A Survey on Deep Generative 3D-aware Image Synthesis

WEIHAO XIA and JING-HAO XUE, University College London, UK

Recent years have seen remarkable progress in deep learning powered visual content creation. This includes deep generative 3D-aware image synthesis, which produces high-fidelity images in a 3D-consistent manner while simultaneously capturing compact surfaces of objects from pure image collections without the need for any 3D supervision, thus bridging the gap between 2D imagery and 3D reality. The field of computer vision has been recently captivated by the task of deep generative 3D-aware image synthesis, with hundreds of papers appearing in top-tier journals and conferences over the past few years (mainly the past two years), but there lacks a comprehensive survey of this remarkable and swift progress. Our survey aims to introduce new researchers to this topic, provide a useful reference for related works, and stimulate future research directions through our discussion section. Apart from the presented papers, we aim to constantly update the latest relevant papers along with corresponding implementations at https://weihaox.github.io/3D-aware-Gen.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Computing methodologies** → **Machine learning**; **Computer vision**; **Image manipulation;**

Additional Key Words and Phrases: 3D-aware image synthesis, deep generative models, implicit neural representation, generative adversarial network, diffusion probabilistic models

ACM Reference format:
Weihao Xia and Jing-Hao Xue. 2023. A Survey on Deep Generative 3D-aware Image Synthesis. *ACM Comput. Surv.* 56, 4, Article 90 (November 2023), 34 pages.
https://doi.org/10.1145/3626193

1 INTRODUCTION

A tremendous amount of progress has been made in deep neural networks that lead to photo-realistic image synthesis. Despite achieving compelling results, most approaches focus on **two-dimensional (2D)** images, overlooking the **three-dimensional (3D)** nature of the physical world. The lack of 3D structure, therefore, inevitably limits some of their practical applications. Recent studies have thus proposed generative models that are 3D-aware. That is, they incorporate 3D information into the generative models to enhance control (especially in terms of multiconsistency) over the generated images. Examples depicted in Figure 1 elucidate that the objective is to produce high-quality renderings which maintain consistency across various views. In contrast to the 2D generative models, the recently developed 3D-aware generative models [13, 33] bridge the gap between 2D images and the 3D physical world. The physical world surrounding us is intrinsically three-dimensional and images depict reality under certain conditions of geometry, material, and

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/W523835/1].
Authors' address: W. Xia and J.-H. Xue, Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK, London, UK; emails: {weihao.xia.21, jinghao.xue}@ucl.ac.uk.

0360-0300/2023/11-ART90
https://doi.org/10.1145/3626193

Fig. 1. The illustrative examples of 3D-aware image synthesis demonstrate the objective of this task, which is to generate high-quality renderings that are consistent across multiple views (top) and potentially provide detailed geometry (bottom). Typically, deep generative 3D-aware image synthesis methods are trained using a collection of 2D images, without depending on target-specific shape priors, ground truth 3D scans, or multi-view supervision. Examples are sourced from [12].

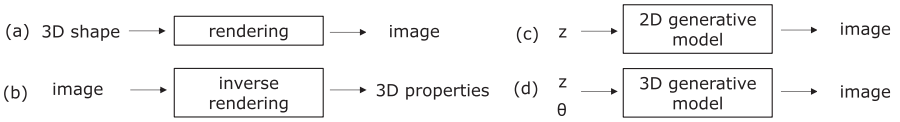


Fig. 2. Comparison of (a) rendering, (b) inverse rendering, (c) 2D generative models, and (d) 3D generative models. 3D generative models learn 3D representations first and then render a 2D image at certain viewpoints. Both 2D and 3D generative models have unconditional and conditional settings. An unconditional generative model maps a noise input z (and a camera pose in 3D models) to a fake image; a conditional model takes additional inputs as control signals, which could be image, text, or a categorical label.

illumination, making it natural to model the image generation process in 3D spaces. As shown in Figure 2, classical rendering (a) renders images at certain camera positions given human-designed or scanned 3D shape models; inverse rendering (b) recovers the underlying intrinsic properties of the 3D physical world from 2D images; 2D image generation (c) is mostly driven by generative models, which have achieved impressive results in photorealistic image synthesis; and 3D-aware generative models (d) offers the possibility of replacing the classical rendering pipeline with effective and efficient models learned directly from images.

Despite striking progress has been made recently in research of deep generative 3D-aware image synthesis, it lacks a timely and systematic review of this progress. In this work, we fill the gap by presenting a comprehensive survey of the latest research in deep generative 3D-aware image synthesis methods. We envision that our work will elucidate design considerations and advanced methods for deep generative 3D-aware image synthesis, present its advantages and disadvantages of different kinds, and suggest future research directions. Figure 3 provides a structured outline and taxonomy of this survey. Figure 4 is a chronological overview of representative deep generative 3D-aware image synthesis methods. We propose to categorize the deep generative 3D-aware image synthesis methods into two primary categories: 3D control of 2D generative models (Section 4) and 3D-aware generative models from single image collections (Section 5). Then, every category is further divided into some subcategories depending on the experimental setting or the specific utilization of 3D information. In particular, 3D control of 2D generative models are further divided into (1) exploring 3D control in 2D latent spaces (Section 4.1), (2) 3D parameters as controls (Section 4.2), and (3) 3D priors as constraints (Section 4.3). Section 5 summarizes methods that target generating photorealistic and multi-view-consistent images by learning 3D representations from single-view image collections. Broadly speaking, this category leverages neural 3D representations to represent scenes, use differentiable neural renderers to render them into the image plane, and optimize the network parameters by minimizing the difference between rendered images and observed images.

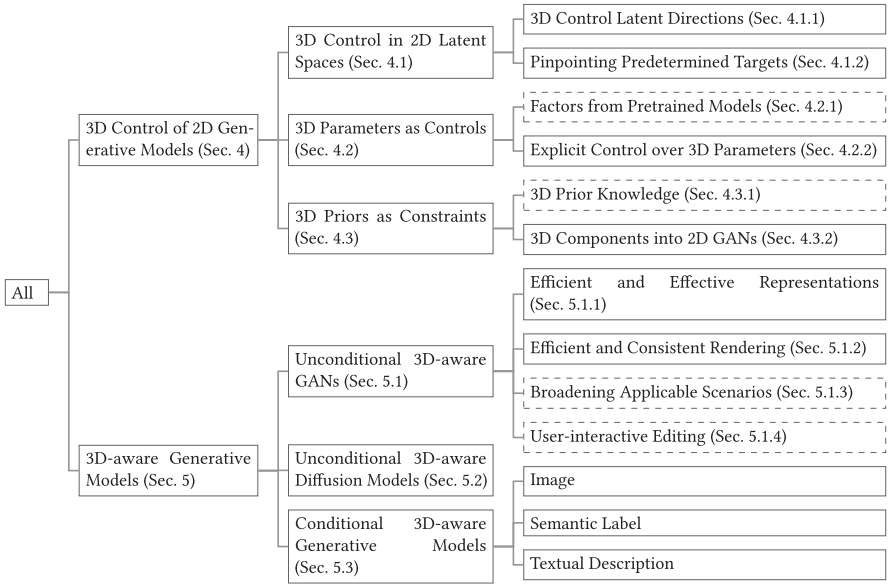


Fig. 3. A systematic taxonomy proposed in this survey of deep generative 3D-aware image synthesis methods. The dashed borders at the third level denote preliminaries, applications, or issues discussed in this subcategory. It should be noted that these methods are not mutually exclusive. For example, a few methods introduce 3D parameters to improve controllability (Section 4.2) while also implementing 3D constraints to improve consistency across multiple views (Section 4.3); EG3D [12] is referenced in Section 5.1.1 and Section 5.1.2 for its approach to 3D-aware representations and rendering algorithms.

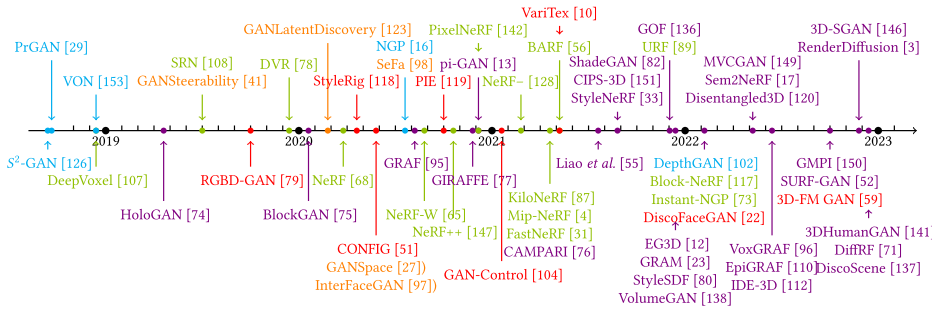


Fig. 4. Chronological overview of representative deep generative 3D-aware image synthesis methods which are categorized by different learning approaches. Methods in lime, orange, red, cyan, and violet, are from Section 2.4, Section 4.1, Section 4.2, Section 4.3, and Section 5, respectively. S^2-GAN [125] and PrGAN [29] are published in 2016 and are not shown in scale. Best viewed in color. The INR-based novel view synthesis methods (refer to Section 2.4) are not the focus of this survey. They are included to offer background information relevant to the topics discussed. The papers represented in this figure are current up to December 2022.

Here, we present a timely up-to-date overview of the growing field of deep generative 3D-aware image synthesis. Considering the lack of a comprehensive survey and an increasing interest and popularity, we believe it necessary to organize one to help computer vision practitioners with this emerging topic. The purpose of this survey is to provide researchers new to the field with a comprehensive understanding of deep generative 3D-aware image synthesis methods and show the superior performance over the status quo approaches. To conclude, we highlight

several open research directions and problems that need further investigation. The scope of this fast-expanding field is rather extensive and a panoramic review would be challenging. We shall present only representative methods of deep generative 3D-aware image synthesis rather than listing exhaustively all literature. This review can therefore serve as a pedagogical tool, providing researchers with the key information about typical methods of deep generative 3D-aware image synthesis. Researchers can use these general guidelines to develop the most appropriate technique for their own particular study. The main technical contributions of this work are as follows:

- Systematic taxonomy. We propose a hierarchical taxonomy for deep generative 3D-aware image synthesis research. We categorize existing models into two main categories: 3D control of 2D generative models and 3D-aware generative models from image collections.
- Comprehensive review. We provide a comprehensive overview of the existing state-of-the-art deep generative 3D-aware image synthesis methods. We compare and analyze the main characteristics and improvements for each type, assessing their strengths and weaknesses.
- Outstanding challenges. We present open research problems and provide some suggestions for the future development of deep generative 3D-aware image synthesis.
- In an attempt to continuously track recent developments in this fast advancing field, we provide an accompanying webpage which catalogs papers addressing deep generative 3D-aware image synthesis, according to our problem-based taxonomy: https://weihaox.github.io/3D-aware-Gen

2 BACKGROUND

This section introduces a few important concepts as the background. In order to formulate deep generative 3D-aware image synthesis, we first clarify how 2D and 3D data are expressed, and how they are converted between each other. Moreover, we introduce two key elements involved in most deep generative 3D-aware image synthesis methods: implicit neural representations and differentiable neural rendering.

2.1 2D and 3D Data, Rendering and Inverse Rendering

The 2D images depict a glimpse into the surrounding 3D physical world with its geometry, materials, and illumination conditions at that moment. Images are composed of an array of pixels (picture elements). The 3D reality can be represented in many different ways, each with its own advantages and disadvantages. There are several examples of such **3D shape representations**, including depth images, point clouds, voxel grids, and meshes. **Depth images** contain distance information between the object and the camera at every pixel. The distance encodes 3D geometry information from a fixed point of view. **Layered depth images (LDIs)** use several layers of depth maps and their associated color values to depict a scene. **Point clouds** comprise vertices in 3D space, represented by coordinates along the x, y, and z axes. These types of data can be acquired by 3D scanners, such as LiDARs or RGB-D sensors, from one or more viewpoints. **Voxel grids** describe a scene or object using a regular grid in 3D space. A voxel (volume element) in 3D space is analogous to a pixel in a 2D image. A voxel grid can be created from a point cloud by voxelization, which groups all 3D points within a voxel. **Meshes** are a collection of vertices, edges, and polygonal faces. In contrast to a point cloud, which only provides vertices locations, a mesh also provides surface information of an object. Nevertheless, deep learning does not provide a straightforward way to process surface information. Instead, many techniques resort to sampling points from the surfaces to create a point cloud from the mesh representation.

As shown in Figure 2(a), images can be obtained by rendering a 3D object or scene under certain viewpoints and lighting conditions. This forward process is called **rendering** (image

synthesis). Rendering has been studied in computer graphics and a wide variety of renderers are available for use. The reverse process, **inverse rendering**, as shown in Figure 2(b), is to infer underlying intrinsic components of a scene from rendered 2D images. These properties include shape (surface, depth, normal), material (albedo, reflectivity, shininess), and lighting (direction, intensity), which can be further used to render photorealistic images. The inverse rendering papers are not classified as 3D-aware image synthesis methods in this survey as they are not deliberately designed for this purpose. 3D-aware image synthesis in this survey include a similar inverse rendering process and a rendering process. In contrast, these methods typically do not produce explicit 3D representations such as meshes or voxels for rendering. They learn **neural 3D representations** (mostly implicit functions), render them into images with **differentiable neural rendering** technique, and optimize the network parameters by minimizing the difference between the observed and rendered images.

2.2 Implicit Scene Representations

In computer vision and computer graphics, 3D shapes are traditionally represented as explicit representations like depths, voxels, point clouds, or meshes. Recent methods propose to represent 3D scenes with neural implicit functions, such as occupancy field [66], signed distance field [82], and radiance field [67]. The **implicit neural representation** (**INR**, neural fields, or coordinate-based representation) is a novel way to parameterize signals across a wide range of domains. Taking images as an example, INR parameterizes an image as a continuous function that maps pixel coordinates to RGB colors. The implicit functions are often not analytically tractable and are hence approximated by neural networks. Here are some popular examples of INR.

Neural Occupancy Field [66, 77, 83] implicitly represents a 3D surface with the continuous decision boundary of a neural classifier. This function approximated with a neural network assigns to every location $p \in \mathbb{R}^3$ an occupancy probability between 0 and 1. Given an observation (*e.g.*, image or point cloud) $x \in \mathcal{X}$ and a location $p \in \mathbb{R}^3$, the representation can be simply parameterized by a neural network f_θ that takes a pair (p, x) as input and outputs a real number which represents the probability of occupancy: $f_\theta : \mathbb{R}^3 \times \mathcal{X} \rightarrow [0, 1]$.

Neural Signed Distance Field [82] is a continuous function that models the distance from a queried location to the nearest point on a shape's surface, whose sign indicates if this location is inside (negative) or outside (positive): $SDF(\boldsymbol{x}) = s, \boldsymbol{x} \in \mathbb{R}^3, s \in \mathbb{R}$. The underlying surface is implicitly described as the zero iso-surface decision boundaries of feed-forward networks $SDF(\cdot) = 0$. This implicit surface can be rendered by raycasting or rasterizing a mesh obtained through marching cubes [63].

Neural Radiance Field [67] **(NeRF)** has attracted growing attention due to its compelling results in novel view synthesis on complex scenes. It leverages an MLP network to approximate the radiance fields of static 3D scenes and uses the classic volume rendering technique [43] to estimate the color of each pixel. This function takes as input a 3D location $\boldsymbol{x}$ and 2D viewing direction $\boldsymbol{d}$, and outputs an directional emitted RGB color $\boldsymbol{c}$ and volume density σ: $f_\theta : (\boldsymbol{x}, \boldsymbol{d}) \rightarrow (\boldsymbol{c}, \sigma)$. It captures 3D geometric details based on pure 2D supervision by learning the reconstruction of given views.

There also exist many other implicit functions proposed to represent a scene, such as neural sparse voxel fields [58], or neural volumes [62].

2.3 Differentiable Neural Rendering

3D rendering is a function that outputs a 2D image from a 3D scene. Differentiable rendering provides a differentiable rendering function, that is, it computes the derivatives of that function in response to different parameters of the scene. Once a renderer is differentiable, it can be

integrated into the optimization of neural networks. One use case for differentiable rendering is to compute a loss in rendered image space and back propagation can be applied to train the network. Driven by the prevalence of NeRF-based methods [67], volume rendering [43] becomes the most commonly used differentiable renderer among the methods that this survey targets. It is naturally differentiable, and the only input required to optimize the NeRF representation is a set of images with known camera poses. Given volume density and color functions, volume rendering is used to obtain the color $C(\mathbf{r})$ of any camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, with camera position $\mathbf{o}$ and viewing direction $\mathbf{d}$ using

$$C(\mathbf{r}) = \int_{t_1}^{t_2} T(t) \cdot \sigma(\mathbf{r}(t)) \cdot \mathbf{c}(\mathbf{r}(t), \mathbf{d}) \cdot dt, \text{ where } T(t) = \exp\left(-\int_{t_1}^{t} \sigma(\mathbf{r}(s))ds\right). \tag{1}$$

$T(t)$ denotes the accumulated transmittance, representing the probability that the ray travels from t_1 to t without being intercepted. The rendered image can be obtained by tracing the camera rays $C(\mathbf{r})$ through each pixel of the to-be-synthesized image.

2.4 INR-based Novel View Synthesis

Novel view synthesis [30, 67, 88, 141] is a long-standing problem that involves rendering frames of scenes from new camera viewpoints. There are existing methods that depend on implicit 3D scene representations. One of the most representative studies in the field is NeRF, which employs neural networks to capture the continuous 3D scene structure within the network weights, resulting in photorealistic synthesis outcomes. These methods usually operate under the **Single-Scene Overfitting (SSO)** experiment. This approach aims to render novel views by learning a deep neural scene representation from multi-view image collections of a specific scene or object. These models are trained per-scene and are primarily designed for tasks of 3D reconstruction, novel view synthesis, or free viewpoint rendering. Its success has inspired various extensions and improvements, leading to a rich body of work in the NeRF family of methods. These variants aim to address limitations, enhance efficiency, and improve the quality and diversity of the generated images. For more details on novel view synthesis, please refer to recent methods [30, 67, 88, 141], surveys [120], and the appendix of this survey.

The focus of this survey, which is on 3D-aware generative models, bears close relevance to 3D novel view synthesis, particularly the INR-based methods. Many 3D-aware generative models derive inspiration from the field of INR-based 3D novel view synthesis. In the following sections, we will demonstrate how these two areas can benefit each other in solving key issues in their respective research fields, in terms of technical implementation and application scenarios.

2.5 Dataset

Similar to 2D models, 3D-aware generative models aim to produce photorealistic and multi-view consistent images. Therefore, the same single-view image datasets are used as in the 2D methods. These datasets predominantly consist of single-view image collections, which are usually unstructured and unannotated. These are typically employed by unconditional generative models and some conditional methods. These image collections can be further categorized into *single* objects and *multiple* objects according to the prominent object numbers in the foreground, and *simple* shape and *variable* shape according to the fineness of the object. Most datasets pertain to a specific category. Current popular categories include human faces (*e.g.*, FFHQ [47] and MetFaces [46]) and bodies (*e.g.*, SHHQ [28] and DeepFashion [60]), scenes with a single salient subject (*e.g.*, LSUN [142] and CompCars [139]), or multiple salient subjects (*e.g.*, CLEVR*n* [74]). Most 3D-aware generative models handle face dataset and they are typically applicable to other datasets without any constraint, but there are other categories of 3D-aware generative

Table 1. Summary of Popular Single-view Image Datasets Organized by Their Major Categories and Roughly Sorted by Their Popularity

dataset	published in	category	# samples	resolution	representative references	keyword
FFHQ [47]	CVPR 2019	Human Face	70k	1024 × 1024	[12, 33, 79, 137]	single, simple-shape
AFHQ [18]	CVPR 2020	Cat, Dog, and Wildlife	15k	512 × 512	[12, 33, 79, 137]	single, simple-shape
SHHQ [28]	ECCV 2022	Human Body	230k	1024 × 512	[140]	single, variable-shape
CompCars [139]	CVPR 2015	Real Car	136K	256 × 256	[33, 73, 74, 137]	single, simple-shape
CARLA [24]	CoRL 2017	Synthetic Car	10k	128 × 128	[11, 13, 23, 119, 137]	single, simple-shape
CLEVR*n* [42]	CVPR 2017	Objects	100k	256 × 256	[74, 76]	multiple, simple-shape
LSUN [142]	2015	Bedroom	300K	256 × 256	[73, 137]	single, simple-shape
CelebA [61]	ICCV 2015	Human Face	200k	178 × 218	[73, 137]	single, simple-shape
CelebA-HQ [45]	ICLR 2018	Human Face	30k	1024 × 1024	[85]	single, simple-shape
MetFaces [46]	NeurIPS 2020	Art Face	1336	1024 × 1024	[33]	single, simple-shape
M-Plants [109]	NeurIPS 2022	Plant	141,824	256 × 256	[109]	single, variable-shape
M-Food [109]	NeurIPS 2022	Food	25,472	256 × 256	[109]	single, variable-shape

models which need inductive bias of specialized domain knowledge. Notably, recent research efforts [92, 108] have been made to extend deep generative 3D-aware image synthesis beyond a single category, aiming to encompass a wide range of categories, such as those found in ImageNet [20]. Open-source tools, such as Self-Distilled StyleGAN [68], are frequently utilized to facilitate the process of data generation. Examples of this include SDIP Dogs and SDIP Elephants.

Table 1 demonstrates a summary of single-view image datasets organized by their major categories and roughly sorted by their popularity how often they are used in studies. The following part provides a succinct overview of the commonly employed datasets in this field of research.

SHHQ (Stylish-Humans-HQ Dataset) [28] , which was recently released, caters to the growing research interest in human body generation. It includes 230K high-quality, real-world, full-body human images, with resolutions ranging from 1024 × 512 up to 2240 × 1920. The SHHQ-1.0 subset, comprising 40K images, lays a solid foundation for extensive research and experimentation in the field of human body generation.

CelebA (CelebFaces Attributes) [61] is a large-scale face attribute dataset consisting of 200K celebrity images with 40 attribute annotations each. CelebA, together with its succeeding CelebA-HQ [45], are widely used in face image generation and manipulation.

FFHQ (Flickr-Faces-HQ) [47] is a high-quality image dataset of human faces crawled from Flickr, which consists of 70,000 high-quality human face images of 1024 × 1024 pixels and contains considerable variation in terms of age, ethnicity, and image background.

AFHQ (Animal-Faces-HQ) [47] consists of 70k high-quality animal face images of 512 × 512 pixels. It includes three domains of cat, dog, and wildlife, each providing 5k images and containing diverse images of various breeds (≥ eight).

CARLA [24] is a synthetic dataset, which contains 10K images which are rendered from Carla Driving simulator [24] using 16 car models with different textures.

LSUN [142] contains approximately one million labeled images for each of 10 scene categories (*e.g.*, bedroom, church, or tower) and 20 object classes (*e.g.*, bird, cat, or bus). The church and bedroom scene images and car and bird object images are commonly used.

Megascans Food (M-Food) and **Plants** (M-Plants) [109] are two *variable-shape* datasets. They are proposed to address two limitations of existing simple-shape benchmarks: (1) they contain low variability of global object geometry, focusing entirely on a single class of objects, like human/cat faces or cars, that do not vary much from instance to instance; (2) they have limited camera pose distribution: for example, FFHQ and Cats are completely dominated by the frontal and near-frontal views. M-Food consists of 199 models of different food items with 128 views per model (25,472 images in total); and M-Plants consists of 1,108 different plant models with 128 views per model (141,824 images in total). Both contain images with 256 × 256 pixels.

CLEVR*n* [74] is a synthetic multiple-object dataset, where n is the number of foreground objects. This dataset is derived from CLEVR [42] by adding a large variety of colours and primitive shapes. In response to demand, the image can be rendered to a desired quantity (*e.g.*, 100k) and resolution (*e.g.*, 64×64 in [74] and 256×256 in [76]). This scheme can be used to generate other synthetic datasets, including SYNTH-CARn and SYNTH-CHAIRn with n foreground objects each. These multiple-object datasets are often used to test models in terms of the independent control over foreground and background [74, 76].

Besides single-view images, some studies [3, 44] use multiview posed images for training. These involve synthetic data rendered from 3D scans, real images captured from the real world, or video frames from category-specific videos, along with corresponding camera parameters. For synthetic data, they use ground-truth camera poses, intrinsics, and bounds to render images from 3D shapes of objects (for example, from ScanNet [19] and ShapeNet [15]). For real data (*e.g.*, videos in the Common Objects in 3D (CO3D) dataset [87]), off-the-shelf software, such as a structure-from-motion package like COLMAP [93], can be used to estimate the camera parameters.

2.6 Evaluation Metrics

There are different dimensions to evaluate 3D-aware image synthesis methods, which can be categorized into two groups: 2D and 3D metrics. 2D metrics evaluate the synthesised images in terms of quality, diversity, and fidelity. 3D metrics access the shape and surface quality, as well as the temporal and multi-view consistency. Model efficiency is taken into account sometimes and is evaluated by model size and training/inference time.

The following part provides a succinct overview of the commonly employed measures. It should be noted that, in the current evaluation landscape, no single metric can comprehensively address all aspects; instead, each typically examines a unique facet. Established metrics (despite their limitations) such as FID [36] and LPIPS [147], which are widely recognized within the research community, are frequently used to assess quality and diversity of generated images. On the other hand, the evaluation of 3D consistencies is less standardized, with various studies proposing their own metrics due to the lack of universally accepted 3D measures in this field. Therefore, in addition to utilizing existing evaluation metrics, the introduction of more reliable and more personalized measures could significantly enhance the assessment of photorealistic and geometric quality of generated images in both general and specific contexts.

2.6.1 Model Efficiency. Two kinds of metrics are commonly employed by current studies to demonstrate the efficiency of their proposed methods: average running time and model complexity. These metrics can be borrowed from deep compression, which evaluates inference runtime, model size and latency. The model complexity is typically assessed by the number of parameters, **floating point operations (FLOPs)**, and **multiply-accumulate operations (MACs)**. Runtime usually means one forward at inference phase. In our case, inference runtime means the time required for rendering an image. However, for deep 3D-aware generative models, we are equally interested in the time required for training. This can be assessed using metrics such as total training time and the number of batches processed per second [150].

2.6.2 Image Similarity. Similarity (or faithfulness) measures the similarity between real images and generated ones. When ground-truth images exist at certain viewpoints, the rendered images are expected to be close to them. The most widely used metrics are **Peak Signal-to-Noise Ratio (PSNR)**, **Structural Similarity (SSIM)** [126], and **Learned Perceptual Image Patch Similarity (LPIPS)** [147]. Pixel-wise reconstruction distances, *e.g.*, mean absolute error, are also used.

PSNR between the ground-truth image and the reconstruction is defined by the maximum possible pixel value of the image and the mean squared error between images.

SSIM measures the structural similarity between images based on independent comparisons in terms of luminance, contrast, and structures. The details can be found in [126].

LPIPS measures the distance between image patches. A lower value means higher similarity between the image patches. A higher value means more differences. LPIPS can therefore also be employed to assess the diversity of images. The most common method for evaluating the diversity of generated images involves randomly selecting two examples from the set of generated images and calculating the average LPIPS distance between them. This measures the degree to which the two images differ from each other, with a higher LPIPS score indicating greater diversity.

2.6.3 Image Quality. These metrics are often used to assess images generated by a generative model, like a **generative adversarial network (GAN)** or a **generative diffusion model (GDM)**. These metrics include but are not limited to **Inception Score (IS)** [91], **Fréchet Inception Distance (FID)** [36], **Kernel Inception Distance (KID)** [6], and **Perceptual Path Length (PPL)** [47]. At present, FID is the predominant metric in the research community for assessing the quality of generated images. In contrast, IS was once popular but has since been less favored due to its inability to accurately reflect the diversity of images and its sensitivity to minor changes in the image set. PPL measures the disentanglement and consistency of the learned latent space in generative models.

FID is defined by the Fréchet distance between features from the real and generated images based on Inception-v3 [115]. Lower FID indicates better perceptual quality.

PPL computes the average distance in feature space between images generated from interpolated points in the latent space. A smaller PPL suggests a smoother and more disentangled latent space, where minor changes in the latent vector lead to coherent and minor changes in the output image.

KID measures the dissimilarity between two probability distributions using samples drawn independently from each distribution. Lower is better.

IS is to measure the quality and diversity of images generated from GAN models. It calculates the statistics of a synthesized image using Inception-v3 Network [115] pretrained on ImageNet [20]. A higher score is better.

2.6.4 Multi-view 3D Consistency. Multi-view 3D consistency is another significantly important aspect in 3D-aware image synthesis. The view-inconsistencies could be caused by shapes and colors. Consistencies in geometry and photometry is basically equivalent to the quality of shape and texture.

Shape Quality is evaluated mostly by calculating differences between the rendered depth map and the pseudo-ground-truth depth, *e.g.*, using MSE [12] or a modified Chamfer distance [79]. For example, given two generated images from two sampled angles of the same scene, Shi et al. [101] uses rotation precision and rotation consistency to evaluate the quality of the depth maps (point cloud). The former is aimed to measure the accuracy of the angle of rotation while the latter targets at the rotation consistency evaluation. In GOF [135], the **mean angle deviation (MAD)** and the **scale-invariant depth error (SIDE)** are used to compare the outputs against the ground-truth depth maps. MAD emphasizes the compactness of surfaces, whereas SIDE emphasizes the accuracy of depths. Some methods use more direct indicators to evaluate the geometry properties of learned surfaces. Xu et al. [135] use average geodesic distance and average curvature between random

points to assess the geometry properties of learned surfaces. The lower these two metrics, the smoother the recovered object surfaces.

Texture Quality could be evaluated by using PSNR and SSIM as image fidelity under different viewpoints. But in most cases, the ground-truth images are not available for evaluation. Several methods use FID to evaluate image quality at different camera poses as part of the multi-view texture quality. A more direct way is to assess multi-view **facial identity consistency (ID)** [12] by calculating the mean Arcface [21] cosine similarity score between pairs of the same face rendered from random camera poses.

Apart from the aforementioned performance measures, pose accuracy [12] is also considered as an important indicator of shape quality and controllability. Poses (pitch, yaw, and roll) are detected with the help of pre-trained face reconstruction models from the generated images and then its L2 errors against the ground-truth poses is computed to determine each model's pose drift.

3 OVERVIEW OF DEEP GENERATIVE 3D-AWARE IMAGE SYNTHESIS

In the following sections, we introduce different kinds of deep generative 3D-aware image synthesis methods. In this survey, we use the terms "generative 3D-aware image synthesis methods" and "3D-aware generative models" interchangeably, since they both fundamentally refer to the same concept related to image synthesis. Table 2 is a detailed comparison of deep generative 3D-aware image synthesis methods. The first (arXiv) draft dates are used to sequence the publications. The publication information can be found in the bibliography.

3.1 Goal and Challenge

To begin with, we provide an overview of deep generative 3D-aware image synthesis to give readers a general understanding of this task, including its goals, challenges, and underlying principles. This serves to prevent readers from becoming overwhelmed with intricate technical details. As previously stated, the task involves learning a model capable of generating images that maintain consistency across multiple viewpoints, without explicitly modeling the subject(s) in 3D. This definition itself implies goals and challenges of the task. The primary goal is to generate 3D-consistent images, while challenges lie in achieving multi-view consistency without explicit 3D shape modeling and solely relying on training datasets comprising unlabeled images without 3D supervision. Several potential issues arise consequently:

- Multiview inconsistency: One common issue is the presence of inconsistent shapes and appearances, where the textures and geometry vary across different views. Another challenge is background sticking, where the foreground subject is not adequately separated from the background.
- Limited camera views: The range of viewpoints provided by current models is often limited, with a lack of capability to generate images from less common perspectives (*e.g.*, looking down from above or from behind). Furthermore, these models may not have the capability for 360-degree image generation.
- Imprecise camera control: Some models struggle to learn fine-grained control over camera parameters, leading to imprecise changes in the obtained images when altering the camera.
- Compromised visual quality: The image quality produced by 3D-aware generative models may not be as high as their 2D counterparts. Additionally, these models often generate images at lower resolutions, typically 256×256 pixels, and rarely exceed resolutions of 1024 pixels.
- Limited application scenarios: Many current methods primarily concentrate on single categories characterized by simple geometry and appearance, such as human faces. These

Table 2. Overview of Deep Generative 3D-aware Image Synthesis Methods

Method	Publication	Dataset / Category	Condition	Geo.	Cont.	Supervision
3D Control in 2D Latent Spaces (Section 4.1)						
GANSteerability [41]	ICLR 2020	ImageNet	N/A	✗	✓	shifted image
GANLatentDiscovery [122]	ICML 2020	Face, ILSVRC	N/A	✗	✓	Unsup.
InterFaceGAN [96]	CVPR 2020	Face	N/A	✗	✓	synthetic image & label
GANSpace [27]	NeurIPS 2020	Face, ImageNet	N/A	✗	✓	Unsup.
SeFa [97]	CVPR 2021	Face, Car, LSUN, ImageNet	N/A	✗	✓	closed-form
LatentCLR [143]	ICCV 2021	Face, Car, Cat, Bird	N/A	✗	✓	Unsup.
3D Parameters as Controls (Section 4.2)						
StyleRig [117]	CVPR 2020	Face	synthesized image	✗	✓	3DMM
DiscoFaceGAN [22]	CVPR 2020	Face	parameters	✗	✓	3DMM
PIE [118]	TOG 2020	Face	real image	✗	✓	3DMM
CONFIG [51]	ECCV 2020	Face	image & parameters	✗	✓	synthetic data
GAN-Control [103]	ICCV 2021	Face	parameters	✗	✓	pseudo param.
3D-FM GAN [59]	ECCV 2022	Face	image & rendering	✗	✓	synthetic data
3D Priors as Constraints (Section 4.3)						
S^2-GAN [125]	ECCV 2016	scene (NYUv2)	Uncon.	✗	✗	normal map
PrGANs [29]	3DV 2017	synthesized from 3D shapes	Uncon.	✓	✗	image and viewpoint
VON [152]	NeurIPS 2018	Chair, Car	Uncon.	✓	✗	2.5D sketch, 3D shape
RGBD-GAN [78]	ICLR 2020	Face, Car	Uncon.	✓	✗	Unsup.
NGP [16]	CGF 2021	Chair, Car	user controls	✓	✓	3D shape, reflectance map
LiftedGAN [98]	CVPR 2021	Face	Uncon.	✓	✓	pseudo multi-view images
DepthGAN [101]	ECCV 2022	Scene (LSUN)	Uncon.	✓	✓	pseudo depth map
3D GANs (Section 5.1 and Section 5.3)						
HoloGAN [73]	ICCV 2019	Face, Cat, Car, LSUN	Uncon.	✗	✗	Unsup.
Liao et al. [55]	CVPR 2020	Multiple Objects	Uncon.	✓	✗	Unsup.
BlockGAN [74]	NeurIPS 2020	Multiple Objects	Uncon.	✗	✓	Unsup.
GRAF [94]	NeurIPS 2020	rendered chair, Face, Cat, bird	Uncon.	✗	✗	Unsup.
pi-GAN [13]	CVPR 2021	Face, Car, CARLA	Uncon.	✓	✗	Unsup.
GIRAFFE [76]	CVPR 2021	Face, Cat, Car, Church, Chair	Uncon.	✗	✓	Unsup.
GOF [135]	NeurIPS 2021	Face, Cat, Car	Uncon.	✓	✗	Unsup.
ShadeGAN [81]	NeurIPS 2021	Face, Cat	Uncon.	✓	✗	Unsup.
CAMPARI [75]	3DV 2021	Face, Cat, Car, Chair	Uncon.	✗	✓	Unsup.
StyleNeRF [33]	ICLR 2022	Face, Cat, Car	Uncon.	✓	✗	Unsup.
GRAM [23]	CVPR 2022	Face, Cat, CARLA	Uncon.	✗	✗	Unsup.
EG3D [12]	CVPR 2022	Face, Cat	Uncon.	✓	✗	Unsup. (posed 2D image)
VolumeGAN [137]	CVPR 2022	Face, Cat, Car, bedroom, CARLA	Uncon.	✓	✗	Unsup.
StyleSDF [79]	CVPR 2022	Face, Cat	Uncon.	✓	✗	Unsup.
Pix2NeRF [11]	CVPR 2022	Face, CARLA, rendered image	image	✗	✗	Unsup.
Sem2NeRF [17]	ECCV 2022	Face, Cat	semantic mask	✗	✗	semantic mask & image
SURF-GAN [52]	ECCV 2022	Face	Uncon.	✗	✗	Unsup.
EpiGRAF [109]	NeurIPS 2022	Face, Cat, variable-shape	Uncon.	✓	✗	Unsup.
IDE-3D [111]	TOG 2022	Face	Uncon.	✓	✓	semantic mask & image
3D-SGAN [145]	ECCV 2022	Human Body	Uncon.	✗	✓	semantic mask & image
DiscoScene [136]	CVPR 2023	Street Scene	Uncon.	✗	✗	Unsup.
3DGP [108]	ICLR 2023	ImageNet, SDIP	Uncon.	✓	✗	Unsup.
3D Diffusion Models (Section 5.2)						
DiffRF [70]	CVPR 2023	Chair	Uncon. /image	✓	✓	Unsup. (posed 2D image)
RenderDiffusion [3]	CVPR 2023	Face, CLEVR, ShapeNet	Uncon. /image	✗	✗	Unsup. (posed 2D image)
HoloDiffusion [44]	CVPR 2023	object (CO3D)	Uncon.	✗	✗	Unsup. (posed 2D video)

"Dataset" means what *Categories* of images are used for training. "Posed images" mean 2D images along with respective extrinsic and intrinsic camera matrices. "Geometry (Geo.)" indicates if the model is available for geometry reconstruction. "Controllability (Cont.)" means if the model has the ability of attribute edit beyond camera pose. "Condition" refers to additional inputs accompanying the latent code, while "Unconditional (Uncon.)" indicates the absence of any extra input. "Unsupervised (Unsup.)" implies that there is no additional training supervision apart from the images.

approaches demonstrate limited capabilities when it comes to composing complex scenes that encompass a variety of objects.

- Expensive training: Training 3D-aware generative models can be resource-intensive and time-consuming, requiring substantial computational power and extensive training time.

Despite significant advancements, these challenges continue to persist in this field. Most methods tend to tackle one or a few aspects, achieving satisfactory performance in those areas while leaving others compromised. The datasets and evaluation metrics introduced previously align with efforts to address these issues.

3.2 Comparisons between Two Primary Categories

In this survey, deep generative 3D-aware image synthesis methods are classified into two main categories: 3D control of 2D generative models (Section 4) and 3D-aware generative models from image collections (Section 5). Then, every category is further divided into some subcategories depending on the experimental setting or the specific utilization of 3D information. In particular, 3D control of 2D generative models are further divided into (1) 3D control in 2D latent spaces, (2) 3D parameters as controls, and (3) 3D priors as constraints. For the category of 3D-aware generative models, most research relied on Generative Adversarial Networks (GANs) [32], thus resulting in a prevalent exploration of 3D-aware GANs [33, 73, 74, 79, 137, 138] (see Section 5.1). Recent trends in the studies are also indicating a growing interest in using diffusion models for 3D generative modeling [3, 14, 44, 49, 70, 133, 144] (see Section 5.2). The conditional 3D-aware generative models [11, 17, 57, 71, 111, 113] are presented in Section 5.3.

Typically, methods in the category of 3D control of 2D generative models exhibit superior image quality and require significantly fewer training resources. In contrast, methods in the second category can generate more consistent multiview images under a larger range of camera movements but with compromised visual quality, and training such models can be resource-intensive and time-consuming. Therefore, improvements for methods in the first category primarily focus on enhancing multiview consistency, while the second category emphasizes the development of efficient and effective representations and rendering processes to improve visual quality and expedite training.

Please note that as this field is relatively nascent, the usage of terminology can be perplexing and might lead to certain confusions or misunderstandings. In certain literature, the term "3D-aware generative model" refers to methods discussed in Section 4, while "3D generative models" are used to describe methods in Section 5. Given that 3D generative models are more aptly suited to 3D tasks, in the context of this survey, we categorize the *methods discussed in Section 4 as those based on a 2D network*, but which introduce various strategies to generate images consistent across multiple views. We refer to the *methods in Section 5 as those utilizing a 3D-aware network design*. We invite readers to focus on understanding the fundamental differences, rather than getting caught up in the specifics of terminology, and to make their own judgments based on the context.

3.3 Relationships with INR-based Novel View Synthesis

This task of deep generative 3D-aware image synthesis, indicated by its nature, is closely related to 3D **novel view synthesis (NVS)** [67, 90, 105–107], particularly the INR methods (*e.g.*, NeRF [67]). Many 3D-aware generative models (see Section 5) draw inspiration from the INR-based NVS methods [30, 40, 65, 67, 72, 88, 141] to address aforementioned challenges. Broadly speaking, both 3D-aware generative models and INR-based NVS methods aim to generate photorealistic and multi-view-consistent images, using a similar pipeline that first learns the implicit neural 3D representation and then renders it from that viewpoint. Particularly, both leverage neural 3D representations to represent scenes, use differentiable neural renderers to render them onto the image plane, and optimize the network parameters by minimizing the discrepancy between rendered and observed images.

However, they are significantly different in training on a multiple-view or single-view image collections, due to their hugely different application scenarios. As with their 2D counterparts, 3D-aware generative models are learned from a collection of single-view images, while 3D novel view synthesis learns a 3D representation from multiple views of a scene. Once trained, 3D-aware generative models generate images from a limited range of viewpoints, unlike the free-view rendering capabilities observed in 3D novel view synthesis (especially when utilizing NeRF-based

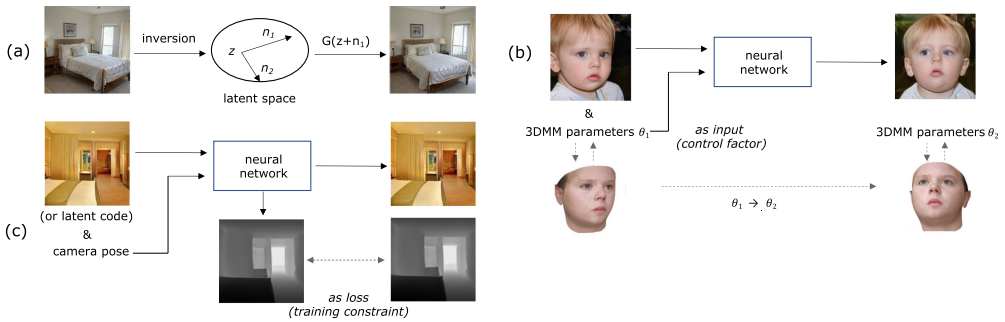


Fig. 5. Comparison of methods aiming to make 2D generative models (mostly GANs) 3D aware. These works are categorized into three groups based on how the 3D control capability is introduced: a) exploring 3D control in the 2D pretrained GAN latent spaces, b) using 3D parameters (*e.g.*, from 3DMM [7]) as input for explicit control, and c) introducing 3D-aware components (*e.g.*, depth estimation module) into 2D GANs and adopting 3D prior knowledge (*e.g.*, depth map) as training constraints.

methods). In contrast, NVS methods excel at synthesizing high-fidelity and detailed images from novel viewpoints, enabling the exploration of previously unseen perspectives.

Considering the intimate relationship between deep generative 3D-aware image synthesis and INR-based NVS, leveraging advancements made in INR-based NVS could potentially broaden the scope of camera movements in 3D-aware image generation. Furthermore, the combination of "free-view rendering and generative modeling" presents a challenging yet promising research avenue that is likely to gain significant attention in future research.

4 3D CONTROL OF 2D GENERATIVE MODELS

Due to the prevalence of 2D generative models, there have been studies aiming to make these pretrained models, especially GANs, 3D aware. These studies, mostly built on the top of a pretrained StyleGAN, can be further categorized into three groups based on how the 3D control capability is introduced: (1) exploring 3D control in 2D latent spaces, (2) adopting explicit control over the 3D parameters, and (3) introducing 3D-aware components into 2D GANs. The same taxonomy can be applied to other generative models. Figure 5 is an illustration of these three categories of methods.

4.1 Exploring 3D Control in 2D Latent Spaces

4.1.1 Discovering 3D Control Latent Directions. It has been demonstrated that pretrained GANs have interpretable directions in their latent spaces. The image generation process is controlled by altering the latent codes $\mathbf{z}$ in the desired directions $\mathbf{n}$ with step α, which is often considered as a linear vector arithmetic $\mathbf{z}' = \mathbf{z} + \alpha\mathbf{n}$. The altered latent codes are then fed into a pretrained GAN $G(\cdot)$ for the edited results: $I' = G(\mathbf{z}')$. The methods in this group are mostly developed for semantic editing, and some have been shown to discover geometric directions as well. They are used to alter pose position or light condition of faces or manipulate geometry (*e.g.*, zoom, shift, or rotation) of natural images. As classified in [132], such directions can be identified through supervised, self-supervised, or unsupervised manners.

Supervised Manner These methods typically sample a large amount of latent codes, synthesize a collection of corresponding images, and annotate them with predefined labels by introducing a pretrained classifier. For example, to interpret the face representation learned by GANs, Shen et al. [96] (May 2020) employ some off-the-shelf classifiers to learn a hyperplane in the latent space serving as the separation boundary and predict semantic scores for synthesized images.

Even though the boundary is searched by solving a bi-classification problem, it can produce continuous face pose changing by moving the latent code. Abdal et al. [1] (Aug 2020) learn a bidirectional mapping between the $\mathcal{Z}$ space and the $\mathcal{W}$ space by using **continuous normalizing flows (CNF)**. Attribute information (including head poses) are injected into the CNF blocks for the desired results. However, such methods rely on the availability of attributes (typically obtained by a face classifier network), which might be difficult to obtain for new datasets and could require manual labeling effort. Jahanian et al. [41] (Jul 2019) use a self-supervised manner to learn these directions without any direct supervision. Sequence of target images are obtained by applying simple augmentations to the source image. Specifically, image shifting is used for camera motion along vertical and horizontal axes, downsampling and central cropping for zooming in and out, and perspective transformation for rotation. With inverted images $G(\mathbf{z})$ and target edits $\texttt{edit}(G(\mathbf{z}), \alpha)$, they learn the direction $\mathbf{n}$ by minimizing the distance between the generated image $G(\mathbf{z}+\alpha\mathbf{n})$ after taking an α-step in the latent direction and the target image $\texttt{edit}(G(\mathbf{z}), \alpha)$.

Unsupervised Manner Some methods [27, 122] aim to discover interpretable directions in the latent space in an unsupervised manner, *i.e.*, without the requirement of paired data. For example, Härkönen et al. [27] (Apr 2020) create interpretable controls for image synthesis by identifying important latent directions based on PCA applied in the latent or feature space. The obtained principal components correspond to certain attributes, and the selective application of the principal components allows for the control of many image attributes. Some of their discovered directions support 3D operations such as rotation or zooming out. This method is considered as "unsupervised" since the directions can be discovered by PCA without using any labels. There is still a need to manually annotate these directions to the target operations and to which layers they should be applied to.

In [97], Shen et al. (Jul 2020) show that latent directions in a pretrained GAN for 3D-aware image synthesis can be directly computed in a closed form without any kind of training or optimization. They propose a **Semantics Factorization (SeFa)** method based on the singular value decomposition of the weights of the first layer of a pretrained GAN. They observe that the semantic transformation of an image, usually denoted by moving the latent code toward a certain direction $\mathbf{n}' = \mathbf{z} + \alpha\mathbf{n}$, is only determined by the latent direction $\mathbf{n}$. Therefore, the directions $\mathbf{n}$ can cause a significant change in the output image $\Delta\mathbf{y}$, *i.e.*, $\Delta\mathbf{y} = \mathbf{y}'-\mathbf{y} = (\mathbf{A}(\mathbf{z}+\alpha\mathbf{n})+\mathbf{b})-(\mathbf{A}\mathbf{z}+\mathbf{b}) = \alpha\mathbf{A}\mathbf{n}$, where $\mathbf{A}$ and $\mathbf{b}$ are respectively the weight and bias of certain layers in G. The obtained formula, $\Delta\mathbf{y} = \alpha\mathbf{A}\mathbf{n}$, suggests that the desired editing with direction $\mathbf{n}$ can be achieved by adding the term $\alpha\mathbf{A}\mathbf{n}$ onto the projected code and indicates that the weight parameter $\mathbf{A}$ should contain the essential knowledge of image variations. The eigenvectors of the matrix $\mathbf{A}^T\mathbf{A}$ should be the desired directions $\mathbf{n}^*$. This gives a closed-form factorization of latent semantics in GANs. This method supports multiple 3D control operations such as car orientation, face and body pose, streetscape and bedroom viewpoint, zoom, shift, as well as rotation on a variety of GANs.

While previous methods predominantly relied on GANs, diffusion models have recently emerged as compelling alternatives. Despite their increasing popularity in image synthesis and editing, the understanding of their latent space is still under exploration. Recently, Kwon et al. [53] introduced an **Asymmetric Reverse Process (Asyrp)** strategy to discover a semantic latent space in frozen pre-trained diffusion models. They coin this semantic latent space for **Denoising Diffusion Models (DDMs)** as "h-space", which has shown its potential in facilitating semantic image editing in a manner akin to GANs. This h-space consists of the bottleneck activations in the DDM's denoiser at each timestep of the diffusion process. Building upon this discovery, Haas et al. [34] delve deeper into understanding the properties of h-space, proposing several innovative methods to discover meaningful semantic directions within it. In a different approach, Brack et al. [9] put forward the Stable Artist, designed to guide the semantic direction in the

latent space of text-conditioned generative diffusion models. Their primary component, **Semantic Guidance (SEGA)**, steers the diffusion process along variable numbers of semantic directions. This offers the ability to make subtle image edits, alter compositions and styles, as well as optimizing overall artistic conception. Beyond these capabilities, SEGA also enables probing of latent spaces to uncover insights into how the model represents learned concepts.

4.1.2 Pinpointing Predetermined Targets. Similar to discovering semantic directions, recent methods [25, 69, 80, 99] have proposed alternative strategies to pinpoint desired edits in the latent space of a pretrained generative model. Unlike the predetermined route provided by the discovered semantic directions, these methods present a different outlook: they acknowledge the goal (or destination) but do not necessitate knowing the specific path to reach it. Such methods are typically formulated as optimization problems, aiming to determine the alterable trajectory from a starting point to the predetermined destination [130, 131]. For instance, the ability to manipulate coarse object position is realized by integrating intermediate constructs, such as "blobs" [26], or heatmaps [123]. These strategies facilitate the modification of either image-aligned semantic attributes, like appearance, or broad geometric characteristics, including object position and pose. In contrast to these methods, a few methods utilize point-based editing, a powerful but less-explored way of controlling GANs. In GANWarping [124], the user is asked to warp a select number of generated images by defining several control points to create customized models. Although the modifications alter the shape of the object, other visual elements such as pose, color, texture, and background are faithfully maintained. However, the realism of the warped images is not ensured. UserControllableLT [25] allows point-based editing by transforming the latent vectors of a GAN. However, this approach only supports editing using a single point being dragged on the image, and it does not handle multiple-point constraints effectively. Additionally, the control is not precise; the target point is often not reached after editing. DragGAN [80], on the other hand, enables users to "drag" any points of an image to precisely reach target points interactively. DragGAN iteratively performs motion supervision and point tracking. The motion supervision directs the handle point towards target position; the point tracking updates the handle point to track the object in the image. Through DragGAN, users can deform an image with precise control over pixel placement, thereby manipulating pose, shape, expression, and layout of various categories such as animals, cars, humans, landscapes, and more. Inspired by the advancements in dragging GANs [25, 80], a handful of studies [69, 99] have begun to explore similar drag-style manipulations on Diffusion models.

4.2 Incorporating 3D Parameters as Controls

Methods using 3D parameters as control factors typically follow a paradigm described as $x' = G(x, \theta)$. Explicit control over the 3D parameters θ gives the edited result x'. Here, θ could be human-interpretable attribute descriptions or a set of parameters from 3D models. Sometimes, the given control factors are intuitively understandable, *e.g.*, (age: 20 years old), (head pose: pitch, yaw, roll), or using the environment map to represent light condition. Most methods in this category incorporate 3D pretrained model parameters into 2D image-based generative models for controllable 3D-aware synthesis. These methods propose solutions to translate controls of 3D face rendering models into GAN-generated processes. Taking face generation as an example, they usually integrate priors from a parametric **3D Morphable Model (3DMM)** [7] as explicit control factors. Table 3 is an overview of methods that incorporate 3D parameters to a 2D generative model.

The models in this section, as well as those in the next, make use of additional 3D models. Their main difference is that the former uses 3D model parameters as input control factors while the

Table 3. Overview of Methods that Incorporate 3D Parameters to a 2D Generative Model

Method	Publication	Control Factor	Supervision ($\theta \leftrightarrows I$)
StyleRig [117]	CVPR 2020	3DMM parameters; θ	$I = \mathcal{R}(\theta)$
DiscoFaceGAN [22]	CVPR 2020	3DMM, SH, angle vector; θ	$\theta = \mathcal{F}(I)$
PIE [118]	TOG 2020	3DMM parameters; θ	$\theta = \text{MoFa}(G(w))$
CONFIG [51]	ECCV 2020	graphic parameters θ	synthetic images with θ
GAN-Control [103]	ICCV 2021	intuitive representation y	$y_i = \mathcal{R}_i(G(z))$
3D-FM GAN [59]	ECCV 2022	a rendering $I_r(\theta)$ (equiv. to θ)	(image, rendering) pairs

latter uses them as supervision signals. There is another series of studies combining implicit 3D representation with 3DMM, either being trained with a reconstruction loss using annotated multi-view datasets [39] or directly imposing 3DMM conditions into 3D NeRF volume and being trained on unannotated single-view images [114]. We focus on leveraging 3D priors for image synthesis based on 2D generative models and will introduce other studies in the remaining sections.

4.2.1 Preliminary: Control Factors from Pretrained Models. Most methods use 3DMM parameters to provide explicit control. 3DMMs are commonly used to represent faces, where faces are parameterized by head rotation ϕ and translation ρ, identity geometry α, expressions β, skin reflectance δ, and scene illumination γ: $\theta = (\phi, \rho, \alpha, \delta, \beta, \gamma) \in \mathbb{R}^m$. The parametric nature of 3DMMs allows navigating and exploring the space of plausible faces. Thus, synthetic images can be rendered based on different parameter configurations. In practice, these 3DMM parameters are first transformed before being used in the network [22, 118]. Besides 3DMM, parameters from other state-of-the-art tools could also be used to provide 3D control factors. In [1], Microsoft Face API predicts pitch and yaw as the head pose. GAN-Control [103] extracts head-pose, expression, illumination, age, and hair color by using several off-the-shelf attribute predictors. DiscoFaceGAN [22] extracts identity, expression, and texture information from 3DMM, approximates scene illumination with **Spherical Harmonics (SH)** [84], and defines face pose as three rotation angles.

For this category, a key question is how to associate these parameters with corresponding images as these methods require supervised training. Except one using existing synthetic data with 3D parameters [51], others use pretrained models to achieve transformations from one direction to another, *i.e.*, $I \rightarrow \theta$ by using attribute predictors [103] or $\theta \rightarrow I$ by synthetic rendering [59, 117], and learn a mapping network for the opposite direction.

4.2.2 Explicit Control over 3D Parameters. With 3D parameters obtained, many methods [51, 59, 103, 117, 118, 121] are developed to incorporate them as input control factors into a 2D generative model for controllable 3D-aware image synthesis. This section demonstrates how these methods introduce 3D parameters and achieve explicit control through data collection, network design, and loss functions.

In MoFA [121], Tewari et al. (Mar 2017) use a CNN to project a face into the 3DMM space, followed by a differentiable renderer to synthesize the reconstructed face. The network is trained on a large collection of face images in a self-supervised manner. Inspired by the computer graphics pipeline, CONFIG [51] (May 2020) uses a set of parameters to represent and control desired factors. Blendshape values control facial expressions, Euler angles control head pose, and environment maps control the illumination. CONFIG has two encoders (E_R and E_S) that encode real face images I_R and the parameters $\theta : \{\theta_1, \ldots, \theta_k\}$ of the synthetic images to a shared latent space $\mathcal{Z}$, which is factorised into elements that each part z_i corresponds to a different facial attribute controlled by θ_i. Each element z_i comes as the i-th parameter of $z \in \mathcal{Z}$ either from θ_i (encoded by E_S) or from a different real face image (encoded by E_R). They adopt a two stage-training scheme to learn a disentangled latent space and produce photorealistic images.

Those based on StyleGANs [47, 48] either use a pretrained StyleGAN model that keeps its weights fixed [117, 118] or make slight modifications to how 3D parameters are incorporated as conditions [22, 103]. StyleRig [117] and PIE [118] are two examples of using a pretrained StyleGAN. StyleRig (Apr 2020) trains a neural network, called RigNet, to inject a subset of parameters into a given StyleGAN latent code w. RigNet is a function rignet$(\cdot,\cdot)$ that maps a pair of StyleGAN code w and subset of 3DMM parameters θ to a new StyleGAN code w', *i.e.* $w' = \text{rignet}(w, \theta)$. Several RigNets are trained, each dealing with a single mode of control (pose, expression, lighting). For self-supervised training, they introduce two key components: a learnable parameter regressor $\mathcal{F}$ and a pretrained differentiable render layer $\mathcal{R}$. $\mathcal{F}$ maps a latent code w to a vector of semantic control parameters θ: $\theta = \mathcal{F}(w)$. $\mathcal{R}$ takes a parameter vector θ as input and generates a synthetic rendering $I_w = \mathcal{R}(\theta)$. StyleRig allows for multiple-attribute editing but only on synthetic facial images rather than real ones. In contrast, PIE (Sep 2020) uses a model-based face auto-encoder to replace $\mathcal{F}$ and $\mathcal{R}$ of StyleRig, facilitating real image editing.

Some methods inherit the main structure of Style-based generators and make slight modifications, mainly different on how the 3D parameters are incorporated as the condition. DiscoFaceGAN [22] (Apr 2020) proposes an unconditional 3D-aware method with controllability on four attributes: identity α, expression β, scene illumination γ, and face pose δ. Their model consists of two networks that learn the mapping (1) $V(\cdot)$ from z-space to θ-space; and (2) $G(\cdot)$ from θ-space to the image space. The latent code z is sampled from a standard normal distribution. The parameters θ is the concatenation of the four control factors $\theta := [\alpha, \beta, \gamma, \delta, \varepsilon]$ and the noise ε, which is the same of z for image diversity. They train four VAEs for α, β, γ, and δ on the θ samples extracted by using an off-the-shelf 3D face reconstruction method from a real image set. Only the decoders are kept after the VAE training and denoted as $V_i, i = 1, 2, 3, 4$, for z-space to θ-space mapping. For training G, they sample $z = [z_1, \ldots, z_5]$ from standard normal distribution, map it to θ, and feed θ to both G and the renderer to obtain a generated face x and a rendered face x', respectively. They apply three types of losses for training: adversarial loss, imitative loss, and contrastive loss. GAN-Control [103] (Jan 2021) builds on the StyleGAN2 [48] architecture. They divide the $\mathcal{Z}$ and $\mathcal{W}$ latent spaces to $N + 1$ separate sub-spaces, in accordance with N control attributes and one residual one for non-concerned information. The original StyleGAN2 architecture is changed from a single mapping network ($w = f(z)$, implemented as an eight-layered MLP) to each control z_i having its own $f(\cdot)$ so that $w_i = f(z_i)$. The combined latent vector (concatenation of the sub-vectors), w, is then fed into the generator G. To enable explicit control over each attribute, they use contrastive learning for disentanglement. Given a set of pretrained attribute predictors $\{\mathcal{R}_i\}_{i=1}^{N}$, they extract intermediate features as the attribute information from sampled images $G(z)$ and use them to calculate the distances during training. To support explicit control during inference, they further train N encoders $\{E_k\}_{k=1}^{N}$, each to map a human-interpretable attribute representation y^k to a latent code w^k. They use the attribute predictors to label the randomly-sampled images as the training data. Different from GAN-Control [103], in 3D-FM GAN [59], Liu et al. (Aug 2022) change StyleGAN G to make it conditional on a given image and a rendering. They estimate the lighting and 3DMM parameters of the face as the 3D parameters θ. These θ are not directly incorporated into G as the explicit control signal but are used instead to generate a rendering $I_r(\theta)$ of the same given image $I(\theta)$, which leads to a paired dataset. The resulting pairs $\{I(\theta), I_r(\theta)\}$ are used for reconstruction training, and $\{I(\theta_i), I_r(\theta_j)\}$ with different attributes of the same identity are for disentangled training.

4.3 Introducing 3D Prior Knowledge as Constraints

This category of studies facilitate the learning of 3D consistency by utilizing one or more kinds of 3D prior knowledge as constraints, such as shape [16, 129, 152], albedo [2, 98], normal [2], and

depth [78, 98, 101]. Both Section 4.2 and this section aim to make a 2D generative model, especially GAN, 3D-aware. Section 4.2 focuses on the methods that incorporate 3D prior knowledge to 2D GANs for explicit control, while this section emphasizes those methods that introduce 3D-aware components into 2D GANs and use 3D prior knowledge as constraints for training. In addition to whether to introduce explicit 3D parameters as inputs, the slight difference between them is also reflected in the different concepts of dataset usage and network structure design. The former is able to control each of the desired attributes because of introducing 3D parameters as control factors but it lacks explicit geometry and texture as holistic 3D supervision, which leads to multi-view inconsistencies. The latter introduces 3D-aware components into 2D generative models (mostly GANs) and uses 3D prior knowledge (*e.g.*, predicted depth from a off-the-shelf depth estimation method) to constrain the training process, resulting in a degree of consistency but in the meantime a lack of fine-grained control. The two types of methods are not mutually exclusive. In addition to introducing 3D parameters to improve controllability, a few methods also implement 3D constraints to improve consistency across multiple views.

4.3.1 Preliminary: 3D Prior Knowledge. Basically, the intrinsic components used to describe the physical world can be used here as priors, including but not limited to shape (surface, depth, and normal), material (albedo, reflectivity and shininess), and lighting (direction, intensity). We introduce in Section 2.1 some common shape representations. Albedo, also referred to as reflection coefficient, is a measure of how reflective a surface is. It is either determined by a value between 0 and 1 or a percentage value. The more reflective a surface is, the higher the albedo value. A surface normal, or simply normal, to a surface at each point is a vector perpendicular to the tangent plane of the surface at that point. Normals represent the curvature of the object and can be used for reflecting light. Depth is the distance between the camera and the object at each pixel. They all contain geometric contextual features. There is a track of studies of intrinsic decomposition, which can be seen as a simplification of inverse rendering for general scenes, aiming to provide interpretable intermediate representations from images. There are also many methods specifically proposed to infer one specific environmental component, such as depth estimation, normal estimation, and light estimation. All these models can be potentially used as the training constraints for this category of methods.

4.3.2 Introducing 3D Components into 2D Models. With the chosen 3D priors, the next important decision for this kind of method is to find a way of introducing 3D-aware components into their models and using 3D priors as training constraints. In S^2-GAN [125], Wang and Gupta (Mar 2016) use a two-stage training for indoor scene synthesis: an unconditional GAN for structure (geometry) generates a surface normal map and the second GAN for style (appearance) takes this surface normal map as condition and outputs an image. VON [152] (Dec 2018) uses shape as the 3D prior and design a GAN based on 3D convolutional neural network to learn the geometry information. An unconditional shape GAN G_s first generates voxel grid shape s from a randomly sampled shape code z_s. The differentiable projection module then projects s to 2.5D sketches $s_{2.5D}$ at a sampled viewpoint z_v. The 2.5D sketches include both the object's depth and silhouette. The texture network G_t finally adds realistic, diverse texture to these 2.5D sketches to generate 2D images: G_t takes $s_{2.5D}$ and another latent code z_t as input and outputs a 2D images. GIS [2] (Sep 2018) and NGP [16] (Feb 2021) utilize more than one 3D prior, such as albedo maps and normal maps, resulting in multiple 2D GANs to learn all the 3D attributes. In the above methods, 3D-aware components are used as intermediates, which are supervised by outputs from pretrained models. There are a few methods that only introduce 3D-aware components into 2D models without using 3D priors from pretrained models to constrain the training. For example, RGBD-GAN [78] (Sep 2019) generates two RGBD images with different camera parameters and then warps them to each

other to ensure 3D consistency. It learns to generate view-consistent images consistent from pure 2D image collections.

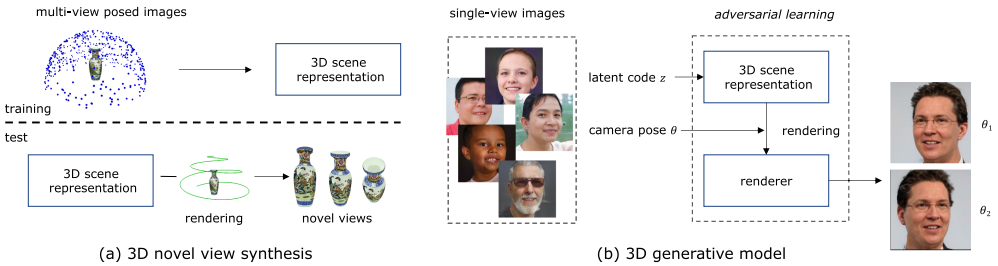
More recently, a few methods [98, 101] are built on top of StyleGAN architectures, either using a pretrained StyleGAN or adapting the vanilla design to their setting. They are also referred to as StyleGAN2-based 2.5D GANs in [12]. LiftedGAN [98] (Nov 2020) equips a pre-trained StyleGAN2 generator with five additional 3D-aware networks, which disentangle the latent space of StyleGAN2 into texture, shape, viewpoint, and lighting. These 3D components are then used for rendering. The proposed model is able to output both the 3D shape and texture, allowing explicit pose and lighting control. To control the viewpoint, DepthGAN [101] (Feb 2022) designs a dual-generator based on StyleGAN. The depth branch G_d takes as the input an uniformly sampled angle θ from range $[\theta_l, \theta_r]$ and a depth latent code, and synthesize a depth image at θ. The rgb branch G_r takes the intermediate feature maps of G_d as the conditions to acquire the geometry information. They use a pre-trained depth prediction model to get the corresponding depth image of each RGB image. A rotation consistency loss is introduced to enhance the multi-view consistency during training. The image synthesized under angle θ_1 is projected to a point cloud and re-projected to the 2D space under θ_2, and compared with the image generated under θ_2. GMPI [149] (Jul 2022) makes a classical 2D GAN, *i.e.*, StyleGAN2, 3D-aware by only introducing (1) a multiplane image style generator branch which produces a set of alpha maps conditioned on their depth; and (2) a pose conditioned discriminator.

Despite impressive image quality, these methods still tend to produce 3D inconsistent faces under large expression and pose variations or scenes under different views due to the lack of a holistic 3D representation. They also inherit inconsistencies introduced by the pretrained model they use. For example, depth estimation methods, especially depth estimated from a single image, are known to suffer from the world-inconsistency. With the advances in differentiable rendering and implicit neural 3D representations, a recent line of work has explored photorealistic 3D-aware face or scene synthesis using *only* 2D image collections as the training data, *without any 3D supervision*. The representative studies are categorized into INR-based NVS methods and 3D-aware generative models. The details of INR-based NVS methods can be found in the appendix, while Section 2.4 and Section 3.3 provide a brief introduction and a thorough discussion of their relationships, respectively. We delve into 3D-aware generative models in the subsequent section.

5 3D-AWARE GENERATIVE MODELS

Inspired by 3D novel view synthesis methods, follow-up works introduce the efficient and expressive neural scene representations, especially INR to the field of 2D generative image synthesis, leading to a new paradigm called *3D-aware generative models* [13, 94]. These methods do not assume a large number of posed images of a single scene. Instead, they learn a model for synthesizing novel scenes by training on single-view images without 3D supervision. As pointed out in Section 3.3, the INR-based novel view synthesis techniques and the methods discussed in this section share common terminology and objectives. As such, they frequently draw inspiration from one another. Both aim to generate multi-view-consistent and photorealistic images, using a similar pipeline that first learns the 3D representation and then renders it from that viewpoint (see Figure 6). It is their application scenarios and training data that differentiate the two kinds of methods. As with their 2D counterparts, 3D-aware generative models generate images from a collection of single views, while 3D novel view synthesis learns a 3D representation from multiple views of a scene.

These 3D-aware generative models follow a similar experimental setting as their 2D counterparts, *i.e.*, generating high quality photorealistic results from single-view image datasets, with an extra goal to ensure 3D consistency across multiple views. In 2D generative models,



(a) 3D novel view synthesis (b) 3D generative model

Fig. 6. The clarification of the defined terminology and the differences between 3D novel view synthesis methods (Section 2.4) and 3D-aware generative models (Section 5). Both aim to generate multi-view-consistent and photo-realistic images, using a similar pipeline that learns the 3D representation first and then renders it. Their differences lie in the application scenarios and training data. 3D novel view synthesis (a) learns a 3D representation from *multiple views* of a scene. 3D generative models (b) learn to generate images from a collection of *single views*. Once trained, 3D-aware generative models generate images under a limited range of viewpoints while NVS methods allow for the free-view rendering of high-fidelity and detailed images.

"unconditional" methods are referred to as those merely inputting latent codes that are sampled from a prior distribution. In this survey, we use "unconditional" 3D-aware generative models to refer to those using latent codes and camera positions as input. In some cases, the camera positions could also be generated from the randomly-sampled latent codes instead of human-understandable control factors. Those taking other inputs, especially image, text, semantic label, or sketch, are the conditional ones.

Much of the earlier research primarily on unconditional generation relied on 3D-aware Generative Adversarial Networks (GANs), thus resulting in a prevalent exploration of 3D-aware GANs [33, 73, 74, 79, 137, 138] (see Section 5.1). However, recent trends in the studies are also indicating a growing interest in using diffusion models for 3D generative modeling [3, 14, 44, 49, 64, 70, 133, 144] (see Section 5.2). We then present conditional 3D-aware generative models [11, 17, 57, 71, 111, 113] in Section 5.3.

5.1 Unconditional 3D-aware GANs

The majority of methods in unconditional 3D-aware generative models belong to the category of GANs. 3D GANs, an outstanding representative of 3D-aware generative models, usually utilize an adversarial framework to learn these representations in an unsupervised manner. The performance details of several representative unconditional 3D-aware GANs [12, 13, 23, 33, 79, 94, 109, 137] on the FFHQ dataset [47] are provided in Table 4, with the FID [36] used as the evaluation metric. Some use generative latent optimization [8, 85] instead of adversarial training [32]. Other kinds of generative models, especially diffusion models [37, 110], which have proven extremely effective in generating high-quality images in recent years, have not yet been widely applied to 3D-ware image synthesis. Only very few recent studies are based on diffusion models for 3D novel view synthesis [128], which are introduced in Section 5.2. These generative models are expected to catch up in the near future.

Like 2D GANs, 3D-aware GANs have recently achieved tremendous breakthroughs in terms of image quality and editability for 2D image synthesis, with the extra goal of 3D consistency by introducing explicit camera control. These methods can be formulated in the form of $I = f(z, \theta)$, where the noise vector z is for appearance and θ means camera pose. Towards editable, high-resolution, and view-consistent image synthesis, proposed methods mainly work on two key components: (1) to learn efficient and expressive representations of geometry and appearance; and

Table 4. The Performance Details of Several Representative Unconditional 3D-aware GANs on the FFHQ Dataset [47], with the FID used as the Evaluation Metric

Method	Publication	Renderer	Upsampler	Resolution	FID on FFHQ
GRAF [94]	NeurIPS 2020	Density, Color	No	128 × 128	46.30
π-GAN [13]	CVPR 2021	Density, Color	No	128 × 128	29.90
StyleNeRF [33]	ICLR 2022	Density, Color	Yes	256 × 256	8.00
				512 × 512	7.80
				1024 × 1024	8.10
StyleSDF [79]	CVPR 2022	SDF, Color	Yes	256 × 256	11.50
				512 × 512	10.07
				1024 × 1024	10.01
VolumeGAN [137]	CVPR 2022	Density, Color	Yes	256 × 256	9.10
GRAM [23]	CVPR 2022	Occupancy, Color	No	256 × 256	14.50
EpiGRAF [109]	NeurIPS 2022	Density, Color	No	512 × 512	9.92
EG3D [12]	CVPR 2022	Density, Color	Yes	256 × 256	4.80
				512 × 512	4.70

(2) to develop accelerated and view-consistent rendering algorithms. Some methods are proposed to facilitate user-interactive editing.

In this section, we first introduce the unconditional 3D-aware GANs based on different neural scene representations. We especially emphasize their efforts towards: (1) learning efficient and expressive geometry and appearance representations (Section 5.1.1); (2) developing accelerated and view-consistent rendering algorithms (Section 5.1.2); (3) broadening the applicable scenarios (Section 5.1.3); and (4) real-time and user-interactive editing (Section 5.1.4).

5.1.1 Efficient and Expressive Representations. Early 3D GANs adopt voxel-based representation. For example, PrGANs [29] (Dec 2016) and VON [152] (Dec 2018) first learn an explicit shape and render images at different viewpoints. They are trained under 3D supervision (as such they are also categorized into Section 4.3). PlatonicGAN [35] (Nov 2018) learns a generative 3D model from an unstructured collection of 2D images. The learned shape and its rendered images are limited to low resolutions and coarse detail due to the computational complexity. The approaches using deep-voxel representations [73, 74] can create finely-detailed images under different poses. HoloGAN [73] (Apr 2019) first uses 3D convolutions to learn a deep-voxel representation (in a canonical pose), then utilizes a rigid-body transformation (3D rotation) to transform this representation to a certain pose, and finally applies a projection unit to render an image. In contrast to HoloGAN, which learns 3D features directly for the entire scene, BlockGAN [74] (Feb 2020) learns 3D features for each object separately. It decomposes a 3D scene into a background and one or more foreground objects, each of which is represented by a noise vector z_i. This design disentangles a scene into separate objects and enables control over camera pose, lighting, and shadow. However, early voxel-based methods [29, 35, 73, 74, 152] fail to synthesize complex scenes and photorealistic details due to the limited grid resolutions. Their reliance on a learned black-box rendering leads to discretization artifacts, degrades view-consistency of the generated images, and makes generalization to unseen camera poses difficult. Liao et al. [55] (Dec 2019) use 3D primitives as abstract object representations and differentiable rendering to project the 3D representations onto the image plane where a 2D generator transforms them into object appearances and composites them into a coherent image.

The limited expressiveness and efficiency of previous methods prevents them from synthesising complex scenes and photorealistic details. Therefore, INRs, especially NeRFs, which have proven

to generate high-fidelity results in novel view synthesis, are introduced to 3D-aware generative models. To avoid the requirement of posed images, an increasing number of methods turn to utilize an adversarial framework to train a generative model for these representations from *unposed* images. The visualization results from COLMAP [93] validate those methods showing greater 3D consistency than the voxel-based representations. GRAF [94] (Jul 2020) uses NeRF to represent the scene and an adversarial framework to train on unposed images. The generator takes camera matrix, camera pose, 2D sampling pattern, and shape/appearance codes as input and predicts an image patch. The discriminator compares the synthesized patch to a real patch extracted from a real image. One significant modification is that GRAF makes NeRF conditioned on two additional latent codes: a shape noise and an appearance noise. A follow-up work pi-GAN [13] (Dec 2020) differs from GRAF in three ways on network architecture and training strategy: (1) pi-GAN uses SIREN [104] as the choice of scene representation rather than a positionally-encoded ReLU MLP [67]; (2) pi-GAN leverages a StyleGAN-inspired mapping network to condition layers in the SIREN on a single input noise code through **feature-wise linear modulation (FiLM)** instead of conditioning on two additional shape/appearance codes; (3) pi-GAN follows ProgressiveGAN [45] where the discriminator grows progressively rather than a patch-based discriminator. Built similarly to pi-GAN, LOLNeRF [85] (Nov 2021), which is capable of single-shot view synthesis of human faces, uses generative latent optimization [8] instead of adversarial training [32]. GIRAFFE [76] improves the BlockGAN [74] framework by replacing the voxel-based representation and 3D-to-2D projection with a NeRF-based compositional 3D scene representation and a neural rendering pipeline. It can rotate, translate, scale each object and change camera poses but with compromised image quality and resolution.

Compared to 2D generative models, these models take many more calculations to render an image (speed) and require much more memory during training to cache intermediate results (memory). Computational constraints limit the rendering resolutions and quality. For high-quality image synthesis (towards 512×512 and beyond), recent methods turn to find more efficient and expressive representations of geometry and appearance, improve the training efficiency, or the combination of both. We focus on the strategies of learning efficient and expressive representations as follows in this subsection, leaving later parts to the next subsection.

CIPS-3D [150] (Oct 2021) adopts a shallow NeRF network (containing only three SIREN blocks) to represent 3D shape and a deep 2D INR network to synthesis high-fidelity appearance. Inspired by recent progress in 3D surface reconstruction, GOF [135] (Nov 2021) combines implicit surfaces and radiance fields. They reinterpret the alpha values in the rendering equation as occupancy representations and reformulate generative radiance fields by predicting alpha values instead of volume densities. EG3D [12] (Dec 2021) proposes to use a tri-plane hybrid 3D representation, formulated from explicit features of StyleGAN2 generator. They align explicit features along three axis-aligned orthogonal feature planes and query any 3D position by projecting it onto each of the three feature planes, resulting in an aggregated 3D feature. The aggregated features are then interpreted as color and density by an additional lightweight decoding network. This tri-plane representation is 3 to 8 times faster than an implicit Mip-NeRF [4] network and only requires a fraction of its memory. A super-resolution module upsamples and refines raw neurally rendered images. VolumeGAN [137] (Dec 2021) explicitly learns a structural representation and a textural representation. It learns a feature volume to represent the underlying structure, which is transformed into a feature field based on a NeRF-style model. The feature field is then aggregated into a 2D feature map as the textural representation. A neural renderer is finally used for appearance synthesis. StyleSDF [79] (Dec 2021) merges a SDF-based 3D representation into the 2D StyleGAN generator. This framework consists of two main components: a backbone conditional SDF volume renderer and a StyleGAN generator. The renderer takes in a latent code and camera

parameters, queries points and view directions within the volume, and projects 3D surface features onto 2D views. To overcome the drawbacks of GIRAFFE and inherit its 3D controllability, a follow-up work GIRAFFE-HD [138] (Mar 2022) leverages a style-based neural renderer, generates the foreground and background independently, and stitches them together to composite a coherent final image. It enforces semantic disentanglement and 3D consistency through training constraints. In contrast to previous methods, VoxGRAF [95] (Jun 2022) adopts a 3D-aware GAN based on a sparse scene representation that allows for efficient rendering. It parameterizes the radiance field on a sparse voxel grid rather than using a coordinate-based MLP and predicts colors and density values on this sparse voxel grid using volume rendering.

5.1.2 Efficient and Consistent Rendering Algorithms. Early studies [73, 74] use simple 3D-to-2D projections for rendering. Such operations fail to produce high-quality images with fine-grained details and are restricted to representing poses in the training dataset. For higher rendering quality, recent methods adopt the state-of-the-art neural volume rendering techniques. GRAF [94] and pi-GAN [13] implement a discretized form of the volume rendering equation and uses the stratified and hierarchical sampling approach introduced by NeRF. The neural volume rendering approach has several advantages over previous 3D-to-2D projections: (1) producing images with fine details and high resolutions and (2) allowing for explicit control over camera pose, focal length, aspect ratio, and other parameters. Despite the advantages, the volume integrations approximated by sampling points along viewing rays are still costly for both training and inference. Some methods render the 3D representation directly at the final image resolution [13, 95]. Due to the high memory and computation cost of volume rendering, direct rendering at target resolution is not efficient and struggles to generate images at high-resolution (512×512 and beyond). Some recently-developed methods make use of a two-stage rendering process [12, 33, 76, 79, 137] or develop efficient volume rendering strategy [23, 81, 150] for high-resolution image generation. Meanwhile, aimed to reduce view-inconsistent artifacts brought by the 2D renderers, they adopt different strategies such as NeRF path regularization [33] and dual discriminators [12].

Two-stage Rendering Process: To high-resolution image generation, some methods adopt a two-stage rendering process. Typically, they first generate a feature map at a low resolution and then employ upsampling in 2D space to progressively increase into the required high resolution. Niemeyer and Geiger [76] improve training and rendering efficiency by combining NeRF with a ConvNet-based renderer. Similarly, Chan et al. [12] perform the majority of the training at a rendering resolution of 64×64 and gradually increase the resolution, pixel-by-pixel, to 128×128, which are fed into a super-resolution module to produce images at the target resolution. However, these pixel-wise learnable upsamplers sacrifice view consistency and impair the quality of the learned 3D geometry due to network designs. In contrast, non-learnable upsamplers that interpolate the feature map with pre-defined lowpass filters (*e.g.*, bilinear interpolation) produce smoother results but lead to non-removable bubble artifacts. In StyleNeRF [33], the two approaches are combined to balance quality and consistency. Despite that upsampler scales the intermediate result to high resolution, it comes with two severe limitations: (1) the texture and shape change as camera moves; (2) the geometry is represented in a low resolution ($\approx 64^3$), both resulting in a compromised multi-view consistency of a generated object. To overcome the above limitations of the two-stage rendering, Skorokhodov et al. [109] drop the upsampler and improve the patch-wise optimization strategy [94] to build a 3D generator. They redesign the discriminator by making it better suited to operating on image patches of variable scales and locations, along with changing the random scale sampling strategy from an annealed uniform to an annealed beta distribution. This allows the proposed EpiGRAF to converge 2–3 times faster than upsampler-based architectures despite the generator modeling geometry in full resolution.

Efficient Rendering Strategy: This part is categorized into two groups of approaches according to their adopted strategies. As for the first group, an expressive and efficient representation allows usage of a simplified sampling strategy [79] or a lightweight decoder [12]. In StyleSDF [79], using SDFs leads to higher view-consistency and expressiveness, even with a simplified volume sampling strategy. They sample N points from N evenly-sized bins of integration intervals instead of stratified sampling [13, 33, 67, 94], which reduces the number of samples by half. VoxGARF [95] accelerates the rendering from the perspective of spatial sparsity, where volume rendering yields a foreground image and an alpha mask. The second group aims to develop efficient sampling strategy [23, 81, 135]. They constrain the point sampling in a reduced space rather than anywhere in the volume. GRAM [23] proposes a manifold predictor $\mathcal{M}$ to predict a reduced space for point sampling and radiance field learning. $\mathcal{M}$ is a light-weight MLP that maps a point $\boldsymbol{x}$ to a scalar value s. The predicted scalar field gives N isosurfaces with a set of predefined levels (constant values $\{l_i\}$): $\mathcal{S}_i = \{\boldsymbol{x} | \mathcal{M}(\boldsymbol{x}) = l_i\}$. The rendering only samples points from intersections between the determined isosurfaces and a camera ray. Pan et al. [81] introduce a light-weighted surface tracking network S to estimate the rendered object surface. This saves rendering computations by just querying points near the predicted surface. The sample region shrinks from the entire volume to a narrow interval around the surface.

Consistent View Regularization: Optimizing the radiance fields from a set of 2D training images can encounter critical degenerate solutions in the absence of geometry constraints, leading to a multi-view inconsistency problem in NeRF-based generative models. Some methods propose multi-view regularizations on colors and shapes to improve photometric and geometric consistencies. Different strategies are adopted to reduce view-inconsistent artifacts brought by the 2D renderers [12, 33, 81, 102, 148]. For example, Gu et al. [33] propose a NeRF path regularization to enforce 3D consistency, which is implemented by sub-sampling high-resolution outputs and comparing them against the low-resolution image generated by NeRF. Chan et al. [12] propose a pose-conditioned dual discriminators with two modifications in traditional GAN discriminators. First, they pass the rendering camera intrinsics and extrinsics matrices to the discriminator as a conditioning label. Second, they take as the input the concatenation of the final result I_r and a low-resolution RGB image I'_r. They interpret the first three feature channels of a neurally rendered feature image as I'_r and bilinearly upsample it to the same size of I_r. This pose-conditioned dual discriminator is used in many follow-up studies. Shi et al. [102] design a geometry-aware discriminator, GeoD, to improve 3D-aware GANs. Besides the real/fake classification, they assign the discriminator an additional geometry branch, aiming to derive the shape-related information (*e.g.*, depth and normal), which is employed as an extra signal to supervise the generator. Pan et al. [81] observe that small variations of shape could lead to similar RGB images that look equally plausible to the discriminator, as the color of many objects is locally smooth. This phenomenon is referred to as shape-radiance (color) ambiguity. To eliminate this problem, they propose a multi-lighting constraint, which is realized by modeling illumination explicitly and rendering with various lighting conditions. To overcome the same issue, MVCGAN [148] (Apr 2022) builds geometry constraints by optimizing multiple views jointly to ensure geometry consistency between views. They minimize re-projection loss between a primary image and a warped image, and integrate a stereo mixup module to encourage the warped image to be similar to a real image. Such scheme guarantees geometry constraints between different views and supports large pose variations. PoF3D [100] (Jan 2023) notice a high sensitivity towards pose priors in existing 3D-aware image synthesis methods and propose a novel approach notably eliminating the need for pose priors. It comprises two principal components: a pose-free generator and a pose-aware discriminator. The pose-free generator maps a latent code to a neural radiance field as well as a camera pose. The pose-aware discriminator, on the other hand, first predicts a camera pose

from the given image and then uses it as the pseudo label for conditional real/fake discrimination. PoF3D demonstrates the potential for high-quality 3D-aware image synthesis without the reliance on 3D pose priors.

5.1.3 Broadening Applicable Scenarios. Previous methods mainly focused on common categories, such as faces, cars, and other single categories with simple geometry and appearance. These methods can only generate a single canonical object and show limited capacity in composing a complex scene containing a variety of objects. However, recent methods have attempted to expand to more complex categories, such as the fine-grained shape [109], human bodies [140, 145], and street scenes [136]. For example, EpiGRAF [109] (Jun 2022) drops the upsampler and improves the patch-wise optimization strategy [94] to build a 3D generator. The generator models the geometry in a full dataset resolution and is able to fit data where the global structure differs a lot between different objects in Megascans [109]. 3D-SGAN [145] (Nov 2022) and 3DHumanGAN [140] (Dec 2022) are two examples of 3D-aware human body synthesis. The human body exhibits a more diverse variety of shapes, poses, and texture variations, presenting far more significant challenges compared with modeling the human face. DiscoScene [136] (Dec 2022) is a 3D-aware generative model for scene synthesis. The proposed model spatially disentangles the entire scene into object-centric generative radiance fields, leveraging only 2D images with global-local discrimination. DiscoScene not only achieves generation fidelity and editing flexibility for individual objects, but also efficiently composes these objects and the background into a complete scene. There are other categories of 3D-aware generative models that require the inductive bias of specialized domain knowledge. For instance, significant progress has been made in the incorporation of 3D-aware image generation with the articulation of the body or face for controllable generative models. This approach combines 3D-aware generative models with deformation fields [5, 38, 112].

Notably, recent research efforts [92, 108] have been made to extend 3D-aware image synthesis beyond a single category, aiming to encompass a wide range of categories, such as those found in ImageNet [20]. Particularly, Skorokhodov et al. [108] (Mar 2023) propose **3D generator with Generic Priors (3DGP)**. Building upon EpiGRAF [109], the generator first produces a tri-plane representation for the scene, given a random latent code z. Subsequently, a shallow 2-layer MLP predicts RGB color and density values from an interpolated feature vector at a 3D coordinate. Images and depths are volumetrically rendered at any given camera position. The model is trained across all 1,000 classes of ImageNet [20], demonstrating the feasibility of multi-categorical 3D synthesis for non-alignable data.

5.1.4 User-interactive Editing. Plenty of approaches are proposed to facilitate user-interactive editing. They can change the background's appearance independent of the foreground, translate or rotate the foreground object in 3D, and change the foreground object's shape and color. For example, some methods [74, 76, 138] use multiple noise vectors z_i to represent the background and each foreground object. Unlike previous studies that learn 3D features directly for the whole scene [73], they learn a 3D feature for each object, render separately, and stitch them together to composite a coherent final image. Such design disentangles a scene into separate objects and enables control over camera pose, lighting, and shadow.

Differently, Kwak et al. [52] (Jul 2022) design a layer-wise **SUbspace in INR NeRF-based generator (SURF-GAN)**. Instead of being used to represent the background and each foreground object, multiple noise vectors in SURF-GAN are injected layer-by-layer into NeRF-based SURF blocks. The interpretable dimensions are captured in layers with sub-modulation vectors. SURF-GAN [52] includes several SURF blocks that take position and view direction as inputs to predict view dependent color. IDE-3D [111] (May 2022) enables local control of the facial shape and texture and supports real-time, interactive editing. It makes two key modifications based on [12] to

enable such interactive disentangled editing. First, they take two (instead of one) codes respectively representing shape and texture, which gives 3D volumes of semantic and texture in the tri-plane representation. Second, they jointly render rgb images and semantic masks (instead of just rgb images) through the volume rendering. The dual discriminator is also changed accordingly to take as input the concatenation of rgb and semantic masks.

Recently, 3D GAN inversion methods have been developed [50, 54, 134], which are based on the previously mentioned 3D GANs and are used for 3D-aware image editing. There are two key distinctions between these 3D GAN inversion methods and 2D GAN inversion methods [132, 151]. Firstly, these 3D methods rely on a 3D-aware GAN model (*e.g.*, EG3D [12]), instead of a 2D GAN (*e.g.*, StyleGANs [47, 48]). Secondly, 3D GAN inversion methods take into account the camera pose. The 3D GAN inversion methods not only achieve realistic and accurate manipulation but also excel in preserving the identity and geometry of the original input. These methods can generate edited 3D shapes from the modified latent codes, leveraging the capabilities of 3D-aware GANs. Additionally, these 3D GAN inversion methods deliver results that are comparable to those of other methods but necessitate significantly fewer computing resources when compared with training a 3D-aware image editing method.

5.2 Unconditional 3D-aware Diffusion Models

Recent trends in research also show a burgeoning interest in employing diffusion models [37, 110] for 3D generative modeling [3, 14, 44, 49, 70, 133, 144]. Diffusion models, which have proven extremely effective in generating high-quality images in recent years, have not yet been widely applied to 3D-ware image synthesis. DiffRF [70] (Dec 2022) introduces a novel method for 3D radiance fields synthesis, utilizing denoising diffusion probabilistic models. It learns multi-view consistent priors from posed image collections, enabling free-view image synthesis and precise shape generation. RenderDiffusion [3] (Nov 2022) introduces a novel image denoising architecture that generates and renders an intermediate three-dimensional representation of a scene in each denoising step. This imposes a robust inductive structure within the diffusion process, yielding a 3D-consistent representation while only requiring 2D supervision. The resulting 3D representation can be rendered from any perspective. HoloDiffusion [44] (Mar 2023) introduces a novel diffusion framework that can be trained end-to-end with only posed 2D images for supervision. Furthermore, it proposes an image formation model that decouples model memory from spatial memory. 3D-aware generative diffusion models are expected to catch up in the near future.

5.3 Conditional 3D Generative Models

The unconditional 3D generative models, primarily GANs (Section 5.1) as well as some methods based on diffusion models (Section 5.2), generally lack the capability to perform precise attribute editing for real images. Similar to 2D counterparts, 3D-aware real image editing could either use pretrained 3D GANs via GAN inversion or train a 3D-aware generative model from scratch with additional inputs.

As an emerging technique to bridge the real and fake image domains, GAN inversion [132, 151] plays an essential role in enabling pretrained GAN models real image editing. It inverts a real image into the latent space of a trained GAN model, which allows us to alter image attributes by varying the inverted code in the latent space (known as latent space traversals). For 3D-aware image editing, some rely directly on 2D GAN inversion and latent space traversal techniques, while others develop tools particularly for 3D GAN inversion. For example, ShadeGAN [81] (Oct 2021) could also be used to reconstruct a given image by performing GAN inversion. Such inversion with this method allows to obtain object properties from the image, such as shape, normal, albedo, and shading. FENeRF [113] (Nov 2021) has attempted to edit the local shape and texture in a

facial volume by using an optimization-based GAN inversion. Lin et al. [57] (Mar 2022) propose a method for multi-view consistent video editing and animation based on 3D GAN inversion. They invert the video frames into the latent space of a pi-GAN by using **pivotal tuning inversion (PTI)** [89] and edit face attributes by using StyleFlow [1]. IDE-3D [111] adopts a hybrid GAN inversion approach. Given a facial image and its semantic label, it obtains texture and semantic latent codes with corresponding encoders and uses them as the initialization for PTI to obtain high-fidelity reconstruction. Editing is performed by drawing on inverted semantic masks.

Other methods target at pose-dependent views by training a 3D network conditioned on a single image. Pix2NeRF [11] (Feb 2022) demonstrates that merely applying learning-based GAN inversion (learning an encoder and keeping the generator fixed during training) is insufficient to obtain an accurate mapping from image to latent space with pi-GAN as the backbone. Instead, they train an encoder jointly with a generator and a discriminator (both of the same architecture and procedure as in [13]). Once trained, given an input image, Pix2NeRF disentangles its pose and content and renders novel views of the content. Sem2NeRF [17] (Mar 2022) takes as input a single-view 2D semantic mask and outputs a NeRF-based 3D representation that can be used to render photorealistic images in a 3D-aware view-consistent manner. AutoRF [71] (Apr 2022) focuses on novel view synthesis of objects without background. This model consists of an encoder that extracts a shape and an appearance code from an object's image, which can be decoded into an implicit radiance field operating in normalized object space and leveraged for novel view synthesis. Object images are generated from real-world imagery by leveraging machine-generated 3D object detections and panoptic segmentation.

6 DISCUSSION

Despite great advances in deep 3D-aware generative image synthesis, challenges remain and its rapid growth is expected to continue. In the following, we provide an overview of future directions, problems to solve, and trends to anticipate. Due to their unique characteristics, some limitations or future trends may only apply to certain categories of methods.

Quality: Unlike traditional graphic rendering, since implicit neural representations do not provide an explicit and holistic 3D shape for rendering, inconsistencies seems inevitable in the generated surface and texture under different viewpoints. Moreover, these strategies are introduced separately rather than endogenously as part of the method. We expect that in the future there will be more endogenous approaches to 3D-aware generative models that generate high-quality, high-resolution, multi-view-consistent images in real time, as well as high-quality 3D geometry.

Speed: There is typically a slow training and inference speed with 3D-aware image synthesis methods. Most attempts to accelerate training and inference time require extra memory for caching trained models or use additional voxel/spatial-tree based scene features. It is expected that future speed-based methods should develop memory-friendly frameworks as well as novel inclusive and learnable scene representations to accelerate training and inference.

Editability: In spite of the promising quality of the produced results, most methods are incapable of editing individual image components. In some methods, latent vectors are used at various points along the pipeline to control composition, shape, and appearance of images. The latent codes enable the model to control small changes in scene content, such as lighting or coloration, per image. Others allow additional input to change aspects of the scene, such as images, texts, semantic labels, or direct control parameters. The editability of deep 3D-aware generative image synthesis methods, however, still has plenty of room for improvement compared with their 2D counterparts.

Forensics: The success of recent generative models has led to many new applications, but also raised ethical and social concerns, such as fraud and fabricated images, videos, and news (known as deepfakes). The ability to detect deepfakes is essential to preventing malicious usage of these models. Recent studies have shown that a classifier can be trained to distinguish deepfakes and generalize to unseen architectures. It may continue to be a cat-and-mouse game in the future, since generated images will become increasingly difficult to detect. Conversely, these images can also be utilized as the training data for identifying fakes.

Generalization: 3D-aware generative image synthesis has been limited to a narrow range of perspectives, falling short of the free-view rendering capabilities seen in NVS. Recent efforts, inspired by progress in INR-based NVS, have expanded the range of camera movements, but still lack the ability to generate 360-degree views. Additionally, while there have been attempts to broaden 3D-aware image synthesis from common categories, which are characterized by simple geometry and appearance, to intricate scenes composed of various objects, these attempts often fall short in terms of controllability and quality. Therefore, there exists a substantial opportunity for future research to broaden the scope of 3D-aware image synthesis to cover a more diverse range of categories applicable in real-world scenarios. However, this expansion will inevitably increase computational complexity, which must be adequately addressed.

Network Design: Most current methods rely on a 2D architecture design that introduces 3D representation, rendering, and multi-view regularization to make a 2D model 3D-aware. The majority of 3D-aware generative models use convolution-powered generative adversarial networks. In recent years, transformer-based 2D image synthesis methods have emerged corresponding to their counterparts in convolutional networks. Developing 3D-aware generative models that are built on transformers with lighter structures and lower computational demands remains to be explored. Since almost all the methods above are based on GANs, developing other generative models that are 3D-aware, especially diffusion models, is also a promising future direction.

Evaluation metrics: It remains to be explored whether there are any reliable metrics that can better evaluate the photorealistic and geometric quality of generated images. Image quality and diversity are mainly measured by general metrics used for generative models. 3D consistency is often evaluated by measuring distances between (pseudo-)depth maps of generated images or comparing similarities of generated face identities at different camera positions. Some use COLMAP reconstruction on their rendering outputs to demonstrate the 3D consistency. However, considering the lack of real reference, these measures only partially reflect the stability of generated results, but cannot reflect the distance from real samples. There is still a lack of effective assessment tools to evaluate the difference between the predicted and expected outcomes in a more reliable and direct manner for deep generative 3D-aware image synthesis.

7 CONCLUSION

This paper presents a comprehensive overview of recent advances in deep 3D-aware generative image synthesis. We propose a systematic taxonomy for deep 3D-aware generative image synthesis methods. Specifically, we categorize the existing approaches into two groups: 3D control of 2D generative models and 3D-aware generative models. We also identify some open problems on this topic to inspire future research. We hope that this timely and up-to-date survey will serve as a starting point for future research to help advance this emerging and challenging field.

REFERENCES

[1] Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. 2021. StyleFlow: Attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. *TOG* 40, 3 (2021), 1–21.

[2] Hassan Abu Alhaija, Siva Karthik Mustikovela, Andreas Geiger, and Carsten Rother. 2018. Geometric image synthesis. In *ACCV*. 85–100.

[3] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J. Mitra, and Paul Guerrero. 2023. RenderDiffusion: Image diffusion for 3D reconstruction, inpainting and generation. In *CVPR*. 12608–12618.

[4] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. 2021. Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*. 5855–5864.

[5] Alexander Bergman, Petr Kellnhofer, Wang Yifan, Eric Chan, David Lindell, and Gordon Wetzstein. 2022. Generative neural articulated radiance fields. *NeurIPS* 35 (2022), 19900–19916.

[6] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401* (2018).

[7] Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*. 187–194.

[8] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. 2017. Optimizing the latent space of generative networks. *arXiv preprint arXiv:1707.05776* (2017).

[9] Manuel Brack, Patrick Schramowski, Felix Friedrich, Dominik Hintersdorf, and Kristian Kersting. 2022. The stable artist: Steering semantics in diffusion latent space. *arXiv preprint arXiv:2212.06013* (2022).

[10] Marcel C. Bühler, Abhimitra Meka, Gengyan Li, Thabo Beeler, and Otmar Hilliges. 2021. VariTex: Variational neural face textures. In *ICCV*. 13870–13879.

[11] Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. 2022. Pix2NeRF: Unsupervised conditional pi-GAN for single image to neural radiance fields translation. In *CVPR*. 3981–3990.

[12] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. 2022. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*. 16102–16112.

[13] Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis. In *CVPR*. 5799–5809.

[14] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. 2023. Generative novel view synthesis with 3D-aware diffusion models. *arXiv preprint arXiv:2304.02602* (2023).

[15] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, and Hao Su. 2015. ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012* (2015).

[16] Xuelin Chen, Daniel Cohen-Or, Baoquan Chen, and Niloy J. Mitra. 2021. Towards a neural graphics pipeline for controllable image generation. In *CGF*, Vol. 40. 127–140.

[17] Yuedong Chen, Qianyi Wu, Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. 2022. Sem2NeRF: Converting single-view semantic masks to neural radiance fields. In *ECCV*. 730–748.

[18] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. StarGAN v2: Diverse image synthesis for multiple domains. In *CVPR*. 8185–8194.

[19] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*. 5828–5839.

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*. 248–255.

[21] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*. 4690–4699.

[22] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. 2020. Disentangled and controllable face image generation via 3D imitative-contrastive learning. In *CVPR*. 5154–5163.

[23] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. 2022. GRAM: Generative radiance manifolds for 3D-aware image generation. In *CVPR*. 10663–10673.

[24] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An open urban driving simulator. In *CoRL*. 1–16.

[25] Yuki Endo. 2022. User-controllable latent transformer for StyleGAN image layout editing. In *Computer Graphics Forum*, Vol. 41. 395–406.

[26] Dave Epstein, Taesung Park, Richard Zhang, Eli Shechtman, and Alexei A. Efros. 2022. BlobGAN: Spatially disentangled scene representations. In *ECCV*. 616–635.

[27] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. GANSpace: Discovering interpretable GAN controls. In *NeurIPS*, Vol. 33. 9841–9850.

[28] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. 2022. StyleGAN-human: A data-centric odyssey of human generation. In *ECCV*. Springer, 1–19.

[29] Matheus Gadelha, Subhransu Maji, and Rui Wang. 2017. 3D shape induction from 2D views of multiple objects. In *3DV*. 402–411.
[30] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. 2021. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction. In *CVPR*. 8649–8658.
[31] Stephan J. Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. 2021. FastNeRF: High-fidelity neural rendering at 200FPS. In *ICCV*. 14346–14355.
[32] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.
[33] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2022. StyleNeRF: A style-based 3D aware generator for high-resolution image synthesis. In *ICLR*.
[34] René Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, and Tomer Michaeli. 2023. Discovering interpretable directions in the semantic latent space of diffusion models. *arXiv preprint arXiv:2303.11073* (2023).
[35] Philipp Henzler, Niloy J. Mitra, and Tobias Ritschel. 2019. Escaping Plato's cave: 3D shape from adversarial rendering. In *ICCV*. 9983–9992.
[36] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, Vol. 30. 6626–6637.
[37] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *NeurIPS*, Vol. 33. 6840–6851.
[38] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. 2023. EVA3D: Compositional 3D human generation from 2D image collections. In *ICLR*.
[39] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. 2022. HeadNeRF: A real-time NeRF-based parametric head model. In *CVPR*. 20342–20352.
[40] Tao Hu, Shu Liu, Yilun Chen, Tiancheng Shen, and Jiaya Jia. 2022. EfficientNeRF: Efficient neural radiance fields. In *CVPR*. 12902–12911.
[41] Ali Jahanian, Lucy Chai, and Phillip Isola. 2020. On the "steerability" of generative adversarial networks. In *ICLR*.
[42] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*. 2901–2910.
[43] James T. Kajiya and Brian P. Von Herzen. 1984. Ray tracing volume densities. *SIGGRAPH* 18, 3 (1984), 165–174.
[44] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy J. Mitra. 2023. HoloDiffusion: Training a 3D diffusion model using 2D images. In *CVPR*. 18423–18433.
[45] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*.
[46] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training generative adversarial networks with limited data. In *NeurIPS*, Vol. 33. 12104–12114.
[47] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*. 4401–4410.
[48] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of StyleGAN. In *CVPR*. 8107–8116.
[49] Seung Wook Kim, Bradley Brown, Kangxue Yin, Karsten Kreis, Katja Schwarz, Daiqing Li, Robin Rombach, Antonio Torralba, and Sanja Fidler. 2023. NeuralField-LDM: Scene generation with hierarchical latent diffusion models. In *CVPR*. 8496–8506.
[50] Jaehoon Ko, Kyusun Cho, Daewon Choi, Kwangrok Ryoo, and Seungryong Kim. 2023. 3D GAN inversion with pose optimization. In *WACV*. 2967–2976.
[51] Marek Kowalski, Stephan J. Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson, and Jamie Shotton. 2020. CONFIG: Controllable neural face image generation. In *ECCV*. 299–315.
[52] Jeong-gi Kwak, Yuanming Li, Dongsik Yoon, Donghyeon Kim, David Han, and Hanseok Ko. 2022. Injecting 3D perception of controllable NeRF-GAN into StyleGAN for editable portrait image synthesis. In *ECCV*. 236–253.
[53] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. 2023. Diffusion models already have a semantic latent space. In *ICLR*.
[54] Yushi Lan, Xuyi Meng, Shuai Yang, Chen Change Loy, and Bo Dai. 2023. Self-supervised geometry-aware encoder for style-based 3D GAN inversion. In *CVPR*. 20940–20949.
[55] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. 2020. Towards unsupervised learning of generative models for 3D controllable image synthesis. In *CVPR*. 5871–5880.
[56] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. 2021. BARF: Bundle-adjusting neural radiance fields. In *ICCV*. 5721–5731.
[57] Connor Z. Lin, David B. Lindell, Eric R. Chan, and Gordon Wetzstein. 2022. 3D GAN inversion for controllable portrait image animation. In *ECCV Workshop*.

[58] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. 2020. Neural sparse voxel fields. In *NeurIPS*, Vol. 33. 15651–15663.
[59] Yuchen Liu, Zhixin Shu, Yijun Li, Zhe Lin, Richard Zhang, and S. Y. Kung. 2022. 3D-FM GAN: Towards 3D-controllable face manipulation. In *ECCV*. 107–125.
[60] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*. 1096–1104.
[61] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *ICCV*. 3730–3738.
[62] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural volumes: Learning dynamic renderable volumes from images. *TOG* 38, 4, Article 65 (2019), 14 pages.
[63] William E. Lorensen and Harvey E. Cline. 1987. Marching cubes: A high resolution 3D surface construction algorithm. *SIGGRAPH* (1987), 347–353.
[64] Wufei Ma, Qihao Liu, Jiahao Wang, Angtian Wang, Yaoyao Liu, Adam Kortylewski, and Alan Yuille. 2023. Adding 3D geometry control to diffusion models. *arXiv preprint arXiv:2306.08103* (2023).
[65] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. 2021. NeRF in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*. 7210–7219.
[66] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy networks: Learning 3D reconstruction in function space. In *CVPR*. 4460–4470.
[67] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*. 405–421.
[68] Ron Mokady, Omer Tov, Michal Yarom, Oran Lang, Inbar Mosseri, Tali Dekel, Daniel Cohen-Or, and Michal Irani. 2022. Self-distilled StyleGAN: Towards generation from internet photos. In *SIGGRAPH*. 1–9.
[69] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. 2023. DragonDiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421* (2023).
[70] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. 2023. DiffRF: Rendering-guided 3D radiance field diffusion. In *CVPR*. 4328–4338.
[71] Norman Müller, Andrea Simonelli, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. 2022. AutoRF: Learning 3D object radiance fields from single view observations. In *CVPR*. 3971–3980.
[72] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *TOG* 41, 4 (2022), 1–15.
[73] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. 2019. HoloGAN: Unsupervised learning of 3D representations from natural images. In *ICCV*. 7587–7596.
[74] Thu Nguyen-Phuoc, Christian Richardt, Long Mai, Yong-Liang Yang, and Niloy Mitra. 2020. BlockGAN: Learning 3D object-aware scene representations from unlabelled images. In *NeurIPS*, Vol. 33. 6767–6778.
[75] Michael Niemeyer and Andreas Geiger. 2021. CAMPARI: Camera-aware decomposed generative neural radiance fields. In *3DV*. 951–961.
[76] Michael Niemeyer and Andreas Geiger. 2021. GIRAFFE: Representing scenes as compositional generative neural feature fields. In *CVPR*. 11453–11464.
[77] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. 2020. Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision. In *CVPR*. 3501–3512.
[78] Atsuhiro Noguchi and Tatsuya Harada. 2020. RGBD-GAN: Unsupervised 3D representation learning from natural image datasets via RGBD image synthesis. In *ICLR*.
[79] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. 2022. StyleSDF: High-resolution 3D-consistent image and geometry generation. *CVPR* (2022), 13503–13513.
[80] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. 2023. Drag your GAN: Interactive point-based manipulation on the generative image manifold. In *SIGGRAPH*.
[81] Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. 2021. A shading-guided generative implicit model for shape-accurate 3D-aware image synthesis. In *NeurIPS*. 20002–20013.
[82] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*. 165–174.
[83] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. 2020. Convolutional occupancy networks. In *ECCV*. 523–540.
[84] Ravi Ramamoorthi and Pat Hanrahan. 2001. An efficient representation for irradiance environment maps. In *SIGGRAPH*. 497–500.
[85] Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. 2022. LOLNeRF: Learn from one look. In *CVPR*. 1558–1567.

[86] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. 2021. KiloNeRF: Speeding up neural radiance fields with thousands of tiny MLPs. In *ICCV*. 14335–14345.
[87] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. 2021. Common objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction. In *ICCV*. 10901–10911.
[88] Konstantinos Rematas, Andrew Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. 2022. Urban radiance fields. In *CVPR*. 12932–12942.
[89] Daniel Roich, Ron Mokady, Amit H. Bermano, and Daniel Cohen-Or. 2022. Pivotal tuning for latent-based editing of real images. *TOG* 42, 1 (2022), 1–13.
[90] Mehdi S. M. Sajjadi and Henning Meyer. 2022. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *CVPR*. 6229–6238.
[91] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training GANs. In *NeurIPS*. 2226–2234.
[92] Kyle Sargent, Jing Yu Koh, Han Zhang, Huiwen Chang, Charles Herrmann, Pratul Srinivasan, Jiajun Wu, and Deqing Sun. 2023. VQ3D: Learning a 3D-aware generative model on ImageNet. *arXiv preprint arXiv:2302.06833* (2023).
[93] Johannes L. Schonberger and Jan-Michael Frahm. 2016. Structure-from-motion revisited. In *CVPR*. 4104–4113.
[94] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. 2020. GRAF: Generative radiance fields for 3D-aware image synthesis. In *NeurIPS*, Vol. 33. 20154–20166.
[95] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. 2022. VoxGRAF: Fast 3D-aware image synthesis with sparse voxel grids. In *NeurIPS*, Vol. 35. 33999–34011.
[96] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020. Interpreting the latent space of GANs for semantic face editing. In *CVPR*. 9240–9249.
[97] Yujun Shen and Bolei Zhou. 2021. Closed-form factorization of latent semantics in GANs. In *CVPR*. 1532–1540.
[98] Yichun Shi, Divyansh Aggarwal, and Anil K. Jain. 2021. Lifting 2D StyleGAN for 3D-aware face generation. In *CVPR*. 6258–6266.
[99] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent Y. F. Tan, and Song Bai. 2023. DragDiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv preprint arXiv:2306.14435* (2023).
[100] Zifan Shi, Yujun Shen, Yinghao Xu, Sida Peng, Yiyi Liao, Sheng Guo, Qifeng Chen, and Dit-Yan Yeung. 2023. Learning 3D-aware image synthesis with unknown pose distribution. In *CVPR*. 13062–13071.
[101] Zifan Shi, Yujun Shen, Jiapeng Zhu, Dit-Yan Yeung, and Qifeng Chen. 2022. 3D-aware indoor scene synthesis with depth priors. In *ECCV*. 406–422.
[102] Zifan Shi, Yinghao Xu, Yujun Shen, Deli Zhao, Qifeng Chen, and Dit-Yan Yeung. 2022. Improving 3D-aware image synthesis with a geometry-aware discriminator. In *NeurIPS*, Vol. 35. 7921–7932.
[103] Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, and Gerard Medioni. 2021. GAN-control: Explicitly controllable GANs. In *ICCV*. 14083–14093.
[104] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. 2020. Implicit neural representations with periodic activation functions. In *NeurIPS*, Vol. 33. 7462–7473.
[105] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. 2021. Light field networks: Neural scene representations with single-evaluation rendering. *NeurIPS* 34 (2021), 19313–19325.
[106] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. 2019. DeepVoxels: Learning persistent 3D feature embeddings. In *CVPR*. 2437–2446.
[107] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. 2019. Scene representation networks: Continuous 3D-structure-aware neural scene representations. *NeurIPS* 32 (2019), 1119–1130.
[108] Ivan Skorokhodov, Aliaksandr Siarohin, Yinghao Xu, Jian Ren, Hsin-Ying Lee, Peter Wonka, and Sergey Tulyakov. 2023. 3D generation on ImageNet. In *ICLR*.
[109] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. 2022. EpiGRAF: Rethinking training of 3D GANs. In *NeurIPS*, Vol. 35. 24487–24501.
[110] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. In *ICLR*.
[111] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. 2022. IDE-3D: Interactive disentangled editing for high-resolution 3D-aware portrait synthesis. In *SIGGRAPH Asia*.
[112] Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. 2023. Next3D: Generative neural texture rasterization for 3D-aware head avatars. In *CVPR*. 20991–21002.
[113] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. 2022. FENeRF: Face editing in neural radiance fields. In *CVPR*. 7672–7682.
[114] Keqiang Sun, Shangzhe Wu, Zhaoyang Huang, Ning Zhang, Quan Wang, and HongSheng Li. 2022. Controllable 3D face synthesis with conditional generative occupancy fields. In *NeurIPS*, Vol. 35. 16331–16343.

[115] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*. 2818–2826.
[116] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. 2022. Block-NeRF: Scalable large scene neural view synthesis. In *CVPR*. 8248–8258.
[117] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. 2020. StyleRig: Rigging StyleGAN for 3D control over portrait images. In *CVPR*. 6141–6150.
[118] Ayush Tewari, Mohamed Elgharib, Mallikarjun B. R., Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. 2020. PIE: Portrait image embedding for semantic control. *TOG* 39, 6 (2020), 1–14.
[119] Ayush Tewari, Mallikarjun B. R., Xingang Pan, Ohad Fried, Maneesh Agrawala, Christian Theobalt, other. 2022. Disentangled3D: Learning a 3D generative model with disentangled geometry and appearance from monocular images. In *CVPR*. 1516–1525.
[120] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Wang Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, and Stephen Lombardi. 2022. Advances in neural rendering. In *Computer Graphics Forum*, Vol. 41. 703–735.
[121] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. 2017. MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCV Workshops*. 1274–1283.
[122] Andrey Voynov and Artem Babenko. 2020. Unsupervised discovery of interpretable directions in the GAN latent space. In *ICML (Proceedings of Machine Learning Research, Vol. 119)*. 9786–9796.
[123] Jianyuan Wang, Ceyuan Yang, Yinghao Xu, Yujun Shen, Hongdong Li, and Bolei Zhou. 2022. Improving GAN equilibrium by raising spatial awareness. In *CVPR*. 11285–11293.
[124] Sheng-Yu Wang, David Bau, and Jun-Yan Zhu. 2022. Rewriting geometric rules of a GAN. *TOG* 41, 4 (2022), 1–16.
[125] Xiaolong Wang and Abhinav Gupta. 2016. Generative image modeling using style and structure adversarial networks. In *ECCV*. 318–335.
[126] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *TIP* 13, 4 (2004), 600–612.
[127] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. 2021. NeRF–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064* (2021).
[128] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. 2022. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628* (2022).
[129] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. 2016. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. *NeurIPS* 29 (2016), 82–90.
[130] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. TediGAN: Text-guided diverse face image generation and manipulation. In *CVPR*. 2256–2265.
[131] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. Towards open-world text-guided face image generation and manipulation. *arxiv preprint arxiv: 2104.08910* (2021).
[132] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. 2022. GAN inversion: A survey. *TPAMI* 45, 3 (2022), 3121–3138.
[133] Jianfeng Xiang, Jiaolong Yang, Binbin Huang, and Xin Tong. 2023. 3D-aware image generation using 2D diffusion models. *arXiv preprint arXiv:2303.17905* (2023).
[134] Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, and Qifeng Chen. 2023. High-fidelity 3D GAN inversion by pseudo-multi-view optimization. In *CVPR*. 321–331.
[135] Xudong Xu, Xingang Pan, Dahua Lin, and Bo Dai. 2021. Generative occupancy fields for 3D surface-aware image synthesis. In *NeurIPS*. 20683–20695.
[136] Yinghao Xu, Menglei Chai, Zifan Shi, Sida Peng, Ivan Skorokhodov, Aliaksandr Siarohin, Ceyuan Yang, Yujun Shen, Hsin-Ying Lee, and Bolei Zhou. 2023. DisCoScene: Spatially disentangled generative radiance fields for controllable 3D-aware scene synthesis. In *CVPR*. 4402–4412.
[137] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 2022. 3D-aware image synthesis via learning structural and textural representations. In *CVPR*. 18430–18439.
[138] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. 2022. GIRAFFE-HD: A high-resolution 3D-aware generative model. In *CVPR*.
[139] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2015. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*. 3973–3981.
[140] Zhuoqian Yang, Shikai Li, Wayne Wu, and Bo Dai. 2022. 3DHumanGAN: Towards photo-realistic 3D-aware human image generation. *arXiv preprint arXiv:2212.07378* (2022).

[141] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*. 4578–4587.
[142] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. 2015. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365* (2015).
[143] Oğuz Kaan Yüksel, Enis Simsar, Ezgi Gülperi Er, and Pinar Yanardag. 2021. LatentCLR: A contrastive learning approach for unsupervised discovery of interpretable directions. In *ICCV*. 14263–14272.
[144] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 2023. 3DShape2VecSet: A 3D shape representation for neural fields and generative diffusion models. *arXiv preprint arXiv:2301.11445* (2023).
[145] Jichao Zhang, Enver Sangineto, Hao Tang, Aliaksandr Siarohin, Zhun Zhong, Nicu Sebe, and Wei Wang. 2022. 3D-aware semantic-guided generative model for human synthesis. In *ECCV*. 339–356.
[146] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. 2020. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492* (2020).
[147] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*. 586–595.
[148] Xuanmeng Zhang, Zhedong Zheng, Daiheng Gao, Bang Zhang, Pan Pan, and Yi Yang. 2022. Multi-view consistent generative adversarial networks for 3D-aware image synthesis. In *CVPR*. 18450–18459.
[149] Xiaoming Zhao, Fangchang Ma, David Güera, Zhile Ren, Alexander G. Schwing, and Alex Colburn. 2022. Generative multiplane images: Making a 2D GAN 3D-aware. In *ECCV*. 18–35.
[150] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. 2021. CIPS-3D: A 3D-aware generator of GANs based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788* (2021).
[151] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. 2020. In-domain GAN inversion for real image editing. In *ECCV*. Springer, 592–608.
[152] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. 2018. Visual object networks: Image generation with disentangled 3D representations. *NeurIPS* (2018), 118–129.

Received 20 December 2022; revised 19 September 2023; accepted 26 September 2023