SIGNeRF: Scene Integrated Generation for Neural Radiance Fields

Jan-Niklas Dihlmann   Andreas Engelhardt   Hendrik Lensch

University of Tübingen

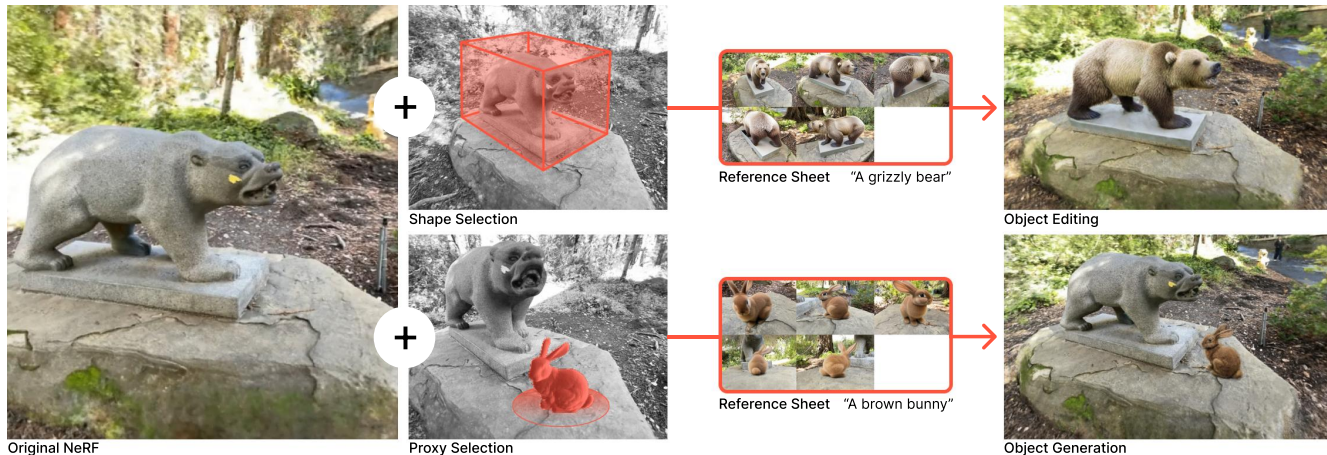{jan-niklas.dihlmann, andreas.engelhardt, hendrik.lensch}@uni-tuebingen.de

Figure 1. **SIGNeRF** – A method that enables NeRF scene editing by either shape selection for object modification or by proxy object placement to insert a synthesized object. The transformed scene is represented by a NeRF trained on a set of edited input images, which are efficiently generated by a diffusion process using a multiview consistent reference sheet. Project page available at https://signerf.jdihlmann.com/

Abstract

Advances in image diffusion models have recently led to notable improvements in the generation of high-quality images. In combination with Neural Radiance Fields (NeRFs), they enabled new opportunities in 3D generation. However, most generative 3D approaches are object-centric and applying them to editing existing photorealistic scenes is not trivial. We propose SIGNeRF, a novel approach for fast and controllable NeRF scene editing and scene-integrated object generation. A new generative update strategy ensures 3D consistency across the edited images, without requiring iterative optimization. We find that depth-conditioned diffusion models inherently possess the capability to generate 3D consistent views by requesting a grid of images instead of single views. Based on these insights, we introduce a multi-view reference sheet of modified images. Our method updates an image collection consistently based on the reference sheet and refines the original NeRF with the newly generated image set in one go. By exploiting the depth conditioning mechanism of the image diffusion model, we gain fine control over the spatial location of the edit and enforce shape guidance by a selected region or an external mesh.

1. Introduction

In this work, we focus on scene-integrated generation for editing existing 3D scenes using generative 2D diffusion models (Fig. 1). We propose an approach to create or modify objects within existing NeRF [27] scenes while preserving the original scenes' structure and appearance. Previous methods such as Instruct-NeRF2NeRF [11] and DreamEditor [57] already demonstrated scene modification with text instructions. However, their pipelines are complex, involve additional training and lack fine-grained control.

We observe that ControlNet [55], an image diffusion model with additional depth guidance is inherently capable of generating coherent and consistent views of an object when applied to image tiles in a grid layout. Using this combined image generation, we construct a *reference sheet* of edits that can subsequently be applied to the original NeRF scene by updating the underlying image dataset. The generation process of the reference sheet is conditioned on the existing NeRF scene and a text prompt that can be tuned
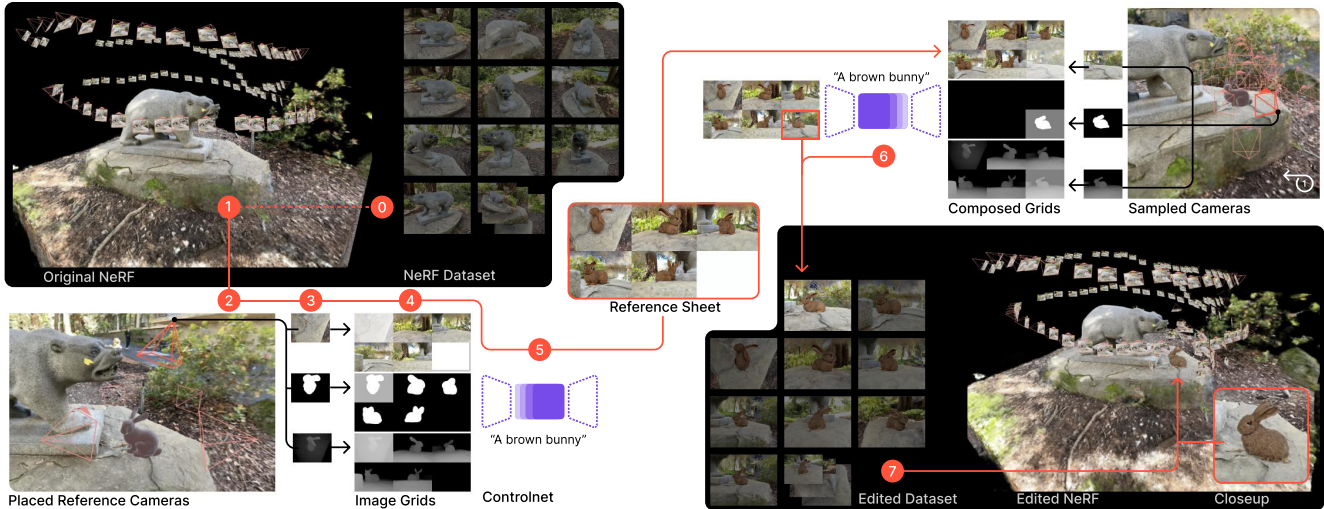
Figure 2. **SIGNeRF pipeline for NeRF scene editing** – here, object generation. First, the original NeRF scene is trained (**1**), and a proxy object is placed into the scene (**2**). After a precise selection, we place reference cameras (5 here) into the scene (**3**), render the corresponding color, depth, and mask images, and arrange them into image grids (**4**). These grids are used to generate the reference sheet with conditioned image diffusion (**5**). To propagate the edits to the entire image set, for each camera, a color, depth, and mask image are rendered and placed into the empty slot of the fixed reference sheet. We generate a new edited image consistent with the reference sheet by leveraging an inpainting mask. The step is repeated for all cameras (**6**). Finally, the NeRF is fine-tuned on the edited images (**7**)

easily to one's liking. This editing capability is further refined by introducing two selection methods in scene space to condition the edits of the diffusion model. A mesh proxy can be placed and composed with the depth data of the existing NeRF to guide the insertion of new objects or a rough bounding box can be used to select existing parts of the scene. This addition grants enhanced control over the position and visual representation of generated entities and preserves the quality of the unedited parts. In summary, our method consists of the following key elements:

- Reference-sheet-based assembly as a simple and easy-to-implement approach to improve consistency in generated multi-view image data when using the ControlNet [55] image diffusion model.
- A modular pipeline to update a NeRF dataset based on a reference sheet that is generated and refined to the user's liking using text prompts. This update scheme can be easily repeated to further increase quality if needed.
- Fine-grained control of the generation process by offering multiple selection modes to constrain edits in scene space.

Compared to existing methods like DreamEditor [57] and Instruct-NeRF2NeRF [11] we introduce a simplified, more modular pipeline that achieves comparable or improved results while adding control to the editing process.

2. Related Work

Text-to-Image Generation   Recent advances in diffusion probabilistic models [13, 44] in combination with the availability of large datasets [41], enabled the generation of high-resolution and diverse images [30, 36, 38]. Fine-tuning these models can significantly enhance the capacity to guide and personalize image generation [14, 37]. In addition, Control-Net [55] has demonstrated that image diffusion models can be fine-tuned to condition them on various secondary inputs, such as depth, poses, or even sketches [47]. Furthermore, editing existing images in a generative manner with instructions is possible as outlined by InstructPix2Pix [3].

Text-to-3D Generation   The field of text-to-3D generation has recently experienced a significant breakthrough, with the combination of image diffusion models and NeRFs [34]. Previously, the emphasis of text-driven 3D generation relied on optimizing meshes [5, 18, 26], point clouds [39], voxels [4], and NeRFs [16], guided by Contrastive Language-Image Pretraining (CLIP) [35] embeddings and supervision. While mesh generation guided by CLIP continues to be an active area of research [9], recent advances in image diffusion models address a significant challenge in text-to-3D generation, namely the scarcity of 3D data. The pioneering work by DreamFusion [34] introduced Score-Distillation-Sampling (SDS), leveraging a pre-trained image diffusion model conditioned on a text input prompt to generate novel 3D objects. Building upon this, Latent-NeRF [25] extended DreamFusion to the latent domain, accelerating the generation process using Stable Diffusion [36].

The generation quality and object variation can be further improved by modifying SDS [15, 22, 48]. Although SDS has proven itself as a powerful tool for 3D object generation, it

requires a large amount of memory and time to train. In contrast, Instruct-NeRF2NeRF [11] proposed that conditioned image diffusion models can be used for 3D object generation without requiring a direct gradient update, making it vastly faster. By iteratively updating the NeRF image set with conditioned generated images, the NeRF can be trained to generate the desired 3D object. The final appearance, however, emerges slowly over time while our method gives an immediate preview.

While there are other promising paths to 3D generation, including the use of 3D diffusion models [20, 28, 31] and synthesizing novel views from single or multiple images to lift an image representation into 3D space [7, 23, 24, 46, 49, 50, 56], this work utilizes 2D image diffusion, given its applicability to NeRF representations.

NeRF Editing   Although NeRFs have proven themselves as a state-of-the-art tool for novel view synthesis and 3D scene reconstruction [2, 27, 29], editing NeRFs is still a challenging task and is the subject of ongoing research. Efforts to manipulate NeRFs were initiated by NeRF-Editing [54], which proposed a method to deform NeRFs by modifying the underlying implicit function. NeRFShop [21] extended the work of cage-based deformation [51], introducing an intuitive selection of objects for both affine and non-affine transformations and object duplication. However, these approaches are limited to simple deformations and do not allow for more complex edits or additions to the scene. In contrast, NeuMesh [52] learns a disentangled neural mesh-based implicit field to edit geometry and appearance, enabling geometry deformation and texture swapping, filling, and painting. It has also been demonstrated that NeRFs can be combined by inserting objects into pre-existing NeRF scenes [19]. Despite the strides made in NeRF editing, the existing solutions provide only a basic level of editing functionality compared to conventional 3D editing software. Generally, obtaining visually pleasing results still requires artistic proficiency and often manual labor.

Generative NeRF Editing   With advances in text-to-3D generation, a new area in NeRF editing emerged, incorporating generative 3D models to modify existing NeRF scenes. Set-the-Scene [6] and Compositional 3D [32] present methods for controlled scene generation using proxy objects and bounding boxes. In addition, Composable 3D Diffusion enables moving generated objects within the composed scene, which consists of multiple NeRFs. On the other hand, SINE [1] allows direct NeRF editing by changing a reference image in 2D space. It uses an editing field to adjust the template NeRF's geometry and appearance to match the image changes. Similarly, VOX-E [42] combines an original NeRF with an edited NeRF generated using text-based SDS. It merges the field by using the attention mask from the image

diffusion model. Yet, these methods often result in overlaying edits and increased memory and training time due to using multiple NeRFs. In contrast, Instruct-NeRF2NeRF [11] uses an Iterative Dataset Update (IDU) strategy to edit NeRF's image dataset with InstructPix2Pix [3], such that the NeRF can be transformed based on editing instructions. Inspired by the work of Instruct-NeRF2NeRF [11], we also update the NeRF image dataset with an image diffusion model. However, unlike Instruct-NeRF2NeRF, we do this in a preprocessing stage using the depth-conditioned ControlNet [55], gaining more control over the generational process. DreamEditor [57] also allows for controlled text-based scene editing by focusing the selection with finetuning an image diffusion model DreamBooth [37] before using SDS. Nevertheless, using SDS and DreamBooth is very time-consuming compared to our approach. Furthermore, Instruct-NeRF2NeRF and DreamEditor can be challenging to control by only semantics and often fail to generate new objects within the scene at specific locations. Blended-NeRF [10] improves positional control by limiting the generation within manually controlled bounding boxes. However, Blended-NeRF results often do not fit the scene and get clipped by its bounding boxes, highlighting the need for further improvements in controlled generative NeRF editing. We introduce a new approach to generative NeRF editing, providing more precise control over the generated edits with a manual selection or proxy object guidance to generate new and fitting objects within an existing scene.

3. Method

SIGNeRF is a method for scene-integrated generation, including edits and object generation within an existing NeRF scene. At the core, we introduce the concept of a reference sheet image grid to maintain multi-view coherence and to gain control over the generation process (Figs. 2 and 4). Given a NeRF model trained on the original image set, a set of views is selected to compose a grid of images. The scene edits are performed on this grid by ControlNet, a conditioned image diffusion model (Sec. 3.1) in one go to produce the reference sheet. In the second step, the reference sheet constrains the generation of an updated version of the full image set. We observe that this two-step procedure already supports generating quite consistent edited views. The edited 3D scene is obtained by finetuning the original NeRF scene with these newly generated views. Optionally, if the multi-view consistency needs to be improved further a second iteration can be performed, where the reference sheet is updated based on the once-edited NeRF to generate a second image set. An overview of the pipeline on consistent grid image generation and subsequent scene-integrated generation is presented in Fig. 2.
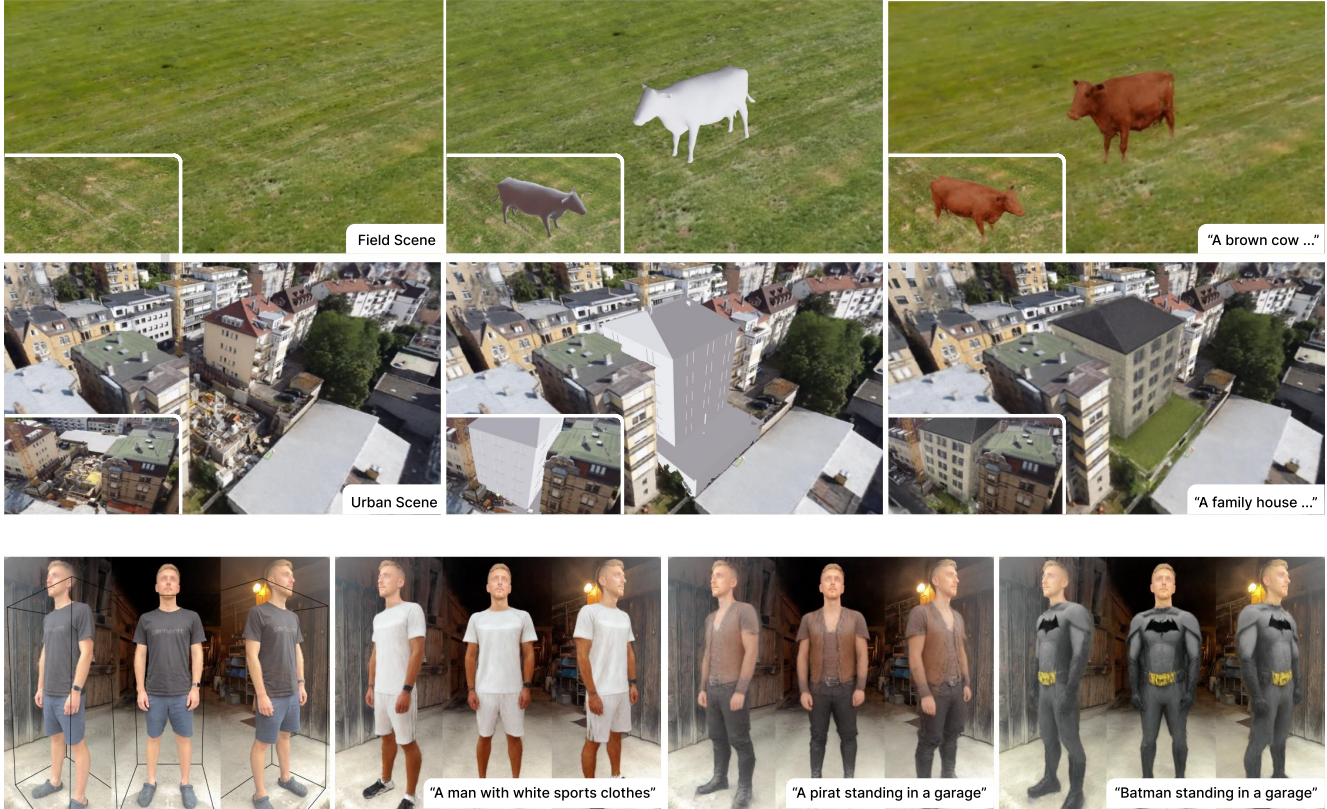
Figure 3. **Object insertion and object modification.** – (top) The cow geometry is centrally placed on a meadow to obtain a photorealistic scene. (middle) Note how occlusions are properly handled when generating the synthetic house based on the inserted proxy. (bottom) Objects can easily be transformed based on a prompt. Due to the more complex surface texture and geometric changes the pirate and the Batman costume required an additional iteration to obtain the same level of consistency compared to the simpler sports clothes.

3.1. Background

NeRF Neural Radiance Fields (NeRFs) [27] implicitly represent a scene by learning a continuous function of volumetric density and color. A 5D coordinate, composed of a spatial location (x, y, z) and a viewing direction (θ, φ), is mapped to a view-dependent emitted radiance (r, g, b) and a volume density σ. Volumetric rendering accumulates the densities and colors at multiple sample locations along the view ray $\mathbf{r}_j$ of a virtual camera C to obtain the final pixel color $\hat{C}(\mathbf{r}_j)$ to either calculate a novel view or to update the neural representation based on the difference to some of the input images $\mathbf{I} = \{\mathbf{I}_1, ..., \mathbf{I}_N\}$.

ControlNet ControlNet [55] is a specific image diffusion model that allows for constraining the image generation process with additional conditions, such as sketches, edge, or depth maps. In the typical process of image-to-image diffusion, an image x is first encoded with an encoder $\mathcal{E}(x) = z$ to produce a latent image z. Then the latent image z is iteratively updated by the diffusion model, which is guided by a

noise predictor UNet ϵ_θ conditioned on the encoded text input $\tau_\theta(\boldsymbol{y})$ and the current time step t. The final latent image z is then decoded by a decoder $\mathcal{D}$ to obtain the generated image $\boldsymbol{y}$. In the case of ControlNet, the image generation process is guided by an additional condition $\boldsymbol{c_f}$, that serves as an additional input to the noise predictor UNet ϵ_θ, (Eq. 1).

$$\epsilon = \epsilon_\theta\left(z_t, t, \boldsymbol{c_f}, \tau_\theta(\boldsymbol{y})\right) \qquad (1)$$

This functionality is used in our pipeline to guide the image-to-image generation process with depth maps.

3.2. Controlled Consistent Generation

The key challenge of 3D generation techniques is to generate consistent views with an image diffusion model. Our approach is based on reference sheet generation which is simpler, faster and features direct control compared to the methods introduced in Sec. 2 which rely on a repetitive cycle of intertwined diffusion and NeRF updates.

Reference Sheet Generation We observe that the image diffusion model ControlNet [55] can already generate multi-

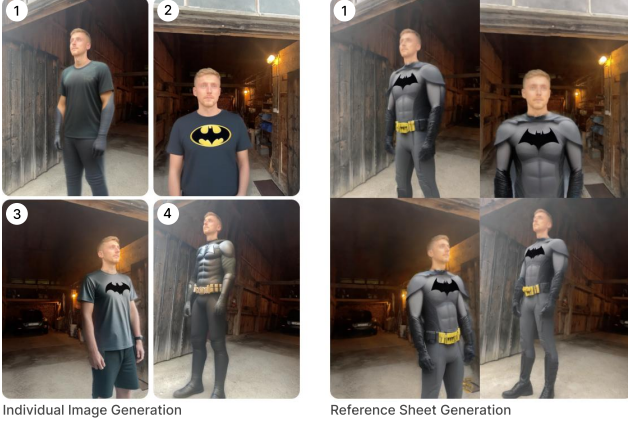Individual Image Generation   Reference Sheet Generation

Figure 4. **Reference Sheet Generation** – Using ControlNet [55] inpainting to edit scene parts image-by-image results in drastically different looks per view (left) although all parameters and the seed are the same. In contrast, we obtain a consistent reference sheet (right) by arranging the input images into a grid, letting ControlNet process the entire sheet in a single generation step.

view consistent images of a scene without the need for iterative refinement. While generating individual views sequentially introduces too much variation to integrate them into a consistent 3D model, arranging them in a grid of images that are processed by ControlNet in one pass significantly improves the multi-view consistency as depicted in Fig. 4. Based on the depth maps rendered from the original NeRF scene we employ a depth-conditioned inpainting variant of ControlNet to generate such a reference sheet of the edited scene. A mask specifies the scene region where the generation should occur.

This step gives a lot of control to the user. Different appearances can be produced by generating reference sheets with different seeds or prompts. The one sheet finally selected will directly determine the look of the final 3D scene.

Image Set Update   Despite the potential of grid generation, we are limited in the number of images we can place in one sheet due to the memory and attention limitations of the image diffusion model. Given that a proper NeRF training might require hundreds of images, an alternative method for generating consistent scene views is necessary. We solve this challenge in a two-step process. First, we generate the reference sheet and subsequently use it to iteratively update the images in the NeRF dataset image-by-image. To ensure that all other views are also consistent with the reference sheet, we originally left one slot empty when producing the reference sheet, e.g. the bottom right corner, and condition this slot with the appropriate original image, mask and depth of the to-be-generated view while keeping the rest of the reference sheet fixed. This way, the appearance of the reference sheet is propagated to all input views.

3.3. Scene Integrated Generation

Based on these ideas we present the full scene-integrated generation pipeline as depicted in Fig. 2 for the specific case of object generation.

1. **Original NeRF Scene** – A NeRF scene S is reconstructed using set of N input images $I_S = \{I_1^S, I_2^S, ..., I_N^S\}$ alongside the corresponding cameras $C_S = \{C_1^S, C_2^S, ..., C_N^S\}$. This NeRF serves as the foundation for the subsequent scene-integrated generation.

2. **Object Selection** – The next step is to mark the 3D region to be edited. For region-based edits, we use a bounding box. To introduce a new object, we use a proxy mesh that is placed in the scene.

3. **Reference Camera Placement** – Around the selected edit region, we place the $M << N$ cameras $C_R = \{C_1^R, C_2^R, ..., C_M^R\}$ for the reference sheet generation. They need to cover a sufficient range of views around the object and have the object properly centered.

4. **Reference Input Image Rendering** – For each reference camera we use the original NeRF scene S to render the RGB images $I_R = \{I_1^R, I_2^R, ..., I_M^R\}$, with corresponding depths D_R and inpainting masks M_R.

5. **Reference Sheet Assembly and Generation** – Each set of M reference input images is arranged in a 2D grid resulting in $\bar{I}_R$, $\bar{D}_R$ and $\bar{M}_R$, respectively. One grid cell remains empty on purpose. The reference sheet R is generated by piping the created grids as inputs and conditions to ControlNet with the selected prompt y:

$$R \leftarrow \text{ControlNet}(\bar{I}_R, \bar{D}_R, \bar{M}_R, y) \qquad (2)$$

6. **Image Set Update** – After producing the desired reference sheet, it is used to generate a new image dataset $I_{\hat{S}} = \{I_1^{\hat{S}}, I_2^{\hat{S}}, ..., I_N^{\hat{S}}\}$. We first create depth maps $D_S = \{D_1^S, D_2^S, ..., D_N^S\}$ and inpainting masks $M_S = \{M_1^S, M_2^S, ..., M_N^S\}$ from the original NeRF scene S and then replace the empty grid cells with the corresponding image or depth map. This results in the following update rule for the image dataset $I_{\hat{S}}$:

$$I_i^{\hat{S}} \leftarrow \text{ControlNet}(R_i, \bar{D}_i^R, M_i^S, y) \quad \forall i \in \{1, 2, ..., N\}$$
$$\text{with } R_i \leftarrow R \oplus \{I_i^S\} \text{ and } \bar{D}_i^R \leftarrow \bar{D}_R \oplus \{D_i^S\} \qquad (3)$$

7. **Finetuning the NeRF Scene** – The final step is to finetune the original NeRF scene S with the generated image dataset $I_{\hat{S}}$ to receive the edited scene $\hat{S}$.

8. **Multiple Iteration (optional)** – While the generated reference sheet always shows a consistent style in all tiles it might happen that the underlying 3D shape is not yet fully consistent. If this leads to visible artifacts Steps 4 to 7 can be optionally repeated once more, this time
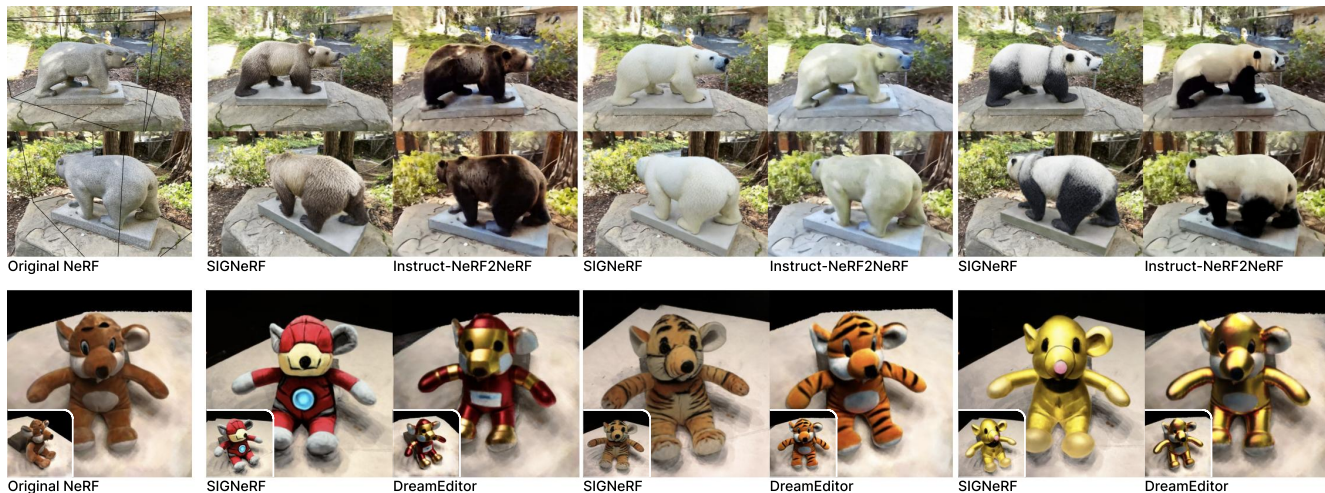
Figure 5. **Qualitative Comparison** – SIGNeRF results are compared to Instruct-NeRF2NeRF [11] (top) and DreamEditor [57] (bottom). For the bear, the generated fur texture with SIGNeRF (left) shows a more distinguished structure and the snout regions is clearly more consistent. Compared to DreamEditor the images are different but the image quality comparable.

starting with the updated NeRF $\hat{S}$ to render the reference input images. The parameters of the ControlNet in Steps 5 and 6 have to be tuned to stick closer to the input in this second iteration.

A key feature of our pipeline is its modularity. Each step can be developed and optimized independently. Compared to the tight iterative NeRF/image diffusion update loops of other approaches, in our pipeline individual steps can easily be exchanged, e.g. to enable different scene modifications, or to repeat some as indicated in the optional Step 8.

Selection Modes For precise control over the generation location, we introduce two basic selection modes: Shape selection and proxy selection (Fig. 1). With shape selection, a region of the scene can be selected by an axis-aligned bounding box. We use this bounding box and generate a per-camera mask by comparing its depth to the rendered NeRF depth. Further, we a combined depth map by clamping the rendered NeRF depth within the point closest and furthest away to the rendering camera within the bounding box.

With the proxy selection mode, one can position an arbitrary mesh within the NeRF scene. Similar to the shape selection, we generate a depth map and a mask for each camera view, but herefore combine the rendered depth map of the NeRF scene with the depth map rendered of the proxy object relative to the camera. In difference, we use the visible part of the proxy object in the rendered view as the mask. As previously described the generated depth map conditions the image diffusion model, while the generated mask is used as an inpainting mask, to blend the generated image with the original NeRF image. The masks can further be dilated in image space to allow control over the to-be-generated area.

Reference Sheet The quality of the reference sheet directly impacts the generation results. One important aspect is the number of reference cameras and their position within the scene. Optimally, we use the fewest cameras necessary to capture the region of interest from all angles. This strategy minimizes generation time while maintaining consistency. Another criterion is the proximity of reference cameras to the original cameras. Generally, it is best to render reference views from a position close to the original camera positions. In cases where the edit region is too distant from any original camera, the reference cameras need to move closer in order to increase the number of pixels in the masks for more details in the generated edits. A standout feature of SIGNeRF is its ability to preview the reference sheet before generating the complete updated image set. Besides choosing the intended appearance we recommend iterative adjustments to the reference sheet by replacing undesired generated images of the image grid until satisfactory results are obtained.

4. Experiments

The proposed pipeline facilitates the generation of new objects within a scene and the editing of existing ones. We assess the quality of the 3D scenes generated by SIGNeRF and compare it to existing methods.

4.1. Experimental Setup

Datasets We use various types of scenes for our experiments, most of them in real-world settings. While some are front-facing scenes, our primary focus is on 360° view scenes due to their inherent challenges and potential in 3D scene generation. Datasets from Instruct-NeRF2NeRF [11], DTU [17] and BlendedMVS [53] are

	Image	NeRF	Background Preserv.	
Method	CLIP T2I Dir. Sim. ↑		PSNR ↑	SSIM ↑
Instruct-N2N [11]	0.1603	0.1600	30.09	0.64
DreamEditor [57]	-	0.1849	-	-
Ours	**0.23**	**0.2125**	**32.22**	**0.83**

Table 1. **Quantitative Evaluation** – Highlighting the CLIP text-to-image directional similarity comparing the diffusion *Image* edit (rendered NeRF view edited with the method-specific diffusion mode) and the trained *NeRF* render (rendered view from the edited NeRF) to the original image with corresponding prompts. The background preservation capability of our method is evaluated using PSNR and SSIM by masking the edited object, comparing the original and the edited NeRF result.

utilized. Additionally, to address the scarcity of realistic NeRF datasets, we created custom scenes using smartphones with PolyCam [12] or drones. Camera parameters were either sourced directly from relevant apps or inferred using COLMAP [40]. The scenes contain 30 to 300 images.

Implementation Details   SIGNeRF is built upon Nerfstudio using Nerfacto [45] as the underlying NeRF implementation. For the image diffusion model, we modified an inpainting version of ControlNet [33, 36, 43, 55] with the SDXL [33]-diffusion backbone to allow for conditioning on masked content. ControlNet scale is set between $[0.4, 1.0]$, guidance values range from $[6, 10]$ and denoising strength varies between $[0.5, 0.95]$. However, these parameters can be adjusted as needed.

4.2. Qualitative Evaluation

Fig. 3 demonstrates SIGNeRF's object generation and editing capabilities (see also the videos in the supplemental material). In the first two rows, novel objects are inserted with fine-grained control of position, orientation and size. The objects are synthesized, and conditioned on a geometry proxy. These objects are integrated seamlessly into the scene, exhibiting fitting lighting and texture properties. Notably in the cow scene, one does not even need strong trackable features for precise placement.

In the last row, an existing object is modified based on a text prompt to generate different appearances. Note that the desired edits only affect the marked object. SIGNeRF is, however, powerful enough to even adjust the geometry where necessary, e.g. short vs. long trousers, while the background is preserved.

Proxy Shape   The impact of the proxy shape and geometric detail is visualized in Fig. 6. The shape of the proxy mesh does influence the final object as ControlNet tries to fit the
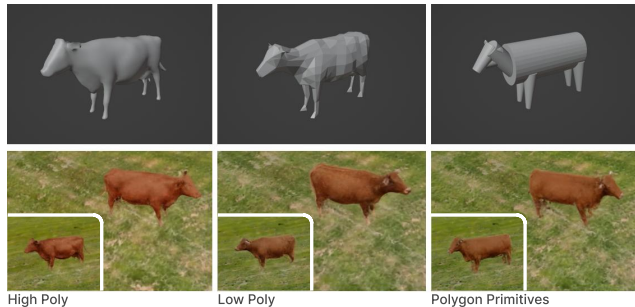

Figure 6. **Influence of the proxy geometry** – The synthetic cow is generated with three different proxy meshes with the prompt "A brown cow". From left to right: High-poly proxy mesh, low-poly proxy mesh, and simple geometric primitives.

depth condition created with the proxy mesh, as apparent when the body of the cow is approximated by a cylinder. Nevertheless, the results indicate that the proxy mesh does not need to be a detailed representation of the object, as the diffusion process added additional geometry details. Both low and high-poly versions yield satisfactory results.

ControlNet Guidance   Further, we studied the impact of the scale parameter of ControlNet [55], which handles the closeness of the generation to the depth condition, on the final NeRF edit. We observe that a low scale strongly reduces the impact of the condition, leading to an edit that deviates largely from the original shape. This allows us to make drastic changes, like the Batman example (Fig. 4, Fig. 3). However, a low scale can also lead to more irritation in the 3D consistency of the generated image, producing artifacts in the NeRF. In these cases, a second iteration of our process (Sec. 3.3), using a low scale in the first and a high scale in the second iterator achieves the desired results. In Fig. 3, the Pirate and the Batman are generated with a second generation iteration, while the white shirt is generated directly.

4.3. Comparison

Scene editing results of SIGNeRF are compared to Instruct-NeRF2NeRF [11] and DreamEditor [57] in Fig. 5. While Instruct-NeRF2NeRF [11] produces washed-out textures and suffers from the Janus effect of showing different faces from different views for the bear, our generated results show a more consistent snout region and more vivid, more structured pelt textures. Here, SIGNeRF is superior in terms of scene preservation, selection precision, generation quality, and color integrity.

In the second row of Fig. 5, it achieves similar results to DreamEditor, while both methods have some artifacts that cannot be directly compared. DreamEditor tends to generate simpler and over-smoothed objects due to the high classifier-free guidance needed for score-destillation-sampling [48].
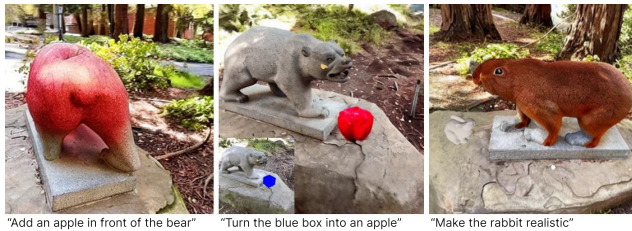
Figure 7. **Instruct-NeRF2NeRF** – Instruct-NeRF2NeRF fails to generate new objects in the scene, even when merging proxy meshes with the NeRF scene. Compare to our results in Fig. 1
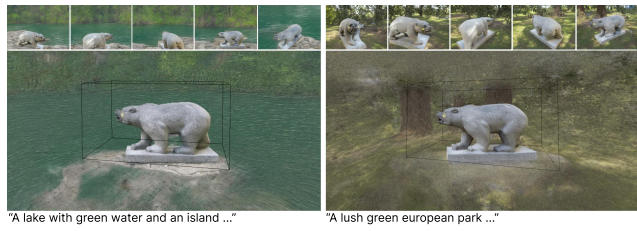

Figure 8. **Limitations** – Trying to modify the background by inverting the object mask yields strong inconsistencies. The reference sheet images (top) show non-overlapping views of the background.

In contrast, SIGNeRF generates more complex and realistic-looking objects but may show consistency artifacts for highly detailed regions.

Both other methods are designed primarily for scene editing, generating new objects semantically independent from the scene is not possible, let alone controlling the object's position, scale, rotation, and shape within the scene. For example, Instruct-NeRF2NeRF fails to generate a rabbit in front of the bear statue (Fig. 7) while SIGNerF embeds it consistently into the scene (Fig. 1). But also for editing tasks, SIGNeRF is superior as it also allows semantically dependent parts to be unedited, e.g. the face in the Batman scene.

Another aspect is the time spent in the generation process. Instruct-NeRF2NeRF and DreamEditor, require more than an hour for a generation that is only fully visible at the end. While SIGNeRF only takes half the time on a single GPU, using one dataset iteration, it additionally provides a preview option with the reference sheet, allowing the user to adjust the output until satisfied before starting the generation process. Furthermore, the image set update (Step 6) can easily be parallelized over all images in our approach while the interlocked NeRF/image updates in the other methods are purely sequential.

4.4. Quantitative Evaluation

Even though the process of generating and editing 3D scenes is inherently subjective, in line with previous works we utilize the CLIP [35] text-to-image directional similarity to provide a quantitative perspective. This metric [8] evaluates the semantic distance between the original image and edited NeRF scene to their corresponding prompt pairs. Table 1 compares the scores averaged over a total of 10 scenes to Instruct-NeRF2NeRF [11] and DreamEditor [57]

Another metric provided in Table 1 accounts for background preservation calculated by masking the edited object and comparing the original and edited NeRF background. Due to their compressive nature image diffusion models are prone to degrading the image quality. Since SIGNeRF employs a masked update strategy, the background is better preserved compared to Instruct-NeRF2NeRF, which uses a repeated iterative update strategy.

4.5. Limitations

Even though we can achieve better quality than Instruct-NeRF2NeRF with masking, we are forced to downscale the images to fit into the reference sheet passed to the image diffusion model, thereby losing some quality for the generated edit. Further, optimal results are obtained when the object occupies the image's center and is positioned close to the camera. As increasing the distance reduces the resolution of the object in the latent space the quality of the generated edit diminishes. This behavior holds true for all image diffusion-based 3D generation methods. Additionally, an off-center object complicates its incorporation into a reference sheet that generates consistent views, thereby making SIGNeRF unsuitable for extended scene modifications (Fig. 8).

5. Conclusion

With SIGNeRF we present a modular pipeline for scene-integrated editing of NeRF scenes. An efficient and easily controllable two-step procedure first generates a tiled reference sheet followed by updating the image set to generate consistent edited views suitable for modifying an existing NeRF representation. The reference sheet additionally shows a preview of the edited scene before generating all images, which is not possible with existing editing methods. SIGN-eRF often achieves consistent 3D generation in a single processing run. In comparison to previous methods, SIGN-eRF is faster and delivers similar or superior editing results without necessitating iterative refinement.

The introduced selection strategies enable generative edits or object insertion within an existing NeRF, even for scenes with complex geometry and appearance.

Acknowledgements

References

[1] Chong Bao, Yinda Zhang, Bangbang Yang, Tianxing Fan, Zesong Yang, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Sine: Semantic-driven image-based nerf editing with prior-guided editing field. pages 20919–20929, 2023. 3

[2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. pages 5470–5479, 2021. 3

[3] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. pages 18392–18402, 2022. 2, 3

[4] Kevin Chen, Christopher B. Choy, Manolis Savva, Angel X. Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. pages 100–116, 2019. 2

[5] Yongwei Chen, Rui Chen, Jiabao Lei, Yabin Zhang, and Kui Jia. Tango: Text-driven photorealistic and robust 3d stylization via lighting decomposition. 2022. 2

[6] Dana Cohen-Bar, Elad Richardson, Gal Metzer, Raja Giryes, and Daniel Cohen-Or. Set-the-scene: Global-local training for generating controllable nerf scenes. *arXiv preprint arXiv:2303.13450*, 2023. 3

[7] Congyue Deng, Chiyu "Max" Jiang, Charles R. Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, and Dragomir Anguelov. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. pages 20637–20647, 2022. 3

[8] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2021. 8

[9] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances In Neural Information Processing Systems*, 35:31841–31854, 2022. 2

[10] Ori Gordon, Omri Avrahami, and Dani Lischinski. Blended-nerf: Zero-shot object generation and blending in existing neural radiance fields. *arXiv preprint arXiv:2306.12760*, 2023. 3

[11] Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. 2023. 1, 2, 3, 6, 7, 8

[12] Chris Heinrich. Polycam: Lidar scanning app for iphone. https://poly.cam/, 2023. Polycam Inc. provides a fast and accurate 3D scanning app leveraging the LiDAR sensor on the iPhone. 7

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. 2022. 2

[15] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv preprint arXiv:2306.12422*, 2023. 2

[16] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. pages 867–876, 2022. 2

[17] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. 6

[18] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. 2022. 2

[19] Verica Lazova, Vladimir Guzov, Kyle Olszewski, Sergey Tulyakov, and Gerard Pons-Moll. Control-nerf: Editable feature volumes for scene rendering and manipulation. pages 4340–4350, 2022. 3

[20] Muheng Li, Yueqi Duan, Jie Zhou, and Jiwen Lu. Diffusion-sdf: Text-to-shape via voxelized diffusion. pages 12642–12651, 2022. 3

[21] Shaoxu Li and Ye Pan. Interactive geometry editing of neural radiance fields. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(1), 2023. 3

[22] 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. 2022. 2

[23] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization, 2023. 3

[24] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *arXiv preprint arXiv:2303.11328*, 2023. 3

[25] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. pages 12663–12673, 2023. 2

[26] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. pages 13492–13502, 2021. 2

[27] 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, 2020. 1, 3, 4

[28] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. pages 4328–4338, 2022. 3

[29] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph. 41, 4, Article 102 (July 2022), 15 pages*, 41(4):1–15, 2022. 3

[30] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021. 2

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

[32] Ryan Po and Gordon Wetzstein. Compositional 3d scene generation using locally conditioned diffusion. *arXiv preprint arXiv:2303.12218*, 2023. 3

[33] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 7

[34] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. 2023. 2

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. pages 8748–8763, 2021. 2, 8

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. pages 10684–10695, 2021. 2, 7

[37] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. pages 22500–22510, 2022. 2, 3

[38] 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, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[39] 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. pages 18603–18613, 2021. 2

[40] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. 36, 2022. 2

[42] Etai Sella, Gal Fiebelman, Peter Hedman, and Hadar Averbuch-Elor. Vox-e: Text-guided voxel editing of 3d objects. *arXiv preprint arXiv:2303.12048*, 2023. 3

[43] Neelay Shah, Tommaso De Rossi, and Mikolaj Czerkawski. Controlnetinpaint: Inpaint images with controlnet. https://github.com/mikonvergence/ControlNetInpaint, 2023. GitHub repository. 7

[44] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. pages 2256–2265, 2015. 2

[45] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. *arXiv preprint arXiv:2302.04264*, 2023. 7

[46] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*, 2023. 3

[47] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. *arXiv preprint arXiv:2211.13752*, 2022. 2

[48] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 2, 7

[49] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 3

[50] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360° views. pages 4479–4489, 2022. 3

[51] Tianhan Xu and Tatsuya Harada. Deforming radiance fields with cages. pages 159–175, 2022. 3

[52] 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. pages 597–614, 2022. 3

[53] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. pages 1790–1799, 2019. 6

[54] Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yuewen Ma, Rongfei Jia, and Lin Gao. Nerf-editing: Geometry editing of neural radiance fields. pages 18353–18364, 2022. 3

[55] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 1, 2, 3, 4, 5, 7

[56] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. pages 12588–12597, 2022. 3

[57] Jingyu Zhuang, Chen Wang, Lingjie Liu, Liang Lin, and Guanbin Li. Dreameditor: Text-driven 3d scene editing with neural fields, 2023. 1, 2, 3, 6, 7, 8