

Received 1 April 2025, accepted 13 April 2025, date of publication 17 April 2025, date of current version 25 April 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3561923

RESEARCH ARTICLE

Zero-Shot 3D Scene Representation With Invertible Generative Neural Radiance Fields

KANGHYEOK KO¹, SUNGYUP KIM², AND MINHYEOK LEE^{1,2}, (Member, IEEE)

¹Department of Intelligent Semiconductor Engineering, Chung-Ang University, Seoul 06974, South Korea

²School of Electrical and Electronics Engineering, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Minhyeok Lee (mlee@cau.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by Korea Government (MSIT) under Grant RS-2024-00337250, and in part by Korea Institute for Advancement of Technology (KIAT) Grant funded by Korea Government (MOTIE) (Advanced Training Program for Smart Sensor Engineers) under Grant P0020967.

ABSTRACT Generative Neural Radiance Fields (NeRFs) have recently enabled efficient synthesis of 3D scenes by training on unposed real image sets. However, existing methods for generating multi-view images of specific input images have limitations, such as requiring camera parameters or additional components for estimating them. In this paper, we propose ZIGNeRF, a novel learning-based approach for zero-shot 3D Generative Adversarial Network (GAN) inversion that generates multi-view images from a single input image without requiring camera parameters. Our method introduces a novel inverter that maps out-of-distribution images into the latent space of the 3D generator without needing additional training steps. We demonstrate the efficacy of ZIGNeRF on multiple real-world image datasets, including Cats, AFHQ, CelebA-HQ, CompCars, and CUB-200-2011. For example, ZIGNeRF achieves an FID of 14.77 for face image generation when trained on the CelebA-HQ dataset. Furthermore, ZIGNeRF is capable of performing 3D operations such as 360-degree rotation and spatial translations by disentangling objects from the background. It can also generate style-mixed images by combining characteristics from two distinct input images, which is a pioneering attempt in 3D-scene synthesis. Our approach opens up new possibilities for flexible and controllable 3D image generation from real-world data.

INDEX TERMS Neural radiance fields, 3D-scene generation, generative adversarial networks, neural rendering, generative artificial intelligence.

I. INTRODUCTION

Generative Adversarial Networks (GANs) [1] have achieved remarkable success in realistic image generation tasks, particularly following the introduction of StyleGAN [2]. Numerous 2D-based generative models have since benefited from a deeper understanding of latent spaces [3], [4], leading to significant advancements in various computer vision tasks, such as conditional image generation and style transfer [5], [6]. However, these 2D-based approaches are limited in their ability to generate novel view images and perform 3D manipulations, such as 360-degree rotations and spatial translations, due to their lack of geometric understanding. Specifically, existing 2D-based methods for novel view

synthesis [7], [8] often encounter 3D-consistency issues when attempting to change the object's pose rather than the viewpoint within a scene.

To address this limitation, researchers have turned to neural radiance fields (NeRF) [9], which encode a scene into a multi-layer perceptron (MLP) for 3D rendering. While conventional NeRF [9] has enabled 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 using multiple sets of multi-view images and camera parameters [9], [10]. To overcome this challenge, researchers have integrated NeRF with generative models, giving rise to generative NeRF. These models [11], [12], [13], [14], [15], [16], [17] can be trained on unposed real-world images, unlike conventional NeRF, which requires multiple images of a single scene [18],

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Sharif¹.

[19], [20]. Furthermore, generative NeRF has been used to obtain conditional samples by incorporating class label information [21] or text encoding [22], [23], [24], [25].

The development of generative NeRF models has ushered in a new era of 3D image creation and manipulation, offering exciting possibilities across various industries. These advanced models learn to represent 3D scenes as continuous functions, allowing for unprecedented flexibility in generating new viewpoints and editing scene elements. For instance, in the film industry, generative NeRFs could revolutionize special effects by enabling directors to effortlessly change camera angles or lighting conditions in post-production without the need for costly reshoots. In e-commerce, these models could enhance product visualization, allowing customers to view items from any angle or even place them in virtual environments. In the field of medical imaging, generative NeRF models are also used generating CT images in multiple contrasts and perspectives from a single contrast scan. This could improve diagnostic capabilities by providing clinicians with additional views without requiring repeated scans, reducing both patient exposure to radiation and overall imaging costs. However, the widespread adoption of generative NeRFs faces challenges when applied to real-world images. Current methods often require additional camera data or time-consuming fine-tuning processes, which can be impractical in many scenarios. For example, in the fashion industry, the ability to quickly generate 3D views of new clothing designs from a single photo could streamline the design process, but only if the technology can work efficiently without requiring specialized camera setups. Overcoming these limitations could unlock the full potential of generative NeRFs across diverse fields, from augmented reality applications in education to more immersive virtual experiences in gaming and social media platforms.

Despite the convenience and intuitiveness of these approaches, they are limited in their ability to edit images and generate 3D representations of specific inputs, such as out-of-distribution or real-world images. To enable more practical applications, generative NeRF models have incorporated optimization-based GAN inversion techniques [26], [27], [28], [29] for the 3D representation of particular input images. However, previous approaches require fine-tuning pre-trained models for specific images [30], [31], [32], [33], which hinders their application to numerous real samples simultaneously and renders the process time-inefficient due to the need for extensive fine-tuning. For example, EG3D [11] and PanoHead [34] facilitate 3D image reconstruction but require fine-tuning steps and camera parameters of the input image during training, which often makes them impractical for real-world datasets. Our approach eliminates this requirement, extending its applicability to a broader range of image datasets.

Additionally, as generative models continue to advance and become widely deployed in real-world settings, protecting the intellectual property (IP) of these models has emerged as a critical consideration. Recent research has focused on devel-

oping compression-resistant watermarking and federated ownership verification strategies [35], [36], [37], [38], [39] to ensure that trained models remain secure from unauthorized usage or modifications. While such watermarking methods primarily target federated or edge AI deployments, their underlying principles reflect the growing awareness of potential IP vulnerabilities in large-scale generative frameworks.

In summary, existing generative NeRF techniques still face several practical limitations:

- They often rely on multiple input views or camera parameters, which are cumbersome or infeasible to obtain in real-world scenarios.
- They typically require extensive fine-tuning steps to handle out-of-distribution images, leading to significant computational overhead.

Therefore, in this paper, we introduce a novel zero-shot approach for reconstructing complete 3D scenes from a single 2D photograph using NeRF. Our primary research objective is to develop a model capable of generating multi-view representations of objects or environments from just one image, even if that image was not part of the training dataset. This zero-shot capability sets our method apart, as it can reconstruct 3D scenes from previously unseen inputs without the need for additional training or fine-tuning. By leveraging advanced 3D GAN inversion techniques, our model produces 3D-consistent renderings of real-world images during inference, maintaining spatial coherence across different viewpoints. A key innovation of our approach is its ability to operate without camera parameters or supplementary components typically required for camera pose estimation. This feature significantly enhances the model's applicability to a wide range of real-world scenarios where such information is often unavailable or difficult to obtain. Our method thus represents a significant step forward in the field of 3D scene reconstruction, offering a powerful tool for applications ranging from virtual reality content creation to automated 3D modeling in computer vision tasks.

Our architectural design consists of two distinct components: the 3D-generation module and the 3D GAN inversion module. The former is based on the principles of GIRAFFE [15], which successfully incorporates 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 is an encoder trained with images synthesized from the generator. This 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 are displayed in fig. 1. Additionally, we demonstrate the suitability of the learning-based inverter design over optimization-based approaches by highlighting the limitations of optimization-based methods for this specific

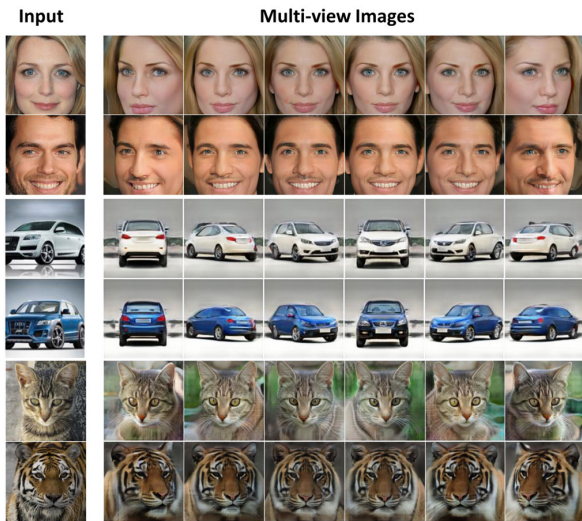


FIGURE 1. Illustration of the 3D reconstruction outcomes with single image inputs achieved using the proposed ZIGNeRF method. The figure showcases the successful application of zero-shot 3D GAN inversion on various real-world image datasets [40], [41], [42].

application, as discussed in section IV-D. The comparison with optimization-based techniques demonstrates their inability to function properly in scenarios where camera parameters are not provided within the input data.

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 a learning-based approach in the field.
- ZIGNeRF exhibits robust 3D feature extraction capabilities and remarkable controllability with respect to input images.
- ZIGNeRF performs 3D scene generation, such as a full 360-degree rotation of real-world car images, a task not fully achieved by many existing generative NeRF models.
- ZIGNeRF can generate 3D scene generation using two input source images, where each image determines the shape and appearance, respectively, which is the first attempt in the domain.

II. RELATED WORKS

A. NEURAL RADIANCE FIELD (NeRF)

NeRF is a groundbreaking approach for creating lifelike 3D scenes from standard 2D images. It models a 3D scene as a continuous function, utilizing an MLP. This MLP transforms spatial coordinates into color (RGB) and density values. Then, it produces new viewpoints of the scene using standard volume rendering methods. Notably, NeRF significantly lowers the computational demands compared to conventional 3D scene modeling methods that use voxels [18], [43], [44], [45], [46]. A limitation of NeRF's training process, which

tailors a single model to each unique scene, is its high demand for a structured set of training images that include the camera's perspectives [18], [47].

B. GENERATIVE NeRF

Generative NeRFs are an extension of the traditional NeRF approach. They aim to train networks to map an internal code to a 3D scene representation using a collection of unposed 2D images, instead of relying on multi-view images with known camera positions. Early models like GRAF [17] and pi-GAN [12] have shown promising results, laying the foundation for further advancements in this area. More recent developments in generative NeRF focus on generating high-resolution images that are consistent in three dimensions. For example, StyleNeRF [14] combines NeRF with a style-based generator to produce high-resolution images. Similarly, EG3D [11] achieves notable results with a hybrid model that improves both computational efficiency and image quality.

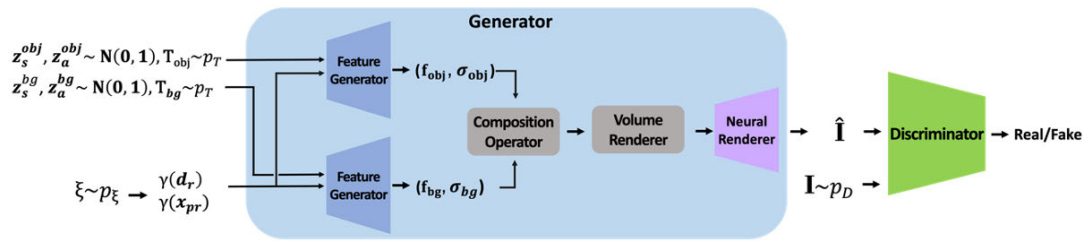
However, practical applications often require specific, targeted samples rather than random ones from the data distribution. Instead of using class-based or text encoding methods common in 2D image generation [48], we utilize GAN inversion as a conditional approach. Traditional conditional generation methods such as class-based or text encoding face two main challenges: they require the training dataset to have specific conditional information (such as labels or text) for each sample, and they do not allow for using real-world 3D images as a conditional input. We address these issues in the existing conditional generative NeRF models by incorporating GAN inversion into generative NeRF, enabling more targeted and practical conditional generation.

C. 3D GAN INVERSION

With the impressive progress in GANs, many researchers have been working on decoding and utilizing their latent spaces. This process, known as GAN inversion, essentially reverses the role of the generator in GANs. The main goal of GAN inversion is to find the latent code for a given image by mapping it back to the generator's latent space. When done effectively, this latent code, obtained via GAN inversion, can accurately reconstruct an image that the pre-trained generator would produce. This also allows for the manipulation of the generated sample by exploring meaningful directions within the latent space [49]. Furthermore, GAN inversion enables the modification of real-world images in this latent space.

Generally, GAN inversion is approached in two ways: through optimization-based and learning-based methods. The optimization-based technique, which is applied to individual samples, starts with a random code and uses gradient descent to find the correct latent code. The learning-based method, on the other hand, trains an encoder to map samples from a dataset to the generator's manifold. In our study, we use the learning-based approach for zero-shot GAN inversion, especially in designing the inverter of ZIGNeRF.

Training process of the 3D generation part



Training process of the Inverter

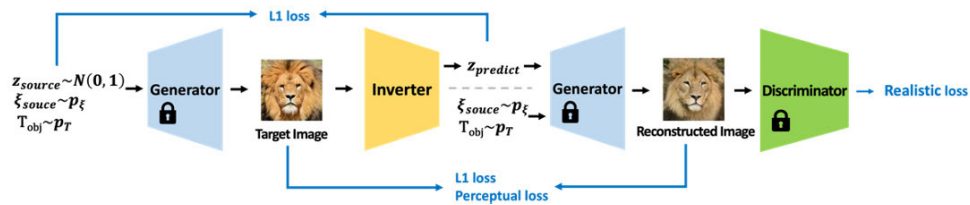


FIGURE 2. Overview of the ZIGNeRF architecture. The 3D generative module is trained to generate photorealistic images that maintain 3D consistency by mapping the latent code and camera pose to a synthesized image. After training the generator, the inverter is trained using the pre-trained generator and discriminator.

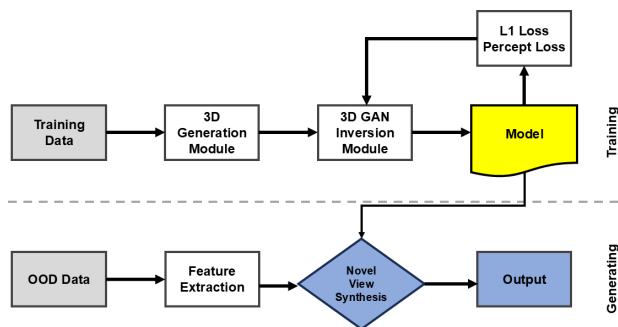


FIGURE 3. A unified view of the proposed ZIGNeRF framework during training (top) and generating (bottom).

Some research has explored 3D GAN inversion with generative NeRFs to create multi-view images from input samples and to manipulate these samples within 3D spaces. Most of these studies have adjusted the pre-trained generator using optimization-based GAN inversion methods. However, this extra step of fine-tuning the generator introduces challenges in terms of flexibility and computational demands. Furthermore, previous methods often needed an additional module to estimate the camera parameters for the input image.

In this study, we introduce a novel inverter specifically designed for zero-shot 3D GAN inversion. This inverter can map images that are not part of the original dataset (out-of-distribution images) into the generator's latent space. Our model is capable of creating 3D representations of real-world images without requiring extra training steps or information about the spatial arrangement of the input image.

Our approach to zero-shot 3D GAN inversion significantly enhances usability, as the trained model can be directly applied to images that were not included in the original training set.

III. METHODS

Our proposed model, ZIGNeRF, aims to generate multi-view images from a single input image that is not part of the original training dataset, a process known as out-of-distribution imaging. This is achieved by integrating the concepts of generative NeRF with GAN inversion techniques. As illustrated in fig.2, our approach consists of two main components: a 3D generation module and a 3D GAN inversion module. The 3D generation module is responsible for generating high-quality 3D images, while the 3D GAN inversion module is designed to map out-of-distribution images into the latent space of the 3D generator.

The 3D generation module is built upon the GIRAFFE [15] architecture, which has demonstrated success in incorporating 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 components. These modifications include the use of residual blocks and spectral normalization to stabilize training and improve the quality of the generated images.

The 3D GAN inversion module, on the other hand, consists of an inverter network that is trained alongside the pre-trained generator. The inverter is designed to map out-of-distribution images into the latent space of the generator, allowing the generator to produce multi-view images of the

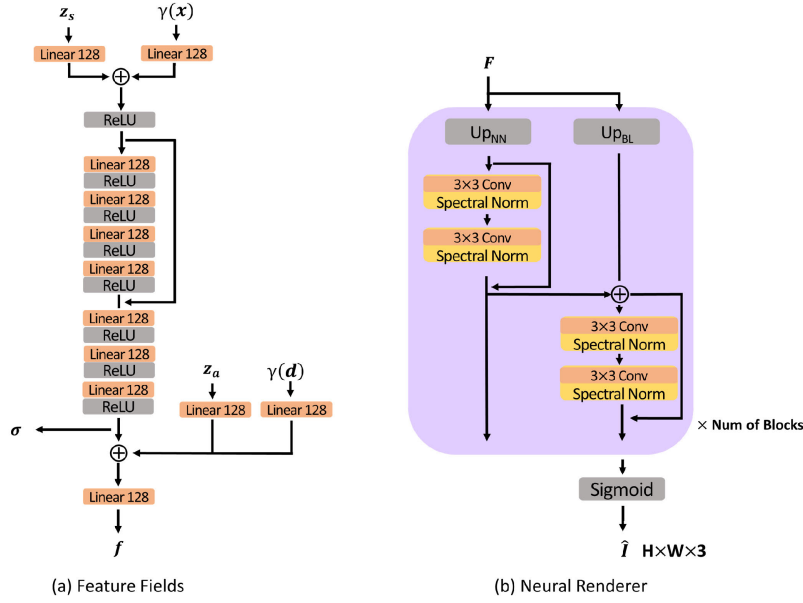


FIGURE 4. Structure of the feature fields and neural renderer components. (a) The feature fields are parameterized using multi-layer perceptrons (MLPs), which map the 3D point $\mathbf{x}$, viewing direction $\mathbf{d}$, and latent codes $\mathbf{z}_s, \mathbf{z}_a$ into a volume density σ and feature $\mathbf{f}$. (b) The neural renderer blocks illustrate the process of transforming the volume-rendered feature image $\mathbf{F}$ into the final synthesized image $\hat{\mathbf{I}}$. The nearest neighbor upsampling operations are denoted as UP_{NN} and UP_{BL} , respectively.

input image. To train the inverter, we use images synthesized by the generator, which inherently contain 3D information. This approach enables the inverter to learn the mapping between real-world images and the generator's latent space, without requiring additional fine-tuning when applied to unseen images during inference.

We present a simple flowchart in fig.3, providing an overview of the rough workflow of our model. In the following subsections, we provide a detailed description of the 3D generation module and the 3D GAN inversion module, as well as the training specifications for our proposed model.

A. 3D GENERATION MODULE

Compositional Generative Neural Feature Field. The core of our 3D generation module is a compositional generative neural feature field, which is inspired by the GIRAFFE [15] architecture. This approach represents a 3D scene as a combination of feature fields, each corresponding to an individual object, with the background treated as an object itself. The structure of the feature fields is parameterized using MLPs, as depicted in fig.4 (a). The MLPs take a three-dimensional point $\mathbf{x} \in \mathbb{R}^3$, a viewing direction $\mathbf{d} \in \mathbb{S}^2$, and latent codes $\mathbf{z} \sim \mathcal{N}(0, 1)$ as input, and output a volume density $\sigma \in \mathbb{R}^+$ and a high-dimensional feature field $\mathbf{f} \in \mathbb{R}^{M_f}$. This representation allows for more flexibility and detail in generating 3D images, particularly in how different elements of the scene are perceived from various viewpoints.

The compositional generative neural feature field enables the representation of complex 3D scenes by decomposing

them into individual objects and a background. This allows for fine-grained control over the appearance and spatial arrangement of each object in the scene. By applying affine transformations to the objects, our model can manipulate their scale, translation, and rotation, providing a high degree of flexibility in generating diverse 3D scenes.

To enable object-level control, we apply affine transformations to the objects in the scene, allowing for manipulation of scale, translation, and rotation:

$$T = \{\mathbf{s}, \mathbf{t}, \mathbf{R}\}, \quad (1)$$

where $\mathbf{s}, \mathbf{t} \in \mathbb{S}$ represent scale and translation parameters, respectively, and $\mathbf{R} \in \text{SO}(3)$ represents rotation. The affine transformation generates a bounding box corresponding to the transformation parameters T for a single object:

$$\tau = \mathbf{R} \cdot \mathbf{s} \mathbf{I} \cdot \mathbf{t}, \quad (2)$$

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

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

$$\mathbf{z} = [\mathbf{z}_s^1, \mathbf{z}_a^1, \dots, \mathbf{z}_s^N, \mathbf{z}_a^N], \quad (4)$$

where $\gamma(\cdot)$ is the positional encoding function [9] applied separately to $\mathbf{x}$ and $\mathbf{d}$, and $C(\cdot)$ is the compositional operator that combines the feature fields from the $N - 1$ objects and the background. Instead of directly outputting the final image, we volume render the composited volume density and feature field. The resulting 2D feature map $\mathbf{F}$, which is fed into the

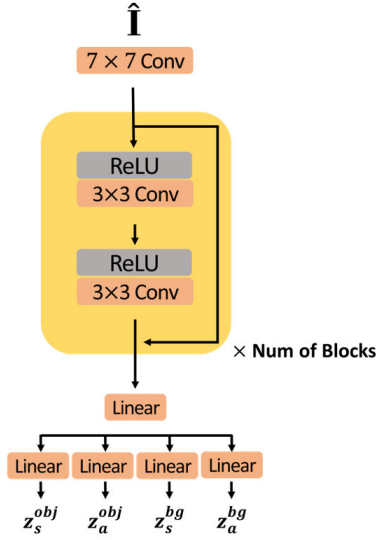


FIGURE 5. Diagram illustrating the architecture of the inverter used in ZIGNeRF.

neural renderer for the final synthesized output, is obtained using the volume rendering function π_v :

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

Neural Renderer with Residual Networks. The neural renderer in our model is responsible for generating the final synthetic image from the volume-rendered feature map. We observed that the original neural renderer in GIRAFFE does not adequately preserve the features, and the learning rates of the decoder and neural renderer are not synchronized, leading to unstable training of the generator.

To address these issues, we propose an improved neural renderer that replaces the 3×3 convolution layer blocks with residual blocks [50] and employs the ReLU activation function instead of the leaky ReLU [51] for faster and more effective rendering. We also adopt spectral normalization [52] as a weight normalization technique to stabilize the neural rendering process. Experimental results demonstrate that our modified neural renderer improves the stability of training and the quality of the generated outputs. The neural renderer, which maps the feature map $\mathbf{F}$ to the final image $\hat{\mathbf{I}} \in \mathbb{R}^{H \times W \times 3}$, is parameterized as:

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

Figure 4 (b) illustrates the architecture of the neural renderer, depicting the transformation of the volume-rendered feature image $\mathbf{F}$ into the final synthesized image $\hat{\mathbf{I}}$.

Discriminator. Similar to the vanilla GAN [1], the discriminator in our model outputs a probability indicating whether the input image is real or fake. We replace the 2D CNN-based discriminator with residual blocks that employ spectral normalization as a weight normalization technique. Figure 6 provides a detailed description of the discriminator architecture. The input image is first processed through a

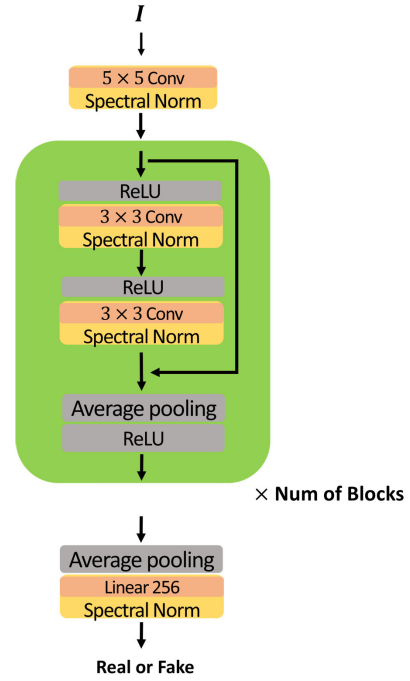


FIGURE 6. Structure of the discriminator network. The input image is passed through residual convolution blocks with spectral normalization, followed by an average pooling operation. The final linear layer with spectral normalization produces the output probability.

series of residual convolution blocks with spectral normalization, followed by an average pooling operation. The output probability is obtained after passing the pooled features through a final linear layer with spectral normalization.

Objectives. The overall objective function for the 3D generation module is defined as:

$$L_{G,D} = L_{GAN} + \lambda L_{R1}, \quad (7)$$

where λ is set to 10. We optimize the network using the GAN objective [1] with the R1 gradient penalty [53].

B. 3D GAN INVERSION MODULE

To map a given image into the latent space of the generator, we introduce a novel inverter network. As illustrated in fig.5, the inverter first applies a 7×7 convolution to the input image and then stacks multiple residual encoder blocks depending on the image resolution (for example, six residual encoder blocks for a 256×256 image). Each residual block consists of two ReLU activations and 3×3 convolutions, enabling deeper feature extraction and thus providing robustness against diverse input distributions, including out-of-distribution images. At the end of the inverter, four linear output layers are used to estimate the latent codes z_s^{obj} , z_a^{obj} , z_s^{bg} , and z_a^{bg} of the input image. This inverter's architecture is designed to mirror key aspects of the generator's feature extraction process, ensuring that the latent representations it produces are well-aligned with the generator's manifold.

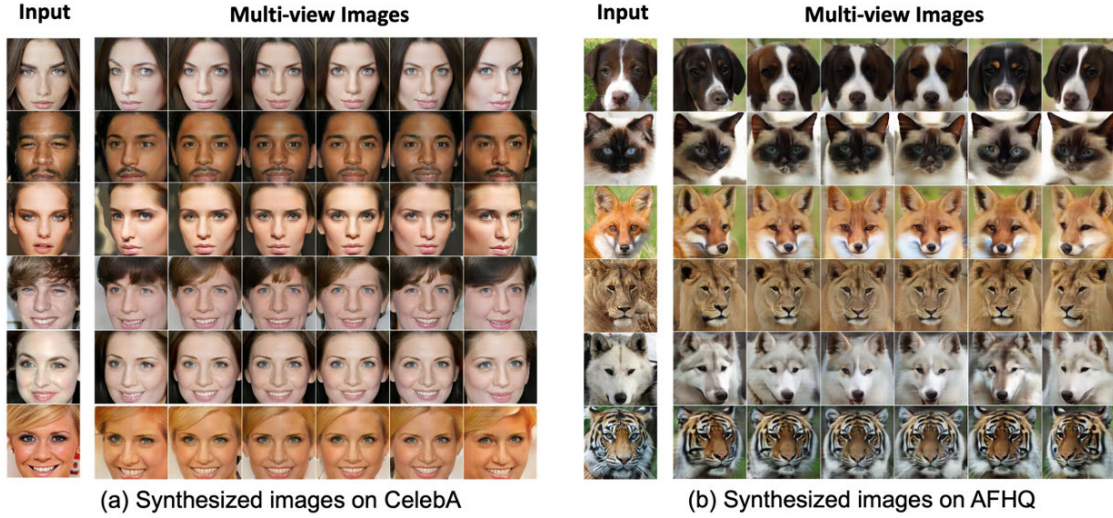


FIGURE 7. Examples of 256^2 multi-view synthesis results on facial datasets: CelebA-HQ [41] and AFHQ [40].



FIGURE 8. Examples of reconstructed images generated from an input car image [42] after applying compositional operations. These results demonstrate the effective separation of the object from the background and the model's ability to provide 3D controllability.

The main challenge in 3D GAN inversion is to map multi-view images of a single object to a unique latent code. To train the inverter, we use the synthesized images $\hat{\mathbf{I}}$ generated by the pre-trained generator as the training data. Since we have access to the source parameters of the generated images, the inverter only needs to estimate the latent code $z^{predict}$ of the input image. The generated training images enable the inverter to learn to extract features from unseen images with varying viewing directions, scales, and rotations. After estimating the latent code, the pre-trained generator reconstructs the input image using $z^{predict}$ and the source parameters, which include the camera pose ξ^{source} and the compositional parameters $T^{source} = \{\mathbf{s}, \mathbf{t}, \mathbf{R}\}$:

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

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

During the training of the inverter, we found that using only the L1 loss between the estimated and ground-truth latent codes in the latent space was insufficient for reconstructing the scene. Therefore, we employ a combination of losses, including a realistic loss calculated using the output of the discriminator, an L1 loss at the image level, and two perceptual losses: the Structural Similarity Index Measure

(SSIM) [54] and the Learned Perceptual Image Patch Similarity (LPIPS) [55]. The inverter is optimized using the following objective function:

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

where $\hat{\mathbf{I}}^{predict}$ is the image reconstructed by the pre-trained generator using the estimated latent code $z^{predict}$. L_{real} represents the realistic loss, L_{latent} and L_{reconst} are the latent-level and image-level L1 losses, respectively, and L_{percept} is the perceptual loss computed using the SSIM and LPIPS metrics.

C. TRAINING SPECIFICATIONS

During training, we randomly sample the latent codes $z_s, z_a \sim \mathcal{N}(0, 1)$ and the camera pose $\xi \sim p_{\xi}$. The hyperparameters λ_1, λ_2 , and λ_3 in the inverter's objective function are set to 10, 100, and 1, respectively. We optimize the model using the RMSProp optimizer [57] with learning rates of 1×10^{-4} , 7×10^{-5} , and 1×10^{-4} for the generator, discriminator, and

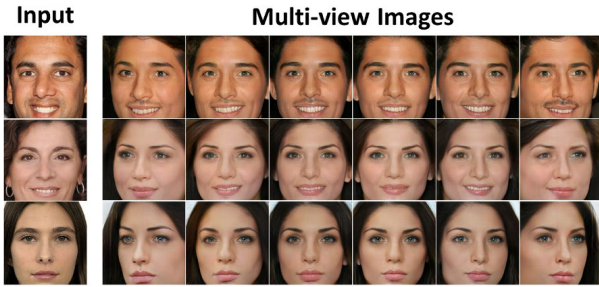


FIGURE 9. Examples of 256^2 synthesized images conditioned on input FFHQ [2] images, generated by the model trained on the CelebA-HQ dataset [41].

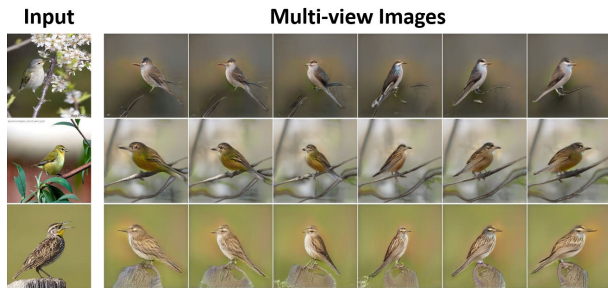


FIGURE 10. Examples of 256^2 multi-view synthesis results on the CUB-200-2011 bird dataset [56].

inverter, respectively. The batch size is set to 32. We train the generator and discriminator for the first 100,000 iterations, followed by 50,000 iterations of training for the inverter while keeping the generator and discriminator fixed.

All training experiments were conducted on a single NVIDIA A6000 GPU. For large-scale datasets such as CelebA-HQ, which contains over 30,000 images, training at a resolution of 128×128 takes approximately 30 hours. When training at higher resolutions (e.g., 256×256), the training time increases approximately threefold. For smaller datasets with fewer images, such as AFHQ or CUB-200-2011, the training time was much lower depending on the dataset complexity and diversity. After training is completed, rendering a single test image takes approximately 1s, demonstrating the feasibility of real-time rendering. This efficiency allows ZIGNeRF to be applicable in real-world production environments, including interactive applications and real-time 3D content generation.

By integrating the 3D generation module and the 3D GAN inversion module, our proposed ZIGNeRF model is capable of generating multi-view images from a single input image, even when the input image is not part of the original training dataset. The 3D generation module ensures high-quality 3D image synthesis, while the 3D GAN inversion module enables the model to map out-of-distribution images into the generator's latent space, facilitating zero-shot 3D reconstruction.

IV. RESULTS

We evaluate the performance of ZIGNeRF in terms of zero-shot feature extraction, 3D controllability, and adaptability. The model is tested on five real-world datasets:

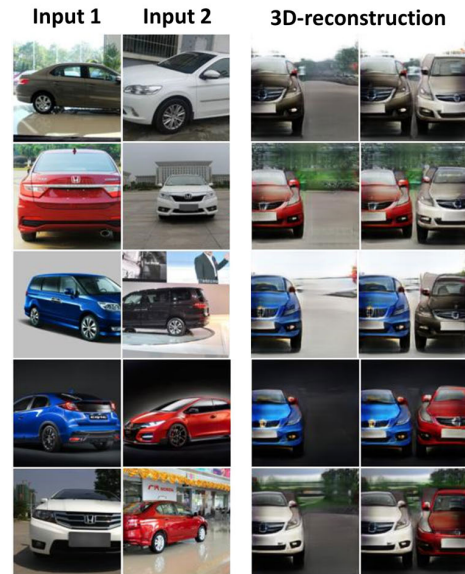


FIGURE 11. Generation of two objects within a single scene. The results demonstrate the model's ability to represent compositional scenes by generating two objects in a single scene. The inverter maps two input images into two sets of latent codes, and the generator, trained on single-object scenes, synthesizes a single scene containing two independent objects.

Cats, AFHQ [40], CelebA [58], CelebA-HQ [41], and CompCar [42]. Two additional datasets, FFHQ [2] and CUB-200-2011 [56], are used to demonstrate the robust adaptation capabilities of the proposed model. It is important to note that 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. Additionally, given real-world images, we do not use any spatial information for inverting input images into latent codes. We begin with a visual validation of the proposed model, examining the similarity between the input image and the reconstructed images, as well as the 3D-consistent controllability. The model is then evaluated using the Fréchet Inception Distance (FID) [59] and Kernel Inception Distance (KID) [60] as a metric. We further highlight the limitations of optimization-based methods with visual evidence. Moreover, we compare our model with an optimization-based method using the Learned Perceptual Image Patch Similarity (LPIPS) [55] metric. Finally, we conduct ablation studies to validate the efficacy of the loss function in optimizing the inverter.

A. CONTROLLABLE 3D SCENE RECONSTRUCTION

We visually demonstrate that our proposed model generates multi-view consistent images corresponding to the input image. Figure 7 showcases 3D reconstruction on CelebA-HQ [41] and AFHQ [40], substantiating that the inverter successfully extracts facial features irrespective of gender or skin color in human faces, and species in animal faces. Figure 8 exhibits the model's controllability and object disentanglement with CompCar [42], indicating that the

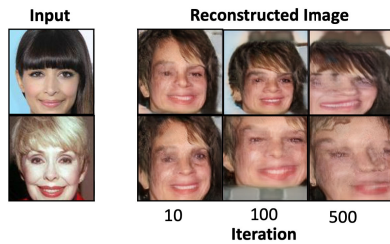


FIGURE 12. Visual evaluation of the optimization-based method.

inverter effectively estimates the latent code of the object and background. 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. 9. Additionally, we demonstrate the application of our model to the CUB-200-2011 dataset [56], as shown in fig. 10. Although CUB-200-2011 has a limited number of 11,788 samples and is characterized by complex backgrounds, our model still produces high-quality outputs from these limited inputs.

B. EXTENDED EXPERIMENTAL RESULTS

In this section, we present additional application results of the proposed model through fig. 11, 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 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 [42] images are synthesized into one image. Similar to 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 by applying the proposed inverter structure, utilizing the CelebA-HQ and AFHQ datasets [41]. 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.

C. 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 table 1 and table 2.

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 [15] model. As an illustration, for the CelebA(HQ) 256^2 dataset, ZIGNeRF achieves a score of 14.98, substantially lower than GIRAFFE's score of 23.14. This is indicative of the model's ability to produce higher-quality images with fewer iterations.

The results presented in table 1 clearly demonstrate the superior performance of ZIGNeRF compared to the baseline GIRAFFE model across all datasets and resolutions. In the unconditional setting, ZIGNeRF consistently achieves lower FID scores, indicating its ability to generate higher-quality images that more closely resemble the real data distribution. For example, on the CompCar dataset at 256^2 resolution, ZIGNeRF obtains an FID score of 22.57, which is significantly lower than GIRAFFE's score of 40.84. This substantial improvement in image quality highlights the effectiveness of the modifications introduced in ZIGNeRF's 3D generation module, such as the use of residual blocks and spectral normalization in the neural renderer.

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 demonstrates 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 GIRAFFE's score of 35.03.

Furthermore, ZIGNeRF's superior performance extends to the conditional setting, where the model generates images based on the latent codes estimated by the inverter. As shown in table 1, ZIGNeRF consistently outperforms GIRAFFE across all datasets and resolutions in the conditional mode. For instance, on the Cats dataset at 128^2 resolution, ZIGNeRF achieves an FID score of 15.06, which is considerably lower than GIRAFFE's score of 24.01. This improvement can be attributed to the effectiveness of ZIGNeRF's 3D GAN inversion module, which accurately maps out-of-distribution images into the generator's latent space, enabling the model to generate high-quality 3D representations of real-world images. These results underscore the importance of the learning-based inverter design in ZIGNeRF, which eliminates the need for additional fine-tuning steps and enables zero-shot 3D reconstruction.

The superior performance of ZIGNeRF in both unconditional and conditional settings highlights the effectiveness of our proposed approach. By integrating a powerful 3D generation module with a learning-based 3D GAN inversion module, our model is able to generate high-quality 3D

TABLE 1. Comparison of the FID scores between the proposed ZIGNeRF model and a baseline model. The models were trained on six different datasets at resolutions of 128^2 and 256^2 .

Method	Models	Cats		CelebA(HQ)		CompCar		AFHQ		CUB-200-2011		FFHQ
		128^2	256^2	128^2	256^2	128^2	256^2	128^2	256^2	128^2	256^2	128^2
Unconditional	GIRAFFE	24.01	21.28	19.45	23.14	38.91	40.84	35.03	38.18	81.16	74.31	64.02
	ZIGNeRF(ours)	12.31	11.21	11.01	14.98	22.67	22.57	12.81	19.96	33.11	29.38	34.77
Conditional	ZIGNeRF(ours)	15.06	16.83	14.77	25.66	25.97	25.41	14.02	28.78	33.73	32.21	34.31

TABLE 2. Comparison of the KID (mean) scores between the proposed ZIGNeRF model and a baseline (e.g., GIRAFFE). The models were trained on four different datasets at a resolution of 128^2 .

Method	CelebA-HQ	AFHQ	FFHQ	CUB-200-2011
GIRAFFE	0.0418	0.0126	0.0492	0.1387
ZIGNeRF (ours)	0.0204	0.0541	0.0315	0.0554

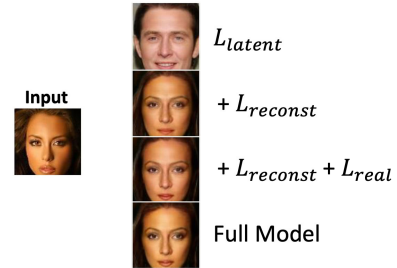
TABLE 3. Quantitative comparison of DreamGS, Hunyuan3D, and our method (ZIGNeRF) on the CUB-200-2011 dataset. We evaluate image quality using three metrics: SSIM, LPIPS with AlexNet and VGG backbones, and PSNR.

Method	SSIM ($\uparrow$)	LPIPS (Alex) ($\downarrow$)	LPIPS (VGG) ($\downarrow$)	PSNR ($\uparrow$)
DreamGS [61]	0.4154	0.7921	0.8057	5.6106
Hunyuan3D [62]	0.3826	0.7937	0.7978	5.8022
Ours	0.6784	0.4229	0.4213	20.6020

representations of real-world images, outperforming baseline methods in terms of image quality and consistency.

Additionally, we present the KID scores in table 2 to provide a complementary assessment of distribution-level similarity between generated and real images. As shown, ZIGNeRF achieves lower KID values than GIRAFFE on CelebA-HQ, FFHQ, and CUB-200-2011, demonstrating a closer match to the underlying data distribution in these datasets. For instance, on FFHQ dataset, ZIGNeRF achieves a KID score of 0.0315, which is considerably lower than GIRAFFE's score of 0.0492. By leveraging this metric, we confirm that ZIGNeRF not only excels in producing visually convincing results but also maintains strong distributional fidelity, further highlighting the robustness and effectiveness of our model for 3D-consistent image generation.

Furthermore, table 3 provides a quantitative comparison between our model (ZIGNeRF) and two state-of-the-art (SOTA) zero-shot image-to-3D methods, DreamGS [61] and Hunyuan3D [62], on the CUB-200-2011 dataset. Specifically, we compare the models using SSIM, LPIPS (with AlexNet and VGG), and PSNR to gauge the visual quality of the generated images. While these metrics are not perfectly aligned with the zero-shot image-to-3D generation objective, given that DreamGS and Hunyuan3D remove the background for 3D rendering, they still offer insight into each model's capacity for reconstructing detailed and perceptually coher-

**FIGURE 13.** Ablation study examining the impact of the loss functions used in training the inverter within ZIGNeRF.

ent images. As evidenced by table 3, ZIGNeRF consistently achieves higher (better) scores in SSIM and PSNR, and lower (better) values in LPIPS, indicating superior image fidelity relative to DreamGS and Hunyuan3D. Although DreamGS and Hunyuan3D focus primarily on object-centric generation, our method preserves both the object and background, thus presenting a more complete 3D scene. Despite the caveat that these traditional 2D-based metrics may not perfectly encapsulate all aspects of zero-shot 3D synthesis, the results clearly affirm ZIGNeRF's strong generative performance and robust zero-shot capabilities. Finally, additional qualitative comparisons with these SOTA models can be found in fig. 16 and section IV-F.

D. COMPARISON WITH OPTIMIZATION-BASED GAN INVERSION APPROACH

We observe the limitations of optimization-based GAN inversion approaches for this specific application. Generally, optimization-based GAN inversion methods tend to restore target images more accurately, albeit at a slower speed. However, in the setting of a 3D-aware generative pipeline, optimization-based methods are contingent upon the availability of camera parameters for training images to align the viewing direction of the restored image. For evaluation, we performed optimization-based GAN inversion using a gradient descent approach with the Adam optimizer. Our experimental evidence, as depicted in fig. 12, shows that the optimization-based method does not properly restore the input images. In contrast, our model operates effectively without the need for camera parameters.

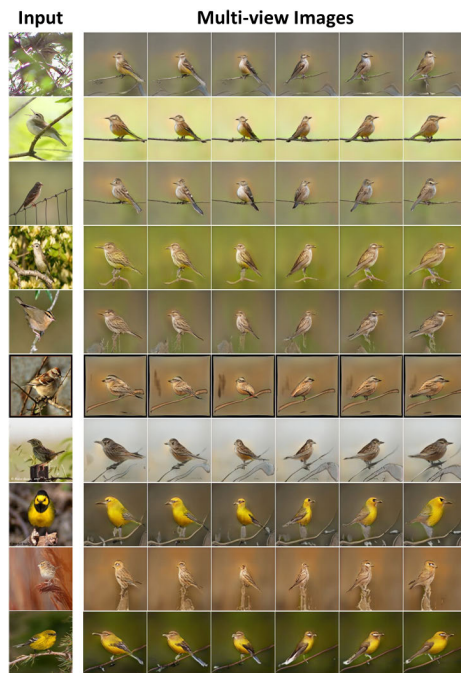
Furthermore, we conduct additional experiments with the LPIPS metric for evaluating input preservation. LPIPS

TABLE 4. Comparison of input preservation using LPIPS (↓) [55] between optimization-based and learning-based GAN inversion methods. Lower scores indicate better performance.

Method	Cats		CelebA(HQ)		CompCar		AFHQ		CUB-200-2011	
	128 ²	256 ²	128 ²	256 ²	128 ²	256 ²	128 ²	256 ²	128 ²	256 ²
Optimization	0.46	0.43	0.34	0.40	0.56	0.53	0.48	0.48	0.75	0.78
Learning(ours)	0.22	0.24	0.23	0.25	0.40	0.43	0.23	0.26	0.59	0.58

TABLE 5. FID scores for the ablation study. The full model is trained with latent loss, reconstruction loss, GAN loss, and perceptual loss.

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

**FIGURE 14.** Additional results using 256² CUB-200-2011 images as input.

measures perceptual similarity using AlexNet pretrained with ImageNet. We compared 1000 pairs of generated images and their corresponding target images. As shown in table 4, our learning-based GAN inversion method is superior to the optimization-based GAN inversion approaches for all datasets.

E. ABLATION STUDY

In the interest of validating the loss function deployed in training the inverter, we undertake an ablation study. The 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 13 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 color, and texture.

Table 5 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 FID experiences a steady enhancement with each loss component incrementally added to the naive model, which originally only employs the latent loss. Initially, the model trained with only the latent loss (L_{latent}) exhibits a relatively high FID score of 80.08. This elevated score indicates a considerable discrepancy between the distribution of the generated images and the real images, suggesting that solely relying on latent loss is insufficient for producing high-fidelity images.

The incorporation of the reconstruction loss ($L_{reconst}$) leads to a significant improvement, as evidenced by the substantial drop in the FID score to 17.82. This demonstrates that aligning the reconstructed image with the input at the image level substantially enhances the quality of the generated images. The reconstruction loss effectively guides the inverter to create images that are not only more visually appealing but also more aligned with the input images' structural and content characteristics.

Further improvement is observed when the GAN loss (L_{real}) is added, reducing the FID score to 15.53. This indicates that the inclusion of the GAN loss contributes to enhancing the realism of the generated images. It pushes the generator to produce outputs that are not only structurally consistent with the input images but also possess a level of realism that closely mimics the distribution of real-world images.

The full model, employing the loss function of ZIGNeRF, achieves the lowest FID score of 14.77. This underscores the synergistic effect of the combined loss functions. The perceptual loss, in particular, plays a crucial role in refining the fine details, such as skin color and texture, which are pivotal for the realistic portrayal of images. It enhances

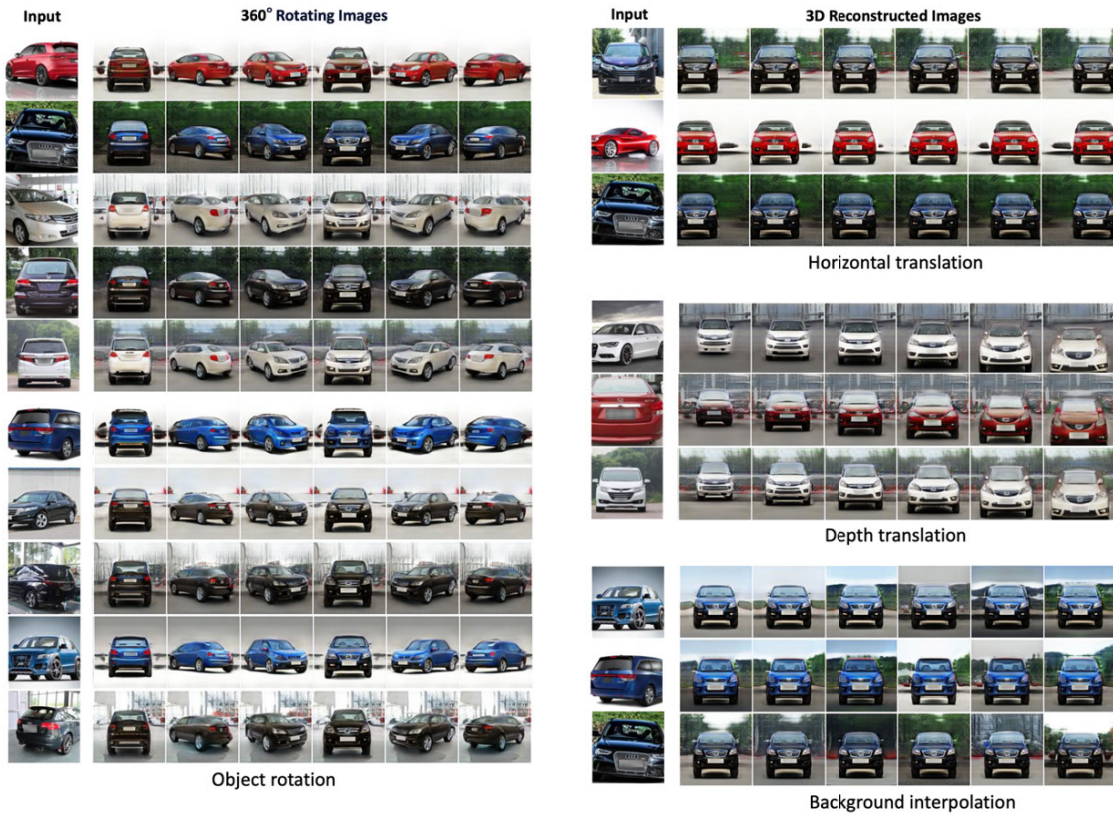


FIGURE 15. Controllable image synthesis using 128^2 CompCars images as input.

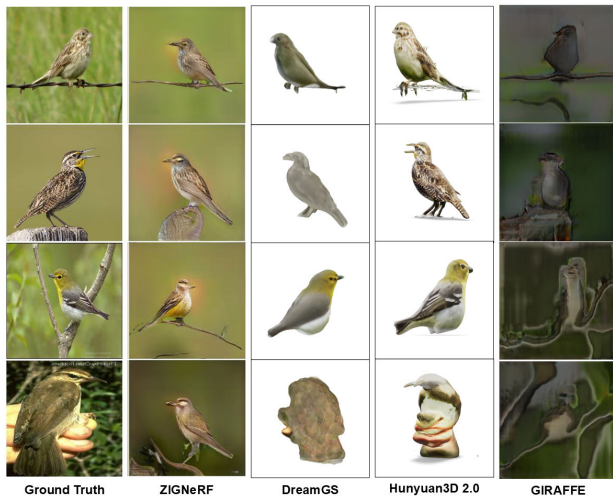


FIGURE 16. Comparison of ZIGNeRF, DreamGS, Hunyuan3D, and GIRAFFE on the CUB-200-2011 dataset. ZIGNeRF and GIRAFFE generate results for unseen directions.

the model's capacity to capture and replicate subtle details that contribute significantly to the perceptual quality of the generated images.

In summary, the stepwise addition of each loss component and the corresponding decrease in the FID score clearly

validate the effectiveness of our composite loss function approach. Each component contributes uniquely to the overall performance, ensuring that the generated images are not only realistic and detailed but also maintain a strong resemblance to the input images.

F. ADDITIONAL QUALITATIVE RESULTS

Figure 14 and 15 deliver additional examples on CUB-200-2011 [56] and CompCar [42] datasets. We embark on a rigorous evaluation of our model using a diverse range of input images sourced from varied datasets. It is worth noting that these results, encompassing distinct categories, are obtained using a single model with different conditional vector inputs, thereby highlighting the large capacity of our model. Moreover, the utilization of the CUB-200-2011 dataset demonstrates the model's adeptness in processing small inputs. By the inherent characteristics of the CUB-200-2011 dataset, which contains complex backgrounds, the model's ability to maintain rendering fidelity under limited conditions is demonstrated. It can also be seen that our model renders well on images containing objects such as borders or branches.

Furthermore, fig.16 presents a comparison among ZIGNeRF, DreamGS [61], Hunyuan3D [62], and GIRAFFE on the CUB-200-2011 dataset. While ZIGNeRF, DreamGS,



FIGURE 17. Multi-view images generated by style mixing two CelebA-HQ input images.

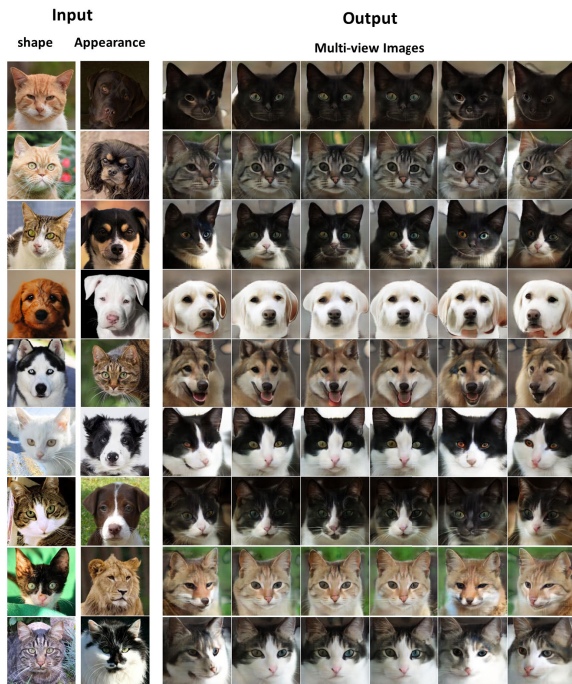


FIGURE 18. Multi-view images generated by style mixing two AFHQ input images.

and Hunyuan3D all aim for zero-shot image rendering, DreamGS leverages a Gaussian Splatting framework, and Hunyuan3D employs both diffusion and transformer architectures. As a result, these methods differ substantially

from each other, making direct comparisons challenging under our current evaluation setup. Nonetheless, the figure offers valuable insight into how each model handles unseen viewpoints. Although the CUB-200-2011 dataset is relatively small, ZIGNeRF demonstrates stronger generalization and image fidelity than DreamGS and GIRAFFE for novel-view synthesis, reinforcing the robustness of our approach in zero-shot 3D reconstruction. Hunyuan3D exhibits strong reconstruction capabilities as well; however, both DreamGS and Hunyuan3D remove the background from the input images and focus solely on rendering the main object, which can be considered a limitation. In contrast, because our model builds upon GIRAFFE as its baseline, ZIGNeRF is able to disentangle the object from the background while preserving key background details that remain visually consistent with the original input image.

The CompCars dataset allows us to experiment with 360-degree image generation using real image inputs representing various car models, colors, and camera poses. It is important to note that a significant advantage of our model is the freedom it provides in the longitudinal movement of objects, along with the capacity to alter the background. This flexibility underpins the model's capacity for highly controllable image synthesis, an attribute that holds immense potential for a wide array of applications.

Figure 17 and Figure 18 exhibit additional style-mixed 3D image synthesis examples on CelebA-HQ and AFHQ datasets. The latent vectors of the style-mixed images are obtained from two distinct source images. The results signify that our inverter successfully disentangles features of source images.

V. CONCLUSION

In this paper, we introduced ZIGNeRF, a novel approach that combines zero-shot 3D GAN inversion with generative NeRFs to enable 3D representation of real-world images. The proposed inverter maps input images onto a latent manifold learned by the generator. During inference, ZIGNeRF generates a 3D reconstructed scene from a single real-world image using the latent code obtained from the inverter, without requiring camera parameters or additional training steps. This addresses a key limitation of existing methods that often necessitate fine-tuning or camera information for 3D reconstruction.

Through extensive experiments on multiple datasets, including CelebA-HQ, AFHQ, CompCars, and CUB-200-2011, we demonstrated that our inverter effectively extracts features from input images with varying poses, validating the 3D controllability and rapid adaptation capabilities of ZIGNeRF. Our model significantly outperforms optimization-based methods in scenarios where camera parameters are not provided, as evidenced by both qualitative and quantitative results.

The ability to generate 3D representations from a single input image has numerous potential applications across various industries. In the field of entertainment and media,

ZIGNeRF could revolutionize visual effects and animation by allowing artists to create 3D models from 2D concept art or photographs, significantly reducing production time and costs. In the domains of augmented reality (AR) and virtual reality (VR), our approach could streamline the creation of immersive environments from real-world data, benefiting applications in education, training, and entertainment. For instance, historical sites could be reconstructed in 3D for virtual tours, or complex machinery could be modeled for interactive training simulations. In the automotive industry, ZIGNeRF could aid in rapid prototyping and design iterations by generating 3D models from sketches or photographs. Furthermore, in the medical imaging domain, ZIGNeRF could be utilized to enhance CT and MRI image reconstruction, enabling the generation of multi-contrast images from a single view or limited-angle data, which could improve diagnostic accuracy and reduce patient exposure to radiation. In remote sensing image change detection (RSICD) tasks, our approach could help overcome challenges such as inconsistencies between different sensor modalities and the need for extensive labeled data by learning cross-domain representations, enabling seamless image translation and synthesis across different sensor types [63].

The novel zero-shot approach introduced in this work represents a significant contribution to the field of 3D image representation, with the potential for broad applications as our pipeline can be generally applied to other existing generative NeRFs. By enabling the creation of 3D models from real-world data without the need for extensive datasets or camera information, ZIGNeRF addresses a critical gap in current 3D reconstruction techniques.

However, real-world deployment of ZIGNeRF presents challenges, primarily due to the inherent computational inefficiency of NeRF-based models, which require extensive sampling and optimization for 3D scene representation. This results in high inference times, making real-time applications impractical, especially on resource-constrained devices. A promising solution is 3D Gaussian Splatting, which replaces dense volumetric sampling with efficient Gaussian primitives, enabling faster and more memory-efficient rendering. Integrating this approach into ZIGNeRF could significantly improve computational efficiency and robustness against extreme object poses and occlusions, making it more viable for real-world applications.

In future work, we plan to extend ZIGNeRF by manipulating the inverted latent code to enable editing of the input image, further enhancing the capabilities of this innovative model. This could open up new possibilities in fields such as digital art, where artists could manipulate 3D representations of their 2D work, or in forensics, where investigators could reconstruct and analyze 3D scenes from limited photographic evidence.

We believe that ZIGNeRF opens up exciting new possibilities for flexible and controllable 3D image generation from real-world data. By addressing the limitations of

existing methods and providing a more accessible approach to 3D reconstruction, our work contributes to the democratization of 3D content creation and has the potential to accelerate innovation across a wide range of industries and applications.

ACKNOWLEDGMENT

An earlier version of this paper was presented at the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2024 [DOI: 10.1109/WACV57701.2024.00491]. Kanghyeok Ko is currently with Samsung Electronics. Sungyup Kim is currently with the School of Computing, Korea Advanced Institute of Science and Technology (KAIST).

(Kanghyeok Ko and Sungyup Kim contributed equally to this work.)

REFERENCES

- [1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–12.
- [2] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.
- [3] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2021, pp. 852–863.
- [4] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116.
- [5] L. Höllein, J. Johnson, and M. Nießner, "StyleMesh: Style transfer for indoor 3D scene reconstructions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6188–6198.
- [6] K. Ko, T. Yeom, and M. Lee, "SuperstarGAN: Generative adversarial networks for image-to-image translation in large-scale domains," *Neural Netw.*, vol. 162, pp. 330–339, May 2023.
- [7] Y. Shen, C. Yang, X. Tang, and B. Zhou, "InterFaceGAN: Interpreting the disentangled face representation learned by GANs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2004–2018, Apr. 2022.
- [8] Y. Shen and B. Zhou, "Closed-form factorization of latent semantics in GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1532–1540.
- [9] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, Jan. 2022.
- [10] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "PixelNeRF: Neural radiance fields from one or few images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4576–4585.
- [11] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. de Mello, O. Gallo, L. Guibas, J. Tremblay, S. Khamis, T. Karras, and G. Wetzstein, "Efficient geometry-aware 3D generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16102–16112.
- [12] E. R. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein, "Pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5795–5805.
- [13] Y. Deng, J. Yang, J. Xiang, and X. Tong, "GRAM: Generative radiance manifolds for 3D-aware image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10663–10673.
- [14] J. Gu, L. Liu, P. Wang, and C. Theobalt, "StyleNeRF: A style-based 3D-aware generator for high-resolution image synthesis," 2021, *arXiv:2110.08985*.

- [15] M. Niemeyer and A. Geiger, "GIRAFFE: Representing scenes as compositional generative neural feature fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11448–11459.
- [16] R. OrEl, X. Luo, M. Shan, E. Shechtman, J. J. Park, and I. Kemelmacher-Shlizerman, "StyleSDF: High-resolution 3D-consistent image and geometry generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13493–13503.
- [17] K. Schwarz, Y. Liao, M. Niemeyer, and A. Geiger, "GRAF: Generative radiance fields for 3D-aware image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2020, pp. 20154–20166.
- [18] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, and M. Zollhöfer, "DeepVoxels: Learning persistent 3D feature embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2432–2441.
- [19] M. Tancik, V. Casser, X. Yan, S. Pradhan, B. P. Mildenhall, P. Srinivasan, J. T. Barron, and H. Kretschmar, "Block-NeRF: Scalable large scene neural view synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8238–8248.
- [20] Q. Xu, Z. Xu, J. Philip, S. Bi, Z. Shu, K. Sunkavalli, and U. Neumann, "Point-NeRF: Point-based neural radiance fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5428–5438.
- [21] K. Jo, G. Shim, S. Jung, S. Yang, and J. Choo, "CG-NeRF: Conditional generative neural radiance fields," 2021, *arXiv:2112.03517*.
- [22] R. Gal, O. Patashnik, H. Maron, A. H. Bermanno, G. Chechik, and D. Cohen-Or, "StyleGAN-NADA: CLIP-guided domain adaptation of image generators," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 1–13, Jul. 2022.
- [23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2021, pp. 8748–8763.
- [24] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.
- [25] T. Wei, D. Chen, W. Zhou, J. Liao, Z. Tan, L. Yuan, W. Zhang, and N. Yu, "HairCLIP: Design your hair by text and reference image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18051–18060.
- [26] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: A styleGAN encoder for image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2287–2296.
- [27] D. Roich, R. Mokady, A. H. Bermanno, and D. Cohen-Or, "Pivotal tuning for latent-based editing of real images," *ACM Trans. Graph.*, vol. 42, no. 1, pp. 1–13, Feb. 2023.
- [28] H. Song, Y. Du, T. Xiang, J. Dong, J. Qin, and S. He, "Editing out-of-domain GAN inversion via differential activations," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel. Cham, Switzerland: Springer, Jan. 2022, pp. 1–17.
- [29] J. Zhu, Y. Shen, D. Zhao, and B. Zhou, "In-domain GAN inversion for real image editing," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, Jan. 2020, pp. 592–608.
- [30] J. Ko, K. Cho, D. Choi, K. Ryoo, and S. Kim, "3D GAN inversion with pose optimization," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 2966–2975.
- [31] C. Z. Lin, D. B. Lindell, E. R. Chan, and G. Wetzstein, "3D GAN inversion for controllable portrait image animation," 2022, *arXiv:2203.13441*.
- [32] J. Xie, H. Ouyang, J. Piao, C. Lei, and Q. Chen, "High-fidelity 3D GAN inversion by Pseudo-multi-view optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 321–331.
- [33] Y. Yin, K. Ghasedi, H. Wu, J. Yang, X. Tong, and Y. Fu, "NeRFInvertor: High fidelity NeRF-GAN inversion for single-shot real image animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 8539–8548.
- [34] S. An, H. Xu, Y. Shi, G. Song, U. Y. Ogras, and L. Luo, "PanoHead: Geometry-aware 3D full-head synthesis in 360°," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 20950–20959.
- [35] H. Nie and S. Lu, "PersistVerify: Federated model ownership verification with spatial attention and boundary sampling," *Knowl.-Based Syst.*, vol. 293, Jun. 2024, Art. no. 111675.
- [36] H. Nie and S. Lu, "Securing IP in edge AI: Neural network watermarking for multimodal models," *Int. J. Speech Technol.*, vol. 54, no. 21, pp. 10455–10472, Nov. 2024.
- [37] H. Nie, S. Lu, M. Wang, J. Xiao, Z. Lu, and Z. Yi, "VeriChroma: Ownership verification for federated models via RGB filters," in *Proc. 30th Eur. Conf. Parallel Distrib. Process. (Euro-Par)*, Madrid, Spain, Aug. 2024, pp. 332–345. [Online]. Available: https://doi.org/10.1007/978-3-031-69766-1_23
- [38] H. Nie, S. Lu, J. Wu, and J. Zhu, "Deep model intellectual property protection with compression-resistant model watermarking," *IEEE Trans. Artif. Intell.*, vol. 5, no. 7, pp. 3362–3373, Jul. 2024.
- [39] H. Nie and S. Lu, "FedCRMW: Federated model ownership verification with compression-resistant model watermarking," *Expert Syst. Appl.*, vol. 249, Sep. 2024, Art. no. 123776.
- [40] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8185–8194.
- [41] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.
- [42] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3973–3981.
- [43] P. Henzler, N. Mitra, and T. Ritschel, "Escaping Plato's cave: 3D shape from adversarial rendering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9983–9992.
- [44] T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y. Yang, "HoloGAN: Unsupervised learning of 3D representations from natural images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7587–7596.
- [45] S. M. Seitz and C. R. Dyer, "Photorealistic scene reconstruction by voxel coloring," *Int. J. Comput. Vis.*, vol. 35, pp. 151–173, Jan. 1999.
- [46] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo magnification: Learning view synthesis using multiplane images," 2018, *arXiv:1805.09817*.
- [47] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An information-rich 3D model repository," 2015, *arXiv:1512.03012*.
- [48] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.
- [49] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9240–9249.
- [50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [51] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," 2015, *arXiv:1505.00853*.
- [52] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*.
- [53] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for GANs do actually converge," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2018, pp. 3481–3490.
- [54] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [55] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
- [56] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200–2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.
- [57] S. Rudner, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*.
- [58] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

- [59] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2017, pp. 6626–6637.
- [60] M. Bilkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *Proc. Int. Conf. Learn. Representations*, Jan. 2018, pp. 1–19.
- [61] J. Tang, J. Ren, H. Zhou, Z. Liu, and G. Zeng, "DreamGaussian: Generative Gaussian splatting for efficient 3D content creation," 2023, *arXiv:2309.16653*.
- [62] Z. Zhao et al., "Hunyuan3D 2.0: Scaling diffusion models for high resolution textured 3D assets generation," 2025, *arXiv:2501.12202*.
- [63] W. Jiang, Y. Sun, L. Lei, G. Kuang, and K. Ji, "Change detection of multisource remote sensing images: A review," *Int. J. Digit. Earth*, vol. 17, no. 1, Dec. 2024, Art. no. 2398051.
- [64] K. Ko and M. Lee, "ZIGNeRF: Zero-shot 3D scene representation with invertible generative neural radiance fields," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 4974–4983.



KANGHYEOK KO received the B.S. degree from the School of Electrical and Electronics Engineering, Chung-Ang University, Seoul, South Korea, in 2022, and the M.S. degree from the Department of Intelligent Semiconductor Engineering, Chung-Ang University, in 2024.



SUNGYUP KIM received the B.S. degree from the School of Electrical and Electronics Engineering, Chung-Ang University, Seoul, South Korea, in 2025. His current research interests include implicit/explicit neural representations and generative adversarial networks.



MINHYEOK LEE (Member, IEEE) received the bachelor's and Ph.D. degrees in electrical engineering from Korea University, in 2015 and 2020, respectively. He has been an Assistant Professor with the School of Electrical and Electronics Engineering, Chung-Ang University, since 2021, and concurrently holds a position with the Department of Intelligent Semiconductor Engineering, since 2023. Prior to these appointments, he was a Research Professor with Korea University, from 2020 to 2021. Over the past three years, he has published over 40 papers in international journals and conferences, primarily focusing on artificial intelligence and generative AI. His research interests include generative artificial intelligence, generative adversarial networks (GANs), diffusion models, neural radiance fields (NeRF), AI-based finance engineering, and bioinformatics.

...