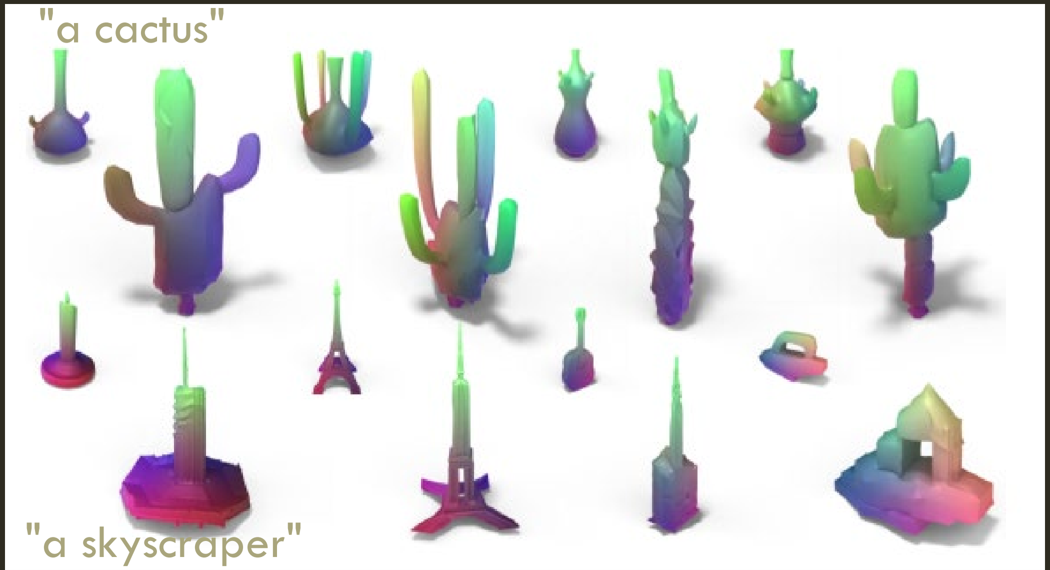
➢ same source + different text prompts



IDENTITY PRESERVATION

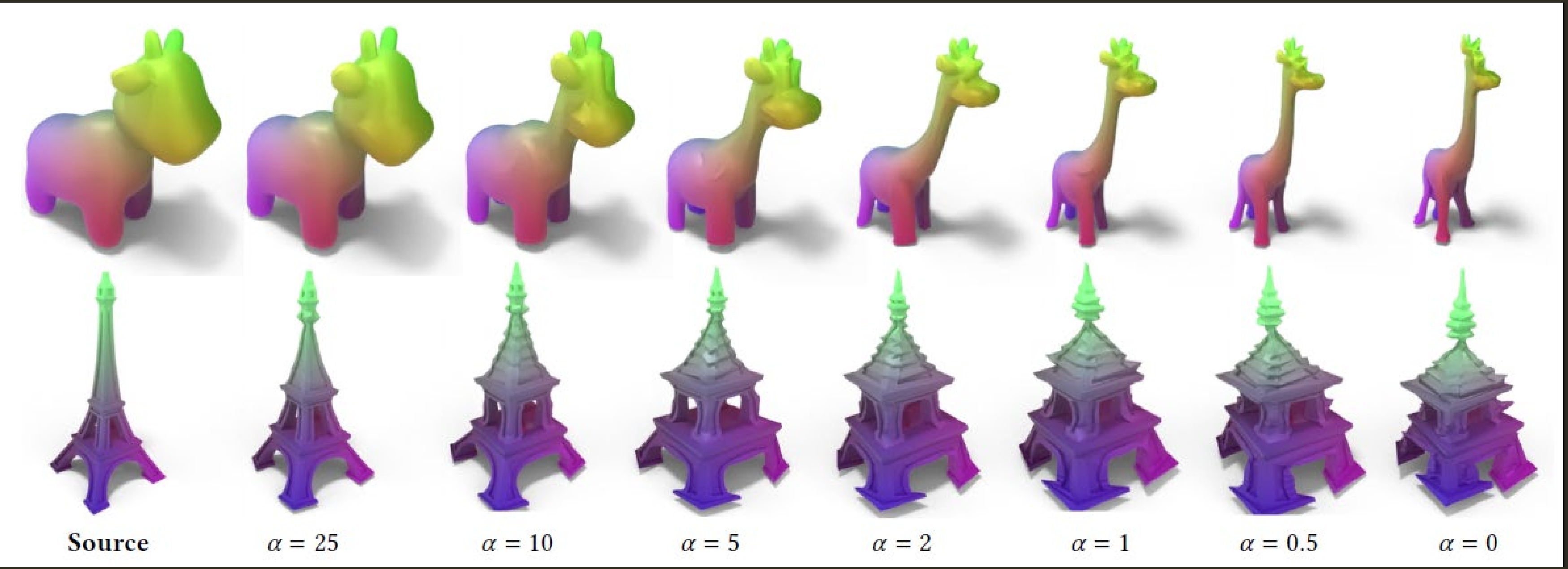
1. Impact of the Input Geometry
2. Jacobian Regularization

➢ different source + same text prompts



$$\mathcal{L}_I(t_j) = \alpha \sum_{i=1}^{|\mathcal{F}|} \|J_i - I\|_2$$

- $\alpha = 25$: the cow and the Eiffel tower do not change meaningfully in accordance to their respective text prompts ("giraffe" and "pagoda")
- $\alpha = 0$: result in some artifacts in the deformed shape.
- Setting α to intermediate values offers the best results.



ABLATION

Viewpoint Consistency Ablation

➜ the qualitative effect of the viewpoint consistency loss (Lvc)

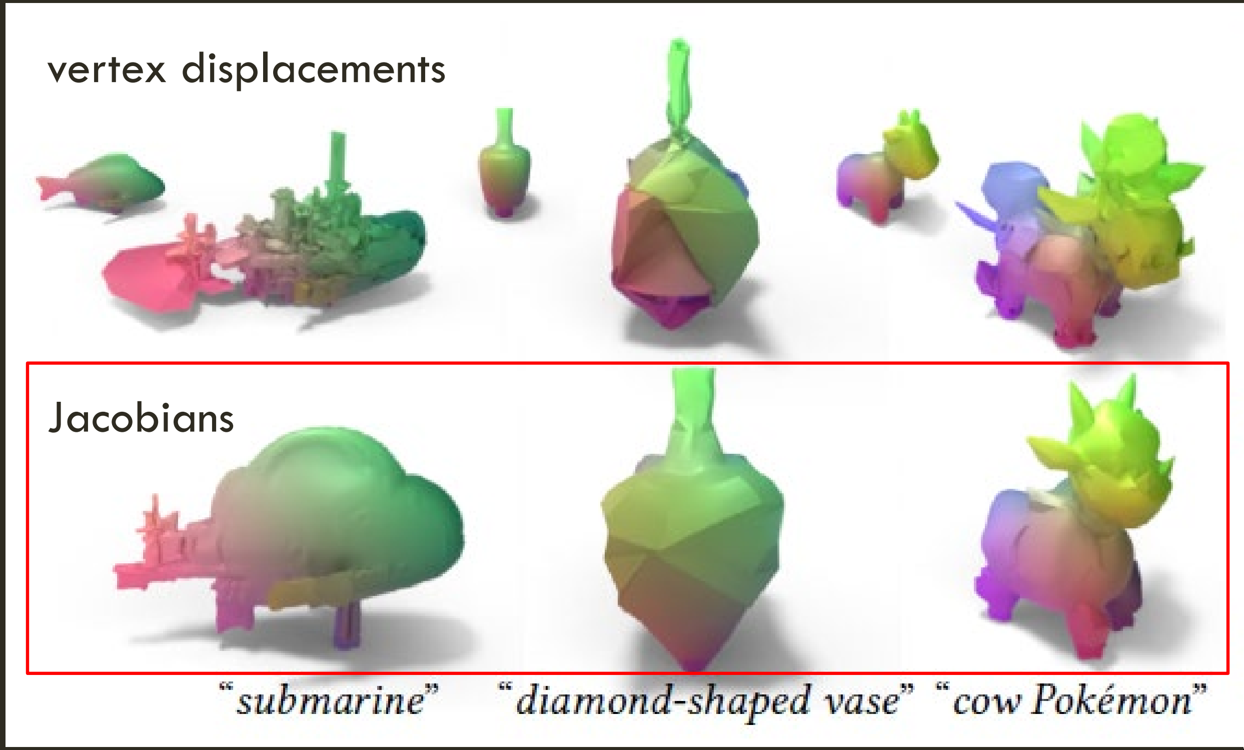
$$\mathcal{L}_{\mathrm{VC}}(v) = \sum_{i=1}^{|\mathcal{R}(\mathcal{M})|} \sum_{\substack{j=1 \\ j \neq i}}^{|\mathcal{R}(\mathcal{M})|} \mathrm{sim}\left(\mathcal{T}_k(P(v, r_i)), \mathcal{T}_k(P(v, r_j))\right)$$

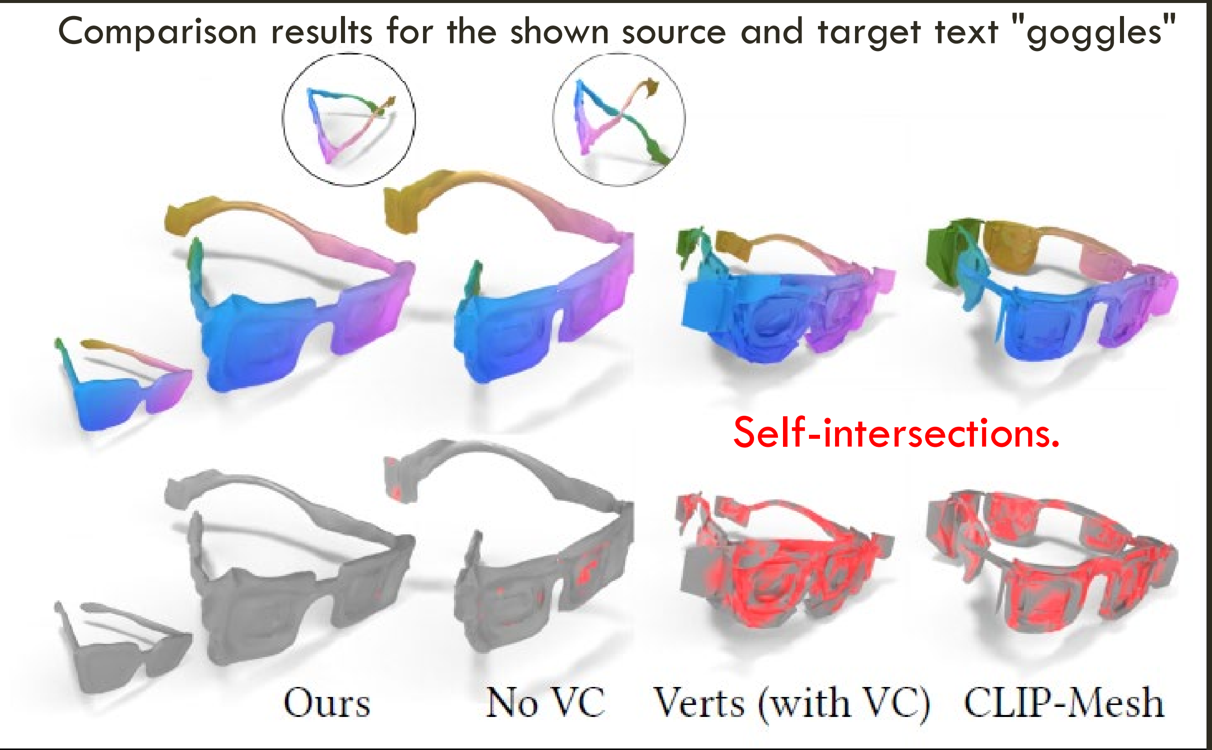
$$\mathcal{L}_{\mathrm{VC}}(\mathcal{M}) = \beta \sum_{v \in \mathcal{V}} \mathcal{L}_{\mathrm{VC}}(v)$$

ABLATION

<u>Deformation Ablation</u> ➜ Effect of Jacobians

1) Surface Quality

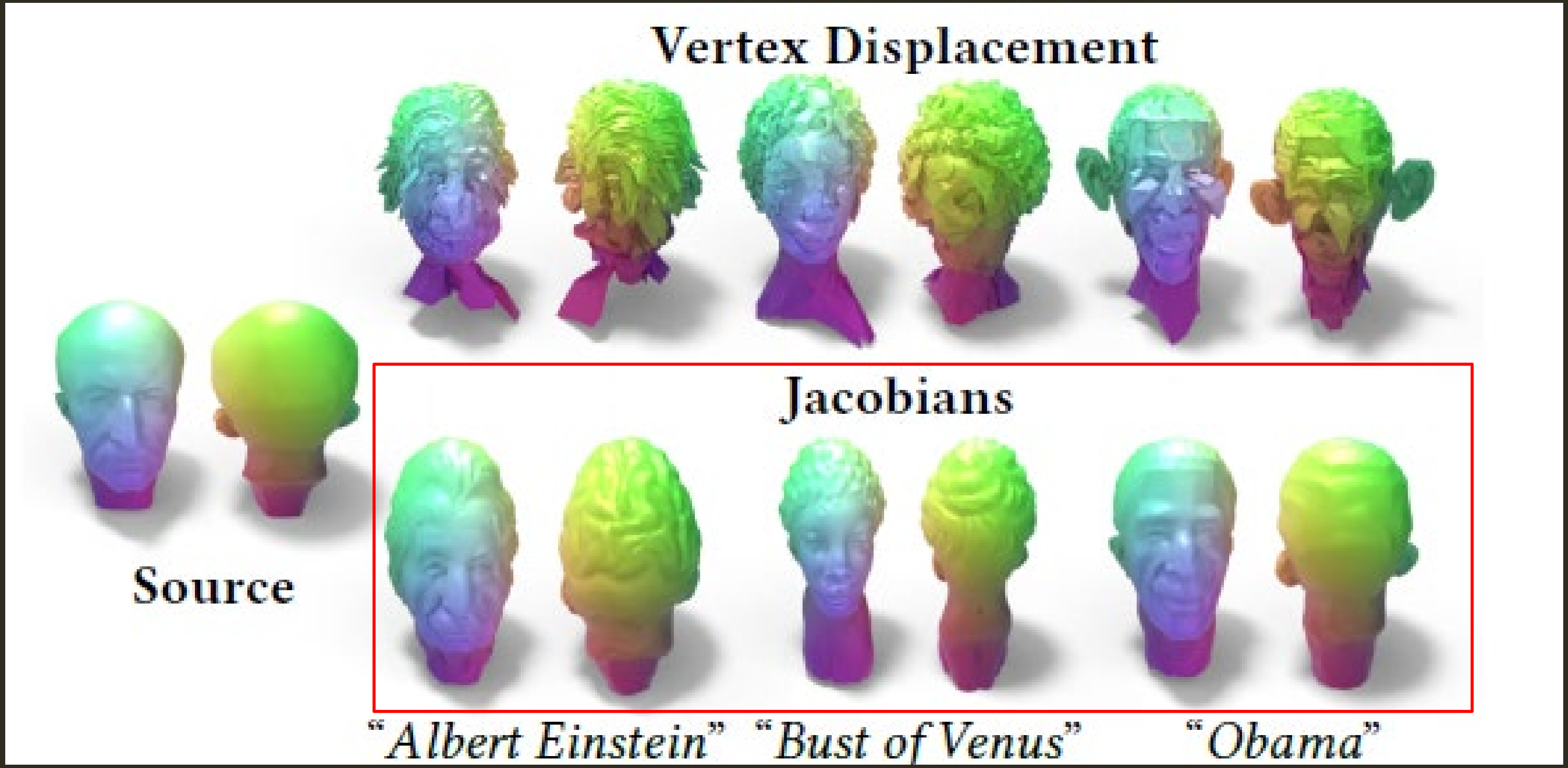




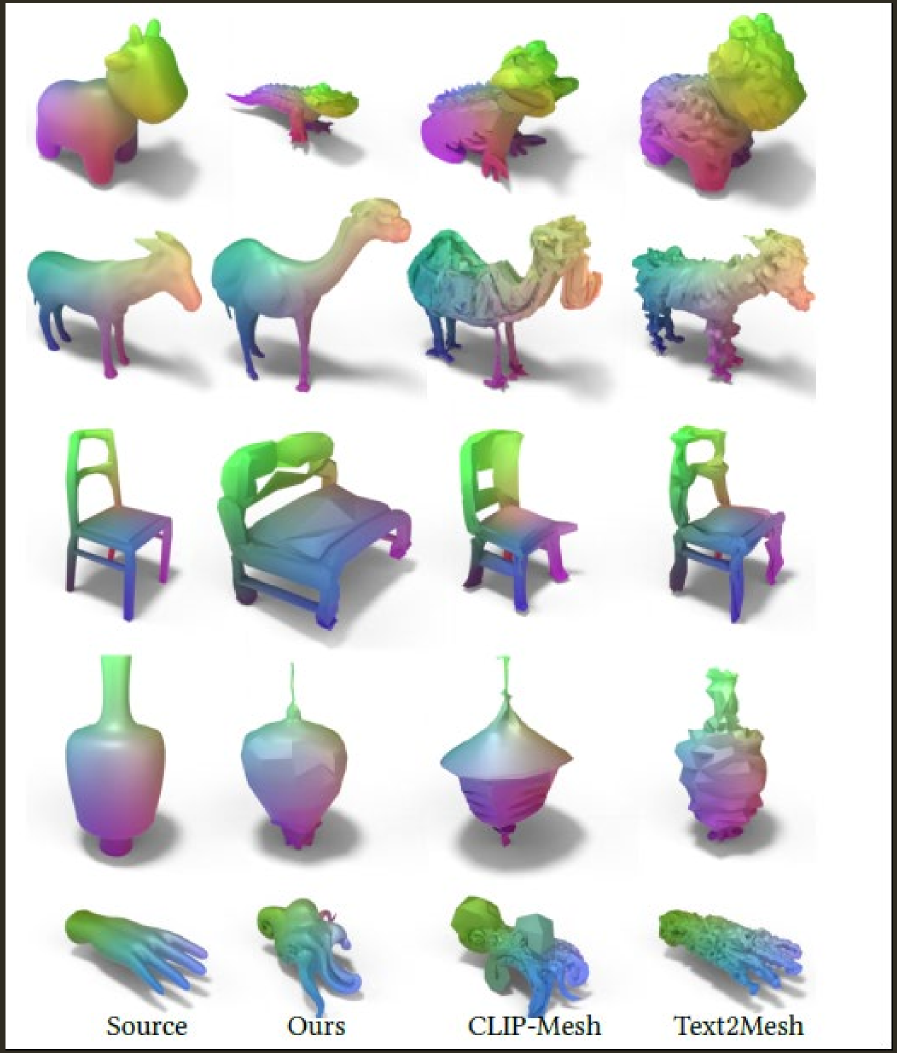
ABLATION

Deformation Ablation ➜ Effect of Jacobians

2) Globally-Coherent Deformations



QUANTITATIVE COMPARISON



- the surface of Dreamfusion meshes has heavy artifacts compared to the smoothness of our deformations obtained through Jacobians.
- Dreamfusion suffers frequently from the Janus problem (see the high heels for instance) which we help alleviate with our View-Consistency loss.

QUANTITATIVE EVALUATION

1) Retrieval Precision
   - the viewpoint consistency loss (L_{VC}) increases the deformation quality of the model

2) Geometric Quality (Self-Intersections)

	CLIP R-Precision (L/14) ↑	Intersections ↓
Ours	55.2%	3.2%
Ours-noVP	51.5%	3.3%
Ours-Verts	55.4%	67.7%
CLIP-Mesh	57.4%	62.8%
Text2Mesh	12.7%	17.3%

Comparison results for the shown source and target text "goggles"

04 CONCLUSION

CONCLUSION

- TextDeformer, a zero-shot text-driven mesh deformation technique.
  - not need to be trained on any 3D dataset or 3D annotations.
  - pre-trained vision-language models trained on billions of visual and language concepts.

- Our work aims to produce high-quality geometry outputs by predicting low-frequency shape changes and high-frequency details through source shape deformations.
  - use per-face Jacobians as a means for predicting smooth mesh deformations enables retaining interesting characteristics of the source shape.
  - an identity regularization term can be controlled by the user to control the magnitude of the deformation.
  - view consistency loss avoids over-fitting geometry to specific salient views, and ensures that the same region is roughly interpreted the same from all viewpoints.

Q & A