

Vox-E: Text-guided Voxel Editing of 3D Objects

Etai Sella¹ Gal Fiebelman¹ Peter Hedman² Hadar Averbuch-Elor¹

¹Tel Aviv University ²Google Research



Figure 1. Given multiview images of an object (left), our technique generates volumetric edits from target text prompts, allowing for significant geometric and appearance changes, while faithfully preserving the input object. The objects can be edited either *locally* (center) or *globally* (right), depending on the nature of the user-provided text prompt.

Abstract

Large scale text-guided diffusion models have garnered significant attention due to their ability to synthesize diverse images that convey complex visual concepts. This generative power has more recently been leveraged to perform text-to-3D synthesis. In this work, we present a technique that harnesses the power of latent diffusion models for editing existing 3D objects. Our method takes oriented 2D images of a 3D object as input and learns a grid-based volumetric representation of it. To guide the volumetric representation to conform to a target text prompt, we follow unconditional text-to-3D methods and optimize a Score Distillation Sampling (SDS) loss. However, we observe that combining this diffusion-guided loss with an image-based regularization loss that encourages the representation not to deviate too strongly from the input object is challenging, as it requires achieving two conflicting goals while viewing only structure-and-appearance coupled 2D projections. Thus, we introduce a novel volumetric regularization loss that operates directly in 3D space, utilizing the explicit

nature of our 3D representation to enforce correlation between the global structure of the original and edited object. Furthermore, we present a technique that optimizes cross-attention volumetric grids to refine the spatial extent of the edits. Extensive experiments and comparisons demonstrate the effectiveness of our approach in creating a myriad of edits which cannot be achieved by prior works¹.

1. Introduction

Creating and editing 3D models is a cumbersome task. While template models are readily available from online databases, tailoring one to a specific artistic vision often requires extensive knowledge of specialized 3D editing software. In recent years, neural field-based representations (e.g., NeRF [29]) demonstrated expressive power in faithfully capturing fine details, while offering effective optimization schemes through differentiable rendering. Their

¹Our code can be reached through our project page at <http://vox-e.github.io/>

applicability has recently expanded also for a variety of editing tasks. However, research in this area has mostly focused on either appearance-only manipulations, which change the object’s texture [46, 48] and style [50, 44], or geometric editing via correspondences with an explicit mesh representation [13, 49, 47]—linking these representations to the rich literature on mesh deformations [19, 40]. Unfortunately, these methods still require placing user-defined control points on the explicit mesh representation, and cannot allow for adding new structures or significantly adjusting the geometry of the object.

In this work, we are interested in enabling more flexible and localized object edits, guided only by textual prompts, which can be expressed through *both* appearance and geometry modifications. To do so, we leverage the incredible competence of pretrained 2D diffusion models in editing images to conform with target textual descriptions. We carefully apply a *score distillation* loss, as recently proposed in the unconditional text-driven 3D generation setting [33]. Our key idea is to regularize the optimization in 3D space. We achieve this by coupling two volumetric fields, providing the system with more freedom to comply with the text guidance, on the one hand, while preserving the input structure, on the other hand.

Rather than using neural fields, we base our method on *lighter* voxel-based representations which learn scene features over a sparse voxel grid. This explicit grid structure not only allows for faster reconstruction and rendering times, but also for achieving a tight *volumetric* coupling between volumetric fields representing the 3D object before and after applying the desired edit using a novel *volumetric correlation loss* over the density features. To further refine the spatial extent of the edits, we utilize 2D cross-attention maps which roughly capture regions associated with the target edit, and lift them to volumetric grids. This approach is built on the premise that, while independent 2D internal features of generative models can be noisy, unifying them into a single 3D representation allows for better distilling the semantic knowledge. We then use these 3D cross-attention grids as a signal for a binary volumetric segmentation algorithm that splits the reconstructed volume into edited and non-edited regions, allowing for merging the features of the volumetric grids to better preserve regions that should not be affected by the textual edit.

Our approach, coined *Vox-E*, provides an intuitive voxel editing interface, where the user only provides a simple target text prompt (see Figure 1). We compare our method to existing 3D object editing techniques, and demonstrate that our approach can facilitate local and global edits involving appearance and geometry changes over a variety of objects and text prompts, which are extremely challenging for current methods.

Explicitly stated, our contributions are:

- A coupled volumetric representation tied using 3D regularization, allowing for editing 3D objects using diffusion models as guidance while preserving the appearance and geometry of the input object.
- A 3D cross-attention based volumetric segmentation technique that defines the spatial extent of textual edits.
- Results that demonstrate that our proposed framework can perform a wide array of editing tasks, which cannot be previously achieved.

2. Related Work

Text-driven Object Editing. Computational methods targeting text-driven image generation and manipulation have seen tremendous progress with the emergence of CLIP [34] and diffusion models [17], advancing from specific domains [1, 39, 32, 12] to more generic ones [30, 4, 11, 22]. Several recent methods allow for performing convincing localized edits on real images without requiring mask guidance [5, 15, 8, 42, 31]. However, these methods all operate on single images and cannot facilitate a consistent editing of 3D objects.

While less common, methods for manipulating 3D objects are also gaining increasing interests. Methods such as LADIS [18] and Changelt3D [2] aim at learning the relations between 3D shape parts and text directly using datasets composed of edit descriptions and shape pairs. These works allow for geometric edits but fail to generalize to out-of-distribution shapes and cannot modify appearance.

Alternatively, several methods have proposed leveraging 2D image projections, matching these to a driving text. Text2Mesh [28] uses CLIP for stylizing 3D meshes based on textual prompts. Tango [10] also styles meshes with CLIP, enabling additionally stylization of lighting conditions, reflectance properties and local geometric variations. TEXTure [35] use a depth-to-image diffusion model for texturing 3D meshes. Unlike our work, these methods focus mostly on texturing meshes, and cannot be used for generating significant geometric modifications, such as adding glasses or other types of accessories.

Neural Field Editing. Neural fields (e.g., NeRF [29]), which can be effectively learned from multi-view images through differentiable rendering, have recently shown great promise for representing object and scenes. Prior works have demonstrated that these fields can be adapted to express different forms of manipulations. ARF [50] transfers the style of an exemplar image to a NeRF. NeRF-Art [44] performs a text-driven style transfer. Distilled Feature Fields [23] distill the knowledge of 2D image feature extractors into a 3D feature field and use this feature field to localize edits performed by CLIP-NeRF [43], which optimizes a radiance field so that its rendered images match

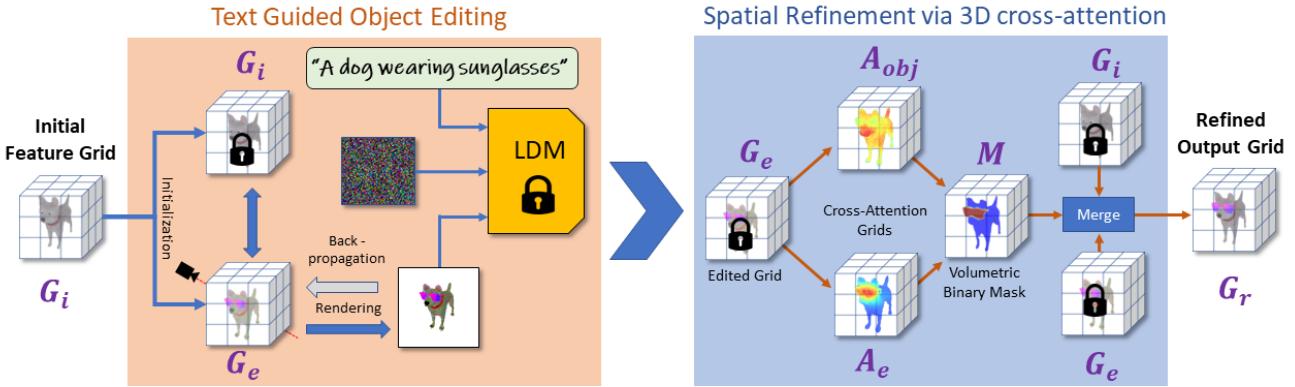


Figure 2. **An overview of our approach.** Given a set of posed images depicting an object, we optimize an initial feature grid (left). We then perform text-guided object editing using a generative SDS loss and a volumetric regularization, optimizing an edited grid G_e . To localize the edits, we optimize 3D cross-attention grids which define probability distributions over the object and the edit regions. We obtain a volumetric mask from these grids using an energy minimization problem over all the voxels. Finally, we merge the initial and edited grid to obtain a refined volumetric grid (right).

with a text prompt via CLIP.

Several works have shown that neural fields can be edited by editing selected 2D images [25, 48]. NeuTex [46] uses 2D texture maps, which can be edited directly, to represent the surface appearance. Other works demonstrated geometric editing of shapes represented with neural fields via correspondences with an explicit mesh representation [13, 49, 47], that can be edited using as-rigid-as-possible deformations [40]. However, these cannot easily allow for modifying the 3D mesh to incorporate additional parts, according to the user’s provided description. Concurrently to our work, Instruct-NeRF2NeRF [14] uses an image editing model to iteratively edit multi-pose images from which an edited 3D scene is reconstructed. Unlike our work which optimizes the underlying 3D representation, they optimize the input images directly. Furthermore, our method is based on grid-based representations rather than neural fields, in particular ReLU Fields [21], which do not require any neural networks and instead model the scene as a voxel grid where each voxel contains learned features. We show that having an explicit grid structure is beneficial for editing 3D objects as it enables fast reconstruction and rendering times as well as powerful volumetric regularization.

Text-to-3D. Following the great success of text-to-image generation, we are witnessing increasing interests in unconditional text-driven generation of 3D objects and scenes. CLIP-Forge [37] uses CLIP guidance to generate coarse object shapes from text. Dream Fields [20], DreamFusion [33], Score Jacobian Chaining [45] and Latent-NeRF [27] optimize radiance fields to generate the geometry and color of objects driven by the text. While DreamFields relies on CLIP, the other three methods instead use a score distillation loss, which enables the use of a pretrained

2D diffusion model. Magic3D [24] proposes a two-stage optimization technique to overcome DreamFusion’s slow optimization. Unlike these works, we focus on the conditional setting. In our case, a 3D object is provided, and the desired edit should preserve the object’s geometry and appearance. Still, we compare with Latent-NeRF in the experiments, as it can use rough 3D shapes as guidance.

3. Method

In this work, we consider the problem of editing 3D objects given a captured set of posed multiview images describing this object and a text prompt expressing the desired edit. We first represent the input object with a grid-based volumetric representation (Section 3.1). We then optimize a *coupled* voxel grid, such that it resembles the input grid on the one hand while conforming to the target text on the other hand (Section 3.2). To further refine the spatial extent of the edits, we perform an (optional) refinement step (Section 3.3). Figure 2 provides an overview of our approach.

3.1. Grid-Based Volumetric Representation

Our volumetric representation is based on the voxel grid model first introduced in DVGO [41] and later simplified in ReLU Fields [21]. We use a 3D grid G , where each voxel holds a 4D feature vector. We model the object’s geometry using a single feature channel which represents spatial density values when passed through a ReLU nonlinearity. The three additional feature channels represent the object’s appearance, and are mapped to RGB colors when passed through a sigmoid function. Note that in contrast to most recent neural 3D scene representations (including ReLU Fields), we do not model view dependent appearance effects, as we found it leads to undesirable artifacts

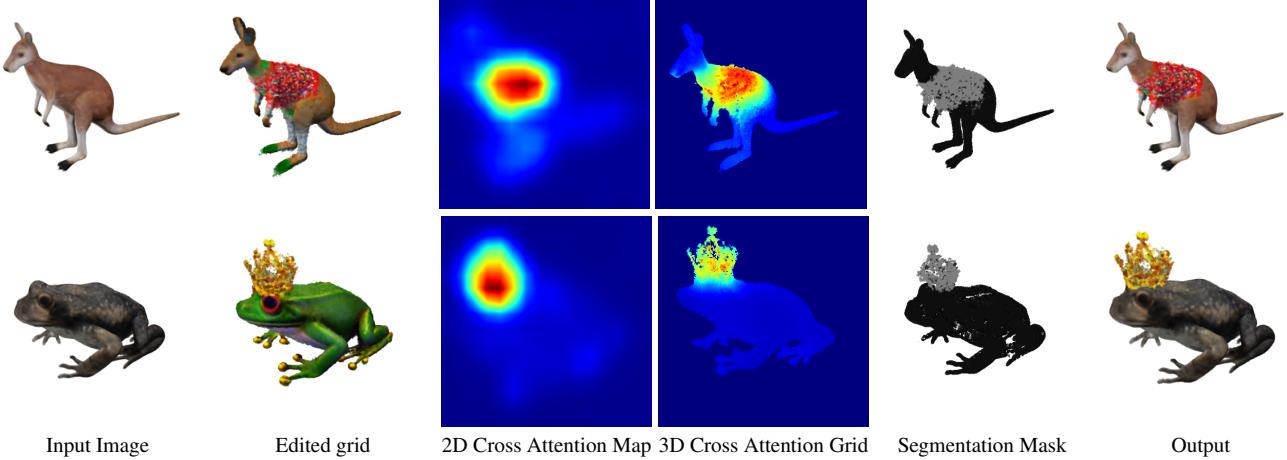


Figure 3. Optimizing 3D cross-attention grids for edit localization. We leverage rough 2D cross-attention maps (third column) for supervising the training of 3D cross-attention grids (fourth column). Provided with cross-attention grids associated with the edit (as demonstrated above for “christmas sweater” and “crown”) and object regions, we formulate an energy minimization problem, which outputs a volumetric binary segmentation mask (fifth column). We then merge the features of the input (first column) and edited (second column) grids using this volumetric mask to obtain our final output (rightmost column). Note that warmer colors correspond to higher activations in the cross-attention maps and edited regions are colored in gray in the binary segmentation mask.

when guided with 2D diffusion-based models.

To represent the input object with our grid-based representation, we use images and associated camera poses to perform volumetric rendering as described in NeRF [29]. However, in contrast to NeRF, we do not use any positional encoding and instead sample our grid at each location query to obtain interpolated density and color values, which are then accumulated along each ray. We use a simple L1 loss between our rendered outputs and the input images to learn a grid-based volume G_i that represents the input object.

3.2. Text-guided Object Editing

Equipped with the initial voxel grid G_i described in the previous section, we perform text-guided object editing by optimizing G_e , a grid representing the edited object which is initialized from G_i . Our optimization scheme combines a generative component, guided by the target text prompt, and a pullback term that encourages the new grid not to deviate too strongly from its initial values. As we later show, our coupled volumetric representation provides added flexibility to our system, allowing for better balancing between the two objectives by regularizing directly in 3D space. Next we describe these two optimization objectives.

Generative Text-guided Objective

To encourage our feature grid to respect the desired edit provided via a textual prompt, we use a Score Distillation Sampling (SDS) loss applied over Latent Diffusion Models (LDMs). SDS was first introduced in DreamFusion [33], and consists of minimizing the difference between noise injected to a generator’s output and noise predicted by a pre-

trained Denoising Diffusion Probabilistic Model (DDPM). Formally, at each optimization iteration, noise is added to a generated image x using a random time-step t ,

$$x_t = x + \epsilon_t, \quad (1)$$

where ϵ_t is the output of a noising function $Q(t)$ at time-step t . The score distillation gradients (computed per pixel) can be expressed as:

$$\nabla_x \mathcal{L}_{SDS} = w(t) (\epsilon_t - \epsilon_\phi(x_t, t, s)), \quad (2)$$

where $w(t)$ is a weighting function, s is an input guidance text, and $\epsilon_\phi(z_t, t, s)$ is the noise predicted by a pre-trained DDPM with weights ϕ given x_t , t and s . As suggested by Lin et al. [24], we use an annealed SDS loss which gradually decreases the maximal time-step we draw t from, allowing SDS to focus on high frequency information after the outline of the edit has formed. We empirically found that this often leads to higher quality outputs.

Volumetric Regularization

Regularization is key in our problem setting, as we want to avoid over-fitting to specific views and also not to deviate too far from the original 3D representation. Therefore, we propose a volumetric regularization term, which couples our edited grid G_e with the initial grid G_i . Specifically, we incorporate a loss term which encourages correlation between the density features of the input grid f_i^σ and the density features of the edited grid f_e^σ :

$$\mathcal{L}_{reg3D} = 1 - \frac{Cov(f_i^\sigma, f_e^\sigma)}{\sqrt{Var(f_i^\sigma)Var(f_e^\sigma)}} \quad (3)$$

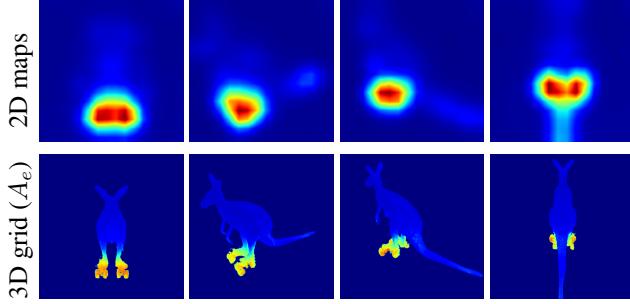


Figure 4. **Cross-attention 2D maps and rendered 3D grids over multiple viewpoints**, obtained for the token associated with the word "rollerskates" (from the "kangaroo on rollerskates" text prompt). While 2D cross-attention may yield inconsistent observations, such as high probabilities over the tail region in the right-most column, our 3D grids can more accurately localize the region of interest (effectively smoothing out such inconsistencies).

This volumetric loss has a significant edge over image space losses as it allows for decoupling the appearance of the scene from its structure, thereby connecting the volumetric representations in 3D space rather than treating it as a multiview optimization problem.

3.3. Spatial Refinement via 3D Cross-Attention

While our optimization framework described in the previous section can mostly preserve the shape and the identity of a 3D object, for local edits, it is usually desirable to only change specific *local* regions, while keeping other regions completely fixed. Therefore, we add an (optional) refinement step which leverages the signal from cross-attention layers to produce a volumetric binary mask M that marks the voxels which should be edited. We then obtain the refined grid G_r by merging the input grid G_i and edited G_e grid as

$$G_r = M \cdot G_e + (1 - M) \cdot G_i. \quad (4)$$

In the context of 2D image editing with diffusion models, the outputs of the cross-attention layers roughly capture the spatial regions associated with each word (or token) in the text. More concretely, these cross attention maps can be interpreted as probability distributions over tokens for each image patch [15, 8]. We elevate these 2D probability maps to a 3D grid by using them as supervision for training a ReLU field. We initialize the density values from the ReLU field trained in Section 3.1 and keep these fixed, while using probability maps in place of color images and optimizing for the probability values in the grid using an L1 loss. As shown in Figures 3 and 4, optimizing for a volumetric representation allows for ultimately refining the 2D probability maps, for instance by resolving over inconsistent 2D observations (as illustrated in Figure 4).

We then convert these 3D probability fields to our binary mask M using a seam-hiding segmentation algorithm

based on energy minimization [3]. Specifically, we extract a segmentation grid that minimizes an energy function composed of two terms: A *unary* term, which penalizes disagreements with the label probabilities, and a *smoothness* term, which penalizes large pairwise color differences within similarly-labeled voxels. We define the label probabilities for voxel cell as the element-wise softmax of two cross-attention grids A_e and A_{obj} , where

- A_e is the cross-attention grid associated with the token describing the edit (e.g. *sunglasses*), and
- A_{obj} is the grid associated with the object, defined as the maximum probability over all other tokens in the prompt.

We compute the smoothness term from local color differences in the edited grid G_e . That is, we sum

$$w_{pq} = \exp\left(\frac{-(c_p - c_q)^2}{2\sigma^2}\right) \quad (5)$$

for each pair of same-labeled neighboring voxels p and q , where c_p and c_q are RGB colors from G_e . In our experiments, we use $\sigma = 0.1$ and balance the data and smoothness terms with a parameter $\lambda = 5$ (strengthening the smoothness term). Finally, we solve this energy minimization problem via graph cuts [7], resulting in the high quality segmentation masks shown in Figure 3.

4. Experiments

We show qualitative editing results over diverse 3D objects and various edits in Figures 1, 5, 6, 8. Please refer to the supplementary material for many fly-through visualizations demonstrating that our results are indeed consistent across different views.

To assess the quality of our object editing approach, we conduct several sets of experiments, quantifying the extent of which these rendered images conform to the target text prompt. We provide comparisons with prior 3D object editing methods in Section 4.2, and comparisons to 2D editing methods in Section 4.3. An additional comparison to an unconditional text-to-3D method is presented in Section 4.4. Results over real scenes are illustrated in Section 4.5. We show ablations in Section 4.6. Finally, we discuss limitations in Section 4.7. Additional results, visualizations, ablations and comparisons can be found in the supplementary material.

Synthetic Object Dataset. We assembled a dataset using freely available meshes found on the internet. Each mesh was rendered from 100 views in Blender. For a quantitative evaluation, we paired each object in our dataset with a number of both local and global edit prompts including:

- "A $\langle object \rangle$ wearing sunglasses".



Figure 5. Results obtained by our method over different objects and prompts (with the inputs displayed on the left). Please refer to the supplementary material for additional qualitative results.

- “A $\langle object \rangle$ wearing a party hat”.
- “A $\langle object \rangle$ wearing a Christmas sweater”.
- “A yarn doll of a $\langle object \rangle$ ”.
- “A wood carving of a $\langle object \rangle$ ”.

We separately evaluate local and global edits, using our spatial refinement step over local edits only. For instance, the first three prompts above are considered local edits (where regions that are not associated with the text prompt should remain unchanged) and the last two as edits that should produce global edits. We provide additional details in the supplementary material.

Runtime. All experiments were performed on a single RTX A5000 GPU (24GB VRAM). The training time for our method is approx. 50 minutes for the editing stage and 15 minutes for the optional refinement stage.

4.1. Metrics

Edit Fidelity. We evaluate how well the generated results capture the target text prompt using two metrics:

CLIP Similarity ($CLIP_{Sim}$) measures the semantic similarity between the output objects and the target text prompts.

	Method	$CLIP_{Sim} \uparrow$	$CLIP_{Dir} \uparrow$
Local	DFF+CN	0.34*	0.05*
	Text2Mesh	0.36*	0.08*
	Latent-NeRF (Sketch / Paint)	0.32 / 0.31	0.01 / 0.01
	Ours	0.36	0.07
Global	DFF+CN	0.32*	0.01*
	Text2Mesh	0.34*	0.03*
	Latent-NeRF (Sketch / Paint)	0.30 / 0.31	0.01 / 0.01
	Ours	0.34	0.02

Table 1. **Quantitative Evaluation.** We compare against the 3D object editing techniques Text2Mesh [28], two variants of Latent-NeRF [27]: SketchShape (Sketch) and Latent-Paint (Paint) and DFF+CN [23, 43], over local (top) and global (bottom) edits.

*Note that Text2Mesh and DFF+CN explicitly train to minimize a CLIP loss, and thus directly comparing them is uninformative over these metrics.

We encode both the prompt and images rendered from our 3D outputs using CLIP’s text and image encoders, respectively, and measure the cosine-distance between these encodings.

CLIP Direction Similarity ($CLIP_{Dir}$) evaluates the quality of the edit in regards to the input by measuring the directional CLIP similarity first introduced by Gal et al. [12]. This metric measures the cosine distance between the direction of the change from the input and output rendered images and the direction of the change from an input prompt (*i.e.* “a dog”) to the one describing the edit (*i.e.* “a dog wearing a hat”).

Edit Magnitude. For ablating components in our model, we use the Frechét Inception Distance (FID) [16, 38] to measure the difference in visual appearance between: (i) the output and input images (FID_{Input}) and (ii) the output and images generated by the initial reconstruction grid (FID_{Rec}). We show both to demonstrate to what extent the appearance is affected by the edit versus the expressive power of our framework.

4.2. 3D Object Editing Comparisons

To the best of our knowledge, there is no prior work that can directly perform our task of text-guided localized edits for 3D objects given a set of posed input images. Thus, we consider Distilled Feature Fields [23] combined with CLIP-NeRF [43] (DFF+CN), Text2Mesh [28] and Latent-NeRF [27] which can be applied in a similar setting to ours. These experiments highlight the differences between prior works and our proposed editing technique.

Distilled Feature Fields [23] distills 2D image features into a 3D feature field to enable query-based local editing of a 3D scenes. CLIP-NeRF edits a neural radiance field by optimizing the CLIP score of the input query and the rendered image. Combining these two methods allows to edit only the relevant parts of the 3D scene. Text2Mesh [28] aims at editing the style of a given input mesh to conform

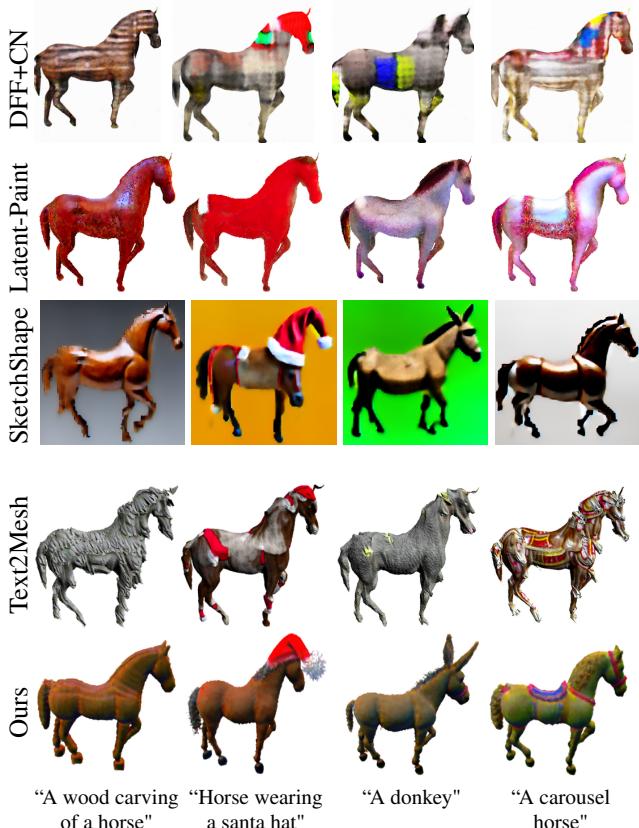


Figure 6. Comparison to other 3D Object editing techniques. We show qualitative results obtained using Text2Mesh [28], two applications of Latent-NeRF [27] (Latent-Paint and SketchShape) and DFF+CN [23, 43] and compare to our method. To accommodate their problem setting, the top three methods are provided with uncolored meshes. Note that the input meshes are visible on the second row from the top (as Latent-Paint does not edit the object’s geometry). As illustrated above, prior methods struggle at achieving semantic localized edits. Our method succeeds, while maintaining high fidelity to the input object.

to a target prompt with a style transfer network that predicts color and a displacement along the normal direction. As it only predicts displacements along the normal direction, the geometric edits enabled by Text2Mesh are limited mostly to small changes. Latent-Paint and SketchShape are two applications introduced in Latent-Nerf [27] which operate on input meshes. SketchShape generates shape and appearance from coarse input geometry, while Latent-Paint only edits the appearance of an existing mesh. Note that Text2Mesh and Latent-NeRF are designed for slightly more constrained inputs than our approach. While our focus is on editing 3D models with arbitrary textures (as depicted from associated imagery), they only operate on uncolored meshes.

We show a qualitative comparison in Figure 6 over an uncolored mesh (its geometry can be observed on the second row from the top as Latent-Paint keeps the input ge-

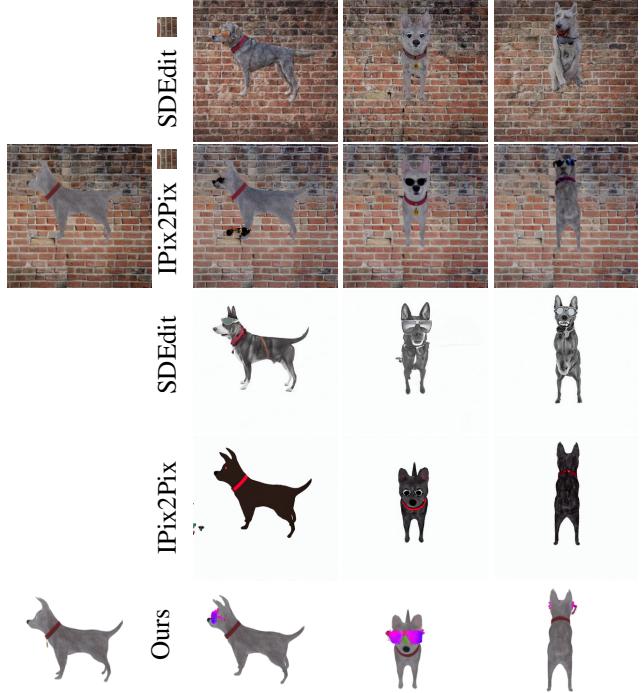


Figure 7. Comparison to 2D image editing techniques. We compare to the text-guided image editing techniques InstructPix2Pix (IPix2Pix) [8] and SDEdit [26] by providing it with images from different viewpoints and a target instruction text prompt (“put sunglasses on the dog” for IPix2Pix and “a dog with sunglasses” for SDEdit and our method). We show one input image on the left, and three outputs on the right (side, front and back views), where the leftmost output corresponds to the input viewpoint. We show two variants, one with added backgrounds (top rows), as we observe that it allows for better preserving the object’s appearance. As illustrated above, 2D techniques cannot easily achieve 3D-consistent edit results (illustrated, for instance, by the sunglasses added on the dog’s back).

ometry fixed). As illustrated in the figure, Text2Mesh cannot produce significant geometric edits (*e.g.*, adding a Santa hat to the horse or turning the horse into a donkey). Even SketchShape, which is designed to allow geometric edits, cannot achieve significant localized edits. Furthermore, it fails to preserve the geometry of the input—although, we again note that this method is not intended to preserve the input geometry. DFF+CN seems generally less suitable for our problem setting, particularly for prompts that require geometric modifications (*i.e.* “A donkey”). Our method, in contrast to prior works, succeeds in conforming to the target text prompt, while preserving the input geometry, allowing for semantically meaningful changes to both geometry and appearance.

We perform a quantitative evaluation in Table 1 on our dataset. To perform a fair comparison where all methods operate within their training domain, we use meshes without texture maps as input for Text2Mesh and Latent-NeRF.

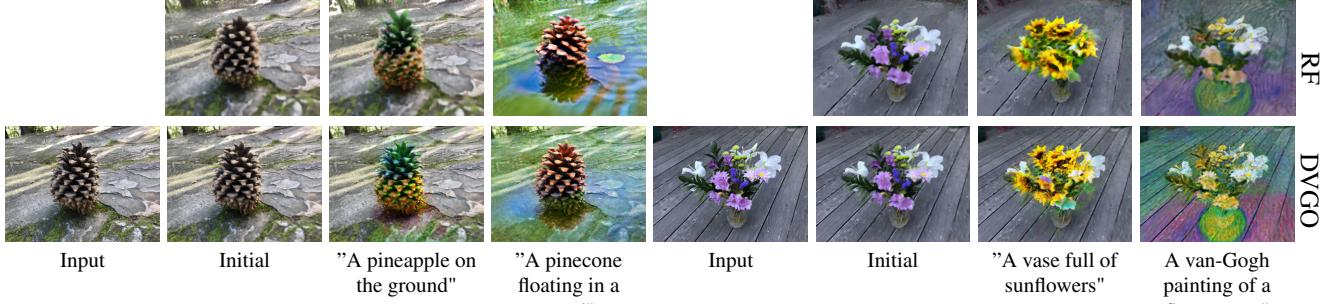


Figure 8. **Editing real scenes with different underlying 3D representations.** We show results obtained when using DVGO [41] (bottom row) and ReLU-Fields (RF, top row). We show samples from the input image dataset (leftmost columns), initial scene reconstructions (second columns), results over local edits (third columns) and results over global edits (rightmost columns).



Figure 9. **Comparison to unconditional text-to-3D generation.** We compare to unconditional text-to-3D methods by comparing to Latent-NeRF [27], providing it with the two target prompts displayed above. We display these alongside our results (LatentNeRF on the left, ours on the right). As illustrated above, unconditional methods cannot easily match an input object, and are also not guaranteed to generate a consistent object over different prompts.

As illustrated in the table, our method outperforms all baselines over both local and global edits in terms of CLIP similarity, but Text2Mesh yields slightly higher CLIP direction similarity. We note that Text2Mesh as well as DFF+CN are advantaged in terms of the CLIP metrics as they explicitly optimize on CLIP similarities and thus their scores are not entirely indicative.

4.3. 2D Image Editing Comparisons

An underlying assumption in our work is that editing 3D geometry cannot easily be done by reconstructing edited 2D images depicting the scene. To test this hypothesis, we modified images rendered from various viewpoints using the diffusion-based image editing methods Instruct-Pix2Pix [8] and SDEdit [26]. We show two variants of these methods in Figure 7, one with added backgrounds, as we observe that it also affects performance. In both cases, as illustrated in the figure, 2D methods often struggle to produce meaningful results from less canonical views (e.g., adding sunglasses on the dog’s back) and also produce highly view-inconsistent results. Concurrently to us, Instruct-NeRF2NeRF [14] explore how to best use these 2D methods to learn view-consistent 3D representations.

4.4. Comparisons to an unconditional text-to-3D model

In Figure 9 we compare to the unconditional text-to-3D model proposed in Latent-NeRF, to show that such unconditional models are also not guaranteed to generate a consistent object over different prompts. We also note that this result (as well as our edits) would certainly look better if fueled with a proprietary big diffusion model [36], but nonetheless, these models cannot preserve identity.

4.5. Real Scenes

In Figure 8, we demonstrate that our method also succeeds in modeling and editing real scenes using the *360° Real Scenes* available by Mildenhall et al. [29]. As illustrated in the figure, we can locally edit the foreground (e.g., turning the pinecone into a pineapple) as well as globally edit the scene (e.g. turning the scene into a Van-Gogh painting). For these more complex and computationally demanding scenes, we also experiment with implementing our method on top of DVGO [41] (bottom row), in addition to ReLU-Fields which we exclusively focus on in all other experiments (top row), as it offers additional features such as scene contraction, a more expressive color feature space and complex ray sampling. These make this underlying representation better suited for editing and reconstructing these real scenes (as illustrated in the columns labeled as ‘Initial’). This experiment also demonstrates that our method is agnostic to the underlying 3D representation of the scene and can readily operate over different grid-based representations.

4.6. Ablations

We provide an ablation study in Table 2 and Figure 10. Specifically, we ablate our volumetric regularization (\mathcal{L}_{reg3D}) and our 3D cross-attention-based spatial refinement module (SR). When ablating our volumetric regularization, we use a single volumetric grid and regularize the SDS objective with an image-based L2 regularization

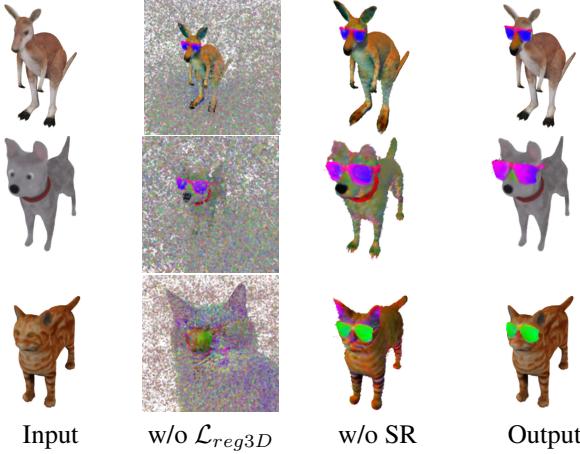


Figure 10. **Qualitative ablation results**, obtained for the target prompt “A <object> wearing sunglasses” over three different objects. Image-space regularization (denoted by “w/o $\mathcal{L}_{\text{reg3D}}$ ”) leads to extremely noisy results. The edited grid before refinement (denoted by “w/o SR”) respects the target prompt, but some of the fidelity to the geometry and appearance of the input object is lost. In contrast, our refined grid successfully combines the edited and input regions to output a result that complies with the target text and also preserves the input object.

$\mathcal{L}_{\text{reg3D}}$	SR	$\text{CLIP}_{\text{Sim}} \uparrow$	$\text{CLIP}_{\text{Dir}} \uparrow$	$\text{FID}_{\text{Rec}} \downarrow$	$\text{FID}_{\text{Input}} \downarrow$
×	×	0.29	0.05	367.53	384.55
✓	✗	0.37	0.08	240.37	288.26
✓	✓	0.36	0.06	119.44	236.32

Table 2. **Ablation study**, evaluating the effect of the volumetric regularizer between our coupled grids ($\mathcal{L}_{\text{reg3D}}$, Section 3.2) and the 3D cross-attention-based spatial refinement module (SR, Section 3.3) over a set of metrics (detailed in Section 4).

loss. More details and additional ablations are provided in the supplementary material, including alternative regularization objectives (such as image-based L1 loss, or volumetric regularization over RGB features) and results using higher order spherical harmonics coefficients.

The benefit of using our volumetric regularization is further illustrated in Figure 10, which shows that image-space regularization leads to very noisy results, and often complete failures (see, for instance, the cat result, where the output is not at all correlated with the input object). Quantitatively, we can also observe that images rendered from these models are of significantly different appearance (as measured using the FID metrics).

Regarding the SR module, as expected, it increases similarity to the inputs (reflected in lower FID scores). This is also clearly visible in Figure 10—for example, geometric differences are apparent by looking at the animals’ paws. The output textures after refinement also are more similar to the input textures. However, we also see that this module slightly hinders CLIP similarity to the edit and text prompt. This is also somewhat expected as we are further constrain-



Figure 11. **Limitations**. Above, we present several failure cases (when provided with rendered images of the uncolored mesh displayed in Figure 6, top row). These likely result from incorrect attribute binding (the horse’s nose turning into a pig’s nose), inconsistencies across views (two horns on the unicorn) or excessive regularization to the input object (carpet on the horse, not below).

ing the output to stay similar to the input, sometimes at the expense of the editing signal.

4.7. Limitations

Our method applies a wide range of edits with high fidelity to 3D objects, however, there are several limitations to consider. As shown in Figure 11, since we optimize over different views, our method attempts to edit the same object in differing spatial locations, thus failing on certain prompts. Moreover, the figure shows that some of our edits fail due to incorrect attribute binding, where the model binds attributes to the wrong subjects, which is a common challenge in large-scale diffusion-based models [9]. Finally, we inherit the limitations of our volumetric representation. Thus, the quality of real scenes, for instance, could be significantly improved by borrowing ideas from works such as [6] (e.g. scene contraction to model the background).

5. Conclusion

In this work, we presented Vox-E, a new framework that leverages the expressive power of diffusion models for text-guided voxel editing of 3D objects. Technically, we demonstrated that by combining a diffusion-based image-space objective with volumetric regularization we can achieve fidelity to the target prompt and to the input 3D object. We also illustrated that 2D cross-attention maps can be elevated for performing localization in 3D space. We showed that our approach can generate both local and global edits, which are challenging for existing techniques. Our work makes it easy for non-experts to modify 3D objects using just text prompts as input, bringing us closer to the goal of democratizing 3D content creating and editing.

6. Acknowledgments

We thank Rinon Gal, Gal Metzer and Elad Richardson for their insightful feedback. This work was supported by a research gift from Meta, the Alon Fellowship and the Yandex Initiative in AI.

References

- [1] Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Trans. Graph.*, 40(3), May 2021.
- [2] Panos Achlioptas, Ian Huang, Minhyuk Sung, Sergey Tulyakov, and Leonidas Guibas. Changeit3d: Language-assisted 3d shape edits and deformations, 2022.
- [3] Aseem Agarwala, Mira Dontcheva, Maneesh Agrawala, Steven Drucker, Alex Colburn, Brian Curless, David Salesin, and Michael Cohen. Interactive digital photomontage. *ACM Trans. Graph.*, 23(3):294–302, 2004.
- [4] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022.
- [5] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kassten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *ECCV*, 2022.
- [6] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022.
- [7] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(11):1222–1239, 2001.
- [8] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. November 2022.
- [9] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *arXiv preprint arXiv:2301.13826*, 2023.
- [10] Yongwei Chen, Rui Chen, Jiabao Lei, Yabin Zhang, and Kui Jia. Tango: Text-driven photorealistic and robust 3d stylization via lighting decomposition. *arXiv preprint arXiv:2210.11277*, 2022.
- [11] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.
- [12] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators, 2021.
- [13] Stephan J Garbin, Marek Kowalski, Virginia Estellers, Stanislaw Szymanowicz, Shideh Rezaeifar, Jingjing Shen, Matthew Johnson, and Julien Valentin. Voltmorph: Real-time, controllable and generalisable animation of volumetric representations. *arXiv preprint arXiv:2208.00949*, 2022.
- [14] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. *arXiv preprint arXiv:2303.12789*, 2023.
- [15] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. 2022.
- [16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. 2017.
- [17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.
- [18] Ian Huang, Panos Achlioptas, Tianyi Zhang, Sergey Tulyakov, Minhyuk Sung, and Leonidas Guibas. Ladis: Language disentanglement for 3d shape editing. *arXiv preprint arXiv:2212.05011*, 2022.
- [19] Takeo Igarashi, Tomer Moscovich, and John F Hughes. As-rigid-as-possible shape manipulation. *ACM transactions on Graphics (TOG)*, 24(3):1134–1141, 2005.
- [20] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *CVPR*, 2022.
- [21] Animesh Karnewar, Tobias Ritschel, Oliver Wang, and Niloy Mitra. Relu fields: The little non-linearity that could. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022.
- [22] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022.
- [23] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in Neural Information Processing Systems*, 35:23311–23330, 2022.
- [24] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440*, 2022.
- [25] Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. Editing conditional radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5773–5783, 2021.
- [26] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jianjun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.
- [27] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022.
- [28] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *CVPR*, 2022.
- [29] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

- [30] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022.
- [31] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *ArXiv*, abs/2302.03027, 2023.
- [32] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021.
- [33] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.
- [34] 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 *ICML*, 2021.
- [35] Elad Richardson, Gal Metzger, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023.
- [36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.
- [37] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *CVPR*, 2022.
- [38] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. <https://github.com/mseitzer/pytorch-fid>, 08 2020. Version 0.2.1.
- [39] Yujun Shen, Jinjin Gu, Xiaou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020.
- [40] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, pages 109–116, 2007.
- [41] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022.
- [42] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022.
- [43] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2022.
- [44] Can Wang, Ruixiang Jiang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Nerf-art: Text-driven neural radiance fields stylization. *arXiv preprint arXiv:2212.08070*, 2022.
- [45] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lift-ing pretrained 2d diffusion models for 3d generation. *ArXiv*, abs/2212.00774, 2022.
- [46] Fanbo Xiang, Zexiang Xu, Milos Hasan, Yannick Hold-Geoffroy, Kalyan Sunkavalli, and Hao Su. Neutex: Neural texture mapping for volumetric neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7119–7128, 2021.
- [47] Tianhan Xu and Tatsuya Harada. Deforming radiance fields with cages. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 159–175. Springer, 2022.
- [48] Bangbang Yang, Chong Bao, Junyi Zeng, Hujun Bao, Yinda Zhang, Zhaopeng Cui, and Guofeng Zhang. Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 597–614. Springer, 2022.
- [49] Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yuewen Ma, Rongfei Jia, and Lin Gao. Nerf-editing: geometry editing of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18353–18364, 2022.
- [50] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields, 2022.