ZIGNeRF: Zero-shot 3D Scene Representation with Invertible Generative Neural Radiance Fields

Kanghyeok Ko Minhyeok Lee*
Chung-Ang University
Seoul, South Korea
{dogworld12, mlee}@cau.ac.kr

Abstract

Generative Neural Radiance Fields (NeRFs) have demonstrated remarkable proficiency in synthesizing multi-view images by learning the distribution of a set of unposed images. Despite the aptitude of existing Generative NeRFs in generating 3D-consistent high-quality random samples within data distribution, the creation of a 3D representation of a singular input image remains a formidable challenge. In this manuscript, we introduce ZIGNeRF, an innovative model that executes zero-shot Generative Adversarial Network (GAN) inversion for the generation of multi-view images from a single out-of-distribution image. The model is underpinned by a novel inverter that maps out-of-domain images into the latent code of the generator manifold. Notably, ZIGNeRF is capable of disentangling the object from the background and executing 3D operations such as 360-degree rotation or depth and horizontal translation. The efficacy of our model is validated using multiple real-image datasets: Cats, AFHQ, CelebA, CelebA-HQ, and CompCars.

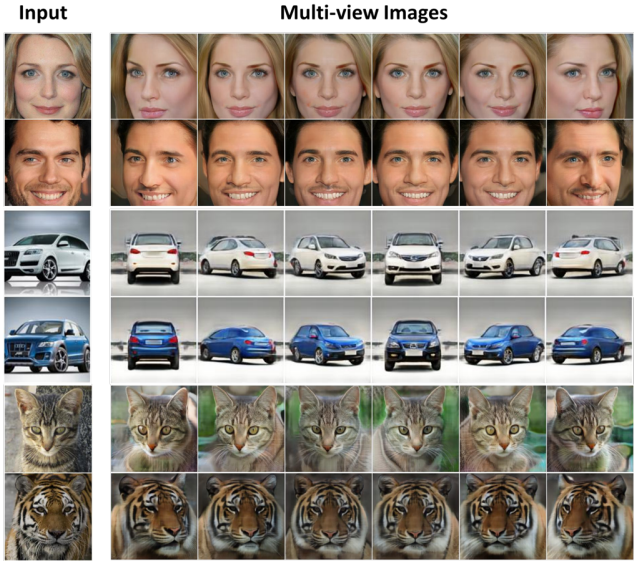

Figure 1. Demonstration of the 3D reconstruction results employing our proposed method, ZIGNeRF. This illustration depicts the successful zero-shot 3D GAN inversion across various real-world image datasets [6, 16, 46].

1. Introduction

The remarkable success of generative adversarial networks (GANs) [9] has spurred significant advancements in realistic image generation with high quality. Particularly, following the emergence of StyleGAN [18], numerous 2D-based generative adversarial network models have benefited from a deeper understanding of latent spaces [17, 19]. Consequently, various computer vision tasks, such as conditional image generation and style transfer [14, 21], have shown substantial progress. However, 2D-based image generation models are constrained in their ability to generate novel view images and 3D manipulation such as 360-degree rotations and spatial translations due to their neglect of geometrical context.

To overcome this challenge, several studies have adopted the neural radiance field (NeRF) [25] approach, which encodes a scene into a multi-layer perceptron (MLP) to provide 3D rendering. Although conventional NeRF [25] has successfully facilitated the development of 3D-aware models and reduced computational costs in novel view synthesis tasks, it remains impractical to train a model overfitted to a single scene with multi-view images [25, 48]. Consequently, various studies have extended NeRF by integrating it with generative models, i.e., generative NeRF. Generative NeRF [2, 3, 7, 10, 28, 29, 35] models can be trained on unposed real-world images, whereas conventional NeRF necessitates multiple images of a single scene [38, 40, 45]. Moreover, generative NeRF has been employed for obtaining conditional samples through techniques such as class

*Corresponding author.

label information [15] or text encoding [8, 30, 31, 42].

Despite the convenience and intuitiveness of these approaches, they possess limitations in image editing and generating 3D representations of specific inputs, such as out-of-domain iamges or real-world images. To enable more practical applications, generative NeRF models have also incorporated optimization-based GAN inversion techniques [32, 33, 39, 51] for the 3D representation of particular input images, including out-of-distribution or real-world images. However, previous approaches have faced a constraint that necessitates fine-tuning on pre-trained models for specific images [20, 22, 43, 47]. This requirement hinders the application of these models to numerous real samples simultaneously and renders the process time-inefficient, as it demands extensive fine-tuning. For example, EG3D [2] and PanoHead [1] facilitate 3D image reconstruction but necessitate fine-tuning steps and camera parameters during training, a constraint that often renders them impractical for real-world datasets. Our approach obviates this requirement, thereby extending its applicability to a broader range of image datasets.

In this study, we propose a novel zero-shot methodology for the generation of multi-view images, derived from input images unseen during the training process. This approach leverages a 3D GAN inversion technique. Notably, our model proffers 3D-consistent renderings of unposed real images during inference, eliminating the need for supplementary fine-tuning.

Our architectural design bifurcates into two distinct components: the 3D-generation module and the 3D GAN inversion module. The former is founded on the principles of GIRAFFE [28], which successfully amalgamates the compositional attributes of 3D real-world scenes into a generative framework. To enhance the precision of 3D real-world reconstruction and improve image quality, we introduce modifications to the GIRAFFE module, specifically in the decoder and neural renderer. The inverter for 3D-aware image reconstruction, on the other hand, is an encoder which is trained with images synthesized from the generator. This strategic approach enables the inverter to accurately map the input image onto the generator's manifold, regardless of the objects' pose. Example results of our model is displayed in Fig. 1. In addition, we show the suitability of the learning-based inverter design over optimization-based approaches by showing the limitation of optimization-based approaches for this specific application, as presented in Fig. 1 of supplementary material.

We subject our model to rigorous evaluation, utilizing five diverse datasets: Cats, CelebA, CelebA-HQ, AFHQ, and CompCars. Additionally, we demonstrate the model's robustness by inputting FFHQ images into a model trained on CelebA-HQ.

The primary contributions of this work are as follows:

- We present ZIGNeRF, a pioneering approach that delivers a 3D-consistent representation of real-world images via zero-shot estimation of latent codes. To our knowledge, this is the first instance of an learning-based approach in the field.
- ZIGNeRF exhibits robust 3D feature extraction capabilities and remarkable controllability with respect to input images. Our model can perform 3D operations, such as a full 360-degree rotation of real-world car images, a feat not fully achieved by many existing generative NeRF models.

2. Related Work

2.1. Neural Radiance Field (NeRF)

NeRF is an influential method for synthesizing photo-realistic 3D scenes from 2D images. It represents a 3D scene as a continuous function using a multi-layer perceptron (MLP) that maps spatial coordinates to RGB and density values, and then generates novel view images through conventional volume rendering techniques. Consequently, NeRF significantly reduces computational costs compared to existing voxel-based 3D scene representation models [12, 27, 36, 38, 50]. However, the training method of NeRF, which overfits a single model to a single scene, considerably restricts its applicability and necessitates multiple structured training images, including camera viewpoints [4, 38].

2.2. Generative NeRF

Generative NeRFs optimize networks to learn the mapping from latent code to 3D scene representation, given a set of unposed 2D image collections rather than using multi-view supervised images with ground truth camera poses. Early attempts, such as GRAF [35] and pi-GAN [3], demonstrated promising results and established the foundation for further research in the generative NeRF domain. Recent works on generative NeRF have concentrated on generating high-resolution 3D-consistent images. The recently proposed StyleNeRF [10] successfully generates high-resolution images by integrating NeRF into a style-based generator, while EG3D [2] exhibits impressive results with a hybrid architecture that improves computational efficiency and image quality.

However, real-life applications frequently necessitate conditional samples that exhibit the desired attribute rather than random samples in data distribution. We adopt GAN inversion as a conditional method, as opposed to class-based or text encoding conditional methods, which are prevalent in 2D generative models [5]. The aforementioned conditional generation techniques, such as class-based or text encoding methods, possess limitations. Firstly, the training dataset must include conditional information, such

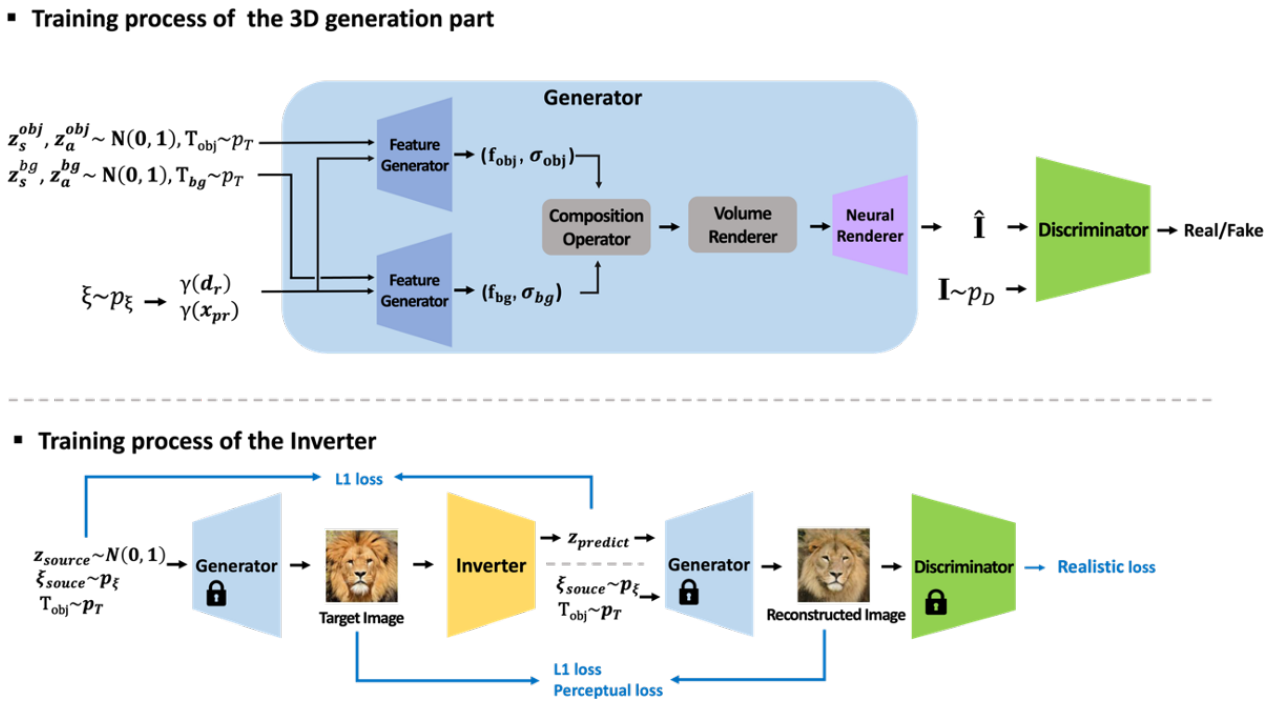

Figure 2. The comprehensive architecture of ZIGNeRF. The 3D generative component is trained to produce photorealistic images consistent with 3D structures by mapping the latent code and camera pose to a synthetic image. Subsequently, the inverter is trained in conjunction with the pre-trained generator and discriminator.

as labels or text corresponding to each sample. Secondly, they cannot provide 3D representation of real-world images as conditional input. We address these limitations in existing conditional generative NeRF models by introducing GAN Inversion into generative NeRF for conditional generation.

2.3. 3D GAN inversion

With the remarkable progress of GANs, numerous studies have endeavoured to understand and explore their latent space to manipulate the latent code meaningfully. GAN inversion represents the inverse process of the generator in GANs. Its primary objective is to obtain the latent code by mapping a given image to the generator's latent space. Ideally, the latent code optimized with GAN inversion can accurately reconstruct an image generated from the pre-trained generator. The output sample can be manipulated by exploring meaningful directions in the latent space [37]. Moreover, real-world images can be manipulated in the latent space using GAN inversion.

Several studies have investigated 3D GAN inversion with generative NeRF to generate multi-view images of input samples and edit the samples in 3D manifolds. Most previous works fine-tuned the pre-trained generator due to the utilization of optimization-based GAN inversion methods. However, additional steps for fine-tuning the generator for GAN inversion impose limitations in terms of adaptability and computational costs.

In this paper, we propose a novel inverter for zero-shot 3D GAN inversion. The proposed inverter can map out-of-distribution images into the latent space of the generator. Our model can generate 3D representations of real-world images without requiring additional training steps. The proposed zero-shot 3D GAN inversion maximizes applicability since the trained model can be directly applied to out-of-distribution images.

3. Method

This work seeks to generate multi-view images from an out-of-distribution image by combining generative NeRF with GAN inversion. The proposed method, graphically delineated in Fig. 2, encompasses two distinct phases: the 3D-generation segment and the inverter for 3D-aware image reconstruction. The first phase involves training the 3D-generation component, an architecture based on GIRAFFE, augmented by enhancements in the neural renderer and the discriminator modules to fortify and expedite the training process. In the second phase, the inverter is trained with the pre-trained generator. The novel inverter is designed to transform out-of-distribution images into latent codes within the generator's latent space. Consequently, the generator can produce multi-view images of the out-of-

distribution image using the latent code derived from the inverter. Throughout the training of the inverter, we utilize the images generated from the generator, imbued with 3D information, as the training dataset. At test time, the inverter executes zero-shot inversion on real-world images, obviating the need for additional fine-tuning for unseen images. The proposed method thereby holds great promise for generating 3D-consistent multi-view images from real-world input images.

3.1. 3D Generation

Compositional Generative Neural Feature Field. Our 3D-generator represents a scene with a compositional generative neural feature field, a continuous function inherited from GIRAFFE, to represent a scene. This is essentially a combination of feature fields, each representing an object in a single scene, with the background also considered an object. In the 3D-generator, a 3D location, $\mathbf{x} \in \mathbb{R}^3$, a viewing direction, $\mathbf{d} \in \mathbb{S}^2$, and latent code, $\mathbf{z} \sim \mathcal{N}(0, 1)$, are mapped to a volume density $\sigma \in \mathbb{R}^+$ and a high-dimensional feature field $\mathbf{f} \in \mathbb{R}^{M_f}$, rather than RGB colour $\mathbf{c} \in \mathbb{R}^3$.

Affine transformation is applied to objects in the scene so that each object can be controlled in terms of poses, which include scale, translation, and rotation:

$$T = \{\mathbf{s}, \mathbf{t}, \mathbf{R}\}, \tag{1}$$

where $\mathbf{s}, \mathbf{t} \in \mathbb{S}$ indicate scale and translation parameters, respectively, and $\mathbf{R} \in$ SO(3) determine rotation. The affine transformation enables object-level control by generating the bounding box corresponding to T of a single object:

$$\tau = \mathbf{R} \cdot \mathbf{sI} \cdot + \mathbf{t}, \tag{2}$$

where $\mathbf{I}$ is the 3 × 3 identity matrix. Compositional generative neural feature field is parameterized with an MLP as follows:

$$C((\sigma_i, \mathbf{f}_i)_{i=1}^N) = C(f_{\theta i}(\gamma(\tau^{-1}(\mathbf{x})), \gamma(\tau^{-1}(\mathbf{d})), \mathbf{z}_i)_{i=1}^N), \tag{3}$$

$$z = [\mathbf{z_s^1}, \mathbf{z_a^1}, ..., \mathbf{z_s^N}, \mathbf{z_a^N}], \tag{4}$$

where $\gamma(\cdot)$ is positional encoding function [25], which is applied separately to x and d, and $C(\cdot)$ is the compositional operator that composites feature field from the N-1 objects and a background. We then volume render the composited volume density and feature field rather than directly output the final image. 2D-feature map, which is fed into neural renderer for final synthesized output, is attained by volume rendering function π_v,

$$\pi_v(C(\sigma, \mathbf{f})) = \mathbf{F}. \tag{5}$$

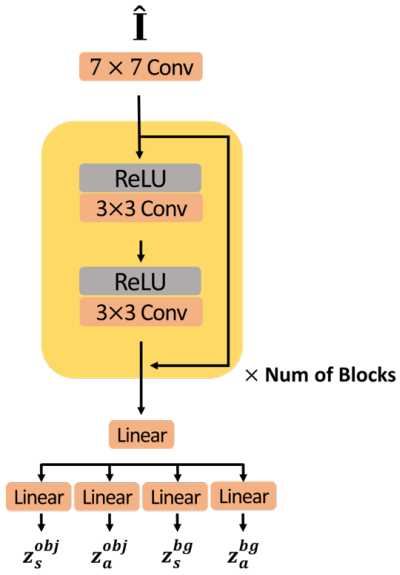

Figure 3. Schematic representation of the architecture of the inverter deployed in ZIGNeRF.

Neural renderer with residual networks. Our model outputs final synthetic image with neural rendering on the output feature map of volume rendering. We observe that the original neural renderer of GIRAFFE does not preserve the feature well. Furthermore, the learning rate of the decoder and the neural renderer is not synchronized; hence the training of the generator is unstable.

We improve the simple and unstable neural renderer of GIRAFFE. Our neural renderer replaces 3×3 convolution layer blocks with residual blocks [11] and employs the ReLU activation rather than leaky ReLU activation [44] for faster and more effective rendering. To stabilize the neural rendering, we adopt spectral normalization [26] as weight normalization. We experimentally verify that the modified neural renderer improves the stability of the training and the quality of the outputs. Our neural renderer, which maps the feature map F to the final image $\hat{I} \in \mathbb{R}^{H \times W \times 3}$, is parameterized as:

$$\pi_\theta(\mathbf{F}) = \hat{\mathbf{I}}. \tag{6}$$

Discriminator. As the vanilla GAN [9], the discriminator outputs probability, which indicates whether the input image is real or fake. We replace the 2D CNN-based discriminator with residual blocks employing spectral normalization as weight normalization.

Objectives. The overall objective function of the 3D-generative part is:

$$L_{\text{G, D}} = L_{\text{GAN}} + \lambda L_{\text{R1}}, \tag{7}$$

where λ = 10. We use GAN objective [9] with R1 gradient penalty [24] to optimize the network.

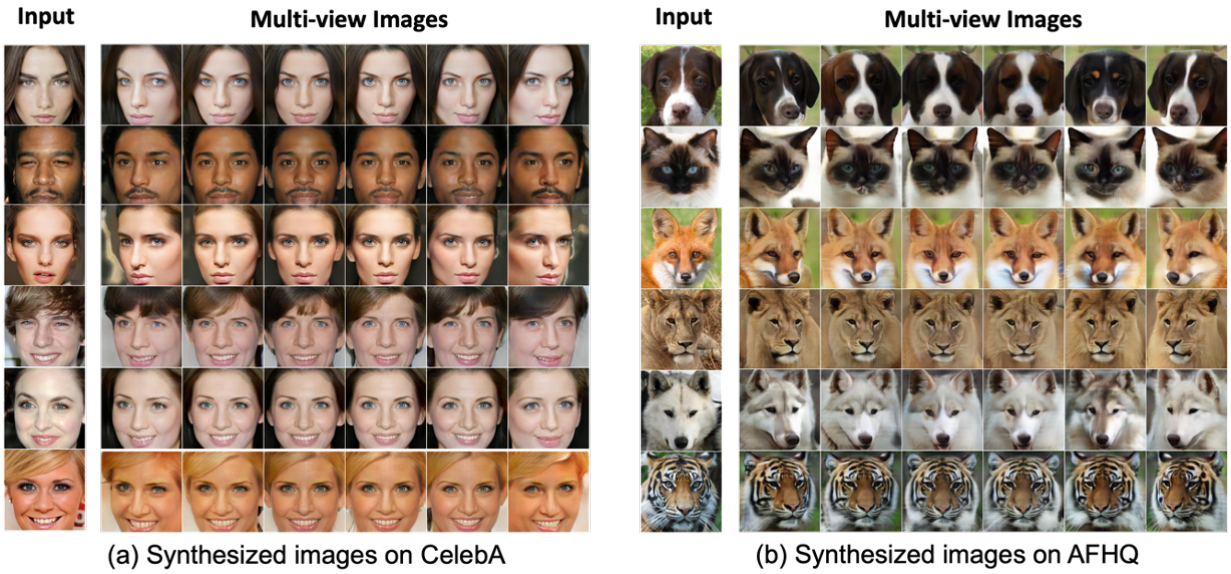

Figure 4. Display of 256^2 multi-view synthesis applied to facial datasets: CelebA-HQ [16] and AFHQ [6].

3.2. Inverter for 3D-aware Image Reconstruction

To invert a given image into latent codes within the generator's latent space, we introduce a novel inverter. This inverter is designed by stacking the residual encoder block with ReLU activations, as depicted in Fig. 3. Four linear output layers are situated at the culmination of the inverter to facilitate output. These residual blocks extract the feature of the input image, and each linear output layer estimates the $\mathbf{z}_{\mathbf{s}}^{\mathbf{obj}}$, $\mathbf{z}_{\mathbf{a}}^{\mathbf{obj}}$, $\mathbf{z}_{\mathbf{s}}^{\mathbf{bg}}$, $\mathbf{z}_{\mathbf{a}}^{\mathbf{bg}}$ of the input image.

The challenge of 3D GAN inversion involves mapping multi-view images of a single object into a unique latent code. To construct a inverter, we opt to use the synthesized image $\hat{\mathbf{I}}$ as the training data. Given that we already possess the source parameters of the generated image, the inverter solely estimates the latent code $\mathbf{z}^{\mathbf{predict}}$ of the input image. The generated training images equip the inverter to extract the feature of unseen images, which vary in viewing direction, scale, and rotation. Following the latent code inference, the pre-trained generator reconstructs the input image using $\mathbf{z}^{\mathbf{predict}}$ and source parameters, which include camera pose, $\boldsymbol{\xi}^{\mathbf{source}}$, and compositional parameter, $\mathbf{T}^{\mathbf{source}} = \{\mathbf{s}, \mathbf{t}, \mathbf{R}\}$:

$$I_\theta(\hat{\mathbf{I}}) = \mathbf{z}^{\mathbf{predict}}, \tag{8}$$

$$G_\theta(\mathbf{z}^{\mathbf{predict}}, \mathbf{T}^{\mathbf{source}}, \boldsymbol{\xi}^{\mathbf{source}}) = \hat{\mathbf{I}}^{\mathbf{reconst}}. \tag{9}$$

As the inverter learns to estimate the latent source code, we found that the L1 loss between the two latent codes in latent space was inadequate for reconstructing the scene. Thus, we opted to employ realistic loss , which is calculated with output of the discrminator and L1 as an image-level loss to generate a plausible image. In addition, we incorporated two perceptual losses, namely the Structural Similarity Index Measure (SSIM) [41] and the Learned Perceptual Image Patch (LPIPS) [49] loss, to conserve the fine details of the source image. The inverter can be optimized using the following function:

$$\begin{aligned} L_I = & L_{\text{real}}(\hat{\mathbf{I}}^{\mathbf{predict}}) \\ & + \lambda_1 L_{\text{latent}}(\mathbf{z}^{\mathbf{source}}, \mathbf{z}^{\mathbf{predict}}) \\ & + \lambda_2 L_{\text{reconst}}(\hat{\mathbf{I}}^{\mathbf{source}}, \hat{\mathbf{I}}^{\mathbf{predict}}) \\ & + \lambda_3 L_{\text{percept}}(\hat{\mathbf{I}}^{\mathbf{source}}, \hat{\mathbf{I}}^{\mathbf{predict}}), \end{aligned} \tag{10}$$

where $\hat{\mathbf{I}}^{\mathbf{predict}}$ indicates the image reconstructed by the pre-trained generator using $\mathbf{z}^{\mathbf{predict}}$. L_{real} denotes the realistic loss, L_{latent} and L_{reconst} represent latent-level and image-level loss, respectively, both utilizing L1 loss. L_{percept} signifies image-level perceptual loss, employing the LPIPS loss and SSIM loss.

3.3. Training Specifications

During the training phase, we randomly sample the latent codes $\mathbf{z}_{\mathbf{s}}, \mathbf{z}_{\mathbf{a}} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, and a camera pose $\boldsymbol{\xi} \sim p_\xi$. The parameters λ_1, λ_2, and λ_3 are set to 10, 100, and 1, respectively, for training the inverter. The model is optimized using the RMSProp optimizer [34], with learning rates of 1×10^4, 7×10^5, and 1×10^4 for the generator, the discriminator, and the inverter, respectively. We utilize a batch size of 32. For the first 100,000 iterations, the generator and the discriminator are trained, and the inverter is trained for the next 50,000 iterations. During the training process of the inverter, the generator and the discriminator remain frozen.

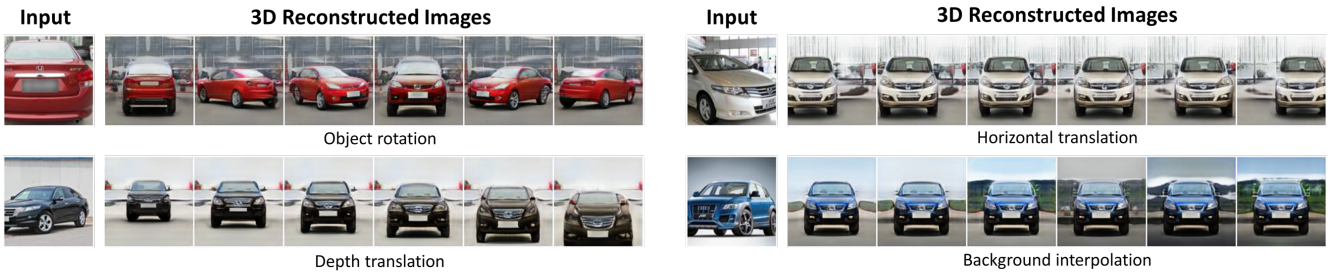

Figure 5. Visualisation of reconstructed images based on an input car image [46], following compositional operations. These illustrations highlight the effective disentanglement of the object from the background and the provision of 3D controllability.

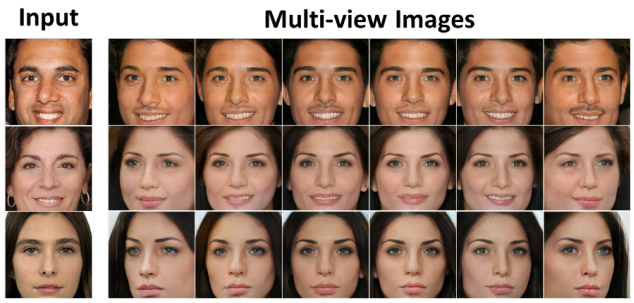

Figure 6. Presentation of 256^2 synthesized images conditioned on input FFHQ [18] images, produced by the model trained on the CelebA-HQ dataset [16].

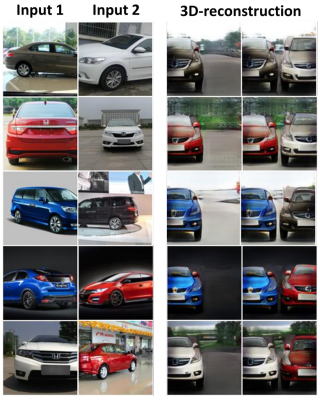

Figure 7. Generating two objects in a single scene. Results exhibit the compositional scene representation by generating two objects in a single scene. The inverter transforms two input images into two sets of the latent codes, and the generator which trained on single-object scenes synthesizes a single scene including two independent objects.

4. Experiments

ZIGNeRF is evaluated concerning zero-shot feature extraction, 3D controllability, and adaptability. We test on five real-world datasets: Cats, AFHQ [6], CelebA [23], CelebA-HQ [16], and CompCar [46]. An additional dataset, FFHQ [18], is used to demonstrate the robust adaptation capabilities of the proposed model. All input images shown in this section were not used during the training process, thereby validating the zero-shot 3D GAN inversion with unseen images. We commence with a visual validation of the proposed model, examining the similarity between the input image and the reconstructed images and 3D-consistent controllability. The model is then evaluated using Fréchet Inception Distance (FID) [13] as a metric. We conclude with ablation studies to validate the efficacy of the loss function in optimizing the inverter.

4.1. Controllable 3D Scene Reconstruction

We visually demonstrate that our proposed model generates multi-view consistent images corresponding to the input image. Figure 4 showcases 3D reconstruction on CelebA-HQ [16] and AFHQ [6], substantiating that the inverter successfully extracts facial features irrespective of gender or skin colour in human faces, and species in animal faces. Figure 5 exhibits the model's controllability and object disentanglement with CompCar [46], indicating that the inverter estimates the latent code of the object and background effectively. Notably, the proposed model can facilitate 3D-consistent 360-degree rotation, a common limitation of generative NeRF methods. We further attest to the robustness of our model by applying it to FFHQ, as shown in Fig. 6.

4.2. Extended Operational Results

In this section, we present the application results of the proposed model through Fig. 7 and Fig. 8, showcasing the generation of two objects within a single scene and style-mixed 3D synthesis.

Our model demonstrates a unique ability to generate multiple objects within a single scene, even when trained on a dataset consisting primarily of single-object scenes. This is accomplished by leveraging multiple decoder segments within our network architecture. Although our empirical exploration has only been executed on one dataset, the theoretical underpinnings suggest a promising generalizability

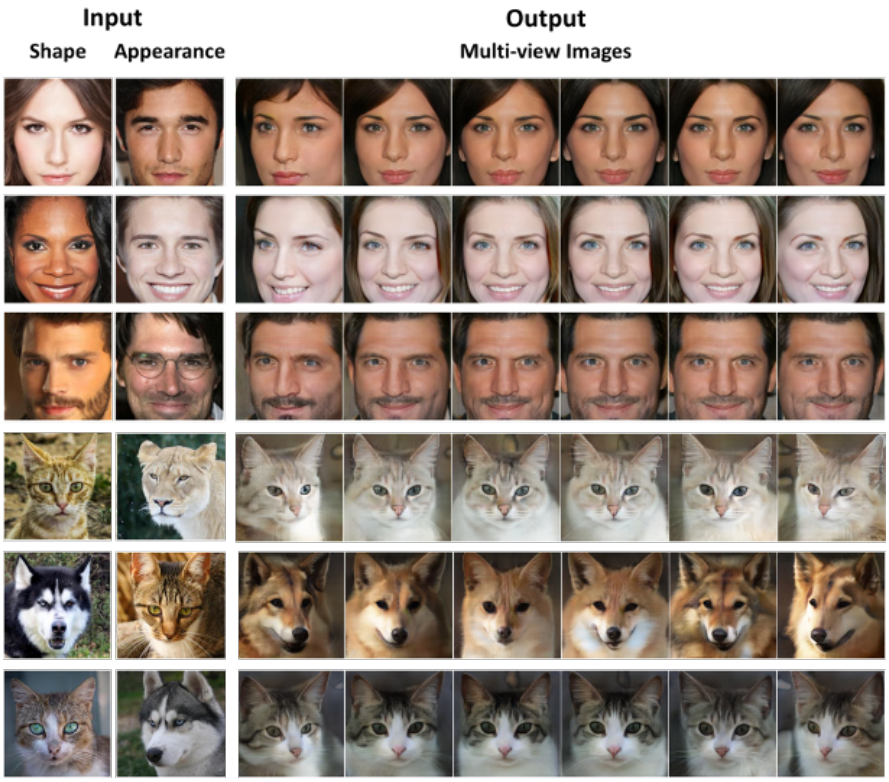

Figure 8. Multi-view images with style mixing of two input images. The inverter extracts the latent codes from two independent input images for generating style mixed images. Each output object is generated by $\mathbf{z_s}$ of the first image and $\mathbf{z_a}$ of the second image.

of this phenomenon. A testament to the robustness of our model is its successful exhibition of zero-shot learning capabilities, as evidenced by an experiment where two CompCars [46] images are synthesized into one image. Like the generation of individual objects, each object within the composite scene retains the ability to undergo transformations such as longitudinal displacement and rotation.

Additionally, we incorporate style mixing in our model with the application of the inverter structure we proposed, utilizing the CelebA-HQ and AFHQ dataset [16]. In the style mixing paradigm that we suggest, our inverter, producing two distinct outputs, generates a shape vector from one image, and an appearance vector from another. These two vectors are subsequently utilized as input for the generator to synthesize a novel object. This process further underscores the model's zero-shot learning capability.

4.3. Quantitative Evaluation

To thoroughly evaluate the efficacy of our proposed model, ZIGNeRF, we conduct experiments in both conditional and unconditional generation modes. The evaluation process involves a random sampling of 20,000 real images alongside 20,000 synthesized images, which is a conventional method to compare generative models. The results are displayed in Tab. 1.

In the context of the unconditional model, we generate samples using random latent codes. The training process entails 100,000 iterations. Notably, our model, ZIGNeRF, significantly outperforms the baseline GIRAFFE [28] model. As an illustration, for the CelebA(HQ) 256^2 dataset, ZIGNeRF achieves a score of 14.98, substantially lower than the GIRAFFE's score of 23.14. This is indicative of the model's ability to produce higher-quality images with fewer iterations.

Turning to the conditional synthesis, the latent codes estimated by the inverter are employed on randomly sampled real images. The training process for the generator is conducted over 100,000 iterations, while the inverter training comprises 50,000 iterations, during which the generator is kept static. When compared to GIRAFFE, ZIGNeRF demonstrate superior performance in conditional samples as well. For instance, in the AFHQ 128^2 dataset, our model attains a score of 14.02, marking a significant improvement over the GIRAFFE's score of 35.03.

4.4. Ablation study

In the interest of validating the loss function deployed in training the inverter, we undertake an ablation study. The

Method	Models	Cats		CelebA(HQ)		CompCar		AFHQ	
		128^2	256^2	128^2	256^2	128^2	256^2	128^2	256^2
Unconditional	GIRAFFE	24.01	21.28	19.45	23.14	38.91	40.84	35.03	38.18
	ZIGNeRF(ours)	**12.31**	**11.21**	**11.01**	**14.98**	**22.67**	**22.57**	**12.81**	**19.96**
Conditional	ZIGNeRF(ours)	**15.06**	**16.83**	**14.77**	**25.66**	**25.97**	**25.41**	**14.02**	**28.78**

Table 1. Comparative analysis of the FID between our proposed ZIGNeRF and a baseline model. The models were trained on four distinct datasets with the resolution of 128^2 and 256^2.

Ablation Losses	FID
L_{latent}	80.08
$+L_{reconst}$	17.82
$+L_{reconst} + L_{real}$	15.53
Full model	14.77

Table 2. FID score of the ablation study. The full model has been trained with latent loss, reconstruction loss, GAN loss, and perceptual loss.

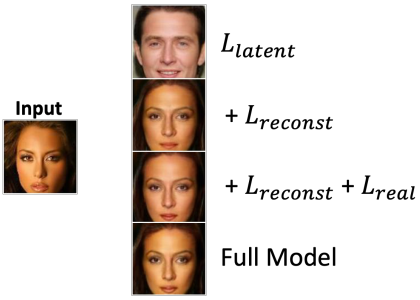

Figure 9. Ablation study of the loss functions employed in the training of the inverter within ZIGNeRF.

study scrutinizes the necessity of each loss component: latent loss, reconstruction loss, GAN loss, and perceptual loss. The imperative nature of each loss function is demonstrated through its incremental addition to the naive model, which is trained solely via latent code comparison. Figure 9 illustrates the individual contribution of each loss function. It is observed that the naive model exhibits limited capability in reconstructing the input image. The reconstruction loss L_{reconst} aligns the reconstructed image with the input at an image-level. The GAN loss L_{GAN} is observed to enhance the realism of the reconstructed image, independent of improving the input-reconstructed image similarity. The full model elucidates that the perceptual loss L_{percept} plays a pivotal role in refining the expression of minute attributes, skin colour, and texture.

Table 2 offers a quantitative testament to the indispensable nature of the loss components used in the training session of the inverter. It is observed that the Fréchet Inception Distance (FID) [13] experiences a steady enhancement with each loss component incrementally added to the naive model, which originally only employs the latent loss.

5. Conclusion

In this paper, we have proposed ZIGNeRF, an innovative technique that manifests a 3D representation of real-world images by infusing a zero-shot 3D GAN inversion into generative NeRF. Our inverter is meticulously designed to map an input image onto a latent manifold, a learning process undertaken by the generator. During testing, our model generates a 3D reconstructed scene from a 2D real-world image, employing a latent code ascertained from the inverter. Rigorous experiments conducted with four distinct datasets substantiate that the inverter adeptly extracts features of input images with varying poses, thereby verifying the 3D controllability and immediate adaptation capabilities of our model.

Our novel approach carries the potential for wide application, given that our pipeline can be generally applied to other existing generative NeRFs. It is worth noting that this zero-shot approach is a pioneering contribution to the field, bringing forth a paradigm shift in 3D image representation. In future work, we envisage extending the proposed method by manipulating the inverted latent code for editing the input image, thereby further enhancing the capabilities of this innovative model.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00251528), and Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0020967, Advanced Training Program for Smart Sensor Engineers).

References

[1] 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*, pages 20950–20959, 2023.

[2] 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.

[3] 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.

[4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020.

[7] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10673–10683, 2022.

[8] 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 Transactions on Graphics (TOG)*, 41(4):1–13, 2022.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[10] 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.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9984–9993, 2019.

[13] 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.

[14] Lukas Höllein, Justin Johnson, and Matthias Nießner. Stylemesh: Style transfer for indoor 3d scene reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6198–6208, 2022.

[15] Kyungmin Jo, Gyumin Shim, Sanghun Jung, Soyoung Yang, and Jaegul Choo. Cg-nerf: Conditional generative neural radiance fields. *arXiv preprint arXiv:2112.03517*, 2021.

[16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[17] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

[18] 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.

[19] 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.

[20] Jaehoon Ko, Kyusun Cho, Daewon Choi, Kwangrok Ryoo, and Seungryong Kim. 3d gan inversion with pose optimization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2967–2976, 2023.

[21] Kanghyeok Ko, Taesun Yeom, and Minhyeok Lee. Superstargan: Generative adversarial networks for image-to-image translation in large-scale domains. *Neural Networks*, 162:330–339, 2023.

[22] Connor Z Lin, David B Lindell, Eric R Chan, and Gordon Wetzstein. 3d gan inversion for controllable portrait image animation. *arXiv preprint arXiv:2203.13441*, 2022.

[23] 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.

[24] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018.

[25] 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.

[26] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[27] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised

learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019.

[28] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021.

[29] 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.

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

[31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[32] 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.

[33] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022.

[34] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[35] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020.

[36] Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. *International journal of computer vision*, 35:151–173, 1999.

[37] 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.

[38] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019.

[39] Haorui Song, Yong Du, Tianyi Xiang, Junyu Dong, Jing Qin, and Shengfeng He. Editing out-of-domain gan inversion via differential activations. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 1–17. Springer, 2022.

[40] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022.

[41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[42] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. Hairclip: Design your hair by text and reference image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18072–18081, 2022.

[43] 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.

[44] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[45] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022.

[46] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3973–3981, 2015.

[47] Yu Yin, Kamran Ghasedi, HsiangTao Wu, Jiaolong Yang, Xin Tong, and Yun Fu. Nerfinvertor: High fidelity nerf-gan inversion for single-shot real image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8539–8548, 2023.

[48] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021.

[49] 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*, pages 586–595, 2018.

[50] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

[51] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 592–608. Springer, 2020.