Reference-Based 3D-Aware Image Editing with Triplanes

Bahri Batuhan Bilecen[1] Yigit Yalin[1] Ning Yu[2] Aysegul Dundar[1]
Bilkent University[1] Netflix Eyeline Studios[2]
{batuhan.bilecen@, yigit.yalin@ug., adundar@cs.}bilkent.edu.tr, ning.yu@scanlinevfx.com

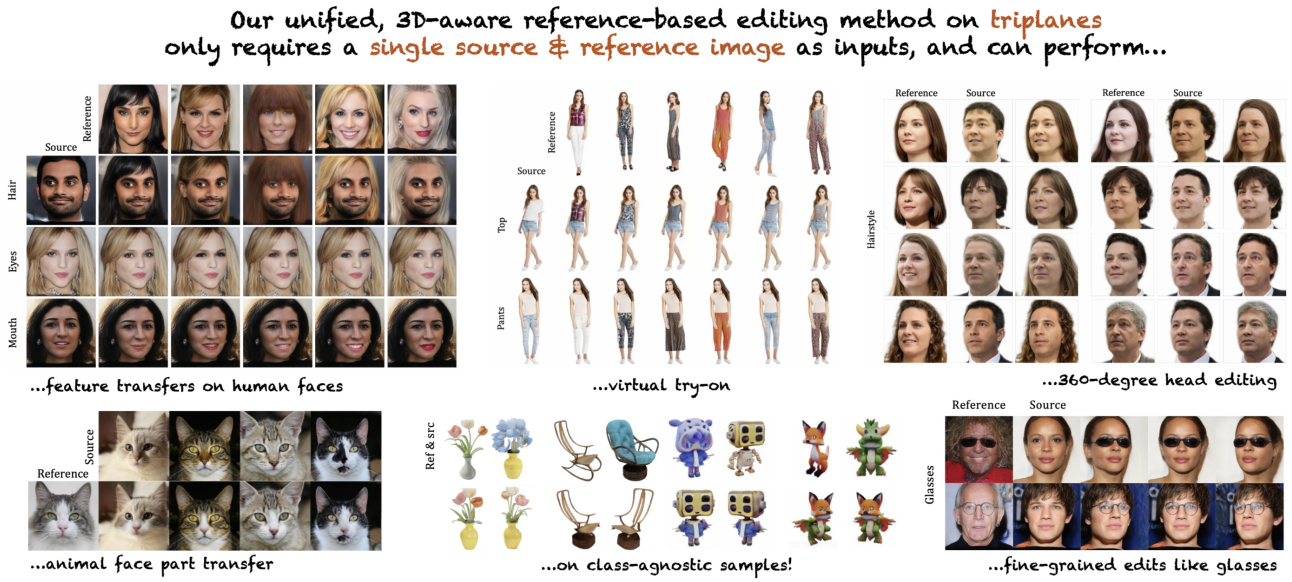

Figure 1. Our approach excels in refer ence-based edits, faithfully reproducing the copied reference parts with a **single** source and reference image. Leveraging 3D-aware triplanes, our edits are versatile and 3D consistent, allowing for rendering from various viewpoints. We show results on human faces, heads, bodies, and extending beyond to animal faces and class-agnostic samples.

Abstract

Generative Adversarial Networks (GANs) have emerged as powerful tools for high-quality image generation and real image editing by manipulating their latent spaces. Recent advancements in GANs include 3D-aware models such as EG3D, which feature efficient triplane-based architectures capable of reconstructing 3D geometry from single images. However, limited attention has been given to providing an integrated framework for 3D-aware, high-quality, reference-based image editing. This study addresses this gap by exploring and demonstrating the effectiveness of the triplane space for advanced reference-based edits. Our novel approach integrates encoding, automatic localization, spatial disentanglement of triplane features, and fusion learning to achieve the desired edits. We demonstrate how our approach excels across diverse domains, including human faces, 360-degree heads, animal faces, partially stylized edits like cartoon faces, full-body clothing edits, and edits on class-agnostic samples. Our method shows state-of-the-art performance over relevant latent direction, text, and image-guided 2D and 3D-aware diffusion and GAN methods, both qualitatively and quantitatively.[1]

1. Introduction

In recent years, the high-fidelity image synthesis performance has been profoundly transformed by the emergence of Generative Adversarial Networks (GANs) [24, 34, 35]. Through adversarial training, they learn to map random distributions to actual data observations, enabling the generation of photo-realistic images from latent codes. The evolution of GANs from 2D into 3D-aware [9, 10, 25, 51] has further boosted this capability, integrating hybrid 3D representations and Neural Radiance Fields (NeRF) [49] into style-based generators, yielding unparalleled success in crafting highly realistic 3D portraits.

The application of GANs extends beyond generation, venturing into real image editing through GAN inversion [3,

[1] three-bee.github.io/triplane_edit

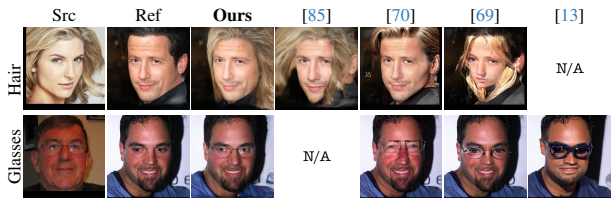

Figure 2. Current methods struggle with 3D consistency [63, 70, 85], faithfulness to the reference [8, 33, 69], and visual artifacts [69, 70]. Our method provides 3D-consistent, reference-based edits from single images, independent of camera poses. N/A indicates the model is incapable of such edits.

56, 62, 74, 76] and manipulation of the latent space [59]. A critical challenge in reference-based image editing via this latent space is striking the optimal balance between retaining essential elements from the input image and incorporating desired attributes from the reference image. This balance is crucial to avoid overshadowing the input image's essence in the editing process, thereby maintaining its identity while still transferring the reference's attributes. This presents itself as a complex problem of disentanglement and fusion, requiring the identification and integration of relevant feature components from both the reference and input images within the latent space to produce the edited output.

Despite the advancements in image editing, there is a noticeable gap in the development of reference-based, 3D-aware image editing techniques. Existing methods in 3D-aware image editing lack reference-based capabilities [23, 39, 76], while current GAN and diffusion reference-based editing approaches do not support 3D-awareness [63, 70, 85] and/or facilitate local editing [7, 8, 13, 53, 59] (Figs. 2 and 6). Addressing this gap, our work focuses on pioneering reference-based, 3D-aware image editing, where we learn spatial disentanglement and fusion within triplane latent spaces [10, 20]. Our primary motivation stems from discovering that triplanes can be manipulated for editing purposes akin to the 2D image domain but offer distinct advantages. Triplanes not only facilitate 3D editing but also alleviate alignment issues inherent in 2D image space. For example, transferring eyeglasses from one person to another in 2D necessitates precise alignment in the image space. Conversely, in the triplane space, stitching is facilitated as images with varying camera parameters can be projected onto the same canonical triplane space. However, ensuring seamless boundaries requires careful attention. To accomplish reference-based image editing, our framework localizes parts within triplanes using masked residual gradients and fuses them using the encoders we train.

Our contributions span the following:

1. We are at the forefront of conceptualizing reference-based 3D-aware image editing as an integrated framework by leveraging the power of triplanes. Our approach includes encoding triplane features, spatial disentanglement with automatic localization of features, and fusion learning for desired image editing.
2. Our work establishes new benchmarks for quantitative and qualitative assessment in reference-based image editing and surpasses the 12 most recent and relevant baseline editing methods. This advancement is quantified by significant improvements in FID for quality and masked pixel-wise metrics for source preservation.
3. We extend our framework based on [10] towards 360-degree human heads [4], animal faces [10], cross-domain edits with cartoon portraits [60], full-body [20] clothing edits, and edits on class-agnostic samples [29, 37, 68], showcasing its versatility and robustness across different triplane domains as shown in Fig. 1. This not only proves the effectiveness of our method but also broadens its potential in creative editing contexts.

2. Related Works

StyleGAN inversion-based editing. It has been shown that well-trained GAN models organize their latent space in a semantically meaningful way that enables edits via latent vector arithmetic. Especially for StyleGAN models [35], many methods are proposed to find interpretable directions in unsupervised ways such as GANSpace [27], StyleFlow [2], StyleSpace [64], StyleCLIP [53] and supervised ways such as InterFaceGAN [59], and recently E3DGE [36], Barbershop [85], HairCLIPv2 [63], and SFE [7]. These methods are combined with real image inversion methods so that an image is projected onto StyleGAN's latent space ($\mathcal{W}$, $\mathcal{W}^+$, $\mathcal{F}$, etc.) and is edited [54, 56]. Recently, EG3D [10] augmented StyleGAN architecture with triplanes that provide efficient 3D-aware representations. For this new domain, new inversion methods are proposed [23, 67, 75, 76]; however, previous editing methods [53, 59] are re-used from 2D StyleGAN literature, which all lack reference-based editing.

Reference-based image-to-image translation. We refer to the models that are end-to-end trained for editing applications as image-to-image translation methods. These methods are trained to change selected attributes of images while preserving the content [12, 21, 30, 40, 41, 47, 58, 66, 77], rather than using the latent codes of StyleGAN. Specifically, they include an encoder for the reference image and another encoder or shared one for the source image [12, 31, 40, 48, 71, 83, 84]. In these models, the reference style can be sampled from a normal distribution, or it can be encoded from a reference image [14, 15, 40]. One major limitation of these works is that they require labeled datasets for each attribute [45]; hence, models like VecGAN [15] and HisD [40] can only achieve a handful of edits (hair color, smile, eyeglasses, bangs, etc.). They also cannot provide editing using images taken from vastly different camera poses and lack faithfulness to the reference.

Text and image-conditioned diffusion models for image editing. There is an abundance of diffusion-based image editing methods [8, 13, 19, 61, 69, 70], leveraging the prior of Stable Diffusion (SD) denoiser [57] and achieving diverse edits. Recently, methods like ControlNet [78], IP [73], and T2I [50] have made editing more controllable. Still, all these SD-based methods lack 3D-aware, faithful-to-reference image editing, especially in the face and body domain. LEDITS++ [8] and InfEdit [69] utilize text prompts and are incapable of image reference-based editing. Paint by Example [70] fine-tunes the denoiser with masked images to carry information from reference images. Nevertheless, [70] is not capable of choosing which region to copy, harming the controllability - more noticeably when camera poses are different. NoiseCLR [13] attempts to find editing directions unsupervised but fails for fundamental face-edit attributes like hair color and style changes. There are 3D-aware denoisers [43, 44, 72] and edit methods [22, 52, 81], but they also become impractical in our context, as they are either fine-tuned with general object datasets [16, 17] failing to perform well in the face domain or lack faithful reference-based editing altogether. There are also methods fusing GAN & diffusion for adding stylization and diversity [5, 26, 39, 60], but controllable reference-based edits are not achieved in those, either.

3. Method

Our main motivation lies in that triplanes can be stitched and blended for editing like in the 2D image domain. However, achieving satisfying results requires carefully designed steps, which will be detailed in this section.

3.1. Localizing parts in the triplane space

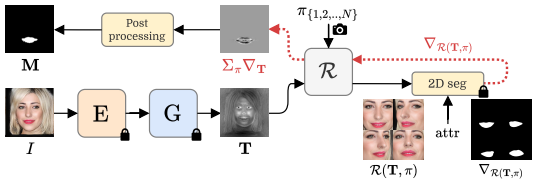

Figure 3. Triplane part localization stage, where $\mathbf{E}$, $\mathbf{G}$, and $\mathcal{R}$ are encoder [6, 76], generator [10], and neural volumetric renderer, respectively. For the 2D segmentation model $\mathcal{S}_{2D}$, we use state-of-the-art off-the-shelf segmentation models [1]. Images other than the input image are zoomed in for visualization purposes.

The most direct way to perform a region transfer from reference to source image would be to mask and copy & paste the region of interest. However, this process becomes intricate in hybrid 3D representations like triplanes, due to the absence of a conventional method for identifying the regions to be masked.

Algorithm 1 Triplane localization and masking

Require: Generator $\mathbf{G}$, encoder $\mathbf{E}$, renderer $\mathcal{R}$, image I, extrinsic matrices $\pi_{1,2,...,N}$, segmentation net $\mathcal{S}_{2D}$, post-processing params (ϵ, γ)
Ensure: Triplane mask $\mathbf{M}$

$\mathbf{T}, \nabla_{\mathbf{T}}, \leftarrow \mathbf{G}(\mathbf{E}(I)), 0$ ▷ *Initialize triplane and gradients*
for each π_i in $\pi_{1,2,...,N}$ **do** ▷ *Triplane localization*
  $I_{\pi_i} \leftarrow \mathcal{R}(\mathbf{T}, \pi_i)$
  $\nabla_{\mathcal{R}(\mathbf{T},\pi_i)} \leftarrow \mathcal{S}_{2D}(I_{\pi_i}, \mathbf{attr})$
  $\nabla_{\mathbf{T}} \leftarrow \nabla_{\mathbf{T}} + \mathbf{autograd}(\mathcal{R}, \nabla_{\mathcal{R}(\mathbf{T},\pi_i)})$
end for

for each c in $\nabla_{\mathbf{T}}$ **do** ▷ *Post-processing*
  $\nabla_{\mathbf{T}_c} \leftarrow \{x \in \nabla_{\mathbf{T}_c} : |\nabla_{\mathbf{T}_c}(x) - \mu_c| \leq \epsilon\}$
  $\nabla_{\mathbf{T}_c} \leftarrow (\nabla_{\mathbf{T}_c} - \min \nabla_{\mathbf{T}_c})/(\max \nabla_{\mathbf{T}_c} - \min \nabla_{\mathbf{T}_c})$
  $\nabla_{\mathbf{T}_c} \leftarrow \{x \in \nabla_{\mathbf{T}_c} : 1 \textbf{ if } \nabla_{\mathbf{T}_c}(x) \geq \gamma \textbf{ else } 0\}$
end for

return $\mathbf{M} \leftarrow \nabla_{\mathbf{T}}$

To address this, we take advantage of the volumetric renderer used in [10] being fully differentiable, and back-propagate the 2D image domain masks to the 3D hybrid triplane domain to calculate gradients on triplanes (Fig. 3, *triplane localization* in Algorithm 1). First, input images are encoded using a pre-trained model [6, 76] to obtain triplane features $\mathbf{T}$. These features are rendered with different camera poses $\pi_{1,2,...,N}$ to create multi-view 2D renderings $\mathcal{R}(\mathbf{T}, \pi)$. An off-the-shelf segmentation network $\mathcal{S}_{2D}$ identifies attributes [1] (e.g., hair, eyes, glasses) in each rendering. The segmentation outputs are assigned as output gradients $\nabla_{\mathcal{R}(\mathbf{T},\pi)}$, and are back-propagated to the triplane domain to accumulate input gradients $\Sigma_\pi \nabla_{\mathbf{T}}$, which localizes the triplane mask.

To convert a gradient mask into a binarized one, we perform mean clipping, normalization, and thresholding with parameters (ϵ, γ) (*post-processing* in Algorithm 1). These parameters are set once and used across all experiments for each attribute and domain, and are in the Supplementary. This localization is done for source and reference images.

For finer granularity and to avoid copying unwanted attributes, such as copying glasses and not the eyes, $\mathcal{W}^+$ directions [59] can be utilized in addition. By computing the difference between triplanes with and without the attribute, $\Delta\mathbf{T}_{\text{attr}} = \mathbf{G}(w) - \mathbf{G}(w - w_{\text{attr}})$ and multiplying with gradient mask $\mathbf{M}$, a more precise mask can be created.

3.2. Implicit fusion by encoding & decoding

After finding suitable masks for reference and source triplanes, $\mathbf{M}_{\text{ref}}$ and $\mathbf{M}_{\text{src}}$ respectively, a naive approach would be to follow Eq. (1):

$$\mathbf{T}_{\text{tmp}} = \mathbf{M}_{\text{ref}} * \mathbf{T}_{\text{ref}} + \mathbf{M}_{\text{src}} * \mathbf{T}_{\text{src}} \quad (1)$$

However, Fig. 8 (V1) reveals only using Eq. (1) distorts the geometry and color consistency, and creates stitching seams

around the editing borders. In addition, in some cases, the contents in $\mathbf{T}_{\text{ref}}$ and $\mathbf{T}_{\text{src}}$ are not enough to ensure the editing looks natural around the region where two masks meet, necessitating the hallucination of additional content to complement the editing. For example, we may want to remove the long hair from the source image and replace it with the short hair from the reference image. In this case, the pixels that correspond to long hair regions need to be inpainted.

Given the above observations and GAN's latent spaces embed natural images, we render the naively fused triplane with the canonical pose π_{can}, re-encode $\mathbf{E}_{\mathcal{W}^+}$, and re-decode via the generator $\mathbf{G}$ to obtain an implicitly fused triplane, shown in Eq. (2):

$$\mathbf{T}_{\text{imp}} = \mathbf{G}(\mathbf{E}_{\mathcal{W}^+}(\mathcal{R}(\mathbf{T}_{\text{tmp}}, \pi_{\text{can}}))) \tag{2}$$

Note that state-of-the-art image inversion methods for EG3D [6, 76] employ both low-rate $\mathcal{W}^+$ and high-rate $\mathcal{F}$ features. The latter is crucial for reconstructing fine image details. *However, our objective in this phase is not to achieve perfect image reconstruction. On the contrary, we aim for the encoder to map the image to a latent space with natural reconstructions.* To achieve this, we disable the high-frequency restoration branch of $\mathbf{E}_{\mathcal{W}^+}$ to prevent encoding visible seams. This allows us to project the edited image onto its nearest representation on $\mathbf{G}$'s manifold, implicitly fusing the masked triplanes. Fig. 4 showcases seamless boundaries across the stitches after this operation.

Note that image details are compromised at this stage, as we solely depend on the low-rate $\mathcal{W}^+$ space. To bring the details back, we only employ $\mathbf{T}_{\text{imp}}$ at the transition regions, as depicted in Eq. (3). For instance, when we transfer the mouth from reference to source, we aim for the triplane features outside the mouth to originate from $\mathbf{T}_{\text{src}}$, the mouth features from $\mathbf{T}_{\text{ref}}$, and features near mouth from $\mathbf{T}_{\text{imp}}$.

$$\begin{aligned}\mathbf{T}_{\text{f}} = \mathcal{E}(\mathbf{M}_{\text{ref}}) * \mathbf{T}_{\text{ref}} + \mathcal{E}(\mathbf{M}_{\text{src}}) * \mathbf{T}_{\text{src}} \\ + (\mathcal{E}(\mathbf{M}_{\text{src}}) - \mathcal{E}(\mathbf{M}_{\text{ref}})) * \mathbf{T}_{\text{imp}}\end{aligned} \tag{3}$$

Here, $\mathcal{E}$ denotes morphological erosion. We also apply Gaussian blurring with parameters (μ, σ) onto the masks to avoid sharp edges. This step is not illustrated in Fig. 4 for brevity, but (μ, σ) are provided in the Supplementary.

Finally, $\mathbf{T}_{\text{f}}$ is rendered from any desired pose π_R, and the final image is obtained, as illustrated in Eq. (4).

$$I_{\text{edited}} = \mathcal{R}(\mathbf{T}_{\text{f}}, \pi_R) \tag{4}$$

3.3. Fine-tuning the image encoder

Although $\mathbf{T}_{\text{imp}}$ obtained in Sec. 3.2 helps tremendously even though a pre-trained encoder is used to obtain it, we notice in some cases where we have skin color inconsistencies, background leakages, and missing high-frequency details around the editing regions (Fig. 8 (V2)). Hence, we fine-tune implicit fusion encoder $\mathbf{E}_{\mathcal{W}^+}$, jointly with the triplane editing pipeline to mitigate the aforementioned effects.

During the fine-tuning phase, we generate renders as ground-truths from various viewpoints of the source and reference images corresponding to different attributes. Then, we employ masked losses to guide $\mathbf{E}_{\mathcal{W}^+}$ in encoding only the visible regions, illustrated in Fig. 5. For instance, if our objective is to transfer hair, we mask the reference renderings to exclude pixels that do not represent hair while doing the opposite for the source ground truth. To ensure that losses do not affect the boundaries, we dilate the segmentation masks. The objective function is provided in Eq. (5):

$$\begin{aligned}\min_{\mathbf{E}} \lambda_{\Phi}\mathcal{L}_{\Phi}(\mathcal{D}(\tilde{\mathbf{M}}_{\text{ref}}) * \mathcal{R}(\mathbf{T}_{\text{f}}, \pi_i), \mathcal{D}(\tilde{\mathbf{M}}_{\text{ref}}) * \mathcal{R}(\mathbf{T}_{\text{ref}}, \pi_i)) \\ +\lambda_{\text{ID}}\mathcal{L}_{\text{ID}}(\mathcal{D}(\tilde{\mathbf{M}}_{\text{src}}) * \mathcal{R}(\mathbf{T}_{\text{f}}, \pi_i), \mathcal{D}(\tilde{\mathbf{M}}_{\text{src}}) * \mathcal{R}(\mathbf{T}_{\text{src}}, \pi_i))\end{aligned} \tag{5}$$

where $\mathcal{L}_{\Phi}$ is the learned perceptual image patch similarity loss (LPIPS) [82], $\mathcal{L}_{\text{ID}}$ is the identity similarity loss [18], $\mathcal{D}$ is dilation operation, and $\tilde{\mathbf{M}}_{\text{src}}$ and $\tilde{\mathbf{M}}_{\text{ref}}$ are the corresponding 2D segmentation masks for the rendered images $\mathcal{R}(\mathbf{T}_{\text{src}}, \pi_i)$ and $\mathcal{R}(\mathbf{T}_{\text{ref}}, \pi_i)$ with poses π_i, respectively.

To achieve identity preservation of the source image, we rely on the ID losses, and to copy the attribute with details, we rely on the LPIPS score from Eq. (5). We omit pixel-wise losses like $\mathcal{L}_2$ since they are highly dependent on the quality of the off-the-shelf 2D segmentation network. During the fine-tuning, we use each subject tuple $(\mathbf{T}_{\text{src}}, \mathbf{T}_{\text{ref}})$ multiple times, rendered from N randomly chosen π_i's.

Compliant with the common encoder training methodology, we use the same training datasets generators are trained with. We employ Ranger optimizer, which is a combination of Rectified Adam [42] with Lookahead [80]. The learning rate is set to $1e^{-4}$, and fine-tuning is done for 1500 steps with a batch size of 2 on a single RTX 4090.

4. Experiments

Evaluation. We present metrics evaluating both the reconstruction and editing qualities of our approach. For editing assessment, we employ the Fréchet Inception Distance (FID) [28], which evaluates realism by comparing the distribution of target images with that of edited images. Specifically, we compute FIDs for adding eyeglasses and hair edits from black to blonde transition using the CelebA [45] dataset. For instance, leveraging ground-truth attribute labels, we add eyeglasses to images without them and compute FIDs between the edited and original images that already have eyeglasses. This procedure is similarly applied to hair edits. For reconstruction evaluation, we mask the edited areas and measure the L2 and Structural Similarity

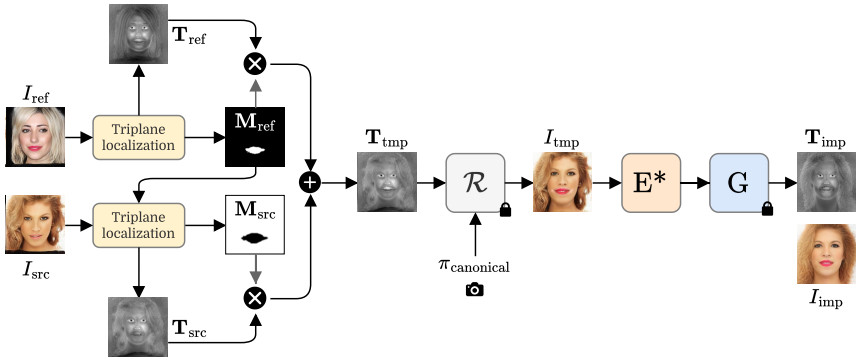

Figure 4. Triplane localization and implicit fusion stages, where $\mathbf{E}^*$ denotes the fine-tuned image encoder that is described in Sec. 3.3. Straightforward stitching in the triplane results in color inconsistency across the boundaries, as shown in I_{tmp} (zoom in for details). Leveraging $\mathbf{E}^*$, we aim to attain seamless boundaries and produce outputs with a natural appearance.

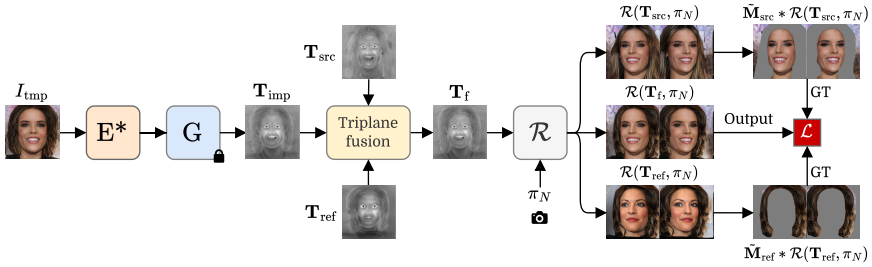

Figure 5. Pipeline for the implicit fusion encoder fine-tuning. We generate masked ground truths for our task by utilizing 2D segmentation networks and via renderings from multiple views. We aim to carry the reference parts in great detail to our source image while preserving the source's identity. Triplane fusion corresponds to Eq. (3).

Index (SSIM) between the input and edited images. For instance, in the eyeglasses edit, we mask the eyeglasses and measure the alteration in the unedited regions. This dual assessment framework ensures a robust evaluation of both the fidelity and quality of our editing approach.

Baselines. We conducted comprehensive comparisons by evaluating our method against a range of latent direction, text and image reference-based, 2D and 3D-aware, GAN, and diffusion-based editing methods. Notably, no existing reference-based editing methods achieve 3D consistency within a single framework.

HiSD and VecGAN++ are reference-based 2D image-to-image methods. InterFaceGAN, StyleCLIP ($\mathcal{W}^+$), SFE ($\mathcal{W}^+/\mathcal{F}$), StyleFusion ($\mathcal{W}$), Barbershop, and HairCLIPv2 ($\mathcal{F}/\mathcal{S}$) operate within the latent spaces of StyleGAN and can be adapted to 3D-aware GANs like EG3D. However, only the last three are reference-based, with two focusing on hair edits. E3DGE is a SDF-based generator, and editing is done via latent directions. For diffusion-based models, LEDITS++ and InfEdit perform text-based editing, while NoiseCLR utilizes pretrained latent edit noise directions. Paint by Example inpaints the masked source image with the reference image but lacks control over which parts of the reference are used.

Table 1. Quantitative scores on CelebA. (✗) indicates the method is not capable of such edits. **First** and second best method are given in **bold** and underlined. Time is measured on Tesla T4.

	Eyeglasses			Hair			Time
	FID ↓	$\mathcal{M}_{\text{SSIM}}$ ↑	$\mathcal{M}_{\mathcal{L}_2}$ ↓	FID ↓	$\mathcal{M}_{\text{SSIM}}$ ↑	$\mathcal{M}_{\mathcal{L}_2}$ ↓	(s)
HiSD [40]	77.56	0.9471	0.0090	94.53	**0.9743**	0.0036	1.1
VecGAN++ [14]	71.47	0.7483	0.0630	80.47	0.9296	0.0090	2.2
Barbershop [85]	✗	✗	✗	**62.80**	0.8756	0.0182	125
HairCLIPv2 [63]	✗	✗	✗	85.75	0.8769	0.0173	180
StyleFusion [33]	✗	✗	✗	84.67	0.8435	0.0198	2.4
InterfaceGAN [59]	88.13	0.9398	0.0104	80.93	0.7888	0.0387	**0.6**
StyleCLIP [53]	80.13	0.8421	0.0476	92.60	0.8716	0.0196	**0.6**
SFE [7]	106.1	0.9341	0.0099	89.49	0.9355	0.0050	5.1
E3DGE [36]	✗	✗	✗	77.86	0.8083	0.0257	1.2
NoiseCLR [13]	107.1	0.7958	0.0440	✗	✗	✗	17.2
LEDITS++ [8]	115.2	0.9645	0.0047	96.56	0.9717	0.0025	25.9
InfEdit [69]	90.33	0.8338	0.1042	105.4	0.7425	0.0613	9.1
Paint by ex. [70]	74.18	0.8828	0.0252	82.38	0.9229	0.0155	9.6
Ours	**66.68**	**0.9818**	**0.0021**	64.59	0.9720	**0.0029**	6.0

Results. We present quantitative and qualitative comparisons with competing methods in Tab. 1 and Fig. 6, respectively. From Tab. 1, it is evident that our method outperforms competing methods significantly in terms of FID and preserves identity better in the non-edited regions.

Observing Fig. 6, our method demonstrates superior per-

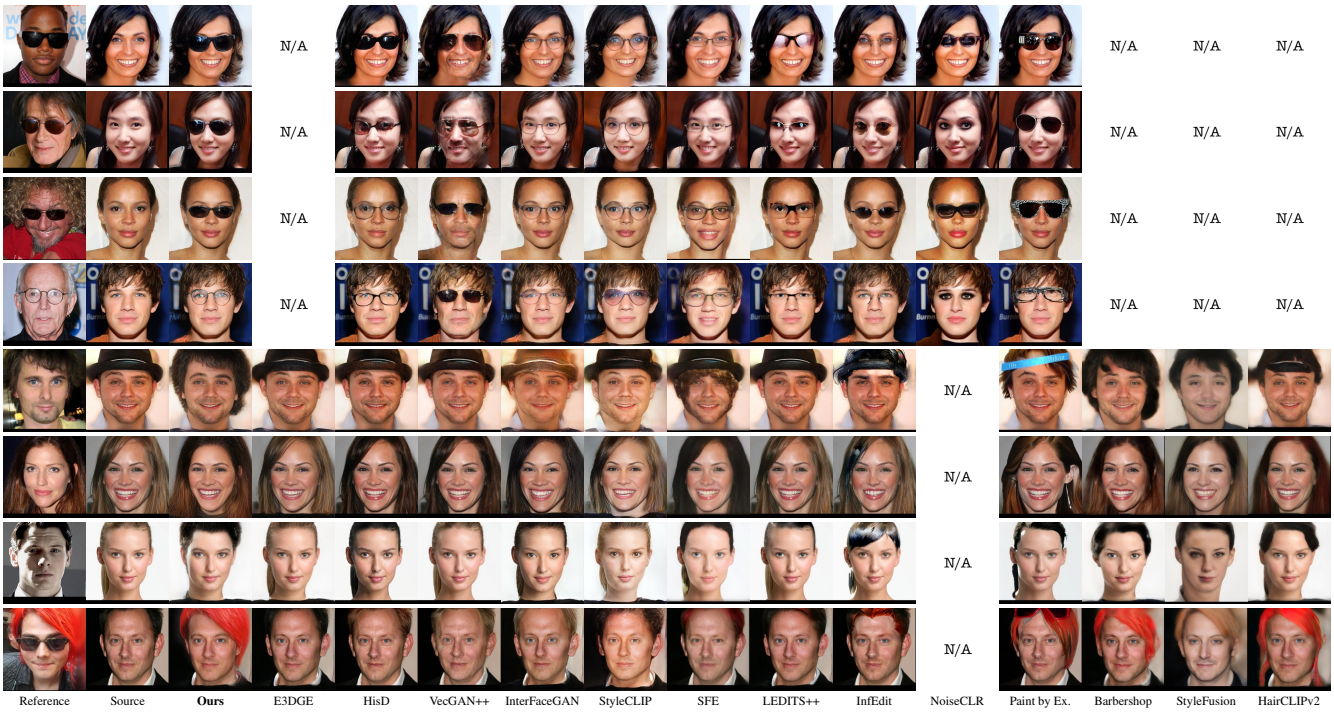

Figure 6. Comparisons with the competing editing methods for glasses addition and hair edits. Ours, HisD, VecGAN++, Barbershop, StyleFusion, HairCLIPv2, and Paint by Ex. use reference images for editing. InterFaceGAN, StyleCLIP, SFE, E3DGE, and NoiseCLR use previously calculated latent directions. LEDITS++ and InfEdit use text prompts. N/A indicates the model is incapable of such edits.

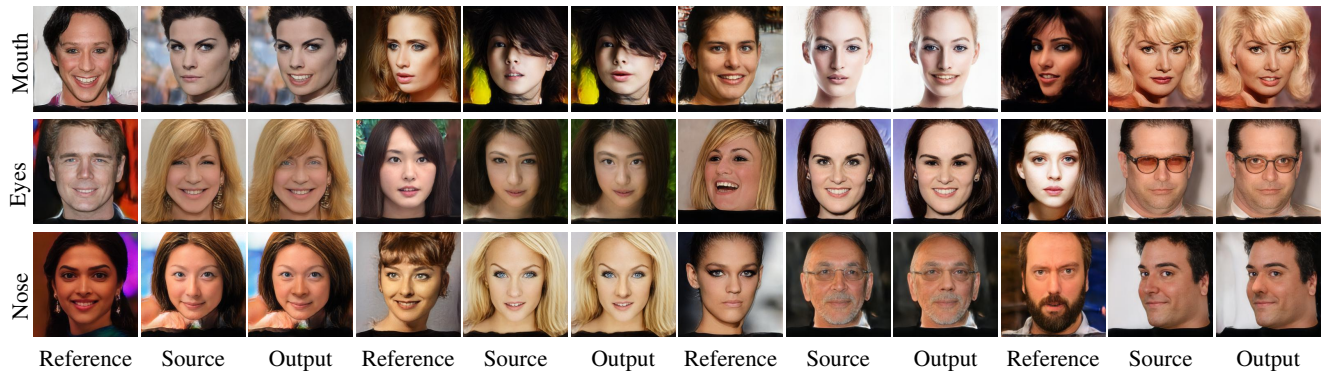

Figure 7. Additional editing examples from the CelebA dataset showcasing our method's ability to seamlessly incorporate features such as lips, eyes, and nose from reference to source, despite pose differences and interference like eyeglasses.

formance in hair and glasses edits compared to competing methods. HisD and VecGAN++ struggle with maintaining fidelity to the reference, particularly with glasses, due to their reliance on low-rate latent spaces. While InterFaceGAN, StyleCLIP, and SFE can add glasses and perform some hair transfers, they falter with uncommon edits like hat removal (row 5) and red hair (row 8) due to the limitations of their $\mathcal{W}^+/\mathcal{F}$ spaces and their non-reference-based approach. E3DGE's pre-trained editing directions yield unsatisfactory results. LEDITS++ and InfEdit, being text-conditioned, fail to accurately reflect the original reference in their edits. NoiseCLR does not effectively explore hairstyle directions, and its glasses modifications are entangled with makeup changes (rows 2 and 4). Paint by Example can transfer some features but often produces severe out-of-domain artifacts (rows 1, 5-8). Barbershop and HairCLIPv2, optimized for hairstyle edits, suffer from geometric inconsistencies (row 8) and fail in some hair edit cases (row 5). Finally, StyleFusion's feature transfer relies on $\mathcal{W}/\mathcal{W}^+$ directions, resulting in the loss of many high-rate details and outputs that do not fully reflect the original images, especially when the features cannot be well-

Table 2. Results of our user study where participants are asked to identify the edited image. Based on this study, we find that our edits are challenging to distinguish.

	FFHQ				AFHQ		
	Eyes	Nose	Mouth	Overall	Eyes	Nose+Mouth	Overall
Original	40%	29%	34%	34%	40%	42%	41%
Ours (edited)	49%	59%	40%	49%	49%	43%	46%
Undecided	11%	13%	26%	16%	11%	15%	13%

Table 3. Quantitative ablation study of our editing. V1 is the post-processing triplane gradients, V2 is the implicit fusion, and V3 fine-tunes the implicit fusion encoder.

			Eyeglasses			Hair		
V1	V2	V3	FID ↓	$\mathcal{M}_{\text{SSIM}}$ ↑	$\mathcal{M}_{\mathcal{L}_2}$ ↓	FID ↓	$\mathcal{M}_{\text{SSIM}}$ ↑	$\mathcal{M}_{\mathcal{L}_2}$ ↓
✗	✗	✗	79.50	0.8451	0.0323	82.42	0.8177	0.0195
✓	✗	✗	74.46	0.9814	0.0022	77.30	0.9674	0.0034
✓	✓	✗	68.19	**0.9822**	**0.0020**	67.04	0.9691	0.0033
✓	✓	✓	**66.68**	0.9818	0.0021	**64.59**	**0.9720**	**0.0029**

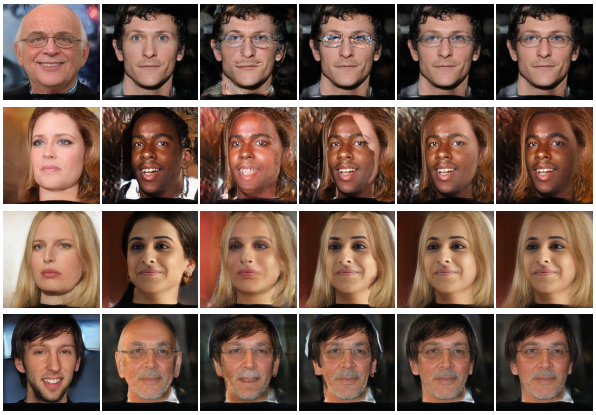

Figure 8. Qualitative ablation study for glasses and hair edits, showing the effects of all fundamental stages of our pipeline.

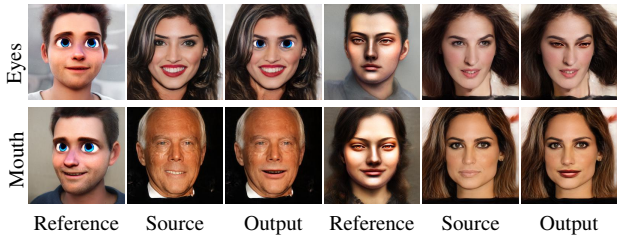

Figure 9. Cross-generator edits with stylizing. Our method achieves copying local parts from stylized images, such as cartoon portraits.

represented in those domains (row 8).

Next, we conduct a user study with 25 participants to evaluate our reference-based edits. Participants are shown original and edited images and asked to identify the edited ones. We utilize outputs of EG3D for both original and edited images to neutralize the influence of encoding on the results, and utilize the same angle for the source and edited

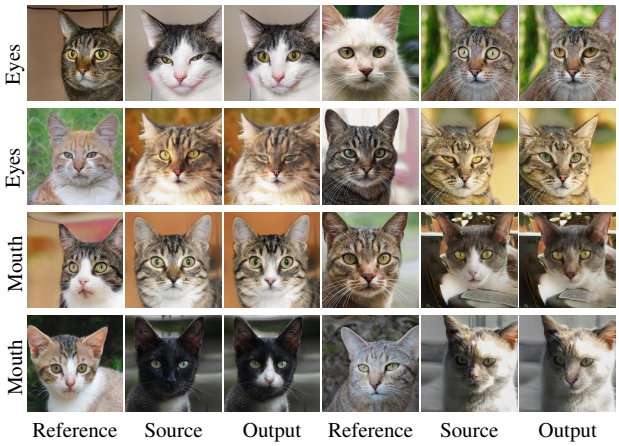

Figure 10. Additional editing examples from AFHQ dataset. The nose and mouth are handled as a single part.

images with random ordering to minimize bias. Participants could also choose "undecided" if they find it difficult to distinguish. The study focus on edits to mouth, eyes, and nose on the FFHQ dataset, and eyes, nose & mouth on the AFHQ dataset. Some participants frequently chose "undecided", while others perform near random chance, in Tab. 2.

Ablation study. We demonstrate the improvements during the development of our pipeline stages, both quantitatively and qualitatively, in Tab. 3 and Fig. 8, respectively.

In our initial ablation study, we apply Eq. (1) to merge the triplanes using a mask calculated via autograd function without any post-processing (No V). Due to the intricate volumetric function affecting many pixels for each value in the triplane, the initial mask fails to stitch images effectively, resulting in blurry outputs. Following the introduction of post-processing (+V1), as described in Sec. 3.1, we successfully achieve clear stitching, as depicted in Fig. 8. This allows us to transfer the hair of one person to our input while aligning the features using the canonical representation of the triplane. However, the resulting output still lacks realism because it combines features from two different images with varying illuminations, identities, and skin colors (row 2). To address the issue of smoothness at stitch boundaries, we follow Sec. 3.2 and perform encoding and decoding via the pretrained encoder and decoder, respectively, on the fused triplane. Since these encoders are trained with real images, they know about real image priors. Despite the input image not being realistic, as shown in Fig. 8 (V1), the encoder successfully encodes its latent to the generator's natural latent space while attempting to preserve the identity in (+V2). However, a pretrained encoder optimized for projecting real images onto the generator's latent space is not optimal for our specific use case. Consequently, we replace the encoder with one trained specifically for this task, elaborated in Sec. 3.3. This specialized encoder (+V3) ensures color preservation (row 2) and enhances editing details, such as the coherence of eyeglass frames, as well as

Figure 11. Extending our method on full-head hair edits on [4].

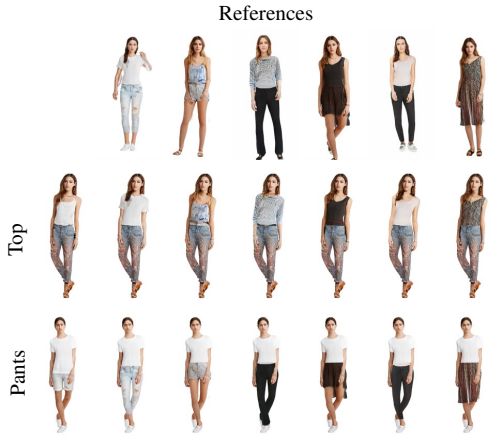

Figure 12. Extending our method on try-on edits on [20].

reducing background leakage (rows 1 and 4).

Cross-generator edits. We also provide novel edits in Fig. 9, where the reference and source triplanes are gathered from stylized and non-stylized generators, respectively. Specifically, we utilize [60] and fine-tune the EG3D backbones via different text prompts [55]. Then, we synthesize stylized triplanes ($\mathbf{T}_{\text{ref}}$) and perform reference-based editing on the triplanes of default EG3D ($\mathbf{T}_{\text{src}}$). The rest of the steps are the same as before, using the original EG3D and encoder we train for EG3D. The results presented in Fig. 9 demonstrate our method's independence from backbones, showcasing its capability to achieve part-based attribute stylization. This differs from [60], which offers global stylization.

Full body and 360-degree head edits. Fig. 7 demonstrates challenging human face reference-based edits on EG3D [10] like transferring lips, eyeglasses, and nose from one person to another. Fig. 10 shows edits on animal face parts for eyes, nose, and mouth. Fig. 12 demonstrates fashion edits on AG3D [20] trained with DeepFashion [46] dataset. Fig. 11 extends human face part edits to full 360-degree hair edits on PanoHead [4]. It is evident that our approach is generalizable to different triplane generators.

While extending our method to different triplane-based generators and datasets, we only changed the 2D segmentation network and the encoder fine-tuning dataset to comply with the generator, when required. We also provide multi-view image results and more examples in Supplementary.

Figure 13. Simultaneous editing results.

Generalizing to class-agnostic edits. Given the importance of large reconstruction models [11, 29, 32, 37, 38, 65, 68, 79], we extend our method to arbitrary object edits using the triplanes of LN3Diff and InstantMesh [37, 68] in Fig. 1, proving the potential capabilities of our method.

5. Conclusion

In conclusion, our work presents a comprehensive framework for reference-based, 3D-aware image editing, leveraging the unique capabilities of triplanes. Through spatial disentanglement and fusion learning, we achieve seamless integration of reference attributes while preserving the identity of the input image. We have shown our method's effectiveness through extensive qualitative and quantitative experiments. Our approach fills a crucial gap by offering a unified and generalizable solution.

Limitations. Our approach relies on the capabilities of EG3D, AG3D, PanoHead, LN3Diff and InstantMesh, which may struggle with background generation and high-quality reconstruction. Consequently, in some instances, rich background details may not be fully presented (row 4 on Fig. 6). However, this can be mitigated by not relying on the generator for background generation.

Future work. Extending our reference-based editing approach beyond triplanes using large reconstruction models [11, 32, 38, 65, 79] is an underexplored path. Specifically, on DiT-based approaches, we believe that the reference and source image tokens can be processed jointly, and the masks created via gradient accumulation can be applied onto self-attention layers for implicit fusion. We think that this can also eliminate the necessity for a canonical space where the source and reference features must be aligned, as the attention mechanism can handle misalignment easily.

References

[1] zllrunning. face-parsing.pytorch. https://github.com/zllrunning/face-parsing.PyTorch/. Accessed: 2024-28-02. 3

[2] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021. 2

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021. 1

[4] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20950–20959, 2023. 2, 8

[5] Qingyan Bai, Zifan Shi, Yinghao Xu, Hao Ouyang, Qiuyu Wang, Ceyuan Yang, Xuan Wang, Gordon Wetzstein, Yujun Shen, and Qifeng Chen. Real-time 3d-aware portrait editing from a single image. In *European Conference on Computer Vision*, 2024. 3

[6] Ananta R. Bhattarai, Matthias Nießner, and Artem Sevastopolsky. Triplanenet: An encoder for eg3d inversion. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 3, 4

[7] Denis Bobkov, Vadim Titov, Aibek Alanov, and Dmitry Vetrov. The devil is in the details: Stylefeatureeditor for detail-rich stylegan inversion and high quality image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9337–9346, 2024. 2, 5

[8] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinario Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8861–8870, 2024. 2, 3, 5

[9] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 1

[10] 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, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 1, 2, 3, 8

[11] Zhaoxi Chen, Jiaxiang Tang, Yuhao Dong, Ziang Cao, Fangzhou Hong, Yushi Lan, Tengfei Wang, Haozhe Xie, Tong Wu, Shunsuke Saito, Liang Pan, Dahua Lin, and Ziwei Liu. 3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion. In *CVPR*, 2025. 8

[12] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[13] Yusuf Dalva and Pinar Yanardag. Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24209–24218, 2024. 2, 3, 5

[14] Yusuf Dalva, Said Fahri Altindis, and Aysegul Dundar. Vecgan: Image-to-image translation with interpretable latent directions. *Proceedings of the European conference on computer vision (ECCV)*, 2022. 2, 5

[15] Yusuf Dalva, Hamza Pehlivan, Oyku Irmak Hatipoglu, Cansu Moran, and Aysegul Dundar. Image-to-image translation with disentangled latent vectors for face editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[16] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022. 3

[17] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 3

[18] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 4

[19] Zheng Ding, Cecilia Zheng, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[20] Zijian Dong, Xu Chen, Jinlong Yang, Michael J Black, Otmar Hilliges, and Andreas Geiger. AG3D: Learning to Generate 3D Avatars from 2D Image Collections. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 8

[21] Aysegul Dundar, Karan Sapra, Guilin Liu, Andrew Tao, and Bryan Catanzaro. Panoptic-based image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8070–8079, 2020. 2

[22] Ziya Erkoç, Can Gümeli, Chaoyang Wang, Matthias Nießner, Angela Dai, Peter Wonka, Hsin-Ying Lee, and Peiye Zhuang. Preditor3d: Fast and precise 3d shape editing. In *CVPR*, 2025. 3

[23] Anna Frühstück, Nikolaos Sarafianos, Yuanlu Xu, Peter Wonka, and Tony Tung. VIVE3D: Viewpoint-independent video editing using 3D-aware GANs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[25] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 1

[26] Yuming Gu, Hongyi Xu, You Xie, Guoxian Song, Yichun Shi, Di Chang, Jing Yang, and Linjie Luo. Diffportrait3d: Controllable diffusion for zero-shot portrait view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10456–10465, 2024. 3

[27] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[28] 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. *Advances in neural information processing systems*, 30, 2017. 4

[29] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3d. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 8

[30] Xianxu Hou, Xiaokang Zhang, Hanbang Liang, Linlin Shen, Zhihui Lai, and Jun Wan. Guidedstyle: Attribute knowledge guided style manipulation for semantic face editing. *Neural Networks*, 145:209–220, 2022. 2

[31] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. *ECCV*, 2018. 2

[32] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions, 2023. 8

[33] Omer Kafri, Or Patashnik, Yuval Alaluf, and Daniel Cohen-Or. Stylefusion: A generative model for disentangling spatial segments, 2021. 2, 5

[34] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1

[35] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 1, 2

[36] Yushi Lan, Xuyi Meng, Shuai Yang, Chen Change Loy, and Bo Dai. Self-supervised geometry-aware encoder for style-based 3d gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20940–20949, 2023. 2, 5

[37] Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy. Ln3diff: Scalable latent neural fields diffusion for speedy 3d generation. In *ECCV*, 2024. 2, 8

[38] Yushi Lan, Shangchen Zhou, Zhaoyang Lyu, Fangzhou Hong, Shuai Yang, Bo Dai, Xingang Pan, and Chen Change Loy. Gaussiananything: Interactive point cloud latent diffusion for 3d generation. In *ICLR*, 2025. 8

[39] Biwen Lei, Kai Yu, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. Diffusiongan3d: Boosting text-guided 3d generation and domain adaption by combining 3d gans and diffusion priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[40] Xinyang Li, Shengchuan Zhang, Jie Hu, Liujuan Cao, Xiaopeng Hong, Xudong Mao, Feiyue Huang, Yongjian Wu, and Rongrong Ji. Image-to-image translation via hierarchical style disentanglement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8639–8648, 2021. 2, 5

[41] Guilin Liu, Aysegul Dundar, Kevin J Shih, Ting-Chun Wang, Fitsum A Reda, Karan Sapra, Zhiding Yu, Xiaodong Yang, Andrew Tao, and Bryan Catanzaro. Partial convolution for padding, inpainting, and image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[42] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*, 2020. 4

[43] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9298–9309, 2023. 3

[44] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[45] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 2, 4

[46] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 8

[47] Yueming Lyu, Jing Dong, Bo Peng, Wei Wang, and Tieniu Tan. Sogan: 3d-aware shadow and occlusion robust gan for makeup transfer. In *Proceedings of the 29th ACM International conference on multimedia*, pages 3601–3609, 2021. 2

[48] Yueming Lyu, Peibin Chen, Jingna Sun, Bo Peng, Xu Wang, and Jing Dong. Dran: detailed region-adaptive normalization for conditional image synthesis. *IEEE Transactions on Multimedia*, 2023. 2

[49] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[50] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 3

[51] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 1

[52] Karran Pandey, Paul Guerrero, Metheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, and Niloy J. Mitra. Diffusion handles: Enabling 3d edits for diffusion models by lifting activations to 3d. *CVPR*, 2024. 3

[53] 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*, pages 2085–2094, 2021. 2, 5

[54] Hamza Pehlivan, Yusuf Dalva, and Aysegul Dundar. Styleres: Transforming the residuals for real image editing with stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1828–1837, 2023. 2

[55] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. 8

[56] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 2

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 3

[58] Wei Shen and Rujie Liu. Learning residual images for face attribute manipulation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4030–4038, 2017. 2

[59] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020. 2, 3, 5

[60] Kunpeng Song, Ligong Han, Bingchen Liu, Dimitris Metaxas, and Ahmed Elgammal. Diffusion guided domain adaptation of image generators. *arXiv preprint https://arxiv.org/abs/2212.04473*, 2022. 2, 3, 8

[61] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18310–18319, 2023. 3

[62] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4): 1–14, 2021. 2

[63] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Weiming Zhang, Gang Hua, and Nenghai Yu. Hairclipv2: Unifying hair editing via proxy feature blending. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2, 5

[64] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021. 2

[65] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 8

[66] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 168–184, 2018. 2

[67] Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, and Qifeng Chen. High-fidelity 3d gan inversion by pseudo-multi-view optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 321–331, 2023. 2

[68] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 2, 8

[69] Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with natural language. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 5

[70] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 5

[71] Guoxing Yang, Nanyi Fei, Mingyu Ding, Guangzhen Liu, Zhiwu Lu, and Tao Xiang. L2m-gan: Learning to manipulate latent space semantics for facial attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2951–2960, 2021. 2

[72] Yunhan Yang, Yukun Huang, Xiaoyang Wu, Yuan-Chen Guo, Song-Hai Zhang, Hengshuang Zhao, Tong He, and Xihui Liu. Dreamcomposer: Controllable 3d object generation via multi-view conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8111–8120, 2024. 3

[73] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. In *arXiv preprint arxiv:2308.06721*, 2023. 3

[74] Ahmet Burak Yildirim, Hamza Pehlivan, Bahri Batuhan Bilecen, and Aysegul Dundar. Diverse inpainting and editing

with gan inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23120–23130, 2023. 2

[75] Fei Yin, Yong Zhang, Xuan Wang, Tengfei Wang, Xiaoyu Li, Yuan Gong, Yanbo Fan, Xiaodong Cun, Ying Shan, Cengiz Oztireli, et al. 3d gan inversion with facial symmetry prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 342–351, 2023. 2

[76] Ziyang Yuan, Yiming Zhu, Yu Li, Hongyu Liu, and Chun Yuan. Make encoder great again in 3d gan inversion through geometry and occlusion-aware encoding. *arXiv preprint arXiv:2303.12326*, 2023. 2, 3, 4

[77] Gang Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Generative adversarial network with spatial attention for face attribute editing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 417–432, 2018. 2

[78] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 3

[79] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *TOG*, 2024. 8

[80] Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 4

[81] Qihang Zhang, Yinghao Xu, Chaoyang Wang, Hsin-Ying Lee, Gordon Wetzstein, Bolei Zhou, and Ceyuan Yang. 3ditscene: Editing any scene via language-guided disentangled gaussian splatting. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[82] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4

[83] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Multimodal image-to-image translation by enforcing bi-cycle consistency. In *Advances in neural information processing systems*, pages 465–476, 2017. 2

[84] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 2

[85] Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Barbershop: Gan-based image compositing using segmentation masks. *ACM Trans. Graph.*, 40(6), 2021. 2, 5