ShapeCraft: Body-Aware and Semantics-Aware 3D Object Design

HANNAH CHA, MICHELLE GUO, MIA TANG, RUOHAN ZHANG, KAREN LIU, and JIAJUN WU, Stanford University, USA

Fig. 1. ShapeCraft generates 3D shapes given text as input. The objects are optimized to fit on various character body shapes.

1 ABSTRACT

For designing a wide range of everyday objects, the designing process should be aware of both the human body and the underlying semantics of the design specification. However, these two objectives present significant challenges to the current AI-based designing tools. In this work, we present a method to synthesize body-aware 3D objects from a base mesh given an input body geometry and either text or image as guidance. The generated objects can be simulated on virtual characters, or fabricated for real-world use. We propose to use a mesh deformation procedure that optimizes for both semantic alignment as well as contact and penetration losses. Using our method, users can generate both virtual or real-world objects from text, image, or sketch, without the need for manual artist intervention. We present both qualitative and quantitative results on various object categories, demonstrating the effectiveness of our approach.

ACM Reference Format:
Hannah Cha, Michelle Guo, Mia Tang, Ruohan Zhang, Karen Liu, and Jiajun Wu. 2024. ShapeCraft: Body-Aware and Semantics-Aware 3D Object Design. *ACM Trans. Graph.* 1, 1 (June 2024), 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Authors' address: Hannah Cha; Michelle Guo; Mia Tang; Ruohan Zhang; Karen Liu; Jiajun Wu, Stanford University, USA.

0730-0301/2024/6-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 INTRODUCTION

Have you struggled to find an everyday object that will fit your body perfectly and match the exact creative concept you have in mind? Recent progress in generative AI models shows promising results in generating 3D objects, which have the potential to facilitate the design process (e.g., help designers rapidly iterate ideas) and enable better customization in industrial design [Chui et al. 2023; Epstein et al. 2023; Makatura et al. 2023]. For designing a wide range of everyday objects, such as glasses, hats, rings, and shoes, the designing process should be aware of both the human *body* and the object *semantics*. For these objects that are designed to be used

[1] M.G., M.T., R.Z., K.L., and J.W. are not enrolled in CS231N.

by humans, being body-aware is essential and the design should be primarily optimized for the interaction between the object and the body that it is designed for. In addition, we want to be able to customize the design, styles, or aesthetics of these objects, i.e., we want the design to be semantically-aware in the sense that it aligns with our design specifications, which can be either text descriptions or visual examples. Therefore, we need to provide tools to address individual differences in the different object categories' demands of body fitness and the underlying semantics of the designs.

Generative AI models, such as Stable Diffusion [Rombach et al. 2022], DALL-E [OpenAI 2023], and DreamFusion [Poole et al. 2022], can generate semantics-aware 3D assets when given design specifications in the format of natural language, although text-to-image models would require another step that converts 2D designs into 3D objects using image-to-3D models [Liu et al. 2023b]. However, these approaches typically optimize objects for semantics-related objectives, such as prompt alignment. However, designing useful objects requires an understanding of the physical interactions between bodies and objects. It is difficult to use text or image to specify design needs for different body shapes and preferred body contacts, hence the resulted designs are not sufficiently body-aware. Additionally, the generated designs are derivatives of datasets that belong to a specific population of certain body shape and size; therefore, to generate designs for characters of all shapes and sizes, we need to explicitly incorporate the awareness of body shapes and contacts within the generative models.

On the other hand, while some previous methods [Blinn et al. 2021; Mezghanni et al. 2022] have been optimizing body contact or functionality for objects, their methods are usually limited to common objects. The optimization process does not consider semantics-related design specifications, such as text or image prompts. Additionally, optimizing functionality for creative objects, especially for individual human bodies and preferred contacts, is significantly more challenging and not well-addressed. Meanwhile, there are works that address both body- and semantics-aware objectives, however, they are limited to specific object categories, such as garments [Sarafianos et al. 2024; Wang et al. 2018].

In this work, we propose a tool to generate both body-aware and semantically-aware, customized 3D designs. The tools can be applied to a wide range of everyday object categories, without relying on object datasets. We build a flexible system that jointly optimizes for multiple objectives. We define semantically-aware design as the process of designing according to a text or visual concept. Personalized, body-aware design is generating a 3D shape that is well-fitted to an individual body or even a specific contact map.

As shown in Figure 1, we showcase a gallery of our generated designs for various digital avatars and multiple object categories. Our qualitative results show that ShapeCraft is effective in generating designs that are simultaneously body-aware and semantics-aware. Additionally, we show that compared to baselines, our joint optimization approach achieves the best results in terms of both objective and subjective metrics.

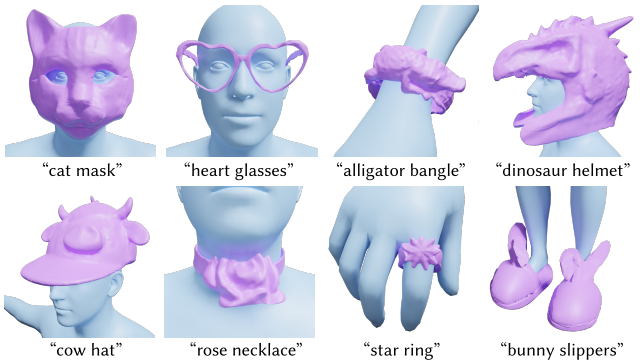

Fig. 2. Our method generates a variety of semantics and body-aware objects from input text prompts.

Fig. 3. Our method can deform the same template mesh into different text-specified geometries that are body-fitting.

3 RELATED WORK

Text or image-conditioned 3D synthesis. Recent works propose to tackle text-conditioned 3D generation either via text-to-3D [Chen et al. 2023a; Lin et al. 2023; Tsalicoglou et al. 2023; Zhu et al. 2023], or image-to-3D [Liu et al. 2023b,a; Qian et al. 2023] where the input image is generated by a text-to-image model such as Stable Diffusion [Rombach et al. 2022] or DALL-E [OpenAI 2023]. Another line of work directly trains 3D diffusion models for various 3D representations, including point clouds [Nichol et al. 2022], meshes [Gao et al. 2022; Liu et al. 2022], or neural fields [Jun and Nichol 2023]. Finally, other works achieve text or image-conditioned mesh generation by deforming a template mesh through text or image guidance [Gao et al. 2023; Michel et al. 2022; Sarafianos et al. 2024].

Compared to text-to-image models, text-to-3D is significantly more challenging, partially due to the lack of large-scale training datasets. However, text-to-3D models can leverage pre-trained 2D models, such as CLIP, to synthesize better objects. Guided by a text prompt (embedded using CLIP), Dreamfields [Jain et al. 2022] synthesize 3D objects leveraging volume rendering. DreamFusion [Poole et al. 2022] and [Wang et al. 2022] distill 2D diffusion models as a differentiable image-based loss. Surface-based differentiable rendering can be used to pass views of explicit 3D objects to CLIP, such as Text2Mesh [Michel et al. 2022] in which they stylize the template mesh while preserving the initial content. CLIP-Mesh [Khalid et al. 2022] generates new 3D objects by deforming a sphere at the vertex level, guided by the input text prompt. Magic3D [Lin et al. 2023] first optimizes a radiance field, extracts the mesh from the radiance field,

and optimizes the mesh via differentiable surface rendering and score distillation. TextDeformer [Gao et al. 2023] leverages differentiable rendering and CLIP, but focuses on the problem of deforming explicit geometry rather than generating it from scratch.

Body-aware 3D synthesis. The design of 3D objects for human-object interaction is an important research topic. For everyday objects, it is important to consider human bodies, poses, and movements when generating 3D designs for humans [Chen et al. 2016; Saul et al. 2010]. To optimize for human interaction, various objective functions and evaluation metrics are defined [Wu et al. 2020]. Several previous works have explored this direction, e.g., in 3D room layout generation [Sun et al. 2024], scene synthesis [Sun et al. 2024; Vuong et al. 2024; Ye et al. 2022; Yi et al. 2023], as well as chairs and other body-supporting surfaces design [Blinn et al. 2021; Leimer et al. 2018, 2020; Zhao et al. 2021; Zheng et al. 2015]. A notable but challenging research direction is garment deformation [Kardash et al. 2022; Li et al. 2023; Sarafianos et al. 2024; Wang et al. 2018]. To deforming 3D objects, one can directly optimize on the 3D space [Jung et al. 2024; Liu et al. 2018; Sorkine et al. 2004], using triplanes [Frühstück et al. 2023] and text-to-mesh methods [Chen et al. 2019; Michel et al. 2022; Mohammad Khalid et al. 2022]. Foundation models can provide guidance or supervision signals for text and image-based stylization [Decatur et al. 2023] and manipulation [Gao et al. 2023] of 3D objects with various deformation methods [Baran et al. 2009; Gao et al. 2018; Groueix et al. 2019; Jacobson et al. 2011; Sumner and Popović 2004; Wang et al. 2015; Yifan et al. 2020; Zhang et al. 2008]. Related effort [Chen et al. 2023b; Richardson et al. 2023; Yeh et al. 2024; Zeng 2023] applied text-to-image generation models to create textures based on the mesh and given text or image.

4 METHOD

Our goal is to design rigid objects that satisfy diverse contact constraints for different body shapes and semantics. Figure 4 shows an overview of our method. It takes in multiple inputs, including a text prompt or image (e.g., generated by text-to-image models or existing images), a template object mesh, a body mesh, and a set of desired contact points. We represent the geometry of the input object using a mesh $\mathcal{M}$ with n vertices $\mathcal{V} \in \mathbb{R}^{n\times 3}$ and m faces $\mathcal{F} \in \{1,\dots,n\}^{m\times 3}$. We aim to optimize a displacement map $\Phi : \mathbb{R}^3 \to \mathbb{R}^3$ across the vertices.

Shape optimization through Jacobians. The design parameterization plays a significant role in the difficult design optimization problem. Naive optimization of the mesh deformation through vertex displacement can result in significant artifacts and is prone to convergence to local minima [Gao et al. 2023]. Inspired by Neural Jacobian Fields [Aigerman et al. 2022], we indirectly optimize the deformation map by optimizing a set of per-triangle Jacobian matrices $J_i \in \mathbb{R}^{3\times 3}$ for every face $f_i \in \mathcal{F}$. The deformation map Φ^* is computed as the mapping with Jacobian matrices that are closest to $\{J_i\}$, solved via the following Poisson optimization problem:

$$\Phi^* = \min_{\Phi} \sum_{f_i\in\mathcal{F}} |f_i| \|\nabla_i(\Phi) - J_i\|_2^2, \tag{1}$$

where $\nabla_i(\Phi)$ denotes the Jacobian of Φ at triangle f_i and $|f_i|$ is the area of that triangle.

4.1 Semantics-Aware Optimization

The user has the option to specify the semantic goals with a text prompt or an input image. Depending on the input modality, our system uses different losses to guide the optimization. We describe the losses for each modality below.

Input text guidance. For text guidance, the goal during the optimization process is to ensure that the resulting object aligns with the text prompt that specifies the desired design outcome. The pre-trained CLIP [Radford et al. 2021] provides a joint text-image feature space, which can be used for this alignment objective. We pass the current deformed mesh $\Phi^*(\mathcal{M})$ to a differentiable renderer $\mathcal{R}$ [Laine et al. 2020] to generate K images from different views:

$$I_k = \mathcal{R}(\Phi^*(\mathcal{M})), \quad k = 1,\dots,K. \tag{2}$$

The images are passed to CLIP to obtain the embeddings of the renders $\texttt{CLIP}(I_k) \in \mathbb{R}^{512}$. We pass the text prompt $\mathcal{P}$ to CLIP to get the language embedding with the same dimension, $\texttt{CLIP}(\mathcal{P}) \in \mathbb{R}^{512}$. Then, we define the text alignment objective to be the negative cosine similarity between the embeddings:

$$\mathcal{L}_s(\mathcal{M}) = \frac{1}{K}\sum_{k=1}^{K} -\text{sim}\left(\texttt{CLIP}(I_k), \texttt{CLIP}(\mathcal{P})\right). \tag{3}$$

Since CLIP operates on 2D images, multi-view consistency is a challenge. Averaging gradients across different views of the object often results in inconsistent artifacts such as incorrect geometry. We adopt the regularization term developed in [Gao et al. 2023], which tackles this problem by utilizing the patch-level deep features of CLIP's vision transformer (ViT). The intuition is that we can split the image into small patches, which are then projected into a higher-dimensional space. For each vertex and each render, we compute the pixel value in that render that contains the vertex. Then, by by associating the pixel value with the nearest corresponding patch center, we obtain a feature vector for that vertex in that render. In this way, we can encourage vertices to have similar deep features across renders from different viewpoints.

Input image guidance. If the user provides an input image $\bar{I}$, the goal of the optimization process is to optimize the shape of the object such that it matches the design in the input image. Inspired by Sarafianos et al. [2024], we use an image-to-3D model [AI 2024] to lift the image to a 3D guidance mesh denoted as $\overline{\mathcal{M}}$. Similar to text guidance, we render the guidance mesh from the K different views:

$$\bar{I}_k = \mathcal{R}(\overline{\mathcal{M}}), \quad k = 1,\dots,K, \tag{4}$$

and compute the cosine similarity of the CLIP embeddings of the guidance mesh renders and the current deformed mesh, averaged across the views:

$$\mathcal{L}_s(\Phi^*(\mathcal{M}), \overline{\mathcal{M}}) = \frac{1}{K}\sum_{k=1}^{K} -\text{sim}\left(\texttt{CLIP}(I_k), \texttt{CLIP}(\bar{I}_k)\right). \tag{5}$$

This loss acts as a soft constraint between the embeddings of the deformed mesh $\Phi^*(\mathcal{M})$ and those of the pseudo-ground truth $\overline{\mathcal{M}}$. For stronger 3D supervision, we use a two-sided Chamfer Distance (CD) loss to measure the distance between two sets of points, $p \in S$

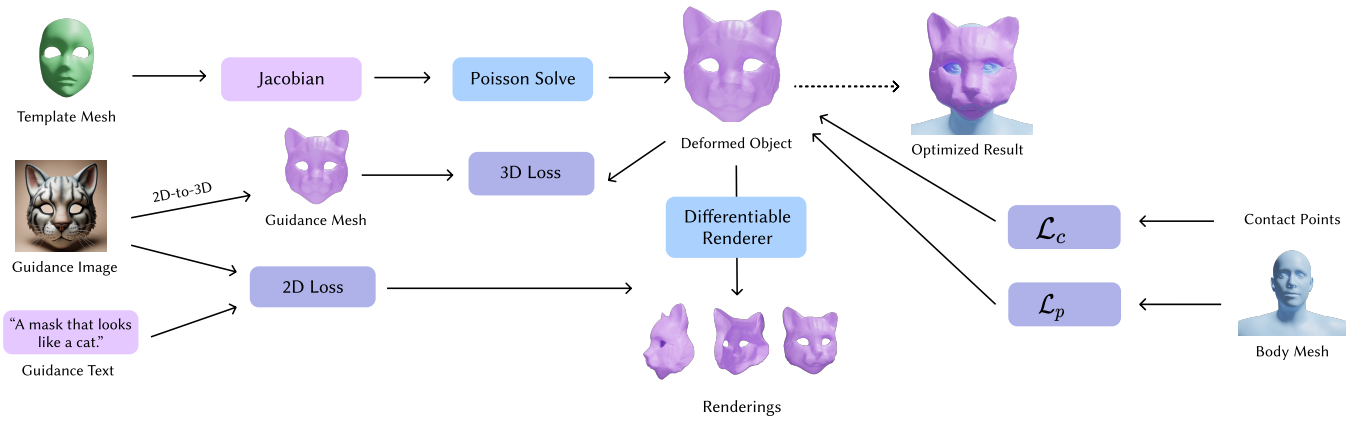

Fig. 4. Method overview. We synthesize body-aware 3D objects from a base mesh given an input body geometry and either text or image as guidance. We propose to use a mesh deformation procedure that optimizes for both semantic alignment as well as contact and penetration losses.

and $\bar{p} \in \bar{S}$, sampled from $\Phi^*(\mathcal{M})$ and $\overline{\mathcal{M}}$, respectively, in each optimization step:

$$\mathcal{L}_{\text{CD}} = \frac{1}{|S|}\sum_{p\in S}\min_{\bar{p}\in\bar{S}}\|p-\bar{p}\|_2^2 + \frac{1}{|\bar{S}|}\sum_{\bar{p}\in\bar{S}}\min_{p\in S}\|\bar{p}-p\|_2^2. \tag{6}$$

For 2D supervision, we use an L1 loss to ensure that the deformed mesh does not deviate too much from the image guidance along each step of the optimization:

$$\mathcal{L}_{\text{2D}} = \frac{1}{K}\sum_{k=1}^{K}|I_k - \bar{I}_k|. \tag{7}$$

4.2 Body-Aware Optimization

A key component of our optimization procedure is to produce objects that will satisfy contact constraints for different body shapes. Inspired by [Ye et al. 2022], given contact vertices $\mathcal{V}_c$, the contact loss $\mathcal{L}_c$ is defined as

$$\mathcal{L}_c(\mathcal{V},\mathcal{V}_c) = \lambda_c \frac{1}{|\mathcal{V}_c|}\sum_{v_c\in\mathcal{V}_c}\min_{v\in\mathcal{V}}||v_c - v||_2^2, \tag{8}$$

where $\lambda_{contact}$ is a tunable weight. This encourages the object to be in contact with the body vertices specified by the input contact vertices. To reduce penetration between the object and the body mesh $\mathcal{M}_b$, we include an additional loss $\mathcal{L}_p$:

$$\mathcal{L}_p(\mathcal{M},\mathcal{M}_b) = \sum_{d_i<D} {d_i}^2, \tag{9}$$

where d_i are signed distances between the object and the body mesh, and D is the penetration distance threshold. In total, the body-aware optimization loss is defined as:

$$\mathcal{L}_b(\mathcal{V},\mathcal{V}_c,\mathcal{M},\mathcal{M}_b) = \lambda_c\mathcal{L}_c(\mathcal{V},\mathcal{V}_c) + \lambda_p\mathcal{L}_p(\mathcal{M},\mathcal{M}_b). \tag{10}$$

In Figure 5, we show the effect of the contact loss $\mathcal{L}_c$ and the penetration loss $\mathcal{L}_p$ during the deformation procedure for the text prompt "a mask that looks like a cat". While semantic optimization severely penetrates the face, integrating contact and penetration losses improve the fit and reduce the penetration, respectively.

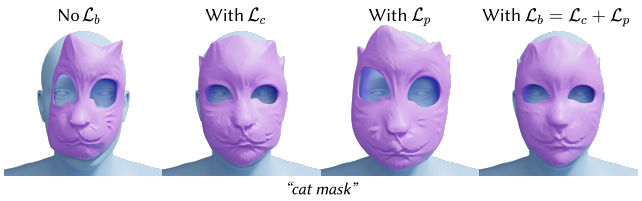

Fig. 5. We show the effect of contact vs. penetration losses on text guided deformation for "cat mask".

4.3 Optimization Problem Statement

In summary, the optimization objective is to optimize the Jacobian matrices J_i according the weighted sum of the semantics-aware and body-aware losses:

$$\mathcal{L}(\mathcal{M}) = \lambda_s\mathcal{L}_s(\Phi^*(\mathcal{M}),\overline{\mathcal{M}}) \tag{11}$$

$$+ \lambda_b\mathcal{L}_b(\mathcal{V},\mathcal{V}_c,\mathcal{M},\mathcal{M}_b) \tag{12}$$

$$+ \alpha\sum_{i=1}^{|\mathcal{F}|}\|J_i - \boldsymbol{I}\|_2. \tag{13}$$

The last term regularizes the predicted Jacobians, where $\boldsymbol{I}$ denotes the identity matrix, and α controls the strength of the deformations. We show the evolution of the mesh deformation process across optimization iterations in Figure 6. During the optimization process, the mesh becomes more semantically aligned with the input guidance, while fitting the body well.

5 EXPERIMENTS AND RESULTS

In our experiments, we seek to answer the following questions:

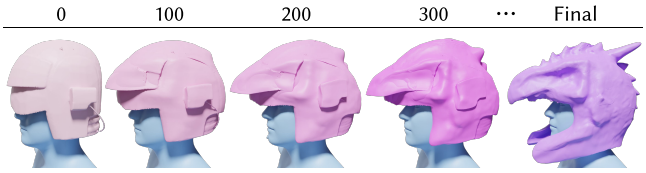

Fig. 6. Evolution of design throughout optimization iterations.

- Can ShapeCraft be used to generate object designs across different semantic targets and body shapes?
- What is the effect of the choice of guidance (text vs. image)?
- How does the body loss affect the design and fit of the object on the body?
- Is joint optimization better than two-stage optimization?
- How does our method compare with baseline methods in terms of semantic alignment and body fit?

5.1 Generality of ShapeCraft

Different object categories and design specifications. As shown in Figure 2, ShapeCraft is a general method that can generate semantically and body-aware everyday objects. Here we intentionally cover a variety of objects that need to be attached to different parts of the body: head (mask, glasses, helmet, visor), neck (necklace), wrist (bangle), finger (ring), and foot (slippers).

Next, we assess whether the same base mesh can deform into multiple target prompts in Figure 3. We find that indeed the same base mesh can be used for different text prompts within the same object category, given that the topology within an object category are often shared across different designs.

Different body shapes. We evaluate our system on different body shapes, ranging from realistic human adults and children to fantasy virtual characters such as dinosaurs and cartoon-looking cows. Figure 7 shows the same object and text prompt designed for different character body shapes. We observe that different body shapes affect the creativity of the optimization due to the amount of free space the object has to deform on the character without penetrating the body, but ShapeCraft is able to optimize for individual body shapes. Given the prompt of "bunny slipper", we show the optimization results for dog, dinosaur, and LEGO character in Figure 7's third row. We observe that the slipper's main feature bunny ears vary across the body shapes with noticeable differences in length and orientation of the ears. For example, the dog's slipper have much more pronounced bunny ears than the LEGO character's. This is due to the variance of character body shapes — the dog's thin leg provides more room for the slipper to grow. Even though the LEGO character's slipper has shorter bunny ears as its rigid leg prevents the ears from growing further, it still managed to be prompt-aligned.

5.2 Justification of Design Choices

Effect of text vs. image guidance. We evaluate the effect of text vs. image guidance in Figure 8. While text guidance occasionally deforms the base mesh into the desired semantics (e.g., "heart glasses", "cat mask"), on most examples it provides limited deformation (e.g., "star ring"), so the results look mostly like the base mesh. In contrast,

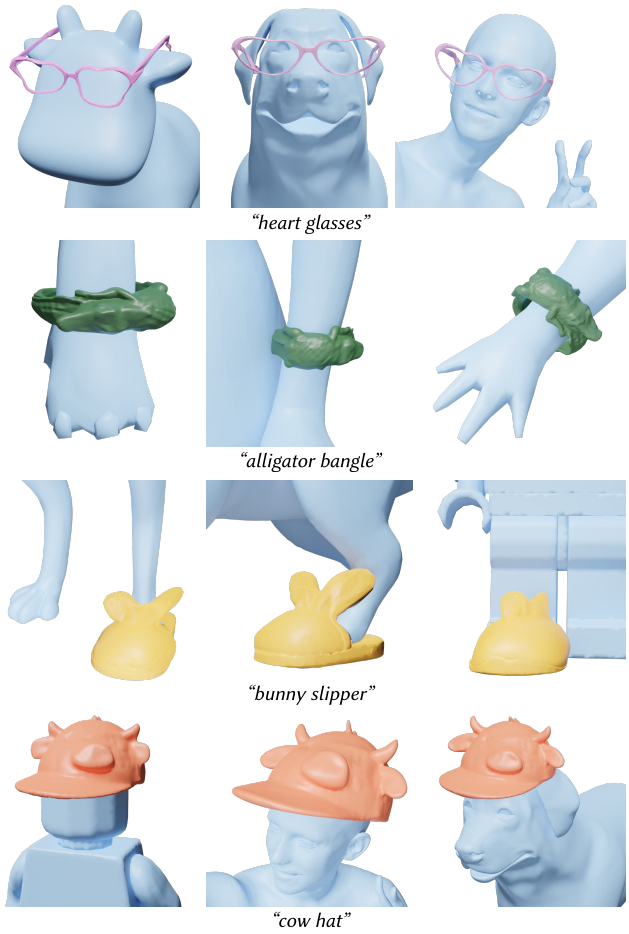

Fig. 7. Our method can customize the same object design for different character body shapes.

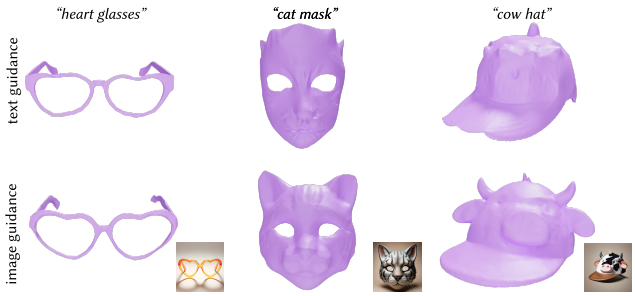

Fig. 8. We evaluate the effect of text vs. image guidance. Image guidance produces stronger control, generating objects that are more prompt-aligned. We show the reference image in the bottom right corner of example of the image guidance row.

image guidance provides a much stronger signal for deformation. This is likely because the CLIP model alone is not as strong as 2D and 3D guidance provided by text-to-image and image-to-3D models.

Fig. 9. With contact and penetration losses, the text-guided deformations are more body fitting.

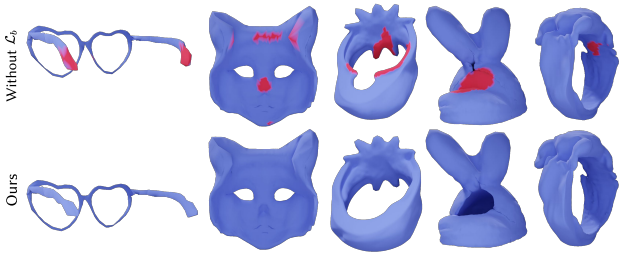

Fig. 10. We visualize penetration maps on objects optimized with (second row) and without body losses (first row). In the penetration maps, the blue regions indicate a positive distance between the mesh and the characters, signifying no penetration. Red regions indicate a negative distance, denoting penetration between the mesh and the interacting character. Without the incorporation of body losses, the generated objects exhibit significant penetration with the character.

Effect of body loss. In Figure 9, we analyze the effect of body losses in the text guidance setting. Even though the base mesh initially fits well on the body, text guidance tends to ignore the body when optimizing for alignment with the prompt, resulting in objects that are not body-aware. When including the body loss, the objects are well fitted on the human.

In Figure 10, we compare objects that are optimized using image guidance, with and without the body loss. We find that including the body loss in the image guidance optimization helps minimize penetrations between the object and the human, making the objects more fit on the body.

Comparison with two-stage optimization. In Figure 11, we analyze alternatives for image-guided 3D generation: (i) the guidance mesh (generated via text-to-image and image-to-3D models), (ii) the guidance mesh with a second body-aware refinement stage. Although making the guidance mesh body-aware reduces penetrations, it cannot make the thin parts of the guidance mesh wider. In contrast, jointly optimizing for both body and semantics from a template mesh mesh results in a fitting bangle that is aligned with the prompt.

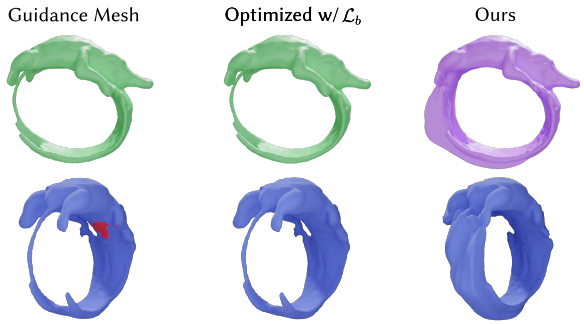

Fig. 11. We show the object mesh (first row) and the penetration map from a different viewpoint (second row). Even though we can apply a body refinement optimization on the guidance mesh to reduce the penetrations, it cannot the fix the thin structure on the object. Jointing optimizing for both body and semantics together results in a more well-formed mesh while also minimizing penetrations.

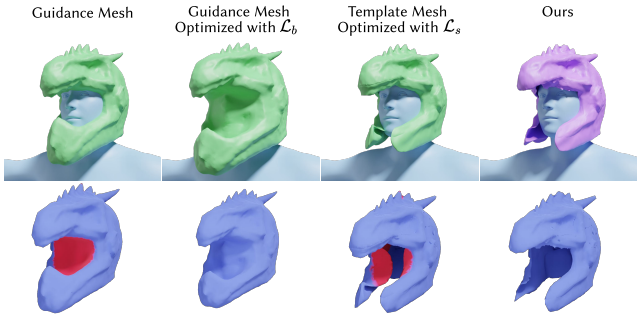

Fig. 12. We show the object mesh (first row) and the penetration map (second row). When the guidance mesh lacks the correct topology, such as missing a hole for the head in the helmet, body refinement cannot fix the issue and ends up enclosing the entire head. Starting from the template mesh and optimizing for semantics alone is also not sufficient; while the topology is correct, the optimizing may introduce penetrations.

In Figure 12, we analyze another example with a "dinosaur helmet". Because the topology of the guidance mesh is incorrect (missing holes for the head), applying a body-aware refinement step optimizes penetrations by enclosing the entire head. Starting from the base mesh and optimizing for semantics alone is also not sufficient; while the topology is correct, the optimization may introduce penetrations. Our method is able to properly generate a helmet that is prompt-aligned while maintaining a hole for the human's head.

5.3 Comparison with Baselines

To provide a quantitative evaluation of ShapeCraft against baselines, we compute prompt alignment, contact distance, and penetration distance on all the different method variants. As shown in Table 1, we find that incorporating body loss significantly reduces penetration, regardless of the Base Mesh. Furthermore, the chamfer distance of the contact points are also well-maintained by using body loss. One interesting observation here is that Guidance mesh without any further edits has the highest CLIP score, which intuitively makes

Table 1. We report quantitative metrics on the prompt alignment, penetration, and chamfer distance of the contact points to the object vertices.

Base Mesh	Input Modality	$\mathcal{L}_b$	CLIP ↑	D_p ↓	D_c ↓
Template	N/A	N/A	0.25	6.7e-2	1.9e-10
Guidance	N/A	N/A	0.28	65.6	2.1e-10
Guidance	N/A	Y	0.27	8.5e-4	5.1e-3
Template	Image	N	0.27	3.7	7.6e-3
Template	Image (Ours)	Y	0.27	7.0e-4	4.6e-3

Table 2. We report results on the user study. We ask participants to rate the prompt alignment, aesthetics, and the perceived comfort of generated objects on the body.

	Align. ↑	Aesth. ↑	Comf. ↑
Template mesh	1.47	5.25	**6.74**
TextDeformer [Gao et al. 2023]	3.19	2.94	2.88
Body-Aware TextDeformer	3.04	3.96	5.17
Ours	**7.78**	**6.40**	5.75

sense since CLIP is better at recognizing common objects in the world, while an "alligator bangle" design has a vast design space.

We also conduct a user study (N=9) asking participants to rate (on a Likert scale of 1-10) the prompt alignment ("how aligned/similar are each of objects to the original prompt?"), aesthetics ("how aesthetic are the following objects?"), and the perceived comfort of the object on the human body ("how comfortable do the following objects look on the human?"). Our method performs the best on prompt alignment and aesthetics, and achieves comparable performance on comfort with the template mesh. We see that the template mesh achieves the lowest score on prompt alignment, which is expected, because the template mesh is only representative of the object category, and not adapted to the creative prompt. We see that TextDeformer itself has the lowest score in comfort which is expected, as the objects were optimized without considering fit to a body. Although Body-Aware TextDeformer achieved a higher score in comfort compared to TextDeformer, its alignment score decreased in comparison. This indicates that the deformation prioritized body-awareness in a way that conflicted with the object's prompt alignment. In contrast, our method remains the most prompt aligned and aesthetic across all methods while also maintaining comfort, showing our method successfully prioritizes both body-awareness and semantic-awareness.

5.4 Applications of ShapeCraft

Fabricated designs. The designs generated by our system are fabricable in the real world. In Figure 13 left, we show the photo of 3D-printed objects — they are directly 3D-printed using ShapeCraft-generated meshes without manual modifications. In Figure 13 right we also show how objects fit on a real human body and characters.

Sketch-guided design. We show a sketch application with ShapeCraft. We ask a user to draw a sketch of an object, and we use ControlNet [Zhang et al. 2023] to convert the sketch into a 2D image. The image is used as an input to our image-guided mesh deformation method. The results are shown in Figure 14.

Fig. 13. We fabricated the objects in the real world with 3D printing. The objects can be worn as accessories on real people and characters.

Fig. 14. We show an application of lifting a sketch into a body-fitting 3D object design.

6 CONCLUSION

In this work, we present ShapeCraft, a 3D object design framework that integrates body and semantic awareness into the generative process. Our method synthesizes body-aware 3D objects from a base mesh using input body geometry and guidance from text or images. The joint optimization for semantic alignment and body-aware losses ensures that the generated objects are both creatively customized and functionally practical. Our evaluations demonstrate the efficacy of ShapeCraft in producing virtual and real-world objects that fit a wide range of body shapes without the need for manual intervention. ShapeCraft not only streamlines the design process but also enables the fabrication of personalized, body-aware objects, thereby enhancing customization and usability in everyday object design.

7 CONTRIBUTIONS AND ACKNOWLEDGEMENTS

H.C., M.G., M.T., and R.Z. helped in the conceptualization of the project. H.C., M.G., and M.T. designed and trained the ShapeCraft, as well as designed and implemented the evaluation framework for ShapeCraft. H.C., M.G., M.T., and R.Z. wrote the paper. K.L., J.W., and R.Z. managed and advised on the project. For this project, Stanford Vision and Learning Lab (SVL)'s GPUs and job scheduling were used. Additionally, this paper was submitted to SIGGRAPH Asia 2024. TextDeformer's code was utilized to perform baseline comparisons. Their source code can be found here: https://github.com/threedle/TextDeformer.

REFERENCES

Tripo AI. 2024. Tripo AI. https://www.tripo3d.ai/.

Noam Aigerman, Kunal Gupta, Vladimir G. Kim, Siddhartha Chaudhuri, Jun Saito, and Thibault Groueix. 2022. Neural jacobian fields: learning intrinsic mappings of arbitrary meshes. *ACM Trans. Graph.* 41, 4, Article 109 (jul 2022), 17 pages. https://doi.org/10.1145/3528223.3530141

Ilya Baran, Daniel Vlasic, Eitan Grinspun, and Jovan Popović. 2009. Semantic deformation transfer. In *ACM SIGGRAPH 2009 papers*. 1–6.

Bryce Blinn, Alexander Ding, R Kenny Jones, Manolis Savva, Srinath Sridhar, and Daniel Ritchie. 2021. Learning Body-Aware 3D Shape Generative Models. *arXiv preprint arXiv:2112.07022* (2021).

Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. 2023b. Text2tex: Text-driven texture synthesis via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 18558–18568.

Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. 2019. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*. Springer, 100–116.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 2023a. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22246–22256.

Xiang'Anthony' Chen, Jeeeun Kim, Jennifer Mankoff, Tovi Grossman, Stelian Coros, and Scott E Hudson. 2016. Reprise: A design tool for specifying, generating, and customizing 3D printable adaptations on everyday objects. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 29–39.

Michael Chui, Eric Hazan, Roger Roberts, Alex Singla, and Kate Smaje. 2023. The economic potential of generative AI. (2023).

Dale Decatur, Itai Lang, Kfir Aberman, and Rana Hanocka. 2023. 3D Paintbrush: Local Stylization of 3D Shapes with Cascaded Score Distillation. *arXiv preprint arXiv:2311.09571* (2023).

Ziv Epstein, Aaron Hertzmann, Laura Herman, Robert Mahari, Morgan R Frank, Matthew Groh, Hope Schroeder, Amy Smith, Memo Akten, Jessica Fjeld, et al. 2023. Art and the science of generative AI: A deeper dive. *arXiv preprint arXiv:2306.04141* (2023).

Anna Frühstück, Nikolaos Sarafianos, Yuanlu Xu, Peter Wonka, and Tony Tung. 2023. VIVE3D: Viewpoint-Independent Video Editing using 3D-Aware GANs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. 2022. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances In Neural Information Processing Systems* 35 (2022), 31841–31854.

Lin Gao, Jie Yang, Yi-Ling Qiao, Yu-Kun Lai, Paul L Rosin, Weiwei Xu, and Shihong Xia. 2018. Automatic unpaired shape deformation transfer. *ACM Transactions on Graphics (ToG)* 37, 6 (2018), 1–15.

William Gao, Noam Aigerman, Thibault Groueix, Vova Kim, and Rana Hanocka. 2023. Textdeformer: Geometry manipulation using text guidance. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. 2019. Unsupervised cycle-consistent deformation for shape matching. In *Computer Graphics Forum*, Vol. 38. Wiley Online Library, 123–133.

Alec Jacobson, Ilya Baran, Jovan Popovic, and Olga Sorkine. 2011. Bounded biharmonic weights for real-time deformation. *ACM Trans. Graph.* 30, 4 (2011), 78.

Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. 2022. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 867–876.

Heewoo Jun and Alex Nichol. 2023. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463* (2023).

Hyunyoung Jung, Seonghyeon Nam, Nikolaos Sarafianos, Sungjoo Yoo, Alexander Sorkine-Hornung, and Rakesh Ranjan. 2024. Geometry Transfer for Stylizing Radiance Fields. In *CVPR*.

Kateryna Kardash, Christos Koutras, and Miguel A Otaduy. 2022. Design of personalized scoliosis braces based on differentiable biomechanics—Synthetic study. *Frontiers in Bioengineering and Biotechnology* 10 (2022), 1014365.

Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Popa Tiberiu. 2022. CLIP-Mesh: Generating textured meshes from text using pretrained image-text models. *SIGGRAPH Asia 2022 Conference Papers* (December 2022).

Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. 2020. Modular Primitives for High-Performance Differentiable Rendering. *ACM Transactions on Graphics* 39, 6 (2020).

Kurt Leimer, Michael Birsak, Florian Rist, and Przemyslaw Musialski. 2018. Sit & Relax: Interactive Design of Body-Supporting Surfaces. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, 349–359.

Kurt Leimer, Andreas Winkler, Stefan Ohrhallinger, and Przemyslaw Musialski. 2020. Pose to Seat: Automated design of body-supporting surfaces. *Computer Aided Geometric Design* 79 (2020), 101855.

Yifei Li, Hsiao-yu Chen, Egor Larionov, Nikolaos Sarafianos, Wojciech Matusik, and Tuur Stuyck. 2023. DiffAvatar: Simulation-Ready Garment Optimization with Differentiable Simulation. *arXiv preprint arXiv:2311.12194* (2023).

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2023. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 300–309.

Hsueh-Ti Derek Liu, Michael Tao, and Alec Jacobson. 2018. Paparazzi: surface editing by way of multi-view image processing. *ACM Trans. Graph.* 37, 6 (2018), 221–1.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023b. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9298–9309.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. 2023a. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. In *The Twelfth International Conference on Learning Representations*.

Zhen Liu, Yao Feng, Michael J Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. 2022. MeshDiffusion: Score-based Generative 3D Mesh Modeling. In *The Eleventh International Conference on Learning Representations*.

Liane Makatura, Michael Foshey, Bohan Wang, Felix HähnLein, Pingchuan Ma, Bolei Deng, Megan Tjandrasuwita, Andrew Spielberg, Crystal Elaine Owens, Peter Yichen Chen, et al. 2023. How Can Large Language Models Help Humans in Design and Manufacturing? *arXiv preprint arXiv:2307.14377* (2023).

Mariem Mezghanni, Théo Bodrito, Malika Boulkenafed, and Maks Ovsjanikov. 2022. Physical simulation layer for accurate 3d modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13514–13523.

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

Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. 2022. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 conference papers*. 1–8.

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. 2022. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751* (2022).

OpenAI. 2023. DALL-E 3. https://openai.com/dall-e-3.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. DreamFusion: Text-to-3D using 2D Diffusion. In *The Eleventh International Conference on Learning Representations*.

Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. 2023. Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors. In *The Twelfth International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. 2023. Texture: Text-guided texturing of 3d shapes. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

Nikolaos Sarafianos, Tuur Stuyck, Xiaoyu Xiang, Yilei Li, Jovan Popovic, and Rakesh Ranjan. 2024. Garment3DGen: 3D Garment Stylization and Texture Generation. *arXiv preprint arXiv:2403.18816* (2024).

Greg Saul, Manfred Lau, Jun Mitani, and Takeo Igarashi. 2010. SketchChair: an all-in-one chair design system for end users. In *Proceedings of the fifth international conference on Tangible, embedded, and embodied interaction*. 73–80.

Olga Sorkine, Daniel Cohen-Or, Yaron Lipman, Marc Alexa, Christian Rössl, and H-P Seidel. 2004. Laplacian surface editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*. 175–184.

Robert W Sumner and Jovan Popović. 2004. Deformation transfer for triangle meshes. *ACM Transactions on graphics (TOG)* 23, 3 (2004), 399–405.

Jia-Mu Sun, Jie Yang, Kaichun Mo, Yu-Kun Lai, Leonidas Guibas, and Lin Gao. 2024. Haisor: Human-aware indoor scene optimization via deep reinforcement learning. *ACM Transactions on Graphics* (2024).

Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. 2023. TextMesh: Generation of Realistic 3D Meshes From Text Prompts. *arXiv preprint arXiv:2304.12439* (2023).

An Dinh Vuong, Minh Nhat Vu, Toan Nguyen, Baoru Huang, Dzung Nguyen, Thieu Vo, and Anh Nguyen. 2024. Language-driven Scene Synthesis using Multi-conditional Diffusion Model. *Advances in Neural Information Processing Systems* 36 (2024).

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. 2022. Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation. *arXiv preprint arXiv:2212.00774* (2022).

Tuanfeng Y Wang, Duygu Ceylan, Jovan Popović, and Niloy J Mitra. 2018. Learning a shared shape space for multimodal garment design. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–13.

Yu Wang, Alec Jacobson, Jernej Barbič, and Ladislav Kavan. 2015. Linear subspace design for real-time shape deformation. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 1–11.

Hongtao Wu, Deven Misra, and Gregory S Chirikjian. 2020. Is that a chair? imagining affordances using simulations of an articulated human body. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 7240–7246.

Sifan Ye, Yixing Wang, Jiaman Li, Dennis Park, C Karen Liu, Huazhe Xu, and Jiajun Wu. 2022. Scene synthesis from human motion. In *SIGGRAPH Asia 2022 Conference Papers*. 1–9.

Yu-Ying Yeh, Jia-Bin Huang, Changil Kim, Lei Xiao, Thu Nguyen-Phuoc, Numair Khan, Cheng Zhang, Manmohan Chandraker, Carl S Marshall, Zhao Dong, et al. 2024. TextureDreamer: Image-guided Texture Synthesis through Geometry-aware Diffusion. *arXiv preprint arXiv:2401.09416* (2024).

Hongwei Yi, Chun-Hao P Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J Black. 2023. MIME: Human-Aware 3D Scene Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12965–12976.

Wang Yifan, Noam Aigerman, Vladimir G Kim, Siddhartha Chaudhuri, and Olga Sorkine-Hornung. 2020. Neural cages for detail-preserving 3d deformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 75–83.

Xianfang Zeng. 2023. Paint3D: Paint Anything 3D with Lighting-Less Texture Diffusion Models. *arXiv preprint arXiv:2312.13913* (2023).

Hao Zhang, Alla Sheffer, Daniel Cohen-Or, Quan Zhou, Oliver Van Kaick, and Andrea Tagliasacchi. 2008. Deformation-driven shape correspondence. In *Computer Graphics Forum*, Vol. 27. Wiley Online Library, 1431–1439.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

Danyong Zhao, Yijing Li, Siddhartha Chaudhuri, Timothy Langlois, and Jernej Barbič. 2021. ERGOBOSS: onomic ptimization of dy-upporting urfaces. *IEEE Transactions on Visualization and Computer Graphics* 28, 12 (2021), 4032–4047.

Youyi Zheng, Han Liu, Julie Dorsey, and Niloy J Mitra. 2015. Ergonomics-inspired reshaping and exploration of collections of models. *IEEE Transactions on Visualization and Computer Graphics* 22, 6 (2015), 1732–1744.

Junzhe Zhu, Peiye Zhuang, and Sanmi Koyejo. 2023. Hifa: High-fidelity text-to-3d generation with advanced diffusion guidance. In *The Twelfth International Conference on Learning Representations*.