Multi-view Consistent Generative Adversarial Networks for Compositional 3D-Aware Image Synthesis

Xuanmeng Zhang[1,2] · Zhedong Zheng[3] · Daiheng Gao[2] · Bang Zhang[2] · Yi Yang[4] · Tat-Seng Chua[3]

Abstract

This paper studies compositional 3D-aware image synthesis for both single-object and multi-object scenes. We observe that two challenges remain in this field: existing approaches (1) lack geometry constraints and thus compromise the multi-view consistency of the single object, and (2) can not scale to multi-object scenes with complex backgrounds. To address these challenges coherently, we propose multi-view consistent generative adversarial networks (MVCGAN) for compositional 3D-aware image synthesis. First, we build the geometry constraints on the single object by leveraging the underlying 3D information. Specifically, we enforce the photometric consistency between pairs of views, encouraging the model to learn the inherent 3D shape. Second, we adapt MVCGAN to multi-object scenarios. In particular, we formulate the multi-object scene generation as a "decompose and compose" process. During training, we adopt the top-down strategy to decompose training images into objects and backgrounds. When rendering, we deploy a reverse bottom-up manner by composing the generated objects and background into the holistic scene. Extensive experiments on both single-object and multi-object datasets show that the proposed method achieves competitive performance for 3D-aware image synthesis.

Communicated by Gang Hua.

✉ Xuanmeng Zhang
    zhangxuanmeng.zxm@gmail.com

    Zhedong Zheng
    zdzheng12@gmail.com

    Daiheng Gao
    daiheng.gdh@alibaba-inc.com

    Bang Zhang
    zhangbang.zb@alibaba-inc.com

    Yi Yang
    yangyics@zju.edu.cn

    Tat-Seng Chua
    dcscts@nus.edu.sg

1   ReLER Lab, AAII, University of Technology Sydney, Ultimo, Australia

2   DAMO Academy, Alibaba Group , Hangzhou, China

3   Sea-NExT Joint Lab, National University of Singapore, Singapore, Singapore

4   Zhejiang University, Hangzhou, China

1 Introduction

We study the problem of compositional 3D-aware image synthesis, aiming at generating images with explicit control over the camera pose and individual object. Different from 2D generative adversarial networks (Brock et al., 2018; Zhu et al., 2017; Choi et al., 2018; Karras et al., 2018; Huang & Belongie, 2017; Karras et al., 2019; Zheng et al., 2019; Karras et al., 2020; Choi et al., 2020), 3D-aware image synthesis models learn 3D scene representations from images, such as voxels (Nguyen-Phuoc et al., 2019, 2020), intermediate 3D primitives (Liao et al., 2020), and neural radiance fields (NeRF) (Schwarz et al., 2020; Chan et al., 2021; Niemeyer & Geiger, 2021; DeVries et al., 2021). Among these approaches, NeRF-based approaches (Schwarz et al., 2020; Chan et al., 2021; Niemeyer & Geiger, 2021; DeVries et al., 2021; Deng et al., 2022b; Gu et al., 2022) have gained a surge of interest due to the superior performance of high-fidelity view synthesis. However, two key challenges remain. (1) Existing approaches (Schwarz et al., 2020; Chan et al., 2021; Niemeyer & Geiger, 2021) do not guarantee geometry constraints between views, hence usually failing to generate multi-view consistent images in some views. (2) As pointed

Fig. 1 Images synthesized by MVCGAN on the CELEBA-HQ (Karras et al., 2018) dataset. We render multi-view images at resolution 512^2 from different viewpoints

by Schwarz et al. (2020), current methods do not work well on scenes that contain multiple objects with complex shapes and diverse backgrounds.

In this paper, we address the first problem by proposing MVCGAN, a multi-view consistent generative model with geometry constraints (see Fig. 1).

Here we present typical failure cases of existing approach (Niemeyer & Geiger, 2021) in Fig. 2. We identify the cause of the inconsistent phenomenon between views: previous methods optimize a single view of the generated image independently while ignoring the geometry constraints between views (see Sect. 3.2.1). To tackle this problem, the proposed MVCGAN takes inspiration from classical multi-view geometry methods (Zhou et al., 2017; Godard et al., 2019) to build geometry constraints across views. Specifically, we perform multi-view joint optimization by enforcing the photometric consistency between pairs of views with re-projection loss and integrating a stereo mixup mechanism into the training process. Therefore, the generator not only learns the manifold of 2D images but also ensures the geometric correctness of the underlying 3D shape. Besides, we notice that NeRF-based generative approaches (Schwarz et al., 2020; Chan et al., 2021; Niemeyer & Geiger, 2021) typically struggle to render high-resolution images with fine details due to the huge computational complexity of NeRF model (Mildenhall et al., 2020). Therefore, we adopt a hybrid MLP-CNN architecture, which contains one MLP-based NeRF model and one CNN-based decoder. Specifically, the MLP-based NeRF model (Mildenhall et al., 2020) renders the geometry of the 3D shape, and the subsequent CNN-based decoder produces fine details for the 2D appearance. The structure can generate photorealistic high-resolution images while alleviating the computation-intensive problem.

We further adopt MVCGAN to multi-object and background-attached scenarios with a compositional framework, MVCGAN+. In specific, MVCGAN+ employs two MVCGAN branches to model the foreground objects and backgrounds separately. Besides, we propose a "decompose and compose" scheme to perform the complex scene generation in a top-down and bottom-up manner. During training, we explicitly incorporate the object masks to decompose the objects and backgrounds from the training images. The disentanglement of objects and backgrounds allows us to impose geometry constraints on the foreground object and the background separately. When rendering the whole scene, we compose the objects and backgrounds via object masks and occlusion relations. In summary, our main contributions are summarized as follows:

1. We identify the cause of the multi-view inconsistency in 3D-aware image synthesis, and propose to incorporate geometry constraints into the generative radiance field for the single-object scene generation.
2. To tackle complex multi-object scenes, we further scale MVCGAN to a compositional framework with top-down and bottom-up manners. To our knowledge, we are among the early attempts to incorporate instance masks into generative radiance fields to tackle complex multi-object scenarios.
3. We demonstrate the effectiveness and scalability of the proposed approach by evaluating on both single-object and multi-object datasets. Extensive experiments substantiate that our method achieves competitive performance for 3D-aware image synthesis.

This paper is an extension of our previous conference version (Zhang et al., 2022). Compared to the preliminary version, this work includes the following new contents. (1) Owing to the inadequate exploration of complex multi-object scenes in current works, we scale MVCGAN (Zhang et al., 2022) to a compositional framework, MVCGAN+, for multi-object 3D-aware image generation. In particular, we model the foreground objects and backgrounds with two separate branches. (2) By incorporating the easily-obtained 2D annotations, i.e., instance masks and bounding boxes, we formulate the multi-object image generation as a "decompose and compose" process. To our knowledge, we are among the first attempts to incorporate instance masks into generative radiance fields to tackle the multi-object generation problems. (3) To further validate the competence of our method, we add more experiments and discussions for ablation studies and visualization results.
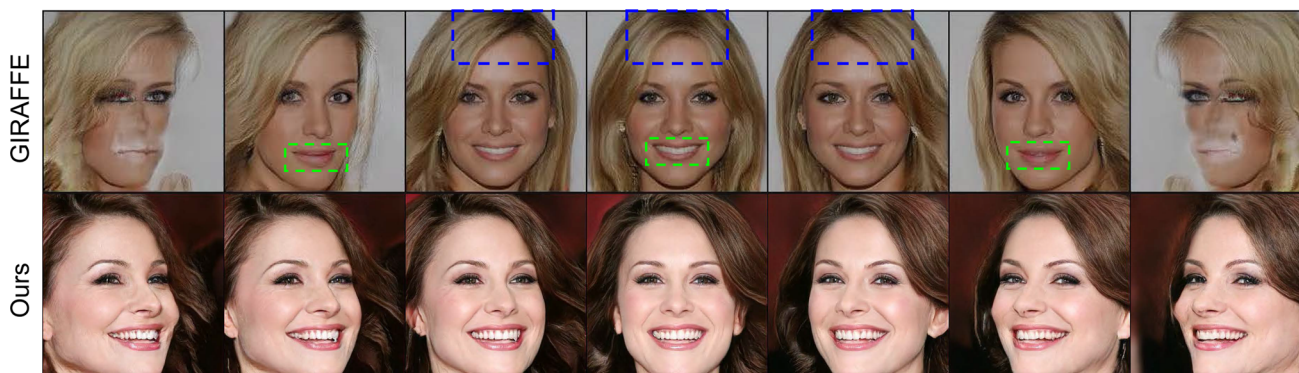
Fig. 2 Typical failure cases. Taking a representative method GIRAFFE (Niemeyer & Geiger, 2021) as an example, the generated images in the first row have obvious inconsistent appearance artifacts between views, such as the direction of hair (blue box) and the opening mouth (green box). Besides, we notice that GIRAFFE (Niemeyer & Geiger, 2021) suffers collapsed results under large pose variations (see the leftmost and rightmost pictures in the first row), which indicates that the model does not learn an appropriate 3D shape. In contrast, our method generates high-quality images with multi-view consistency (see the second row) (Color figure online)

2 Related Work

2.1 Multi-view Geometry

A large number of approaches reconstruct 3D structures with multi-view geometry constraints as supervision signals, such as COLMAP (Schonberger & Frahm, 2016) and ORB-SLAM (Mur-Artal et al., 2015). In recent years, some deep learning techniques (Zhou et al., 2017; Godard et al., 2019; Yao et al., 2018) also combine traditional approaches (Chen & Williams, 1993; Collins, 1996; Szeliski & Golland, 1999) to address 3D vision problems. Inspired by the classical multi-view geometry methods (Chen & Williams, 1993; Debevec et al., 1996; Andrew, 2001; Seitz & Dyer, 1996; Zhou et al., 2017; Godard et al., 2019), we explicitly involve the geometry constraints in the training process for learning a reasonable 3D shape.

2.2 Neural Radiance Fields

Recently, using volumetric rendering and implicit function to synthesize novel views of a scene has gained a surge of interest. Mildenhall et al. (2020) represent complex scenes as Neural Radiance Fields (NeRF) for novel view synthesis by optimizing an implicit continuous volumetric scene function. Due to the simplicity and extraordinary performance, NeRF (Mildenhall et al., 2020) has been extended to plenty of variants, e.g., faster training (Yu et al., 2022a), faster inference (Yu et al., 2021a; Reiser et al., 2021; Garbin et al., 2021; Rebain et al., 2021; Lindell et al., 2021), pose estimation (Yen-Chen et al., 2021; Lin et al., 2021; Jeong et al., 2021; Meng et al., 2021; Wang et al., 2021), generalization (Chibane et al., 2021; Chen et al., 2021; Yu et al., 2021b; Trevithick & Yang, 2021; Liu et al., 2022), video (Xian et al., 2021; Dynamic view synthesis, 2021; Li et al., 2021, a; Peng et al., 2021), and depth estimation (Wei et al., 2021).

2.3 3D-Aware Image Synthesis

Generating photorealistic and editable image content is a long-standing problem in computer vision and graphics. In the past years, generative adversarial networks (GAN) (Goodfellow et al., 2020) have demonstrated impressive results in synthesizing high-resolution images of high quality from unstructured image collections (Zhu et al., 2017; Brock et al., 2018; Choi et al., 2018; Karras et al., 2018; Huang & Belongie, 2017; Karras et al., 2019; Zheng et al., 2019; Karras et al., 2020; Choi et al., 2020). Despite the tremendous success, most of the methods typically only learn the manifold of 2D images while ignoring the 3D representation of the scene. In recent years, several recent works have investigated how to incorporate 3D representation into generative models (Alhaija et al., 2018; Nguyen-Phuoc et al., 2019; Zhu et al., 2018; Liao et al., 2020; Nguyen-Phuoc et al., 2020; Henderson et al., 2020; DeVries et al., 2021). Nguyen-Phuoc et al. (2019) combine a strong inductive bias about the 3D world with deep generative models to learn disentangled representations of 3D objects. Nguyen-Phuoc et al. (2019) provides control over the pose of generated objects through rigid-body transformations of the learned 3D features. Schwarz et al. (2020) propose GRAF, generative radiance fields for 3D-aware image synthesis from unposed 2D images. pi-GAN (Chan et al., 2021) adopts a SIREN-based neural implicit representation with periodic activation functions as the backbone of the generator. GIRAFFE (Niemeyer & Geiger, 2021) represents scenes as compositional generative neural feature fields. ShadeGAN (Pan et al., 2021) models the illumination to

regularize the training process. Combining the occupancy representation with radiance fields, Xu et al. (2021) introduce Generative Occupancy Fields (GOF) to shrink the sample region of the volume rendering process. StyleNeRF (Gu et al., 2022) integrates NeRF (Mildenhall et al., 2020) to the StyleGAN-like generator (Karras et al., 2019, 2020) to close the gap between 2D and 3D GANs. Zhou et al. (2021) extend CIPS (Anokhin et al., 2021) to CIPS-3D, a 3D-aware generator that composes of NeRF and implicit neural representation network. StyleSDF (Or-El et al., 2022) achieves high-resolution image genearation and 3D surface modeling by integrating the SDF-based 3D representation into the 2D style-based generative model (Karras et al., 2019, 2020). Recently, Chan et al. (2022) introduce a novel tri-plane representation with 3D inductive bias, resulting in a more efficient and expressive 3D GAN framework, EG3D. VolumeGAN (Xu et al., 2022) learns a structural and tex-tural representation with a 3D feature volume and neural renderer respectively. Deng et al. (2022b) reduce the number of sampling points by learning generative 2D manifolds (GRAM), while GRAM-HD (Xiang et al., 2022) achieves better results by performing super-resolution in the 3D space. VoxGRAF (Schwarz et al., 2022) explores sparse voxel grid representations to accelerate training. Skorokhodov et al. (2022) redesign the patch-based discriminator to improve the optimization scheme of 3D generative adversarial networks. However, these methods typically optimize a single view of the generated scene independently and ignore the underlying geometry constraints across views.

3 Methodology

Our goal is to generate photorealistic high-resolution images with explicit control over the camera pose while maintaining multi-view consistency. We now present the main components of the proposed method. First, we briefly review the background of NeRF-based generative adversarial networks (Schwarz et al., 2020; Niemeyer & Geiger, 2021; Chan et al., 2021) and identify the limitations of previous methods (see Sect. 3.1). Second, we analyze the cause of the multi-view inconsistency problem and present Multi-View Consistent Generative Adversarial Networks (MVCGAN) for single object generation (see Fig. 5 for an overview). At last, based on MVCGAN, we further introduce a compositional framework (MVCGAN+) for multi-object image generation in Sect. 3.3.

3.1 Preliminaries

Neural Radiance Fields. Neural radiance field (NeRF) synthesizes novel views of the scene by optimizing a fully-connected network using a set of input views. The MLP
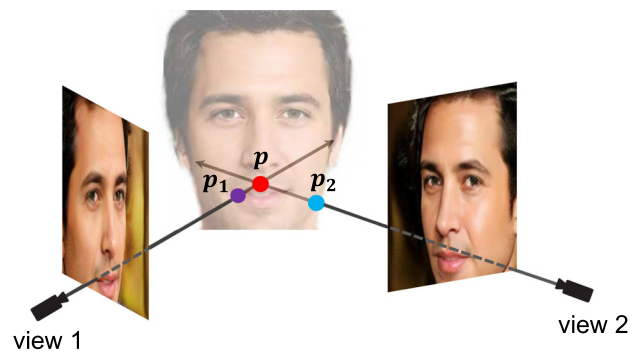

Fig. 3 Visualization of shape-radiance ambiguity. For illustration, we assume p (the red dot) is the location of correct geometry, p_1 (the violet dot) and p_2 (the blue dot) are incorrect geometries. In the absence of geometry constraints, the model can fit incorrect geometry p_1 in view 1 and p_2 in view 2 independently to simulate the effect of the correct geometry p (Color figure online)

network maps a continuous 5D coordinate (3D location $\mathbf{x}$ and 2D viewing direction $\mathbf{d}$) to an emitted color $\mathbf{c}$ and volume density σ (Mildenhall et al., 2020):

$$(\gamma(\mathbf{x}), \gamma(\mathbf{d})) \rightarrow (\mathbf{c}, \sigma), \tag{1}$$

where γ indicates the positional encoding mapping function. To render the neural radiance field from a viewpoint, Mildenhall et al. (2020) use classic volume rendering to accumulate the output colors $\mathbf{c}$ and densities σ into an image.

Generative Radiance Fields. Generative neural radiance fields aim to learn a model for synthesizing novel scenes by training on unposed 2D images. Schwarz et al. (2020) adopt an adversarial framework to train a generative model for radiance fields (GRAF). The generative radiance field is conditioned on a shape code z_s and an appearance code z_a:

$$(\gamma(\mathbf{x}), \gamma(\mathbf{d}), z_s, z_a)) \rightarrow (\mathbf{c}, \sigma). \tag{2}$$

Following GRAF (Schwarz et al., 2020), Niemeyer and Geiger (2021) introduce a compositional generative neural feature field (GIRAFFE). Inspired by StyleGAN (Karras et al., 2019), Chan et al. (2021) instead propose periodic implicit generative adversarial networks (pi-GAN) with feature-wise linear modulation (FiLM) conditioning.

Limitations. We notice two limitations of existing approaches (Schwarz et al., 2020; Niemeyer & Geiger, 2021; Chan et al., 2021). First, they do not guarantee geometry constraints between different views. Consequently, they usually suffer from collapsed results under large pose variations or have obvious inconsistent artifacts across views. Second, these approaches mostly cannot tackle the scene that contains multiple objects and complex backgrounds.
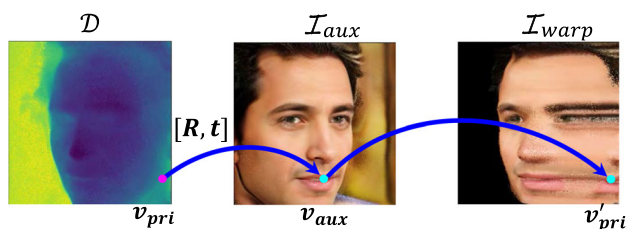
Fig. 4 Illustration of the warping process. For each pixel v_{pri} in the primary image $\mathcal{I}_{pri}$, we first calculate the location of v_{aux} (the corresponding pixel of v_{pri} in the auxiliary image $\mathcal{I}_{aux}$) based on the depth value $\mathcal{D}(v_{pri})$ and camera transformation matrix $[R, t]$. Then we can reconstruct the pixel v'_{pri} of the warped image $\mathcal{I}_{warp}$ from the primary view using the value of pixel v_{aux}. We observe that the warped image has a wrong appearance, which verifies the incorrect geometry shape learned by the model

3.2 MVCGAN for Single-Object Image Generation

3.2.1 Image-Level Multi-view Joint Optimization

Shape-radiance Ambiguity. In this part, we analyze the cause of the multi-view inconsistency problem in NeRF-based generative models. We observe that optimizing the radiance fields from a set of 2D training images can encounter critical degenerate solutions in the absence of geometry constraints. This phenomenon is referred to as shape-radiance ambiguity (Zhang et al., 2020), in which the model can fit the training images with inaccurate 3D shape by a suitable choice of radiance field at each surface point (see Fig. 3). To better illustrate the shape-radiance ambiguity, we warp the rendered images from view 1 to view 2 based on the underlying depth and camera transformation matrix $[R, t]$ (see the details of warping process in Fig. 4 and Eq. 4). We find the warped image shows a wrong appearance, which verifies the assumption of degenerate solutions to the learned 3D shape. To avoid the shape-radiance ambiguity, NeRF (Mildenhall et al., 2020) requires a large number of posed training images from different input views for the scene. However, generative radiance fields have neither annotated camera poses nor sufficient multi-view images in the training dataset. Consequently, the generative model can synthesize reasonable images in some views but produce poor renderings in other views (see Fig. 2).

Warping Process. To alleviate the shape-radiance ambiguity (Zhang et al., 2020), we propose to establish multi-view geometry constraints (Chen & Williams , 1993; Debevec et al., 1996; Andrew, 2001; Seitz & Dyer, 1996; Zhou et al., 2017; Godard et al., 2019) via the warping process between views. First, following pi-GAN (Chan et al., 2021), we adopt a style-based generator that contains a synthesis network G_s (a SIREN-based (Sitzmann et al., 2020; Chan et al., 2021) generative radiance field) and a mapping network G_m (a simple MLP network with ReLU) (see Fig. 5). Given a latent code

$z \in \mathbb{R}^{256}$ in the input latent space $\mathcal{Z}$, the mapping network $G_m : \mathcal{Z} \rightarrow \mathcal{W}$ can produce the intermidiate latent $w \in \mathbb{R}^{256}$, which controls the synthesis network G_s at each layer. Second, instead of only optimizing a single view independently, we aim to optimize multiple views jointly to maintain the 3D consistency across views. As shown in the left of Fig. 5, we randomly sample two camera poses, i.e., the primary pose ξ_{pri} and the auxiliary pose ξ_{aux}, from the pose distribution p_ξ. Taking ξ_{pri} and ξ_{aux} as input, the generative model G_s synthesizes two views of the generated images separately: the primary image $\mathcal{I}_{pri}$ and the auxiliary image $\mathcal{I}_{aux}$. Then we can build geometry constraints between ξ_{pri} and ξ_{aux} via image warping, which reconstructs the primary view by sampling pixels from the auxiliary image $\mathcal{I}_{aux}$. Specifically, for each point v_{pri} in the primary image $\mathcal{I}_{pri}$, we first find the corresponding pixel v_{aux} in the auxiliary image $\mathcal{I}_{aux}$ through the stereo correspondence, and then reconstruct the pixel v'_{pri} of the warped image $\mathcal{I}_{warp}$ from primary view using the value of v_{aux} (see Fig. 4). Next, we present a detailed calculation procedure of the warping process. The stereo correspondence is calculated based on the depth map $\mathcal{D}$ of the primary image and camera transformation matrix from ξ_{pri} to ξ_{aux}. The depth can be rendered in a similar way as rendering the color image (Mildenhall et al., 2020; Deng et al., 2022a). Given the pixel v_{pri} from the primary view, the depth value $\mathcal{D}(v_{pri})$ is formulated as:

$$\mathcal{D}(v_{pri}) = \sum_{i=1}^{N} T_i (1 - exp(-\sigma_i \delta_i)) d_i,$$

$$T_i = exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j), \tag{3}$$

where N is the number of samples in the camera ray, $\delta_i = d_{i+1} - d_i$ is the distance between adjacent sample points and σ_i is the volume density of sample i (refer to Mildenhall et al. (2020); Deng et al. (2022a) to see more details). With the depth value $\mathcal{D}(v_{pri})$, we can obtain the homogeneous coordinates h_{pri} of pixel v_{pri} in the primary camera coordinate system through perspective projection. Then the projected coordinates h_{aux} in the auxiliary view can be calculated as:

$$h_{aux} = K[R, t]\mathcal{D}(v_{pri})K^{-1}h_{pri}, \tag{4}$$

where the camera intrinsics K are known parameters and the camera transformation matrix $[R, t]$ can be calculated from the primary pose ξ_{pri} and the auxiliary pose ξ_{aux}. Finally, we can reconstruct the pixel v'_{pri} in the warped image $\mathcal{I}_{warp}$ from the primary view using the value of pixel v_{aux} (located in h_{aux} of $\mathcal{I}_{aux}$).

Image-level Joint Optimization. After obtaining the warped image $\mathcal{I}_{warp}$, we perform image-level multi-view joint optimization by enforcing the photometric consistency and
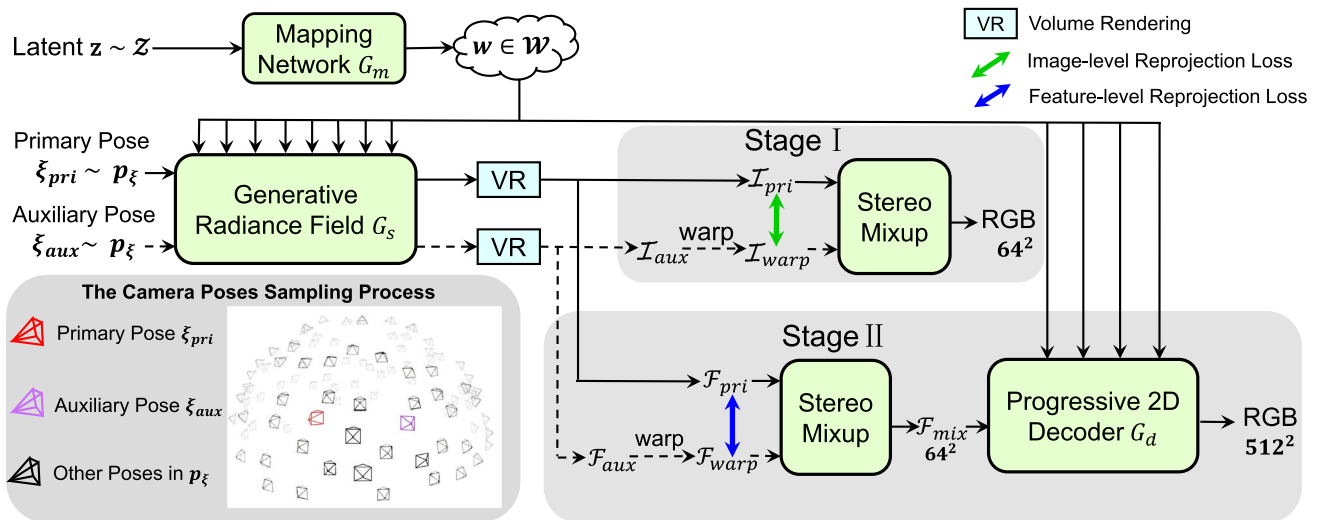
Fig. 5 The generator of MVCGAN. During training, the generative radiance field network G_s takes primary pose ξ_{pri} and auxiliary pose ξ_{aux} as input. The mapping network G_m maps the input latent z to intermediate latent w, which conditions both the generative radiance field network G_s and the progressive 2D decoder G_d. In Stage I, we directly render primary image $\mathcal{I}_{pri}$ and auxiliary image $\mathcal{I}_{aux}$ with the color and density output from G_s. Then we perform image-level multi-view joint optimization and output a low-resolution RGB image (64^2). In Stage II, we instead use volume rendering to accumulate 2D feature maps at low resolution (64^2), and then perform multi-view optimization at the feature level. The progressive 2D decoder G_d upsamples 2D feature map $\mathcal{F}_{mix}$ to a high-resolution RGB image (128^2, 256^2, 512^2) for fine 2D details. During inference, only the primary pose is required without auxiliary pose (the dotted lines do not participate in inference)

employing a stereo mixup module (see Fig. 6). To satisfy the geometry constraints between views, we enforce the photometric consistency across views by minimizing the re-projection loss between the primary image $\mathcal{I}_{pri}$ and the warped image $\mathcal{I}_{warp}$. Following the common practice in image reconstruction (Wang et al., 2004; Zhao et al., 2016; Pillai et al., 2019; Zhou et al., 2017; Godard et al., 2019; Lyu et al., 2021), we formulate the image-level re-projection loss as the combination of L1 (Zhao et al., 2016) and SSIM (Wang et al., 2004):

$$
\begin{aligned}
\mathcal{L}_{ir} = {} & (1 - \mu)||\mathcal{I}_{pri} - \mathcal{I}_{warp}||_1 + \\
& \frac{\mu}{2}(1 - SSIM(\mathcal{I}_{pri}, \mathcal{I}_{warp})),
\end{aligned} \tag{5}
$$

where SSIM is a perceptual metric of image structural similarity and $\mu = 0.85$ empirically. In addition to being similar to the primary image, the warped image should also look like a real image. A straightforward method is introducing two discriminators. One is to compare the warped image $\mathcal{I}_{warp}$ with an arbitrary real image sampled from the training dataset, and the other one compares the primary image $\mathcal{I}_{pri}$. However, introducing extra modules can increase the computation complexity. Inspired by the *mixup* strategy (Zhang et al., 2018), we instead propose a stereo mixup module to optimize both $\mathcal{I}_{pri}$ and $\mathcal{I}_{warp}$ by constructing a virtual mixed image:

$$
\mathcal{I}_{mix} = \eta\mathcal{I}_{pri} + (1 - \eta)\mathcal{I}_{warp}, \tag{6}
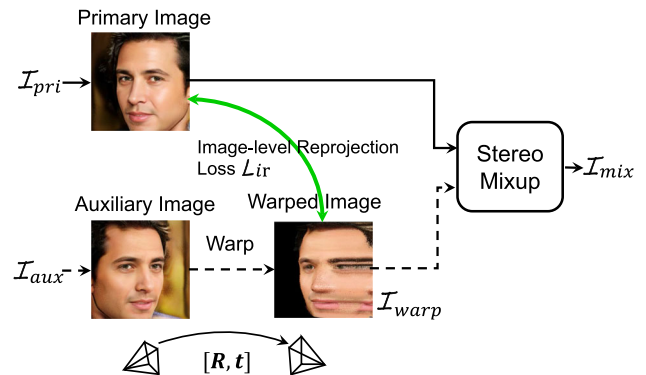$$

Fig. 6 Image-level multi-view joint optimization. We enforce the photometric consistency between the primary image $\mathcal{I}_{pri}$ and the warped image $\mathcal{I}_{warp}$ by minimizing the image-level re-projection loss $\mathcal{L}_{ir}$. Besides, we integrate a stereo mixup module to encourage the warped image to be similar to a real image. The dotted line does not participate in the inference stage

where η is a dynamic number randomly sampled from the range of [0, 1] in every training iteration, and $\mathcal{I}_{mix}$ is the input of the discriminator. It is worth noting that the auxiliary pose is introduced to construct the geometry constraints, and is thus only required in the training process. In the inference stage, the generative model only takes the primary pose ξ_{pri} and latent code z as input to generate the primary image directly.

3.2.2 Feature-Level Multi-view Joint Optimization

In practice, we also encounter one practical challenge: NeRF-based generative models (Schwarz et al., 2020; Niemeyer & Geiger, 2021; Chan et al., 2021) typically struggle to render high-resolution images with fine details due to the huge computational of NeRF (Mildenhall et al., 2020) model. To render images with both fine 2D details and correct 3D shape, we design a two-stage training strategy and extend multi-view optimization to the feature level. We begin training at a low resolution (64^2) in Stage I, and then increase to high resolutions (128^2, 256^2, 512^2) in Stage II (see Fig. 5). In Stage I, we directly render primary and auxiliary images with the color and density output from the generative radiance field network G_s. With the guidance of geometry constraints, we perform image-level multi-view joint optimization to enhance the geometric reasoning ability of the model. In Stage II, to alleviate the computation-intensive problem of rendering high-resolution images, we instead train the model via feature-level multi-view optimization for better visual quality. First, we adopt a hybrid MLP-CNN architecture to disentangle the geometry of the 3D shape from fine details of 2D appearance. Then we generalize volume rendering (Niemeyer & Geiger, 2021) to the feature level by rendering 2D primary feature map $\mathcal{F}_{mix}$ at low resolution (64^2):

$$\mathcal{F}_{pri} = \sum_{i=1}^{N} T_i (1 - exp(-\sigma_i \delta_i)) f_i, \qquad (7)$$

where $f_i \in \mathbb{R}^{256}$ is the feature before the final layer of G_s, and other symbols are defined in Eq. 3. The auxiliary feature map $\mathcal{F}_{aux}$ is rendered in the same way as $\mathcal{F}_{pri}$, and the warped feature map $\mathcal{F}_{warp}$ can be obtained through the warping process. Second, we perform multi-view feature-level joint optimization on low-resolution feature maps (64^2). To enforce the geometry consistency in the feature space, we take the implicit diversified Markov Random Fields (MRF) loss (Wang et al., 2018) as the feature-level re-projection loss:

$$\mathcal{L}_{fr} = L_{mrf}(\mathcal{F}_{pri}, \mathcal{F}_{warp}), \qquad (8)$$

which can encourage the model to capture high-frequency geometry details (Feng et al., 2021). Then the stereo mixup mechanism is also applied to the 2D feature maps: $\mathcal{F}_{mix} = \eta \mathcal{F}_{pri} + (1 - \eta)\mathcal{F}_{warp}$. Third, we increase the resolution with a style-based 2D decoder (Karras et al., 2019) G_d, which takes $\mathcal{F}_{mix}$ as input and then upsamples to high-resolution RGB image (see Fig. 7). The 2D decoder G_d is conditioned by the mapping network G_m through adaptive instance normalization (AdaIN) (Huang & Belongie, 2017; Dumoulin et al., 2020; Karras et al., 2019). As training progresses, we
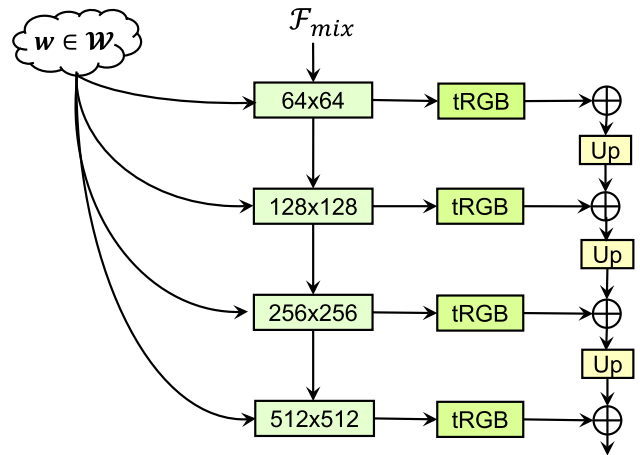

Fig. 7 Progressive 2D decoder G_d. During training, the decoder takes the stereo mixup feature $\mathcal{F}_{mix}$ (produced by $\mathcal{F}_{pri}$ and $\mathcal{F}_{warp}$) as input at low resolution (64^2). Then the intermediate latent w conditions the decoder at each layer. Here tRGB denotes the 1x1 convolutions which convert the high-dimensional features to RGB images, and Up denotes the bilinear upsampling operation

adopt the progressive growing strategy to grow the generator for higher resolution (Karras et al., 2018). When new layers are added to G_d, we use skip connections to fade the inserted layers in smoothly to stabilize and speed up the training process (Karras et al., 2018, 2020).

3.3 MVCGAN+: Towards Multi-Object Generation

While remarkable results have been achieved on 3D-aware image generation, existing methods (Schwarz et al., 2020; Chan et al., 2021; Deng et al., 2022b; Gu et al., 2022; Xu et al., 2021; Pan et al., 2021; Chan et al., 2022) mostly focus on the scene with a single object in the center, and do not work well on multi-object scenes. At present, only GIRAFFE (Niemeyer & Geiger, 2021) considers the compositional properties of scenes and allows for multi-object image generation. However, GIRAFFE (Niemeyer & Geiger, 2021) learns the compositional generative feature fields in an unsupervised manner, which is infeasible to decompose the scene into individual objects precisely. The lack of appropriate supervision makes GIRAFFE (Niemeyer & Geiger, 2021) can only be verified on simple synthetic data, i.e., CLEVR (Johnson et al., 2017), while more realistic scenes with complex geometry shapes and diverse textures still remain unexplored.

To extend to scenarios with multiple objects and backgrounds, we further propose MVCGAN+, a two-branch framework with extra supervision (see Fig. 8 for an overview). We formulate the multi-object scene generation as a "decompose and compose" process. During training, MVCGAN+ learns the whole scene via a top-down decomposition man-
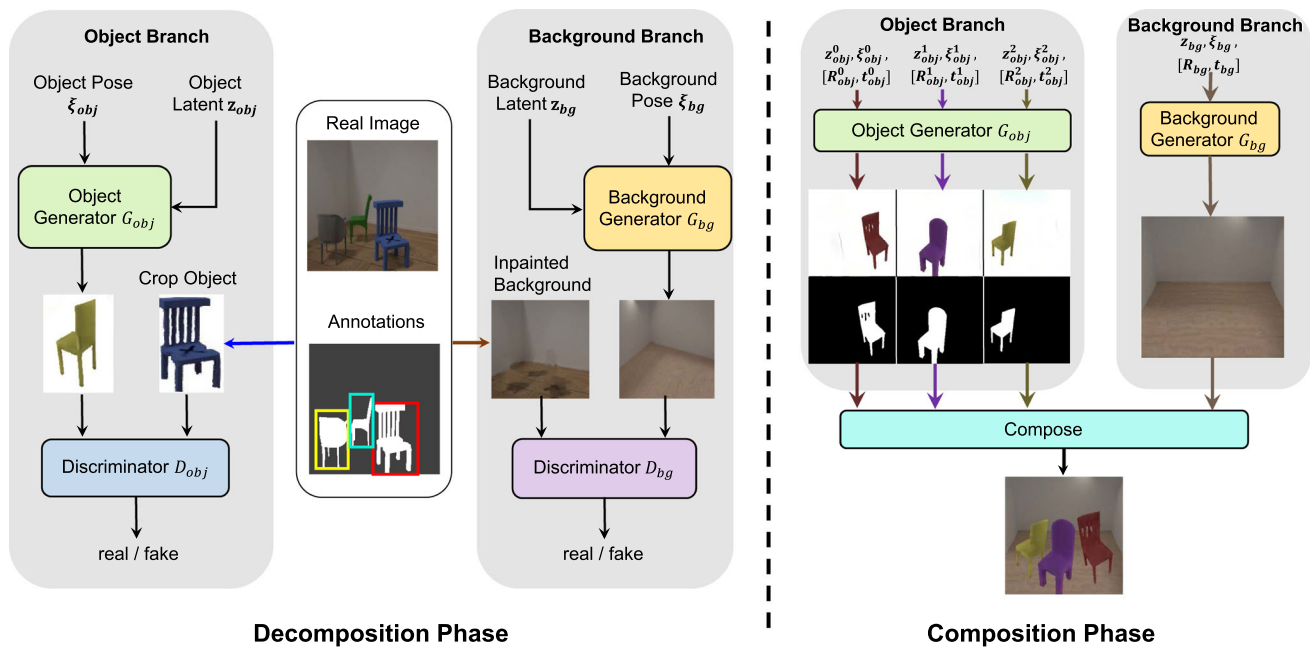
Fig. 8 An overview of MVCGAN+. I. Decomposition Phase. We adopt a "top-down" strategy to train the object branch and the background branch. Specifically, we decompose the real images into foreground objects and backgrounds via masks and bounding boxes. Then we impose multi-view constraints to optimize the object generator G_{obj} and the background generator G_{bg} individually. Two discriminators, i.e., D_{obj} and D_{bg}, are employed to perform adversarial training on generated images and real images. II. Composition Phase. We deploy a reverse "bottom-up" manner for rendering. We first generate foreground objects images and background images with the object branch and the background branch respectively. Then the whole image can be composed with object masks and occlusion relations

ner. Specifically, we incorporate easily-accessed 2D annotations, i.e., object bounding boxes and instance masks, into training to disentangle objects and backgrounds. MVC-GAN+ contains one object branch with G_{obj} and D_{obj}, one background branch with G_{bg} and D_{bg}. For the object branch, we randomly select a single object from the whole scene and crop the corresponding patch with the masked backgrounds as the real object image (see Fig. 8). We encourage the object generator G_{obj} to model the foreground object while leave the background region with empty space. One problem remains that the content of unbounded and occluded scenes, e.g., masked backgrounds, can locate at any distance of the ray. Due to the inherent ambiguity of 2D-to-3D correspondence, the object generator can generate arbitrary geometry outside the target object regions. Consequently, there may exist some semi-transparent materials floating in the space and cause cloudy and foggy artifacts when viewed from another angle.

Therefore, we add the sum of the color weights along the ray on the accumulated color to suppress the low-density areas:

$$\hat{\mathbf{c}}(\mathbf{r}) = \mathbf{c}(\mathbf{r}) + \left(1 - \sum_{i=1}^{N} T_i (1 - exp(-\sigma_i \delta_i))\right) * c_{white},$$
$$(9)$$

where $\mathbf{c}(\mathbf{r})$ is the accumulated color of ray $\mathbf{r}$ by volume rendering, $c_{white} = 1$ is the color of the white background (the value of white color equals to 1 in the normalized color space), $\sum_{i=1}^{N} T_i (1 - exp(-\sigma_i \delta_i))$ is the sum of weights of sampled color along the ray r (see more details in Eq. 3 and Eq. 5 of the original NeRF paper (Mildenhall et al., 2020)), and other symbols are defined in Eq. 3.

For the background branch, we follow NeRF++ (Zhang et al., 2020) to use an additional network G_{bg} to model the complex backgrounds. As shown in Fig. 8, we remove all the foreground objects and fill the holes with the image inpainting methods (Telea, 2004). Considering the layout and geometries of the background environment are relatively simple, we can easily inpaint the occluded areas by searching the patches with similar textures from surrounding regions. In this way, MVCGAN+ models the objects and backgrounds individually by leveraging the information of object bounding boxes and instance masks. The disentanglement of the objects and backgrounds allow us to impose the multi-view geometry constraints on the object and the background branch separately.

In the composition phase, to compose the generated objects and backgrounds into a coherent scene, we first perform object arrangements and then reason about the geometry

Table 1 Quantitative comparison

	CELEBA-HQ		FFHQ		AFHQv2	M-Plants	M-Food	Room-chairs
	256^2	512^2	256^2	512^2	256^2	256^2	256^2	256^2
GRAF (Schwarz et al., 2020)	47.5	57.7	67.2	71.2	75.8	129.3	138.0	Fail
pi-GAN (Chan et al., 2021)	39.7	41.8	38.1	39.9	42.0	76.7	85.3	Fail
GOF (Xu et al., 2021)	51.6	57.8	60.9	64.7	49.6	83.6	87.2	Fail
ShadeGAN (Pan et al., 2021)	52.6	53.1	54.5	58.3	52.4	84.9	93.4	Fail
GIRAFFE (Niemeyer & Geiger, 2021)	36.0	36.2	34.6	37.7	29.2	52.2	84.3	162.9
GRAM (Deng et al., 2022b)	23.6	OOM	29.8	OOM	18.5	188.6	178.9	Fail
Ours	**11.8**	**12.9**	**13.7**	**13.4**	**17.1**	**32.9**	**30.2**	**69.7**

Quantitative comparisons with best results are given in bold
We calculate FID between 20,000 generated and real images. "OOM" represents the out-of-memory error, and "Fail" denotes the model fail to converge

relations between foreground objects and the backgrounds. For the object placement, we follow GIRAFFE (Niemeyer & Geiger, 2021) to transform the coordinate of the object-centric space to the scene space with the rotation matrix R_{obj} and the translation vector t_{obj}:

$$k(\mathbf{x}) = R_{obj}\mathbf{x} + t_{obj}, \tag{10}$$

where $k(\mathbf{x})$ is the transformed coordinate, t_{obj} is the object location in the scene space. We generate the holistic image by performing alpha composition:

$$\mathcal{I}_{final} = \mathcal{I}_{fg} \cdot \mathcal{M} + (1 - \mathcal{M}) \cdot \mathcal{I}_{bg}, \tag{11}$$

where $\mathcal{I}_{fg}$ is the rendered foreground object image, and $\mathcal{I}_{bg}$ is the rendered background image. The foreground object mask $\mathcal{M} = \sum_{i=1}^{N} T_i(1 - exp(-\sigma_i\delta_i))$ is generated by G_{obj} according to the accumulated density. For the overlapping areas between objects, we reason about the occlusion relations by combing 3D spatial locations of objects and depth values. Specifically, for every pixel in the render image, the object closest to the camera location will occlude other objects as well as the backgrounds.

4 Experiments

4.1 Datasets

We conduct experiments on both single-object and multi-object datasets.

Single-object Datasets. For the single-object datasets, we report results on five high-resolution image datasets: CELEBA-HQ (Karras et al., 2018), FFHQ (Karras et al., 2019), AFHQv2 (Choi et al., 2020), M-Plants (Skorokhodov et al., 2022), and M-Food (Skorokhodov et al., 2022). CELEBA-HQ (Karras et al., 2018) consists of 30,000 high-quality images of human face. Flickr-Faces-HQ (FFHQ) (Karras et al., 2019) is a widely-used human face dataset that contains 70,000 high-quality images. Animal Faces-HQ (AFHQv2) (Choi et al., 2020) contains 15,000 high-quality animal face images. Here we choose the cat face images in the AFHQv2 (Choi et al., 2020) dataset to conduct experiments. Megascans Plants (M-Plants) (Skorokhodov et al., 2022) dataset consists of 141,824 plant images, while Megascans Food (M-Food) (Skorokhodov et al., 2022) contains 25,472 food images.

Multi-object Datasets.

For multi-object scenes, existing datasets, e.g., CLEVR (Johnson et al., 2017), multi-dSprites (Matthey et al., 2017), Object Room (Burgess & Kim, 2018), Tetrominoes (Kabra et al., 2019), and CATER (Girdhar & Ramanan, 2019), typically contain objects with the simplest geometric shapes and plain backgrounds. Take a representative dataset CLEVR (Johnson et al., 2017) as an example, the scene contains 3 kinds of objects, i.e., cube, sphere, and cylinder, all of which are geometric primitives that have standard and symmetrical geometries. In this paper, we conduct experiments on a more complex and realistic dataset Room-chair (Yu et al., 2022b), which contains indoor scenes with chairs, walls, and floors. Specifically, we adopt the script (Yu et al., 2022b) to render 32,000 images at a resolution of 256^2. To render chairs with diverse shapes, we choose 649 chair models from ShapeNet (Chang et al., 2015) library. For the backgrounds, we use 50 types of floors with different textures and materials, e.g., wooden floors. Each image contains a random number of chairs with a maximum number of 4. Besides, we also render the instance masks and obtain the object bounding boxes as the annotations. It is worth noting that the geometry of the chair is much more complex than other objects in ShapeNet (Chang et al., 2015) like cars and bowls, because chairs have many thin and fine structures such as backrests and legs. To our knowledge, Room-chair is the most challenging multi-object dataset we can find.
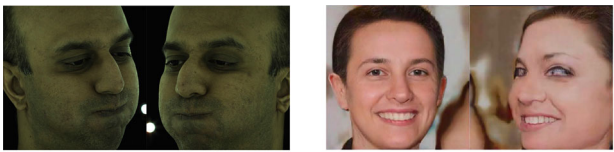
| images with ground-truth multi-view (Face-IDS=0.87) | images with corrupted multi-view artifacts (Face-IDS=0.32) |

Fig. 9 The face identity preservation score (Face-IDS) of images

4.2 Training Details

We use a progressive growing convolutional discriminator D_ϕ to compare the fake image produced by generator G_θ and real image $\mathcal{I}$ sampled from the training data with distribution $p_\mathcal{D}$. For single-object generation, we train MVC-GAN using a non-saturating GAN objective with R_1 gradient penalty (Mescheder et al., 2018) and the proposed geometry-constrained objective L_{re} as the total loss:

$$
\begin{aligned}
\mathcal{V}(\theta, \phi) = & \mathbf{E}_{z \sim \mathcal{Z}, \xi_{pri} \sim p_\xi, \xi_{aux} \sim p_\xi} [f(D_\phi(G_\theta(z, \xi_{pri}, \xi_{aux})))] \\
& + \mathbf{E}_{\mathcal{I} \sim p_\mathcal{D}} [f(-D_\phi(\mathcal{I})) - \lambda ||\nabla D_\phi(\mathcal{I})||^2] + \mathcal{L}_{re},
\end{aligned}
\tag{12}
$$

where $f(t) = -log(1 + exp(-t))$, $\mathcal{L}_{re} = \mathcal{L}_{ir}$ for Stage I (see Eq. 5), $\mathcal{L}_{re} = \mathcal{L}_{fr}$ for Stage II (see Eq. 8), and $\lambda = 10$. We employ Adam optimizer (Kingma & Ba, 2015) with $\beta_1 = 0$, $\beta_2 = 0.9$, and the batch size of 56 for optimization. The initial learning rate is set to 6.0×10^{-5} for the generator and 2.0×10^{-4} for the discriminator, and decay over training to 1.5×10^{-5} and 5.0×5^{-5} respectively.

For the multi-object generation, we train the object generator G_{obj} and the background generator G_{bg} using the same Adam optimizer, the learning rate, and the batch size as the single-object generation. The main difference between the multi-object and the single-object generation is that we sample camera pose from different distributions due to the different scenes of training datasets (please refer to Sect. 1 of Appendix for the specific camera pose distribution of each dataset).

4.3 Comparison with SOTA

For quantitative comparison, we report Frechet Inception Distance (FID) (Heusel et al., 2017) to evaluate image quality. We compare our approach against five state-of-the-art 3D-aware image synthesis methods: GRAF (Schwarz et al., 2020), pi-GAN (Chan et al., 2021), GOF (Xu et al., 2021), ShadeGAN (Pan et al., 2021), GIRAFFE (Niemeyer & Geiger, 2021), and GRAM (Deng et al., 2022b). As shown in Table 1, our method consistently outperforms other methods (Schwarz et al., 2020; Niemeyer & Geiger, 2021; Chan et al., 2021; Xu et al., 2021; Pan et al., 2021) on both

single-object and multi-object datasets (Karras et al., 2018, 2019; Choi et al., 2020; Skorokhodov et al., 2022) by a large margin. Especially, on the Room-Chair dataset (Yu et al., 2022b), we observe most methods cannot handle the multi-object scenarios and fail to learn an appropriate generative model for the scene. In contrast, the extension MVCGAN+ can effectively render the compositional scenes with disentangled objects and backgrounds, outperforming GIRAFFE by a clear margin. To further demonstrate the effectiveness of the proposed method, we also visualize the generated images on single-object and multi-object datasets for qualitative comparison. As illustrated in Figs. 10 and 11, we render images from a wide range of viewpoints. On single-object datasets, we observe that GRAF (Schwarz et al., 2020), GIRAFFE (Niemeyer & Geiger, 2021) and pi-GAN (Chan et al., 2021) either fail to synthesize reasonable results under large view variations or have obvious multi-view inconsistent artifacts. For multi-object scenarios, we note that GIRAFFE (Niemeyer & Geiger, 2021) suffers from collapsed results when the viewpoint changes. By comparison, our method achieves the best performance both in visual quality and multi-view consistency. Please refer to the appendix and the supplementary material[1] for more visualization results.

4.4 Ablation Studies

Image-level and Feature-level Optimization. We conduct ablation studies to help understand the individual contributions of image-level and feature-level multi-view joint optimization. From Fig. 12a, we observe that the generated images maintain the multi-view consistency under pose variations (FID = 22.5), indicating that image-level optimization can guide the model to learn a reasonable 3D shape. With feature-level optimization (see Fig. 12b), our approach can further improve the visual quality of generated images (FID = 13.7). As shown in Fig. 12, we note that the images generated by feature-level optimization have more delicate details, such as clear wrinkles, the highlight on the forehead, and the shadow of the cheeks.

Multi-view Consistency. On the human face dataset (Karras et al., 2019), we take inspiration from Lin et al. (2022) to adopt the face identity preservation score (Face-IDS) to evaluate the multi-view consistency of generated images. For the portrait image animation and attribute-editing task (Lin et al., 2022; Wu et al., 2022; Deng et al., 2020), the face identity preservation score (Face-IDS) can reflect how well the identity is preserved for the target image compared to the source image. Here we use Face-IDS to evaluate the multi-view consistency by measuring the similarity between different views. We first generate 1000 faces and render each face from two

---

[1] https://youtu.be/k207rGznpEk.

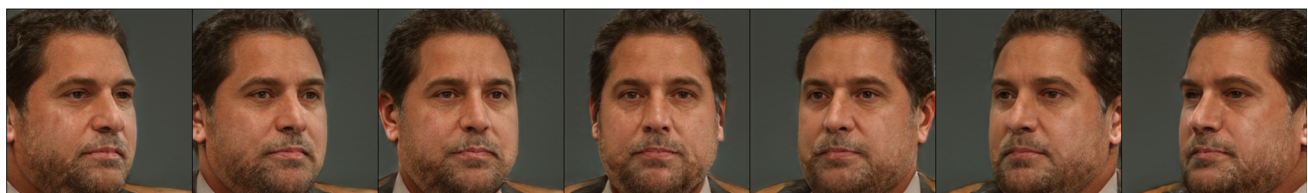(a) Results on FFHQ (Karras et al., 2019).
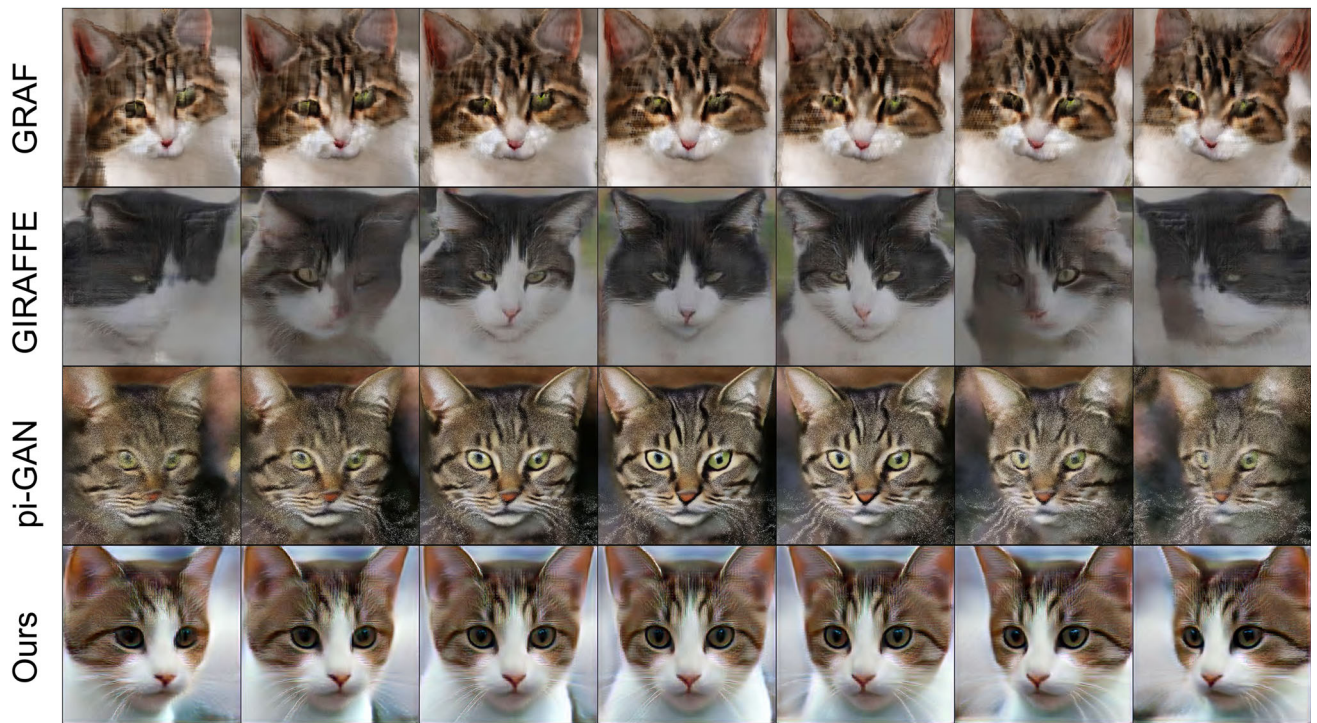
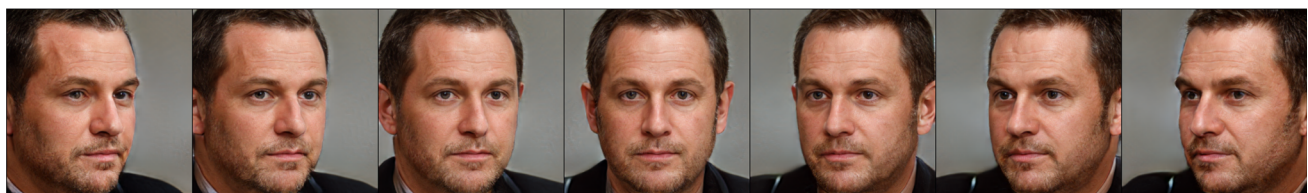(b) Results on AFHQv2 (Choi et al., 2020).

Fig. 10 Qualitative comparison at 512^2 resolution on single-object datasets

Fig. 11 Qualitative comparison at 256^2 resolution under the multi-object setting on Room-Chair (Yu et al., 2022b). We render the scenes from different camera view points

(a) With image-level multi-view join optimization (FID=22.5).

(b) With feature-level multi-view join optimization (FID=13.7).

Fig. 12 Ablation study on FFHQ (Karras et al., 2019) at 256^2 resolution
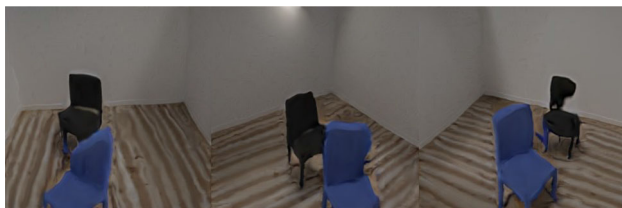
Fig. 13 Without the decomposition phase, the generated images will have poor object qualities and cannot disentangle objects and backgrounds

random camera poses. Then, for each image pair of the same generated face, we calculate the cosine similarity of the predicted embeddings with a pertrained ArcFace network (Deng et al., 2019). The ArcFace similarity score has values between -1 and 1 (greater value means more similar, see Fig. 9 for examples). Finally, we compute the mean score of 1000 faces as the face identity preservation score (Face-IDS).

As shown in Table 2, our method achieves the best face identity preservation score (multi-view consistency). We further conduct experiments to study whether increasing the number of auxiliary poses can improve the multi-view consistency or not. From Table 2, we observe that using more auxiliary poses leads to degenerated performance: the face identity preservation score (Face-IDS) decreases to 0.58 and 0.51 for 2 and 3 auxiliary poses respectively. We suspect the performance drop is caused by two reasons. First, since both the primary and auxiliary poses are randomly sampled from the camera pose distribution, sampling more poses cannot bring performance gain. Second, increasing the number of auxiliary poses brings much more GPU memory consumption, because the model has to perform volume rendering many times for one iteration. Consequently, we need to reduce the batch size to 8 and adjust the weight of the R_1 gradient penalty (Mescheder et al., 2018) (λ in Eq. 12). However, the decreased batch size affects the training stability and makes the model hard to converge, while the increased R_1
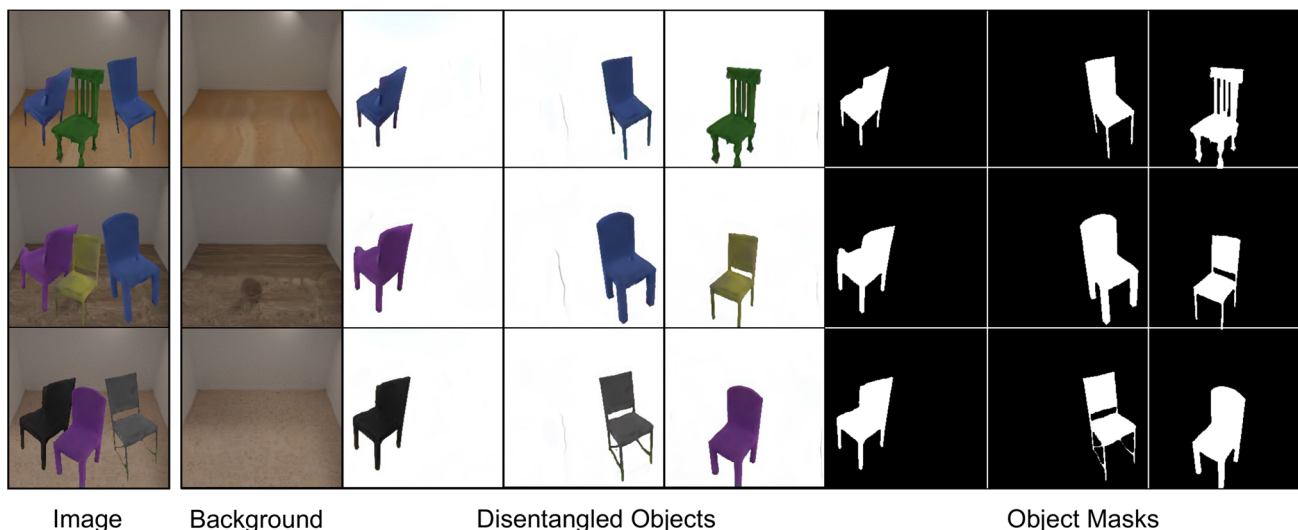
Fig. 14 Scene Decomposition. The generated images can be decomposed into individual objects and backgrounds

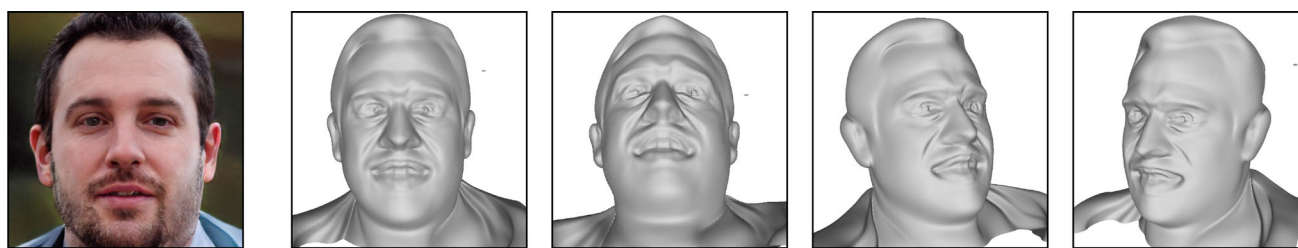| Image | Background | Disentangled Objects | Object Masks |

Fig. 15 Visualization of extracted 3D meshes with single-view 3D reconstruction (Lorensen & Cline, 1987)

Table 2 Quantitative evaluation of multi-view consistency

Method	Face-IDS
GRAF (Schwarz et al., 2020)	0.41
pi-GAN (Chan et al., 2021)	0.50
GOF (Deng et al., 2022b)	0.48
ShadeGAN (Pan et al., 2021)	0.45
GIRAFFE (Niemeyer & Geiger, 2021)	0.55
Ours (1 auxiliary poses)	0.62
Ours (2 auxiliary poses)	0.57
Ours (3 auxiliary poses)	0.51

On FFHQ, we calculate the average face identity preservation score (ID) of generated images at 256^2 resolution

regularization weight can hurt the overall performance (Karras et al., 2021; Mescheder et al., 2018).

Markov Random Fields Loss.

Previous papers (Feng et al., 2021) found ID-MRF loss can better capture high-frequency details than L1 loss in the 3D face reconstruction (Feng et al., 2021) and image reconstruction task (Wang et al., 2018). Therefore, we adopt Implicit Diversified Markov Random Field (ID-MRF) loss (Wang et al., 2018) to enforce the geometry consistency between views. We also conduct experiments to compare the effect of ID-MRF and L1 loss on multi-view consistency. Since there is no ground truth for the generated image, we adopt the face identity preservation score (Face-IDS) as the quantitative metric of multi-view consistency.

When using the vanilla L1 loss, we observe that the model still achieves similar multi-view consistency (face identity preservation score = 0.61) as the ID-MRF loss (face identity preservation score = 0.62). It seems that ID-MRF loss has no obvious advantages over L1 loss. We suspect that the problem is in the quantitative metric of multi-view consistency, because the face identity preservation score may not be able to capture these high-frequency details such as the wrinkles visualized in the Fig. 9 of Feng et al. (2021). As mentioned in the last paragraph (**Multi-view Consistency**), we compute the face identity preservation score (Face-IDS) with the Arcface cosine similarity (Deng et al., 2019). But the extracted embedding may lose the high-frequency and fine-grained details due to the pooling operation of the ArcFace network (Deng et al., 2019). Therefore, the simple L1 loss can obtain a similar face identity preservation score as ID-MRF loss.
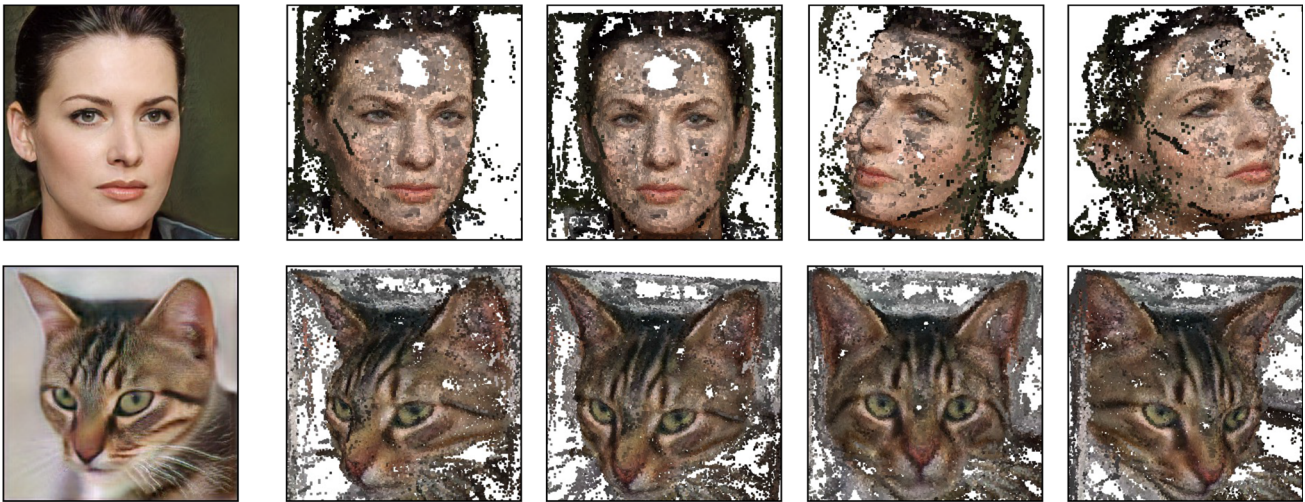
Decompose and Compose.

Fig. 16 Visualization of the COLMAP reconstruction (Schonberger & Frahm, 2016) from synthesized multi-view images
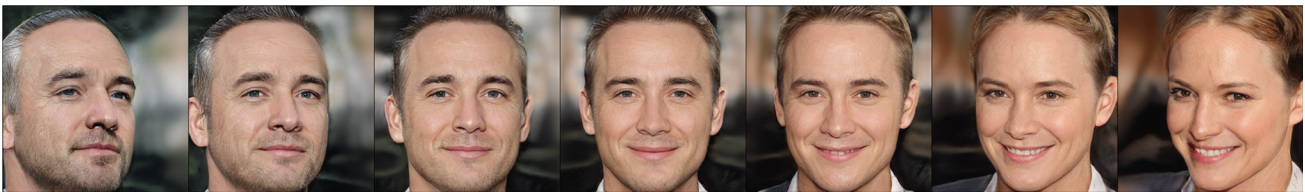

Fig. 17 Style interpolation. We perform linear interpolation simultaneously in both the intermediate latent and camera pose space. We can observe that the transition results are smooth and consistent

The "decompose and compose" paradigm is essential in compositional image generation. If we directly generate multiple objects using the single object method and then compose them into a whole scene without the decomposition phase, the generated image will have poor object qualities and cannot disentangle objects and backgrounds (see the Fig. 13). This problem mainly comes from the discriminator, which plays a critical role in the training process of GANs. If there is no decomposition phase, we need to perform adversarial training with a scene-level discriminator between the rendered scenes and real images. In this case, the model will pay more attention to the global coherence of the whole scene, and neglect the supervision of individual objects. For a single object in the scene, the scene-level discriminator can provide weak learning signals on the object radiance field, because the object region only occupies a small proportion of the whole image. The inadequate training of the single object can lead to degenerated object quality. More importantly, the scene-level discriminator can not disentangle objects and backgrounds, making the background generator easily overfit the whole scene. In contrast, we deploy the decomposition phase to train the object branch and background branch individually. On the one hand, using two discriminators (the object discriminator D_{obj} and the background discriminator D_{bg}) can provide sufficient supervision for objects, leading to a better quality of the generated objects. On the other hand, the disentanglement of objects and backgrounds allows us to control them separately, such as moving and rotating each object or the background.

Scene Decomposition. We also investigate the disentanglement of foregrounds and backgrounds of MVCGAN+. As shown in Fig. 14, our method can decompose foreground objects and backgrounds from the holistic scene. The disentanglement allows us to control each object and the background individually. We can perform scene editing such as adding, moving, deleting, rotating, and changing individual objects or backgrounds. Please refer to the supplementary video for more visualization results.

3D Representation. To better illustrate the learned 3D representation, we visualize the underlying 3D shape from generated images with 3D reconstruction methods (Schonberger & Frahm, 2016; Lorensen & Cline, 1987). For the single-view 3D reconstruction, we adopt the marching cubes algorithm (Lorensen & Cline, 1987) to extract the underlying geometry of the generated image (see Fig. 15 for the visualized 3D meshes). To further demonstrate the multi-view consistency of our method, we also perform multi-view 3D reconstruction (Schonberger & Frahm, 2016) to recover the 3D shape from generated multi-view images. Specifically, we first render images of a single instance from 35

Fig. 18 Style mixing. The source A and B images are generated from input latent codes z_A and z_B. The images in the red box are generated by applying the w_B (the intermediate latent corresponding to z_B) to G_s and w_A (corresponding to z_A) to G_d. The images in the green box are generated by applying the w_A to G_s and w_B to G_d (Color figure online)

views, and then perform dense 3D reconstruction by running COLMAP (Schonberger & Frahm, 2016) with default parameters and no known camera poses. The results in Figs. 15 and 16 validate the correctness of the 3D shape learned by our model.

Style Interpolation. We also conduct style interpolation experiments to investigate the intermediate latent w learned by the mapping network G_m. Given two generated images, we perform linear interpolation both in the intermediate latent space $\mathcal{W}$ and the camera pose space. As illustrated in Fig. 17, the smooth transition of both pose and appearance demonstrates that our model learns semantically meaningful intermediate latent space $\mathcal{W}$.

Shape-detail Disentanglement. Besides, we design a style mixing experiment to study what kinds of representations the generative radiance field G_s and progressive 2D decoder G_d learned respectively. Specifically, we input two latent codes z_A and z_B into the mapping network G_m, and obtain the corresponding intermediate latent w_A, w_B in $\mathcal{W}$ space. Then we can generate style mixing images by applying w_A and w_B to control the different parts of the generator (G_s and G_d). As shown in Fig. 18, we observe that controlling G_s changes the 3D shape (identity and pose) while controlling G_d changes 2D appearance details (colors of skins, hair, and beard). The results verify that the hybrid MLP-CNN architecture can disentangle the geometry of the 3D shape from fine details of the 2D appearance.

5 Conclusion

We present a multi-view consistent generative model (MVC-GAN) for compositional 3D-aware image synthesis. The key idea underpinning the proposed method is to enhance the geometric reasoning ability of the generative model by intro-

ducing geometry constraints. Besides, we adapt MVCGAN to more complex and multi-object scenes. Extensive experiments on single-object and multi-object datasets demonstrate that the proposed method achieves state-of-the-art performance for 3D-aware image synthesis.

Appendix

A Implementation Details

A.1 Network Architectures

Generative Radiance Field. The generative radiance field network G_s is a 8-layer SIREN-based MLP with periodic activation functions (Sitzmann et al., 2020). The dimension of the hidden layers is 256.

Mapping Network. The mapping network G_m is a 4-layer MLP network with leakyReLU as the activation function. The dimension of the hidden layers is 256. We sample the input latent code z from a 256-dimensional standard Gaussian distribution.

Progressive 2D Decoder. The progressive 2D decoder G_d is a fully-convolution neural network, which decreases the feature dimension from 256 (at 64^2) to 32 (at 512^2).

Discriminator. The discriminator D_ϕ is a progressive growing convolutional network, which uses eight layers for 64^2 and fourteen layers for 512^2.

A.2 Datasets

We conduct experiments on both single-object and multi-object high-resolution image datasets: CELEBA-HQ (Karras et al., 2018), FFHQ (Karras et al., 2019), AFHQv2 (Choi et al., 2020), M-Food (Skorokhodov et al., 2022), M-Plants (Skorokhodov et al., 2022), and Room-Chair (Yu et al., 2022b).

Fig. 19 The images are rendered from 35 camera poses at resolution 256^2

CELEBA-HQ. CELEBA-HQ[2] (Karras et al., 2018) consists of 30,000 high-quality images of human face at 1024^2 resolution. During training, we sample the pitch and yaw of the camera pose from a Gaussian distribution with the horizontal standard deviation of 0.3 radians and the vertical standard deviation of 0.155 radians.

FFHQ. Flickr-Faces-HQ (FFHQ)[3] Karras et al. (2019) is a large scale human face dataset which contains 70,000 high-quality images at 1024^2 resolution. The images contain various styles with different ages, ethnicity, and background. Besides, the humans in the images wear different accessories such as earrings, sunglasses, hats, and eyeglasses. In the training stage, we sample the pitch and yaw of the camera pose from a Gaussian distribution with the horizontal standard deviation of 0.3 radians and the vertical standard deviation of 0.155 radians.

AFHQv2. Animal Faces-HQ (AFHQv2)[4] Choi et al. (2020) contains 15,000 high-quality animal face images at 512^2 resolution. The dataset has three categories: cat, dog, and wildlife, with each category providing 5,000 images. Following previous works (Schwarz et al., 2020; Niemeyer & Geiger, 2021; Chan et al., 2021), we conduct experiments on the cat face images to make a fair comparison. During training, we sample the pitch and yaw of the camera pose from a uniform distribution with the horizontal standard deviation of 0.4 radians and the vertical standard deviation of 0.2 radians.

M-Plants. Megascans Plants (M-Plants) dateset[5] (Skorokhodov et al., 2022) consists of 141,824 plant images rendered from 1,108 models at 256^2 resolution. During training, we sample the pitch and yaw of the camera pose from a uniform distribution with the horizontal standard deviation of 2π radians and the vertical standard deviation of π radians.

---

[2] https://github.com/tkarras/progressive_growing_of_gans.

[3] https://github.com/NVlabs/ffhq-dataset.

[4] https://github.com/clovaai/stargan-v2.

[5] https://rethinking-3d-gans.github.io/.

Fig. 20 Images synthesized by MVCGAN on CELEBA-HQ (Karras et al., 2018) at resolution 512^2

M-Food. Megascans Food (M-Food)[6] (Skorokhodov et al., 2022) contains 24,472 food images at 256^2 resolution. There are a variety of of food items in the dataset, such as apples, oranges, mushrooms, and biscuits. In the training process, we sample the pitch and yaw of the camera pose from a uniform distribution with the horizontal standard deviation of 2π radians and the vertical standard deviation of π radians.
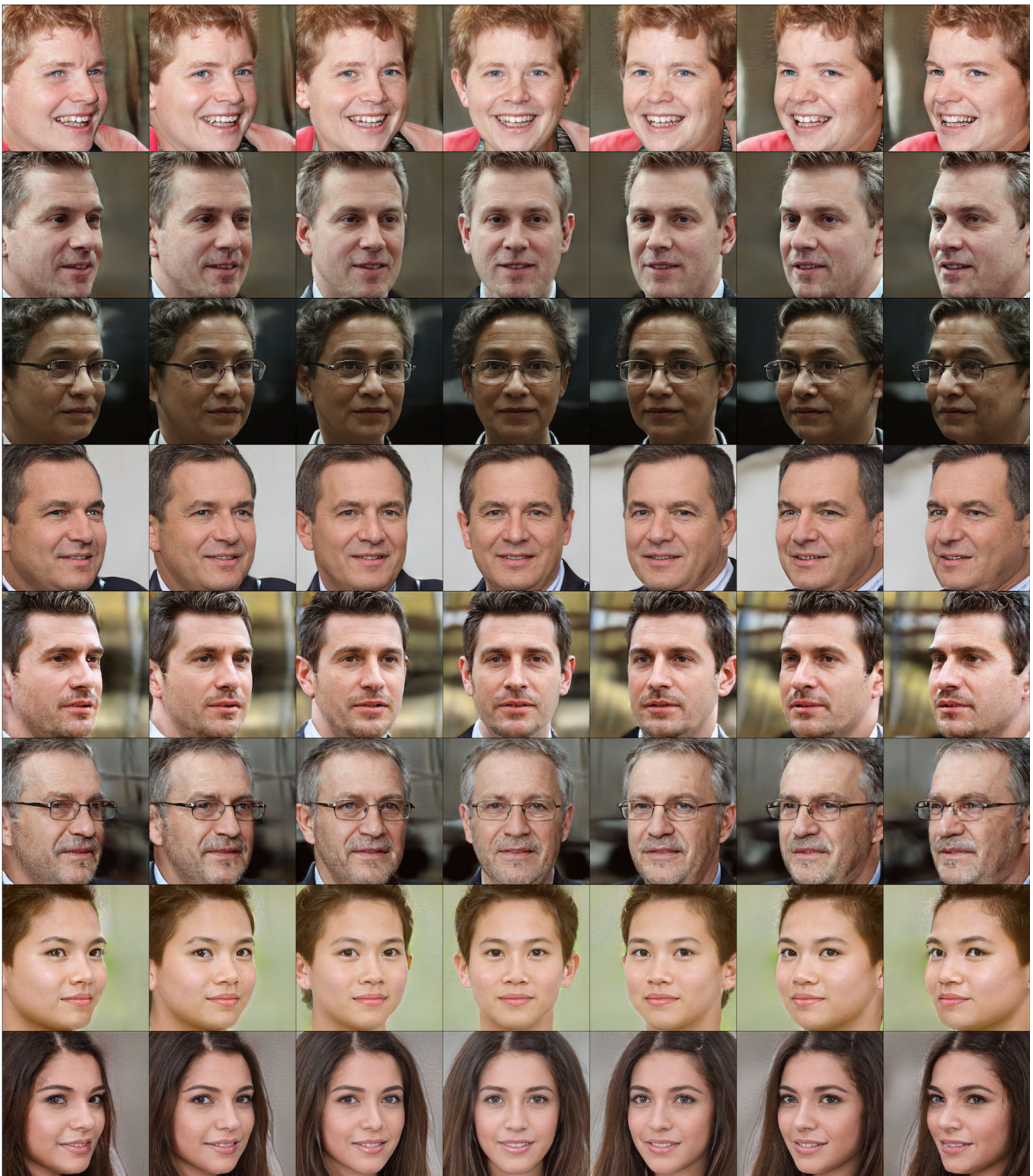
Fig. 21 Images synthesized by MVCGAN on FFHQ (Karras et al., 2019) at resolution 512^2

Room-Chair. Room-Chair (Yu et al., 2022b) is a multi-object indoor scene dataset with random number of chairs and various of backgrounds. We follow the script[7] (Yu et al., 2022b) to

render 32,000 images at 256^2 and the corresponding instance masks. In the training process, we sample the pitch and yaw of the camera pose from a uniform distribution with the horizontal standard deviation of 2π radians and the vertical standard deviation of 0.3 radians.
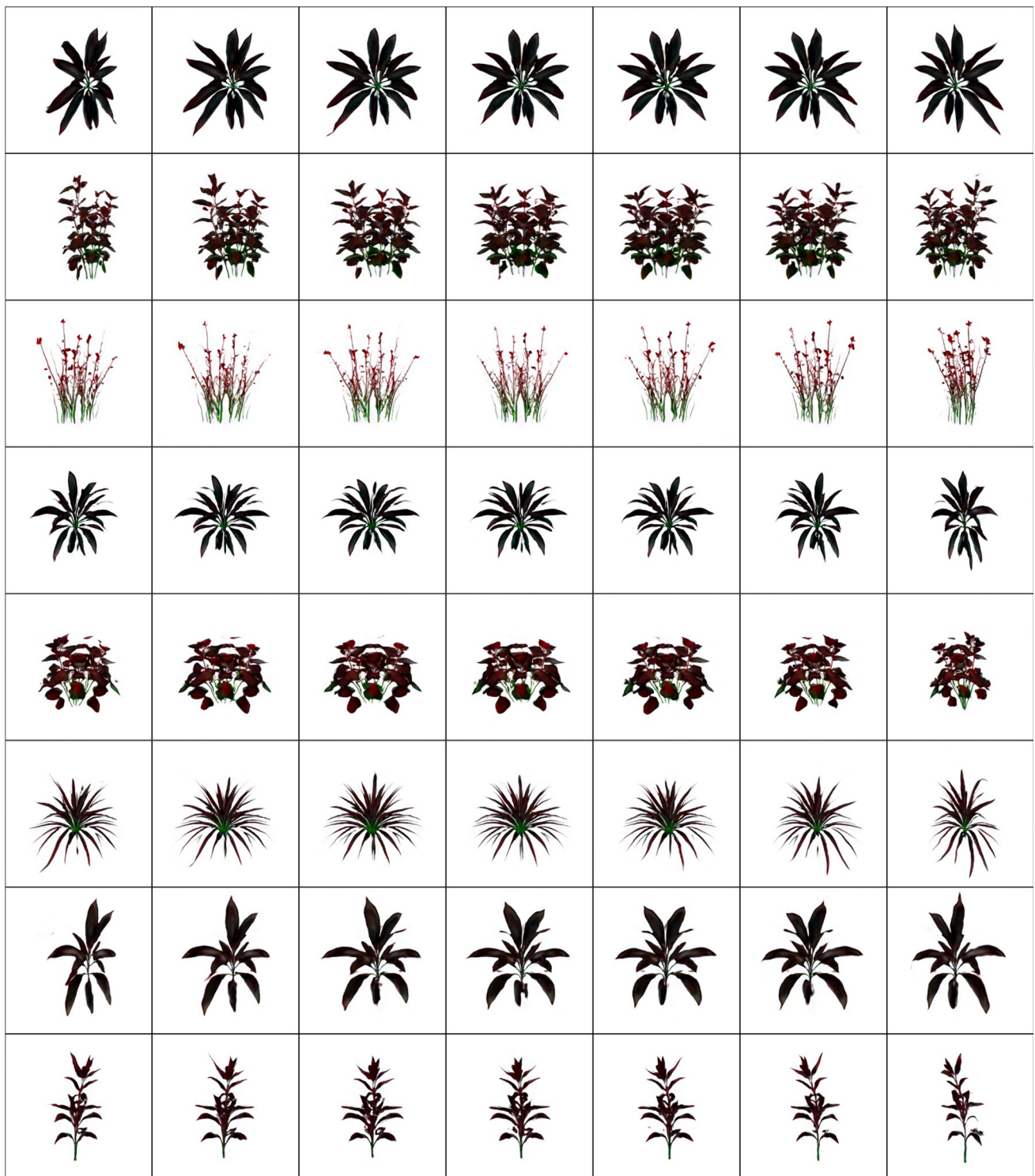
---

[7] https://github.com/KovenYu/uORF.

Fig. 22 Images synthesized by MVCGAN on M-Plants (Skorokhodov et al., 2022) at resolution 256^2

B Additional Results

We provide additional results to show the multi-view consistency and the quality of the generated images.

As shown in Fig. 19, we render images of a single instance from 35 views images. We also provide more generated
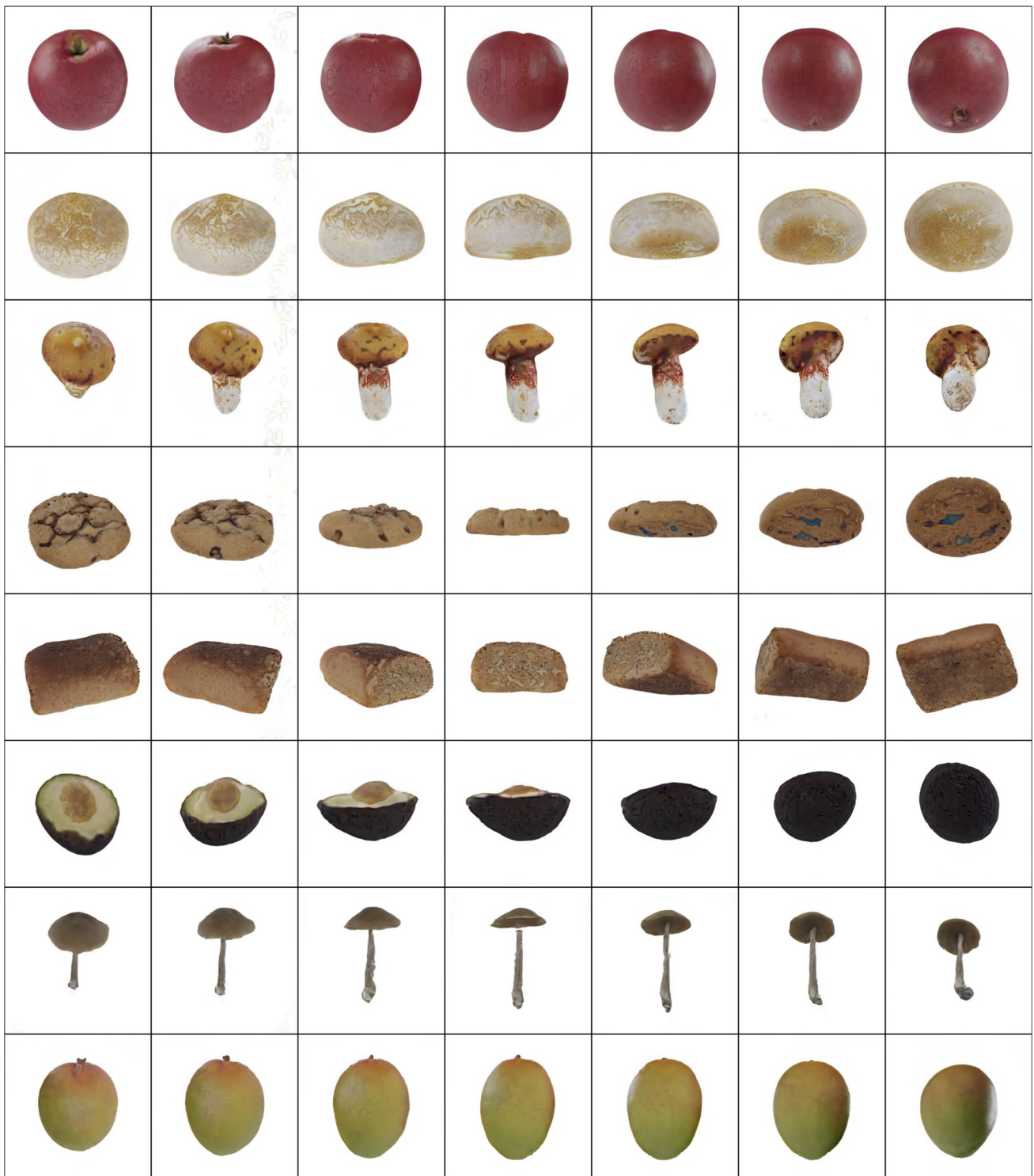
Fig. 23 Images synthesized by MVCGAN on M-Food (Skorokhodov et al., 2022) at resolution 256^2

images in Figs. 20, 21, 22, 23, and 24. Please also refer to the supplementary video[8] for more results.

C Discussion

C.1 Comparison to StyleGAN3

StyleGAN3 (Karras et al., 2021) can also produce multi-view images with random latent walk. Here we compare the proposed method and StyleGAN3 as follows.

---

[8] https://youtu.be/k207rGznpEk.

Fig. 24 Images synthesized by MVCGAN+ on Room-Chair (Yu et al., 2022b) at resolution 256^2

1. The fundamental difference is that our method represents the scene in 3D space, while StyleGAN3 (Karras et al., 2021) operates in the 2D domain. To generate an image, we first query the 3D represention of the scene (neural radiance fields), and then use volume rendering to synthesis image from a specific viewpoint. Every image generated by the proposed model has a underlying 3D represetnion. Therefore, we can extract the underlying geometry of the generated image and export as meshes or pointclouds (see Figs. 15 and 16.

2. Our method is more controllable. Our method explicitly disentangles the camera pose from the latent code, while StyleGAN3 encodes both the camera pose and the identity into the latent code. Therefore, we can generate images from the same identity from different views, or generate different identity from the same viewpoint. Besides, the proposed method also support other camera operations, e.g., rotate, translate, zoom-in, and zoom-out (see the supplementary video). In contrast, the random latent walk process of StyleGAN3 (Karras et al., 2021) is arbitrary and uncontrollable. Since the identity and the camera pose are coupled in the the latent code, changing the latent code can change both the camera pose and the identity. As shown in Video 1a and Video 1b on the project page[9] of StyleGAN3 Karras et al. (2021), we observe the mouth and expression also change in different views.

References

(2021) Dynamic view synthesis from dynamic monocular video. In *ICCV* (pp. 5712–5721).

Alhaija, HA., Mustikovela, S. K., Geiger, A., & Rother, C. (2018). Geometric image synthesis. In *ACCV* (pp. 85–100).

Andrew, A. M. (2001). Multiple view geometry in computer vision. *Kybernetes*.

Anokhin, I., Demochkin, K., Khakhulin, T., Sterkin, G., Lempitsky, V., & Korzhenkov, D. (2021) Image generators with conditionally-independent pixel synthesis. In *CVPR* (pp. 14278–14287).

Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale gan training for high fidelity natural image synthesis. In *ICLR*.

Burgess, C., & Kim, H. (2018). 3d shapes dataset.

Chan, E. R., Lin, C. Z., Chan, M. A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L. J., Tremblay, J., Khamis. S, et al. (2022). Efficient geometry-aware 3d generative adversarial networks. In *CVPR* (pp. 16123–16133).

Chan, E. R., Monteiro, M., Kellnhofer, P., Wu, J., & Wetzstein, G. (2021). pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR* (pp. 5799–5809).

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H. et al. (2015). Shapenet: An information-rich 3d model repository. arXiv:1512.03012

Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., & Su, H. (2021). Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *ICCV* (pp. 14124–14133).

Chen, S. E., & Williams, L. (1993). View interpolation for image synthesis. In *Conference on Computer graphics and interactive techniques*.

Chibane, J., Bansal, A., Lazova, V., & Pons-Moll, G. (2021). Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In *CVPR* (pp. 7911–7920).

Choi, Y., Choi, M., Kim, M., Ha, JW., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR* (pp. 8789–8797).

Choi, Y., Uh, Y., Yoo, J., & Ha, J. W. (2020). Stargan v2: Diverse image synthesis for multiple domains. In *CVPR* (pp. 8188–8197).

Collins, R. T. (1996). A space-sweep approach to true multi-image matching. In *CVPR* (pp. 358–363).

Debevec, P. E., Taylor, C. J., & Malik, J. (1996). Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Conference on Computer graphics and interactive techniques* (pp. 11–20).

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *CVPR* (pp. 4690–4699).

Deng, K., Liua, A., Zhu, J. Y., & Ramanan, D. (2022a). Depth-supervised nerf: Fewer views and faster training for free. In *CVPR* (pp. 12882–12891).

Deng, Y., Yang, J., Chen, D., Wen, F., & Tong, X. (2020). Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *CVPR* (pp. 5154–5163).

Deng, Y., Yang, J., Xiang, J., & Tong, X. (2022b). Gram: Generative radiance manifolds for 3d-aware image generation. In *CVPR* (pp. 10673–10683).

DeVries, T., Bautista, M. A., Srivastava, N., Taylor, G. W., & Susskind, J. M. (2021). Unconstrained scene generation with locally conditioned radiance fields. In *ICCV* (pp. 14304–14313).

Dumoulin, V., Shlens, J., & Kudlur, M. (2020). *A learned representation for artistic style*.

Feng, Y., Feng, H., Black, M. J., & Bolkart, T. (2021). Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG), 40*(4), 1–13.

Garbin, S. J., Kowalski, M., Johnson, M., Shotton, J., & Valentin, J. (2021). Fastnerf: High-fidelity neural rendering at 200fps. In *ICCV* (pp. 14346–14355).

Girdhar, R., & Ramanan, D. (2019). Cater: A diagnostic dataset for compositional actions and temporal reasoning. In *ICLR*.

Godard, C., Mac Aodha, O., Firman, M., & Brostow, G. J. (2019). Digging into self-supervised monocular depth estimation. In *ICCV*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial nets. *Communications of the ACM, 63*(11), 139–144.

Gu, J., Liu, L., Wang, P., & Theobalt, C. (2022). Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. In *ICLR*.

Henderson, P., Tsiminaki, V., & Lampert, C. H. (2020). Leveraging 2d data to learn textured 3d mesh generation. In *CVPR* (pp. 7498–7507).

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems, 30*

Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV* (pp. 1501–1510).

Jeong, Y., Ahn, S., Choy, C., Anandkumar, A., Cho, M., & Park, J. (2021) Self-calibrating neural radiance fields. In *ICCV* (pp. 5846–5854).

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C.,& Girshick, R. (2017) Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR* (pp. 2901–2910).

Kabra, R., Burgess, C., Matthey, L., Kaufman, R. L., Greff, K., Reynolds, M., & Lerchner, A. (2019). Multi-object datasets.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of gans for improved quality, stability, and variation. In *ICLR*.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T. (2021). Alias-free generative adversarial networks. In *NeurIPS*.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *CVPR* (pp. 4401–4410).

---

[9] https://nvlabs.github.io/stylegan3/.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *CVPR* (pp. 8110–8119).

Kingma, D. P., & Ba, J. (2015) Adam: A method for stochastic optimization. In *ICLR*.

Li, T., Slavcheva, M., Zollhoefer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., & Lv, Z. (2021a). Neural 3d video synthesis. arXiv:2103.02597

Li, Z., Niklaus, S., Snavely, N., & Wang, O. (2021b). Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6498–6508).

Liao, Y., Schwarz, K., Mescheder, L., & Geiger, A. (2020). Towards unsupervised learning of generative models for 3d controllable image synthesis. In *CVPR* (pp. 5871–5880).

Lin, C. H., Ma, W. C., Torralba, A., Lucey, S. (2021). Barf: Bundle-adjusting neural radiance fields. In *ICCV* (pp. 5741–5751).

Lin, C. Z., Lindell, D. B., Chan, E. R., & Wetzstein, G. (2022). 3d gan inversion for controllable portrait image animation. arXiv:2203.13441

Lindell, D. B., Martel, J. N., & Wetzstein, G. (2021). Autoint: Automatic integration for fast neural volume rendering. In *CVPR* (pp. 14556–14565).

Liu, Y., Peng, S., Liu, L., Wang, Q., Wang, P., Theobalt, C., Zhou, X., & Wang, W. (2022). Neural rays for occlusion-aware image-based rendering. In *CVPR* (pp. 7824–7833).

Lorensen, W. E., & Cline, H. E. (1987). Marching cubes: A high resolution 3d surface construction algorithm. *ACM Siggraph Computer Graphics, 21*(4), 163–169.

Lyu, X., Liu, L., Wang, M., Kong, X., Liu, L., Liu, Y., Chen, X., & Yuan, Y. (2021). Hr-depth: High resolution self-supervised monocular depth estimation. In *AAAI*.

Matthey, L., Higgins, I., Hassabis, D., & Lerchner, A. (2017). dsprites: Disentanglement testing sprites dataset.

Meng, Q., Chen, A., Luo, H., Wu, M., Su, H., Xu, L., He, X., & Yu, J. (2021). Gnerf: Gan-based neural radiance field without posed camera. In *ICCV* (pp. 6351–6361).

Mescheder, L., Geiger, A., & Nowozin, S. (2018). Which training methods for gans do actually converge? In *ICML* (pp. 3481–3490).

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV* (pp. 405–421). Springer.

Mur-Artal, R., Montiel, J. M. M., & Tardos, J. D. (2015). Orb-slam: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics, 31*(5), 1147–1163.

Nguyen-Phuoc, T., Li, C., Theis, L., Richardt, C., & Yang, Y. L. (2019). Hologan: Unsupervised learning of 3d representations from natural images. In *CVPR* (pp. 7588–7597).

Nguyen-Phuoc, T. H., Richardt, C., Mai, L., Yang, Y., & Mitra, N. (2020). Blockgan: Learning 3d object-aware scene representations from unlabelled images. *Advances in Neural Information Processing Systems, 33*, 6767–6778.

Niemeyer, M., & Geiger, A. (2021). Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR* (pp. 11453–11464).

Or-El R., Luo, X., Shan, M., Shechtman, E., Park, J. J., & Kemelmacher-Shlizerman, I. (2022). Stylesdf: High-resolution 3d-consistent image and geometry generation. In *CVPR* (pp. 13503–13513).

Pan, X., Xu, X., Loy, C. C., Theobalt, C., & Dai, B. (2021). A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. In *NeurIPS*.

Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., & Zhou, X. (2021). Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR* (pp. 9054–9063).

Pillai, S., Ambruş, R., & Gaidon, A. (2019). Superdepth: Self-supervised, super-resolved monocular depth estimation. In *ICRA* (pp. 9250–9256).

Rebain, D., Jiang, W., Yazdani, S., Li, K., Yi, K. M., & Tagliasacchi, A. (2021). Derf: Decomposed radiance fields. In *CVPR* (pp. 14153–14161).

Reiser, C., Peng, S., Liao, Y., & Geiger, A. (2021). Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 14335–14345).

Schonberger, J. L., & Frahm, J. M. (2016). Structure-from-motion revisited. In *CVPR* (pp. 4104–4113).

Schwarz, K., Liao, Y., Niemeyer, M., & Geiger, A. (2020). Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems, 33*, 20154–20166.

Schwarz, K., Sauer, A., Niemeyer, M., Liao, Y., & Geiger, A. (2022). Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. In *NeurIPS*.

Seitz, S. M., & Dyer, C. R. (1996). View morphing. In *Conference on computer graphics and interactive techniques* (pp. 21–30).

Sitzmann, V., Martel, J., Bergman, A., Lindell, D., & Wetzstein, G. (2020). Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems, 33*, 7462–7473.

Skorokhodov, I., Tulyakov, S., Wang, Y., & Wonka, P. (2022). Epigraf: Rethinking training of 3d gans. In *NeurIPS*.

Szeliski, R., & Golland, P. (1999). Stereo matching with transparency and matting. *International Journal of Computer Vision, 32*(1), 45–61.

Telea, A. (2004). An image inpainting technique based on the fast marching method. *Journal of Graphics Tools, 9*(1), 23–34.

Trevithick, A., & Yang, B. (2021). Grf: Learning a general radiance field for 3d scene representation and rendering. In *ICCV* (pp. 15182–15192).

Wang, Y., Tao, X., Qi, X., Shen, X., & Jia, J. (2018). Image inpainting via generative multi-column convolutional neural networks. *Advances in neural information processing systems, 31*.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600–612.

Wang, Z., Wu, S., Xie, W., Chen, M., & Prisacariu, V. A. (2021). NeRF: Neural radiance fields without known camera parameters. arXiv:2102.07064

Wei, Y., Liu, S., Rao, Y., Zhao, W., Lu, J., & Zhou, J. (2021). Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *ICCV* (pp. 5610–5619).

Wu, Y., Deng, Y., Yang, J., Wei, F., Chen, Q., & Tong, X. (2022). Anifacegan: Animatable 3d-aware face image generation for video avatars. In *NeurIPS*.

Xian, W., Huang, J. B., Kopf, J., & Kim, C. (2021). Space-time neural irradiance fields for free-viewpoint video. In *CVPR* (pp. 9421–9431).

Xiang, J., Yang, J., Deng, Y., & Tong, X. (2022). Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. arXiv:2206.07255

Xu, X., Pan, X., Lin, D., & Dai, B. (2021). Generative occupancy fields for 3d surface-aware image synthesis. *Advances in Neural Information Processing Systems, 34*, 20683–20695.

Xu, Y., Peng, S., Yang, C., Shen, Y., & Zhou, B. (2022). 3d-aware image synthesis via learning structural and textural representations. In *CVPR* (pp. 18430–18439).

Yao, Y., Luo, Z., Li, S., Fang, T., & Quan, L. (2018). Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV* (pp. 767–783).

Yen-Chen, L., Florence, P., Barron, J. T., Rodriguez, A., Isola, P., & Lin, T. Y. (2021). inerf: Inverting neural radiance fields for pose estimation. In *IROS* (pp. 1323–1330).

Yu, A., Fridovich-Keil, S., Tancik, M., Chen, Q., Recht, B., & Kanazawa, A. (2022a). Plenoxels: Radiance fields without neural networks. In *CVPR* (pp. 5501–5510).

Yu, A., Li, R., Tancik, M., Li, H., Ng, R., Kanazawa, A. (2021a). Plenoctrees for real-time rendering of neural radiance fields. In *ICCV* (pp. 5752–5761).

Yu, A., Ye, V., Tancik, M., & Kanazawa, A. (2021b). pixelnerf: Neural radiance fields from one or few images. In *CVPR* (pp. 4578–4587).

Yu, H. X., Guibas, L. J., Wu, J. (2022b). Unsupervised discovery of object radiance fields.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). mixup: Beyond empirical risk minimization. In *ICLR*.

Zhang, K., Riegler, G., Snavely, N., & Koltun, V. (2020). Nerf++: Analyzing and improving neural radiance fields. arXiv:2010.07492

Zhang, X., Zheng, Z., Gao, D., Zhang, B., Pan, P., & Yang, Y. (2022). Multi-view consistent generative adversarial networks for 3d-aware image synthesis. In *CVPR* (pp. 18450–18459).

Zhao, H., Gallo, O., Frosio, I., & Kautz, J. (2016). Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging, 3*(1), 47–57.

Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., & Kautz, J. (2019). Joint discriminative and generative learning for person re-identification. In *CVPR* (pp. 2138–2147).

Zhou, P., Xie, L., Ni, B., & Tian, Q. (2021). Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. arXiv:2110.09788

Zhou, T., Brown, M., Snavely, N., & Lowe, D. G. (2017). Unsupervised learning of depth and ego-motion from video. In *CVPR* (pp. 1851–1858).

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV* (pp. 2223–2232).

Zhu, J. Y., Zhang, Z., Zhang, C., Wu, J., Torralba, A., Tenenbaum, J., & Freeman, B. (2018). Visual object networks: Image generation with disentangled 3d representations. *Advances in Neural Information Processing Systems, 31*.