Local 3D Editing via 3D Distillation of CLIP Knowledge

Junha Hyung[1,2†] Sungwon Hwang[1] Daejin Kim[3] Hyunji Lee[1] Jaegul Choo[1]

[1]KAIST AI [2]Kakao Enterprise Corp. [3]Scatter Lab

{sharpeeee, shwang.14, alee6868, jchoo}@kaist.ac.kr, daejin@scatterlab.co.kr

(a) Original 3D model (b) Locally-edited 3D model (c) Rendered mask

Figure 1. Our local editing NeRF (LENeRF) enables users to edit specific areas of 3D assets based on textual prompts by estimating a 3D mask for tri-plane features. For instance, given an original 3D radiance field (a), users can define their desired area to edit (underlined text prompt, e.g., "eyes"). LENeRF then generates a 3D mask, which is employed for feature fusion, allowing for targeted modifications that adhere to the editing prompt (e.g., "blue eyes") (b). Additionally, as illustrated in (c), the 3D mask itself can be rendered and visualized for further analysis.

Abstract

3D content manipulation is an important computer vision task with many real-world applications (e.g., product design, cartoon generation, and 3D Avatar editing). Recently proposed 3D GANs can generate diverse photorealistic 3D-aware contents using Neural Radiance fields (NeRF). However, manipulation of NeRF still remains a challenging problem since the visual quality tends to degrade after manipulation and suboptimal control handles such as 2D semantic maps are used for manipulations. While text-guided manipulations have shown potential in 3D editing, such approaches often lack locality. To overcome these problems, we propose Local Editing NeRF (LENeRF), which only requires text inputs for fine-grained and localized manipulation. Specifically, we present three add-on modules of LENeRF, the Latent Residual Mapper, the Attention Field Network, and the Deformation Network, which are jointly used for local manipulations of 3D features by estimating a 3D attention field. The 3D attention field is learned in an unsupervised way, by distilling the zero-shot mask generation capability of CLIP to the 3D space with multi-view guidance. We conduct diverse experiments and thorough evaluations both quantitatively and qualitatively.[1]

1. Introduction

3D content editing has many real-world applications including but not limited to product design, cartoon generation, and 3D Avatar editing. However, it often necessitates the use of sophisticated tools with complex interfaces, which can be difficult for novice users and labor-intensive even for seasoned professionals. While explicit 3D repre-

[†]This work was done during an internship at Kakao Enterprise Corp.

[1]We will make our code publicly available.

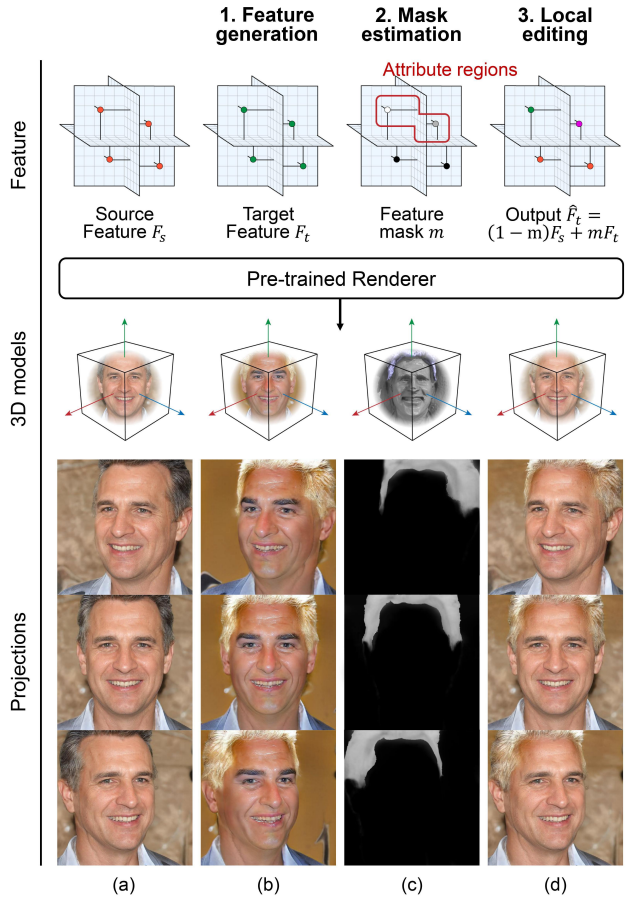

Figure 2. Concept figure of LENeRF. Our method enables local editing of 3D assets by generating the target feature and estimating a 3D mask which guides the model on where to make changes at the feature level. Note that the mask is estimated for tri-plane features, not for raw RGB outputs.

sentations such as voxels and meshes are commonly used for 3D generation and editing [17,31,54], they are memory-intensive and lack photorealism. In contrast, recent advances in Neural Radiance Fields (NeRF) [33] have shown promising progress in representing 3D environments using implicit representations [14, 21, 33, 37] combined with volume rendering techniques that enable high-quality novel view synthesis. NeRF-based 3D GANs [4, 5, 11, 16, 36, 43, 50,52] have made further progress towards generating a category of 3D aware contents with a single model, extending the per-scene optimization scheme of NeRF.

Several studies [29, 44, 45, 48] have attempted to address the challenges of NeRF editing, yet certain limitations persist. Works such as Edit-NeRF [29] and CLIP-NeRF [48] have pioneered NeRF manipulations, but they are constrained to low-resolution synthetic datasets and lack the capability to perform localized editing. Opposed to translation [9, 53] or style transfer [13] tasks, editing typically demands a certain degree of localization. However, achieving this with text-only control proves to be a challenging objective. Alternative methods [44, 45] that rely on semantic masks for editing face their own limitations: 2D guidance is not ideal for 3D editing and lacks the descriptiveness required for fine-grained editing. Furthermore, these approaches require inversion steps and are difficult to generalize across different domains, as they depend on the availability of labeled semantic masks.

To overcome the existing limitations, we propose Local Editing NeRF (LENeRF), which focuses on the important aspects of 3D editing: photorealism, multi-view consistency, usability, diversity, and locality. With LENeRF, high-resolution photo-realistic radiance fields can be edited while maintaining their quality and multi-view consistency. One notable advantage of LENeRF is its text-only editing, making it more usable than other methods. This allows our approach to be applied to any domain by leveraging the multi-modal embedding space of Contrastive Language Image Pre-training (CLIP) [39]. Additionally, our method achieves real-time editing as it does not require any test-time optimization process.

Our proposed approach exhibits particularly robust performance in *local* 3D editing. This is achieved through a unique method of editing features in the 3D space independently by granting position-wise freedom to the features. The naive approach of directly manipulating the latent code often results in *global* changes to the 3D content, because features in the 3D space are spatially entangled with each other as the entire radiance field is conditioned with a single latent code. To address this issue, we propose to generate a 3D mask on the region of interest with a masking prompt (e.g., "hair") and manipulate the features inside the region while leaving the rest unchanged. Inspired by the previous approach which introduces the explanation method for capturing the regions of the interest [25], we estimate 3D masks in an unsupervised fashion by using 3D distillation of the 2D CLIP model. Although the CLIP model is not 3D-aware and the 3D GAN lacks text-conditioned mask generation capability, our method enables the collaboration of two pre-trained models to generate a text-conditioned 3D mask, as demonstrated in Figure 1 (c).

LENeRF comprises three add-on modules, namely Latent Residual Mapper (LRM), Attention Field Network (AFN), and Deformation Network (DN) as depicted in Figure 3. LRM generates a latent code that produces a target feature field. AFN generates a soft 3D mask indicating our region of interest. The source feature field is distorted using DN and subsequently interpolated with the target field to synthesize the final feature field. LENeRF is trained with CLIP guidance [38, 39], and AFN is additionally trained with CLIP-generated zero-shot pseudo labels.

The main contributions of our paper are as follows:

- We introduce Local Editing NeRF (LENeRF), a 3D content editing framework capable of localized, photo-realistic editing using a convenient real-time *text-based* interface.
- Our method consists of add-on modules and does not require any domain-specific labels, allowing the method to be generalized to other models and domains.
- Our proposed technique involves a novel 3D distillation of CLIP knowledge, specifically an unsupervised approach that utilizes the 3D GAN and CLIP models jointly to generate 3D masks.
- We present diverse quantitative and qualitative results, along with various applications such as sequential editing, real image editing, and out-of-distribution editing.

2. Related Work

3D-Aware GANs While Generative Adversarial Networks (GANs) [15] have demonstrated their ability to generate high-quality 2D images, generating 3D-aware content with GANs remains a challenging task. Several approaches have attempted to integrate GANs and 3D scene representations [34, 35]. Recent progress in 3D GANs [4, 5, 11, 16, 27, 36, 43, 50, 52] that employ NeRF [33] have achieved success in generating high-quality, multi-view consistent images. Building upon the progress made by previous research in this field, our work seeks to enhance its applicability by introducing a novel method for text-guided local editing.

NeRF Manipulations To enable local editing in 3D assets, EditNeRF [29] proposes a conditional NeRF that enables users to modify the color and shape of a specific region using scribbles. However, the editing ability of EditNeRF is limited to adding or removing local parts of 3D objects. Semantic mask guided 3D editing approaches [22, 29, 44, 46] have made some progress in 3D editing. Among these works, FENeRF [45] is trained to render 3D images and their corresponding semantic masks, allowing editing to be performed by inverting the model with edited semantic masks to modify 3D models. However, using a 2D mask for 3D editing is suboptimal since fine-grained attributes such as expression or texture cannot be manipulated using masks. Also, these methods require segmentation maps for both training and editing, which is not applicable to certain domains where segmentation maps are not available. Meanwhile, there is a line of research utilizing CLIP [39] for 3D generation and editing [3, 19, 20, 26, 32, 42, 48] that leverages the ease of manipulation using a text. However, the generated and edited content are not photorealistic and cannot provide fine-grained control. On the other hand, our work focuses on high-quality, localized manipulations that are well-suited for real-world applications.

3. Preliminaries

NeRF NeRF [33] is an implicit representation of 3D space using a MLP Ψ, as $(\mathbf{c}, \sigma) = \Psi(\mathbf{x}, \mathbf{d})$, where $\mathbf{x} = (x, y, z)$ is a point in 3D space, $\mathbf{d}$ is a viewing direction, and resulting $(\mathbf{c}, \sigma)$ are color and volume density, respectively. The color and density values along the same ray are integrated to estimate a single pixel given camera extrinsic and pose as:

$$\hat{C}(\mathbf{r}) = \int_{k_n}^{k_f} T(k)\sigma(\mathbf{r}(k))c(\mathbf{r}(k), \mathbf{d})\, dk, \tag{1}$$

where k_n and k_f are near and far bounds, $\mathbf{r}(k) = \mathbf{o} + k\mathbf{d}$ denotes a ray projected from the camera centered at position $\mathbf{o}$, and $T(m) = \exp(-\int_{k_n}^{m} \sigma(\mathbf{r}(m))dm)$ is an accumulated transmittance along the ray. The parameters of Ψ are optimized to produce $\hat{C}(\mathbf{r})$, which are expected to be close to the ground truth pixel value.

StyleGAN Latent Code The expressiveness of the StyleGAN latent code w can be increased by inputting k different latent codes, denoted by $\mathbf{w}$ ($\mathbf{w} \in \mathcal{W}^k \subsetneq \mathbb{R}^{k\times 512}$)[2], into each of the k StyleGAN layers. This can further be extended to $\mathcal{W}^k_*$ as introduced in [47]. Unlike $\mathcal{W}^k$, whose distribution is bounded by the mapping network, $\mathcal{W}^k_*$ can reside outside the range of the mapping network. We choose to utilize $\mathcal{W}^k_*$ in our pipeline to enable a wider range of attribute manipulation.

EG3D EG3D [4] is a 3D-aware GAN based on NeRF. Feature vectors of a radiance field are sampled from tri-plane generated by StyleGAN2 [24] and summed to a tri-plane feature vector $F = g_\theta(\mathbf{x}, \mathbf{w})$ given a 3D position $\mathbf{x} \in \mathbb{R}^3$ and a latent code $\mathbf{w}$ that modulates the StyleGAN2 layers, and g_θ denotes the function of the whole process. F is further processed by a small decoder to a M_f-dimensional feature vector. Then a set of sampled features $\{F_{ij}\}_{j=1}^{N_s}$ along the set of camera rays $\{\mathbf{r}_i\}_{i=1}^{H_V \times W_V}$ is aggregated with volumetric rendering using Eq. 1, where N_s is the number of sampled points per ray, and H_V and W_V are the height and the width of the rendered feature image, respectively. Finally, the feature image is decoded into the high-resolution image by a super-resolution module π^{SR}:

$$\pi^{SR} : \mathbb{R}^{H_V \times W_V \times M_f} \rightarrow \mathbb{R}^{H \times W \times 3} \tag{2}$$

4. LeNeRF Framework for 3D Local Editing

In this section, we present Local Editing NeRF (LENeRF) which enables *localized* editing of 3D scenes using a text prompt for manipulation, t_{edit} (e.g., *blue eyes*), and a prompt specifying the region of interest, t_{mask} (e.g., *eyes*). Unlike previous methods that perform global editing

[2]For simplicity, we use the notations ($\mathcal{W}^k$, $\mathcal{W}^k_*$) from [47].

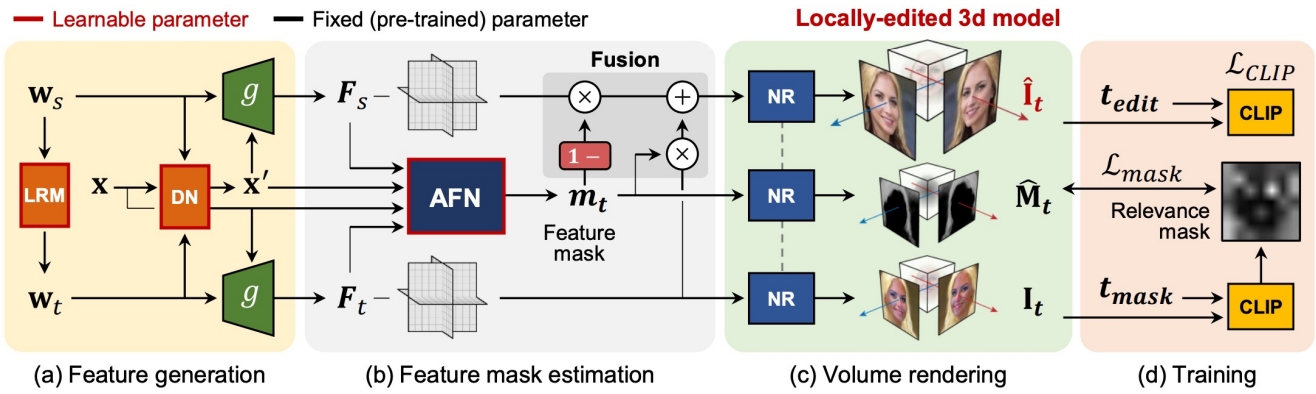

Figure 3. The overall architecture and the training objective of LeNeRF. (a) The target latent code and the deformation field are generated by the latent residual mapper (LRM) and the deformation network (DN), respectively. (b) The tri-plane features for the source and target are encoded by the pretrained generator and the soft mask for combining both features is estimated by our attention field network (AFN). (c) By using the neural renderer (NR) which incorporates the volume rendering and a super-resolution (not applied to rendered masks), we obtain the rendered results for the source and target features, mask, and the locally-edited features, which is the final output of LeNeRF. (d) In the training process, only LRM, DN, and AFN are trained for local editing, while other parameters for the generator and neural renderer are fixed.

by updating the latent codes of 3D GAN [48], we propose generating a 3D mask using CLIP [39] supervision and performing localized editing with the generated mask through a *position-wise* transformation of feature fields.

As shown in Fig. 3, LENeRF consists of frozen pretrained NeRF generator [4] along with trainable add-on modules: Latent Residual Mapper (Section 4.1), Attention Field Network (Section 4.2), and Deformation Network (Section 4.3). The Latent Residual Mapper (LRM) maps the source latent code $\mathbf{w}_s \in \mathcal{W}^k$ to the target latent code $\mathbf{w}_t \in \mathcal{W}^k_*$, which in turn conditions the source and target feature vectors F_s and F_t respectively (Fig. 3 (a)).

Then the 3D soft mask corresponding to the regions specified by t_{mask} is estimated by the Attention Field Network (AFN) (Fig. 3 (b)). The Deformation Network handles geometric manipulations by deforming the source feature field $\{F_s\}$, and the source and target features are fused to produce the final feature fields. The modules are trained using the proposed CLIP loss $\mathcal{L}_{CLIP^+}$, and the AFN is trained additionally with the mask loss $\mathcal{L}_{mask}$ via pseudo-labels generated with relevance map aggregation conditioned on the mask prompt t_{mask} (Fig. 3 (d), Section 4.4). Once trained, LeNeRF performs 3D editing in real-time.

4.1. Latent Residual Mapper (LRM)

Inspired by CLIP-guided 2D image editing techniques [1, 12, 38, 49, 51], we train the Latent Residual Mapper (Figure 3 (a)), a mapper function that generates the target latent code $\mathbf{w}_t$ given the source latent code $\mathbf{w}_s$. The mapper is trained using CLIP guidance, and once trained, manipulation can be performed in real-time without the need for any inversion steps [38].

Instead of using $\mathbf{w}_t$ directly for generating the final radiance field, we utilize the mapper to produce tri-plane features $\{F_t\}$, which are composited with the source tri-plane feature field $\{F_s\}$ to generate the final feature field. This composition involves interpolating two features based on an estimated 3D mask (Section 4.2).

The features in the 3D space are spatially entangled because the entire feature field is conditioned with a single latent code. By utilizing more than one latent code (in our case, two), we can grant position-wise freedom to the features, which enables the localized editing of the source features based on the 3D mask and the target features. Specifically, given the source latent code $\mathbf{w}_s \in \mathcal{W}^k$, the mapper outputs the target latent code $\mathbf{w}_t \in \mathcal{W}^k_*$ which is used to produce the final feature field, described as:

$$\begin{aligned} M(\mathbf{w}_s) &= (M^1(\mathbf{w}_s^1), ..., M^N(\mathbf{w}_s^N)) \\ &= (\Delta\mathbf{w}^1, ..., \Delta\mathbf{w}^N), \end{aligned} \tag{3}$$

$$\mathbf{w}_i = (\mathbf{w}_s^1 + \Delta\mathbf{w}^1, ..., \mathbf{w}_s^N + \Delta\mathbf{w}^N), \tag{4}$$

where $M^i, i \in 1, 2, ..., N$ are fully-connected networks and $\mathbf{w}^n$ denotes the subset of the latent code $\mathbf{w}$ that modulates the n-th group of StyleGAN2-like backbone layers.

We design our mapper to estimate the residual values of latent codes, which enables easier regularization and training. LRM is trained to minimize the $\mathcal{L}_{CLIP^+}$ objective given a target prompt t_{edit}. See Section 4.4 for more details.

4.2. Attention Field Network (AFN)

The Attention Field Network (AFN) parameterizes the 3D attention field that serves as a soft 3D mask for interpo-

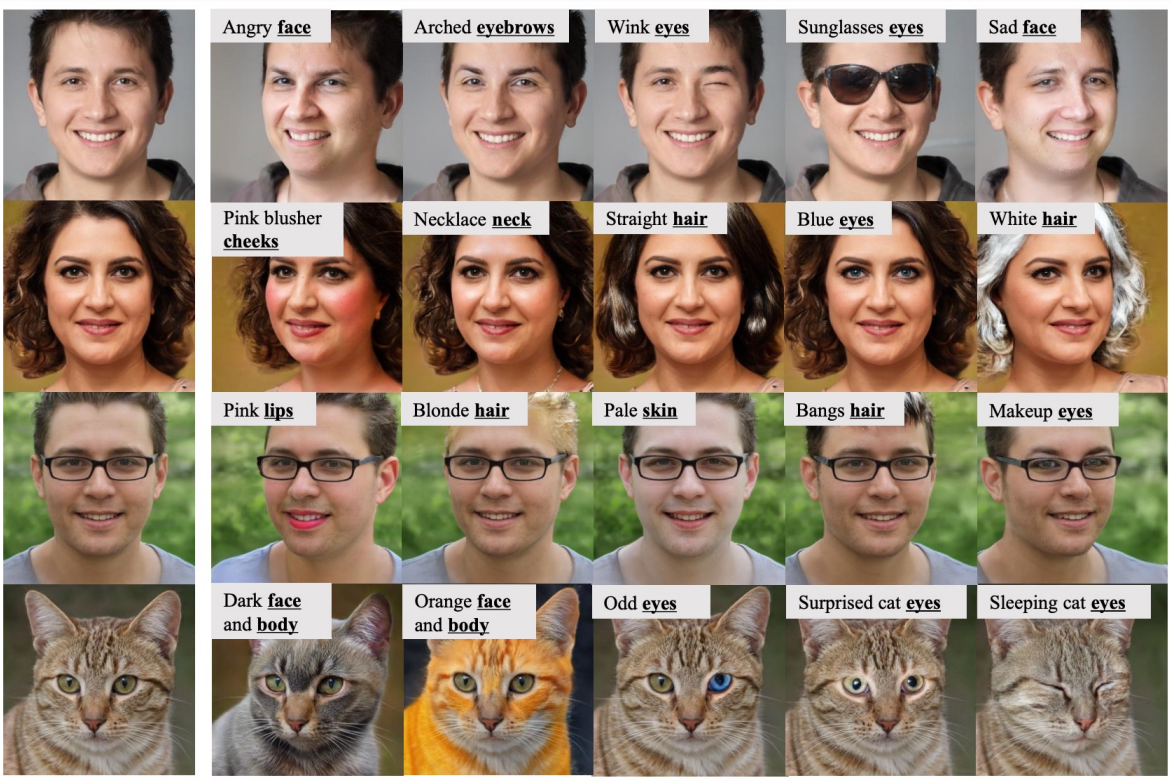

Figure 4. Curated Examples of LENeRF.

lation between the source and the target features. It is conditioned on both the source and target features, 3D position, and the corresponding latent codes, allowing the model to take into account both point semantics and global context:

$$A : \mathbb{R}^{2M_f+6} \times \mathbb{R}^{2d} \rightarrow [0,1], \\ m = A(F_s \oplus \mathbf{x}, F_t \oplus \mathbf{x}, \mathbf{w}_s, \mathbf{w}_t), \tag{5}$$

$$\hat{F} = (1-m) * F_s + m * F_t, \tag{6}$$

where F_s, $F_t \in \mathbb{R}^{M_f}$ are the source and target tri-plane feature vectors, each conditioned by $\mathbf{w}_s$, $\mathbf{w}_t \in \mathbb{R}^d$ respectively, and $m \in \mathbb{R}$ is the value of soft mask at $\mathbf{x}$.

For a position $\mathbf{x}$ inside the region of interest described by t_{mask}, AFN is trained to produce a high mask value m so that the estimated feature $\hat{F}$ is dominated by the target feature F_t. For a position $\mathbf{x}$ outside of our region of interest, low mask values are estimated to ensure that the source feature remains unchanged. Utilizing an estimated 3D soft mask enables 3D editing focused on desired regions, while avoiding unintended changes in other areas. Note that we do not use any domain-specific mask prediction model or semantic labels for AFN training. Instead, we show that it is possible to distill CLIP's capability to generate zero-shot 2D masks into 3D using our proposed loss function, as explained in Section 4.4.

4.3. Handling Large Deformations

Deformations are necessary for manipulations that involve large geometric changes (e.g., opening mouth), as interpolating between features representing different semantics (e.g., teeth and lips) often leads to artifacts (please refer to the supplement for figures). The deformation field for the source radiance field is estimated by the Deformation Network T conditioned with latent codes ($\mathbf{x}' = T(\mathbf{x}, \mathbf{w}_s, \mathbf{w}_t) + \mathbf{x}$), and Eq. 5 can be re-written as:

$$m = A(F_s \oplus \mathbf{x}', F_t \oplus \mathbf{x}, \mathbf{w}_s, \mathbf{w}_t). \tag{7}$$

4.4. Training

Generator The estimated features (Eq. 7) are decoded into a radiance field and rendered into a 2D image $\hat{I}_t$ via volume rendering (Eq. 1) followed by the super-resolution module π^{SR} (Eq. 2) given a randomly sampled camera pose $\mathbf{v} \sim \mathcal{Z}_v$, where $\mathcal{Z}_v$ is the camera pose distribution. Similarly, the source image I_s and the raw target image I_t are rendered using the same process with F_s and F_t.

CLIP Loss We use pretrained CLIP [39] text encoder E_T and image encoder E_I to optimize our modules so that the estimated image $\hat{I}_t$ matches t_{edit}. However, a direct approach to maximize the cosine similarity of $E_I(\hat{I}_t)$ and $E_T(t_{edit})$) frequently leads to degenerate solutions. This

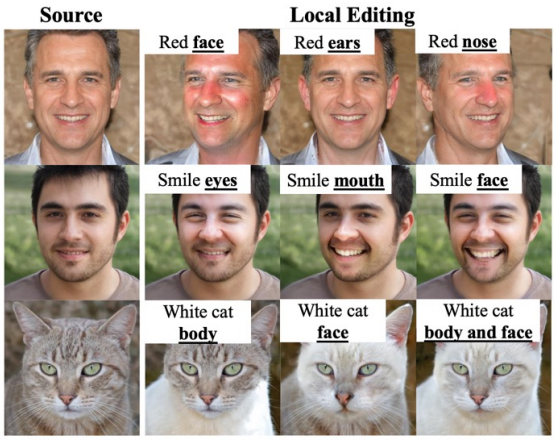

Figure 5. Results of partial editing.

is because the CLIP embedding space includes false positives, where degenerate images result in a high similarity score with the target text embedding. Therefore, we instead propose a robust $\mathcal{L}_{CLIP}$ loss that utilizes contrastive loss together with image augmentations:

$$\mathcal{L}_{CLIP} = -\log \frac{\sum_{s^+ \in S^+} e^{q \cdot s^+}}{\sum_{s \in S^+ \cup S} e^{q \cdot s}}, \tag{8}$$

where S^+ is a set of positive samples[3] and S^- is a set of negative samples[4]. Here, $q = E_I(aug(\hat{I}_t)) - E_I(aug(I_s))$, t_{src} is a neutral text describing the source image (e.g., "Photo of a person"), and aug refers to a random image augmentation. $\mathcal{L}_{CLIP}$ maximizes the mutual information between q and the positive samples while minimizing that of the negative samples. Comparing the directions of the CLIP embeddings and using multiple cross-domain positive samples lead to stable and robust optimizations, along with the negative samples inhibiting lazy manipulations.

Also, like [38], we incorporate an identity preserving loss [40] $\mathcal{L}_{id}$, using a pretrained ArcFace [10] model, and a L_2 norm of the manipulation step $M(\mathbf{w}_s)$ to preserve the attributes of the source radiance fields. The total loss for training the Latent Residual Mapper and the Deformation Network is $\mathcal{L}_{CLIP^+} = \mathcal{L}_{CLIP} + \lambda_{L2}\|M_(\mathbf{w}_s)\|_2 + \lambda_{id}\mathcal{L}_{id}$.

3D Distillation of CLIP's Knowledge As we lack ground-truth 3D semantic masks to supervise the 3D attention field, we render them into 2D and supervise them using generated 2D pseudo-masks. This is necessary since training AFN directly with $\mathcal{L}_{CLIP^+}$ often leads to the generation of irrelevant and poorly localized attention fields, resulting in severe artifacts (Please see supplement). This is not surprising, given that the target text prompt t_{edit} does not always

[3] $S^+ := \{E_T(t_{edit}) - E_I(aug(I_s)), E_T(t_{edit}) - E_T(t_{src})\}$
[4] $S^- := \{E_T(t_{edit}) - E_I(aug(I_s))\}$

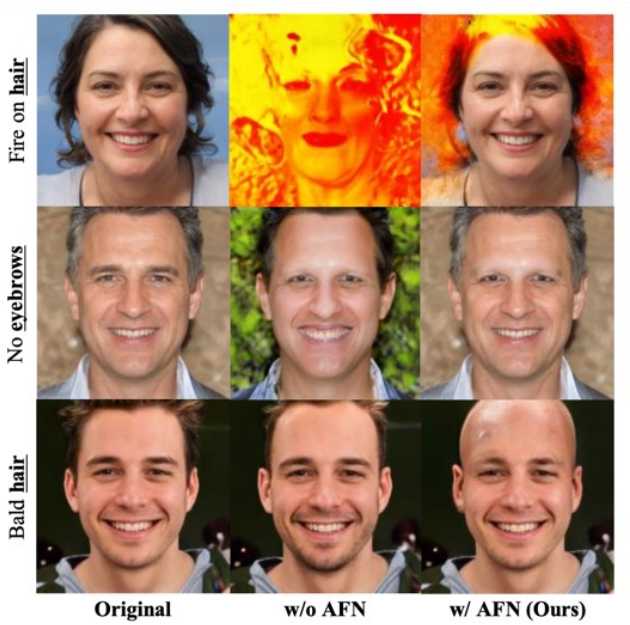

Figure 6. Results of out-of-distribution cases.

align with the region of interest, and the CLIP signal tends to have a global effect on the radiance fields. Furthermore, we presume that additional complexity to the model without adequate regularization easily leads to degenerate solutions.

Therefore we first introduce a separate masking prompt t_{mask} solely describing our region of interest. Next, we utilize t_{mask} together with the CLIP model to generate a 2D relevance mask of the source or target image that serves as a pseudo-label for AFN training. Inspired by transformer visualization methods [2, 6], we aggregate relevance maps across the transformer heads and layers to build a relevance mask $\mathbf{M} \in \mathbb{R}^{H_V \times W_V}$ with respect to the model output y, given t_{mask} (More details in the supplementary).

Unlike [6], where the transformer output y_t is the logit value of class t, our output y is the cosine similarity between the embedding vector $E_t(t_{mask})$ and $E_I(I)$. Generally, it is sufficient to use the source image I_s for the mask generation, but for large deformations, we estimate two masks from both the source and intermediate images I_s and I_t and perform a $\max$ operation.

The estimated relevance mask $\mathbf{M}$ is used as a pseudo-label to guide the training of AFN by minimizing the MSE loss $\mathcal{L}_{mask}$ between the volume-rendered attention field $\hat{\mathbf{M}}_t$ and $\mathbf{M}$. While the generated pseudo-label $\mathbf{M}$ is low-dimensional and often inaccurate, the semantic information contained in the input features to AFN allows for training even with noisy guidance (Figure 3 (d)).

Mask Regularization We utilize total variation regularization [41] (Eq. 9) to regularize the 3D attention field for spa-

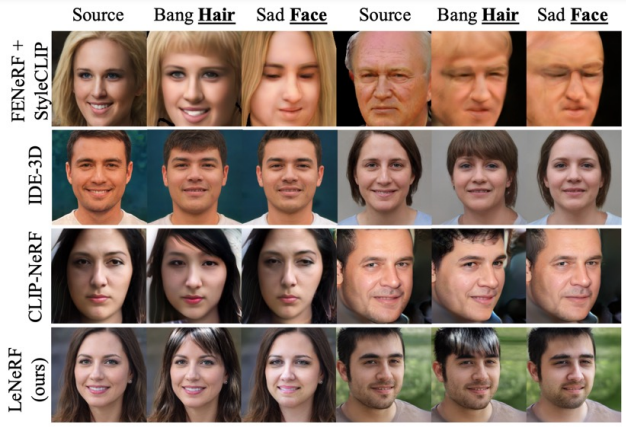

Figure 7. Results of LENeRF and various baselines.

tial smoothness, where $\Delta_x^2 m$ denotes the squared difference between $m_{i+1,j,k}$ and $m_{i,j,k}$, and analogously for $\Delta_y^2 m$ and $\Delta_z^2 m$:

$$\mathcal{L}_{tv} = \frac{1}{N} \sum_{m \in \mathcal{A}} \sqrt{\Delta_x^2 m + \Delta_y^2 m + \Delta_z^2 m}. \tag{9}$$

Also, the sparsity regularization (Eq. 10) is used in order to focus on the region of interest:

$$\mathcal{L}_{sparsity} = - \sum_{\mathbf{x}_i \in K_{top}} \log m_i - \sum_{\mathbf{x}_i \in K_{bottom}} log(1 - m_i), \tag{10}$$

where K_{top} refers to the set of coordinates with the top-k m values, and K_{bottom} refers to the bottom-k. In summary, the objective function for AFN A is

$$\begin{aligned} \mathcal{L}_{AFN} = \lambda_{mask} \mathcal{L}_{mask} + \lambda_{tv} \mathcal{L}_{tv} \\ + \lambda_{sparsity} \mathcal{L}_{sparsity} + \lambda_{CLIP^+} \mathcal{L}_{CLIP^+}. \end{aligned} \tag{11}$$

Please refer to the supplement for additional descriptions, hyperparameters, and implementation details.

5. Experiments

Datasets LeNeRF utilizes pretrained EG3D [4] generator trained on FFHQ [23], AFHQv2 CATS [9], and ShapeNet Cars [8, 28]. However, please note that the datasets are not used when training LENeRF.

Baselines We compare LENeRF against three state-of-the-art NeRF editing methods: CLIP-NeRF [48][5], FENeRF [45], and IDE-3D [44]. While CLIP-NeRF is a text-guided editing model, FENeRF and IDE-3D are not directly comparable to our method since semantic masks are used as a control handle. Therefore we create two new baselines

[5]Note that the full code of CLIP-NeRF has not been released at the time of writing. Therefore, the results are our reproduced CLIP-NeRF.

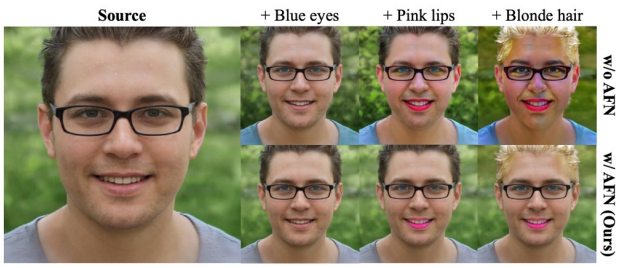

Figure 8. Result of sequential editing.

based on the two models (FENeRF+StyleCLIP and IDE-3D+StyleCLIP) which can edit radiance fields with text guidance. Specifically, we invert the images generated by the two methods to StyleGAN latent codes using e4e [47] encoder. StyleCLIP [38] is used to edit the inverted images given a target text prompt. Face semantic maps are then extracted from the edited images using a pretrained face parsing network of SofGAN [7], which is used for the input guidance. Note that only FFHQ results can be reported for FENeRF and IDE-3D since there are no available semantic labels and pretrained parsing networks for other datasets.

Qualitative evaluations Fig. 4 provides curated results of our model, showing the quality, diversity, and view consistency of LENeRF. Fig. 7 presents qualitative comparisons against other three methods. FENeRF+StyleCLIP and CLIP-NeRF fail to generate high-quality renderings, and the quality of images degrades even further after editing. IDE-3D+StyleCLIP synthesizes high-fidelity images before editing, but it is difficult to edit texture or expression using semantic mask guidance. Also, semantic mask-based methods require either inversion steps which cannot be done in real-time or an encoding step which degrades identity preservation. LENeRF, however, can perform localized editing in real-time and preserve the quality and the identity of original content.

Fig. 5 shows partial editing results where we use the same attribute prompt t_{edit} while varying the mask prompt t_{mask}. This enables highly localized editing without unwanted changes or color-bleeding artifacts. Fig. 8 demonstrates sequential editing results where we perform different manipulations sequentially. Multiple changes to the source latent code destroy the initial content gradually when trained without AFN (bottom) whereas, the identity is preserved in LENeRF (top) by manipulating only the local parts of the radiance field. Fig. 6 shows that LENeRF is robust to out-of-distribution editing (e.g., fire on hair). LENeRF has the representation capability to generate samples that are outside of the latent space by fusing the two radiance fields, thereby reflecting the target prompts while preserving the identity.

Quantitative evaluations Table 1 provides quantitative

Table 1. PSNR, R-precision, and FID performance of various baselines and LENeRF on FFHQ and Cats.

	FFHQ				Cats	
	PSNR ↑	R-pre ↑	FID ↓	ΔFID ↓	FID ↓	ΔFID ↓
CLIP-NeRF	7.08	0.21	41.5	+19.8	18.6	+12.7
FeNeRF + SC	5.33	0.19	37.4	+31.2	-	-
IDE3D + SC	12.44	0.61	4.56	+3.29	-	-
LeNeRF w/o AFN	10.81	0.66	**4.37**	+4.90	**2.71**	+4.17
LeNeRF (Ours)	**20.86**	**0.78**	**4.37**	**+2.21**	**2.71**	**+1.94**

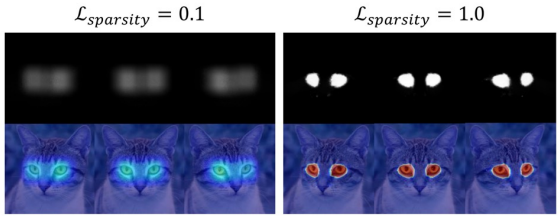

Figure 9. Results of the effect of $\lambda_{sparsity}$ on the mask.

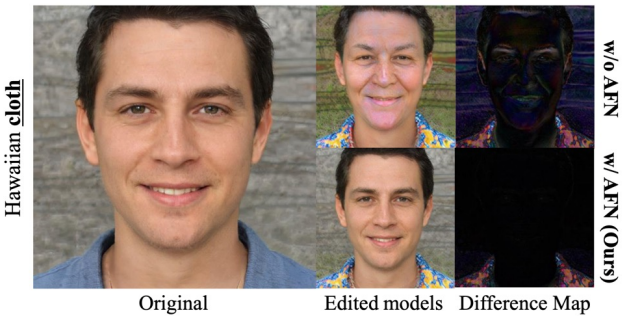

Figure 10. Results on comparison of pixel-wise difference maps.

Figure 11. Importance of our objective functions $\mathcal{L}_{CLIP+}$, $\mathcal{L}_{mask}$, and Deformation Network (DN).

measures of LENeRF and various baselines. We quantify the unintended changes by computing the PSNR between the source and the manipulated images outside of the region of interest specified by t_{mask}. The ground-truth mask of the region is detected by a pretrained face parsing network [7]. We assess the quality of the source images and manipulated images using Fréchet Inception Distance (FID) [18], and compare the differences to evaluate post-editing degradation. Also for FFHQ, we report R-precision using a pretrained attribute classifier [30] to measure how well the text prompt is reflected in the manipulated results. We generate 30k samples with randomly sampled text prompts to calculate FID, and sample 1K images per text prompt to calculate PSNR and R-precision.

Among the models, LENeRF shows the best FID score, demonstrating that it produces high-quality results after the manipulation. Also, LENeRF have the smallest FID difference between the images before and after the edit, indicating the robustness of the model. Moreover, LENeRF shows the highest PSNR and R-precision; it has the best performance in achieving the desired manipulations while minimizing unintended changes. Please refer to the supplement for more details.

User study We ask users to evaluate LENeRF along with various baselines in the range from 1 to 10 regarding 1) fidelity, 2) locality, 3) identity preservation, and 4) how well the text prompt is reflected in the results. LENeRF outperforms all baselines by a large margin on each criterion, and the scores are in the supplement.

Ablation study Fig. 9 shows the effect of controlling $\lambda_{sparsity}$. A larger weight on $\mathcal{L}_{sparsity}$ produces a sharper 3D mask whereas a smaller weight leads to a smoother mask. Fig. 10 visualizes the pixel-wise difference map to show the effect of AFN and the radiance fusion technique on the locality. Generating a 3D mask helps on minimize the unintended changes outside of the region of interest and reduces color-bleeding artifacts that are incorporated with the target prompt t_{edit}. Fig. 11 shows the importance of the DN and our objective functions $\mathcal{L}_{CLIP+}$ and $\mathcal{L}_{mask}$. Minimizing the naive CLIP loss results in degenerate solutions. Training AFN without $\mathcal{L}_{mask}$ fails to estimate a 3D mask, resulting in severe artifacts. Finally, LENeRF without DN leads to the interpolation of unrelated features and cannot handle geometric manipulations.

6. Conclusion

Conclusion While texts provide rich and descriptive information for editing, they often lack locality. LENeRF focuses on this aspect and proposes a point-wise feature manipulation technique via 3D mask generation and radiance fusion. We show that 3D-aware knowledge of NeRF and rich multi-modal information from CLIP can be combined to create a robust 3D mask without any additional datasets. LENeRF generates high-fidelity and find-grained 3D editing results.

Acknowledgments. This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2019-0-00075, Artificial Intelligence Graduate School Program (KAIST) and No. 2021-0-01778, Development of human image synthesis and discrimination technology below the perceptual threshold).

References

[1] Rameen Abdal, Peihao Zhu, John Femiani, Niloy J. Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In Munkhtsetseg Nandigjav, Niloy J. Mitra, and Aaron Hertzmann, editors, *SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference, Vancouver, BC, Canada, August 7 - 11, 2022*, pages 48:1–48:9. ACM, 2022. 4

[2] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015. 6

[3] Zehranaz Canfes, M. Furkan Atasoy, Alara Dirik, and Pinar Yanardag. Text and image guided 3d avatar generation and manipulation. *CoRR*, abs/2202.06079, 2022. 3

[4] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16102–16112. IEEE, 2022. 2, 3, 4, 7

[5] Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. Pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5799–5809. Computer Vision Foundation / IEEE, 2021. 2, 3

[6] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 782–791. Computer Vision Foundation / IEEE, 2021. 6

[7] Anpei Chen, Ruiyang Liu, Ling Xie, Zhang Chen, Hao Su, and Jingyi Yu. Sofgan: A portrait image generator with dynamic styling. *ACM Trans. Graph.*, 41(1):1:1–1:26, 2022. 7, 8

[8] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5939–5948. Computer Vision Foundation / IEEE, 2019. 7

[9] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8185–8194. Computer Vision Foundation / IEEE, 2020. 2, 7

[10] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):5962–5979, 2022. 6

[11] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. GRAM: generative radiance manifolds for 3d-aware image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10663–10673. IEEE, 2022. 2, 3

[12] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Trans. Graph.*, 41(4):141:1–141:13, 2022. 4

[13] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2414–2423. IEEE Computer Society, 2016. 2

[14] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas A. Funkhouser. Local deep implicit functions for 3d shape. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 4856–4865. Computer Vision Foundation / IEEE, 2020. 2

[15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 3

[16] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 2, 3

[17] Benoît Guillard, Edoardo Remelli, Pierre Yvernay, and Pascal Fua. Sketch2mesh: Reconstructing and editing 3d shapes from sketches. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 13003–13012. IEEE, 2021. 2

[18] 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. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017. 8

[19] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: zero-shot text-driven generation and animation of 3d avatars. *ACM Trans. Graph.*, 41(4):161:1–161:19, 2022. 3

[20] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 857–866. IEEE, 2022. 3

[21] Chiyu "Max" Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, and Thomas A. Funkhouser. Local implicit grid representations for 3d scenes. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19,*

2020, pages 6000–6009. Computer Vision Foundation / IEEE, 2020. 2

[22] Kacper Kania, Kwang Moo Yi, Marek Kowalski, Tomasz Trzcinski, and Andrea Tagliasacchi. Conerf: Controllable neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18602–18611. IEEE, 2022. 3

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4401–4410. Computer Vision Foundation / IEEE, 2019. 7

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 3

[25] Daejin Kim, Mohammad Azam Khan, and Jaegul Choo. Not just compete, but collaborate: Local image-to-image translation via cooperative mask prediction. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2021. 2

[26] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *CoRR*, abs/2205.15585, 2022. 3

[27] Adam R. Kosiorek, Heiko Strathmann, Daniel Zoran, Pol Moreno, Rosalia Schneider, Sona Mokrá, and Danilo Jimenez Rezende. Nerf-vae: A geometry aware 3d scene generative model. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5742–5752. PMLR, 2021. 3

[28] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 7

[29] Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. Editing conditional radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 5753–5763. IEEE, 2021. 2, 3

[30] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3730–3738. IEEE Computer Society, 2015. 8

[31] Éloi Mehr, Ariane Jourdan, Nicolas Thome, Matthieu Cord, and Vincent Guitteny. Disconet: Shapes learning on disconnected manifolds for 3d editing. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3473–3482. IEEE, 2019. 2

[32] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13482–13492. IEEE, 2022. 3

[33] 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. In *ECCV*, 2020. 2, 3

[34] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7587–7596. IEEE, 2019. 3

[35] Thu Nguyen-Phuoc, Christian Richardt, Long Mai, Yong-Liang Yang, and Niloy J. Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3

[36] Michael Niemeyer and Andreas Geiger. GIRAFFE: representing scenes as compositional generative neural feature fields. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11453–11464. Computer Vision Foundation / IEEE, 2021. 2, 3

[37] Jeong Joon Park, Peter Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 165–174. Computer Vision Foundation / IEEE, 2019. 2

[38] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, October 2021. 2, 4, 6, 7

[39] 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. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 2, 3, 4, 5

[40] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: A stylegan encoder for image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2287–2296. Computer Vision Foundation / IEEE, 2021. 6

[41] Leonid I. Rudin and Stanley J. Osher. Total variation based image restoration with free local constraints. In *Proceed-*

ings 1994 International Conference on Image Processing, Austin, Texas, USA, November 13-16, 1994, pages 31–35. IEEE Computer Society, 1994. 6

[42] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18582–18592. IEEE, 2022. 3

[43] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: generative radiance fields for 3d-aware image synthesis. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2, 3

[44] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. IDE-3D: interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *CoRR*, abs/2205.15517, 2022. 2, 3, 7

[45] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. Fenerf: Face editing in neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7672–7682, June 2022. 2, 3, 7

[46] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. Fenerf: Face editing in neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7662–7672. IEEE, 2022. 3

[47] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *arXiv preprint arXiv:2102.02766*, 2021. 3, 7

[48] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3825–3834. IEEE, 2022. 2, 3, 4, 7

[49] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Towards open-world text-guided face image generation and manipulation. *CoRR*, abs/2104.08910, 2021. 4

[50] Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. GRAM-HD: 3d-consistent image generation at high resolution with generative radiance manifolds. *CoRR*, abs/2206.07255, 2022. 2, 3

[51] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jiahui Zhang, Shijian Lu, Miaomiao Cui, Xuansong Xie, Xian-Sheng Hua, and Chunyan Miao. Towards counterfactual image manipulation via CLIP. In João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni, editors, *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 3637–3645. ACM, 2022. 4

[52] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. CIPS-3D: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *CoRR*, abs/2110.09788, 2021. 2, 3

[53] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251. IEEE Computer Society, 2017. 2

[54] Matthias Zwicker, Mark Pauly, Oliver Knoll, and Markus H. Gross. Pointshop 3d: an interactive system for point-based surface editing. *ACM Trans. Graph.*, 21(3):322–329, 2002. 2