Generative Multiplane Neural Radiance for 3D-Aware Image Generation

Amandeep Kumar[1] ✉ Ankan Kumar Bhunia[1] Sanath Narayan[2] Hisham Cholakkal[1]
Rao Muhammad Anwer[1,3] Salman Khan[1] Ming-Hsuan Yang[4,5,6] Fahad Shahbaz Khan[1,7]
[1]Mohamed bin Zayed University of AI [2]Technology Innovation Institute [3]Aalto University
[4]University of California, Merced [5]Yonsei University [6]Google Research [7]Linköping University

Abstract

We present a method to efficiently generate 3D-aware high-resolution images that are view-consistent across multiple target views. The proposed multiplane neural radiance model, named GMNR, consists of a novel α-guided view-dependent representation (α-VdR) module for learning view-dependent information. The α-VdR module, facilitated by an α-guided pixel sampling technique, computes the view-dependent representation efficiently by learning viewing direction and position coefficients. Moreover, we propose a view-consistency loss to enforce photometric similarity across multiple views. The GMNR model can generate 3D-aware high-resolution images that are view-consistent across multiple camera poses, while maintaining the computational efficiency in terms of both training and inference time. Experiments on three datasets demonstrate the effectiveness of the proposed modules, leading to favorable results in terms of both generation quality and inference time, compared to existing approaches. Our GMNR model generates 3D-aware images of 1024×1024 *pixels with* 17.6 *FPS on a single V100. Code : https://github.com/VIROBO-15/GMNR*

1. Introduction

The advances in generative adversarial networks (GANs) [21] have resulted in significant progress in the task of high-resolution photorealistic 2D image generation [27, 28, 30]. The problem of generating 3D-aware images that render an object in different target views has received increasing interest in the recent years. Learning such 3D-aware image generation is challenging due to the absence of 3D geometry supervision or multi-view inputs during training. Furthermore, the synthesized 3D-aware images are desired to be of high-resolution, generated at extrapolated views (*i.e.*, large non-frontal views) and consistent across camera views.

In the absence of 3D supervision, existing 3D-aware image generation approaches [8, 13] typically rely on learning the 3D geometric constraints by using either implicit [47, 42, 23, 43] or explicit [35, 51] 3D-aware inductive biases and a rendering engine. While implicit representations, *e.g.*, neural radiance fields [40] (NeRF), possess the merits of better handling complex scenes along with memory efficiency, their slow querying and sampling generally negatively affects the training duration, inference time as well as the 3D-aware generation of high-resolution images. On the other hand, explicit representations, *e.g.*, voxel grid [48], are typically fast but have large memory footprint leading to scaling issues at higher resolutions. These issues are recently addressed [65] by utilizing multiplane images (MPI) as an explicit representation to transfer the knowledge learned by a 2D GAN to 3D-awareness. In this way, existing 2D GANs, *e.g.*, StyleGAN [28], can be extended to obtain the alpha maps conditioned on the plane's depth, followed by conditioning the discriminator on a target pose for training the image synthesis model.

While the aforementioned scheme of avoiding a volumetric rendering of pixels enables efficient training and inference, it may lead to inaccurate rendering of object shapes at extrapolated views due to fewer multiplanes during training. Moreover, inconsistent artifacts across different views can occur since such a scheme optimizes the warping from canonical pose to a single target pose. A straightforward way to overcome this issue is to increase the resolution in the disparity space, *i.e.*, more planes. This can likely help in reducing these artifacts resulting in improved rendering at extrapolated viewing angles. However, this will result in significantly increasing the training time as well as memory overhead. In this work, we show how to collectively address the above issues without any significant degradation of training and inference speed.

Contributions: We propose an efficient approach named Generative Multiplane Neural Radiance (GMNR), that learns to synthesize 3D-aware and view-consistent high-resolution images across difference camera poses. To this end, we introduce a novel α-guided view-dependent representation module (α-VdR) that enables the generator to bet-

✉ amandeep.kumar@mbzuai.ac.ae

Figure 1. Generated examples using our proposed 3D-aware view-consistent GMNR approach. For each example, we show the generated canonical view along with the rendered images at two different target poses. Our GMNR efficiently synthesizes 3D-aware high-resolution (512×512 in row 2; 1024×1024 in rows 1 and 3) scenes with detailed geometry along with consistent rendering across multiple views at a speed of 17.6 frames per second (1024×1024 pixels) on a single Tesla V100 GPU.

ter learn view-dependent information during training. Our α-VdR employs a linear combination of learnable *image-specific* viewing direction and *image-agnostic* position coefficients along with an $\boldsymbol{\alpha}$-guided pixel sampling technique to compute the view-dependent representation efficiently. The proposed sampling technique ensures that a balanced set of valid pixel locations from each multiplane is considered when computing the view-dependent representation, resulting in 3D-aware high-resolution images with diminished artifacts in the target poses. Moreover, we employ a view-consistency loss for enforcing photometric similarity across multiple rendered views. Consequently, our GMNR generates 3D-aware high-resolution images that are view consistent across different camera poses while maintaining the computational efficiency at inference.

Extensive qualitative and quantitative experiments are conducted on three datasets: FFHQ [29], AFHQv2-Cats [11] and MetFaces [26]. Our GMNR performs favorably against existing works published in literature. When generating images of 1024×1024 pixels on FFHQ dataset, GMNR outperforms the best existing approach [65] by reducing the FID from 7.50 to 6.58, while operating at a comparable inference speed of 17.6 frames per second (FPS) on a single tesla V100. Fig. 1 shows 3D-aware high-resolution generated scenes from our GMNR exhibiting detailed geometry and consistent rendering across multiple views.

2. Preliminaries

Problem Statement: In this work, the goal is to learn a 2D GAN for generating 3D-aware high-resolution images that are view-consistent, such that the generated images identically encapsulate the synthesized objects at different target camera poses p_t. Here, multiplane images are generated to capture the 3D information and are then utilized to render a view-consistent 2D image at a target camera pose p_t. The multiplane images consist of a set of L fronto-parallel planes $i \in \{1, \cdots, L\}$, each of size $H \times H \times 4$, *i.e.*, each plane i comprises an RGB image $C_i \in \mathbb{R}^{H\times H\times 3}$ and an alpha map $\boldsymbol{\alpha}_i \in [0,1]^{H\times H\times 1}$. The distance between the camera and a plane i is denoted by depth $d_i \in \mathcal{R}$. Next, we describe our baseline framework for generating view-consistent 3D-aware high-resolution images.

2.1. Baseline Framework

Our baseline model is motivated by the recent multi-plane image generation method, GMPI [65], since it focuses on computationally efficient generation of 3D-aware images. GMPI extends the StyleGANv2 [30] network with a branch for obtaining alpha maps and a differentiable renderer for generating 3D images at different target camera poses. Moreover, the base GMPI framework reuses the same color-texture across all planes, in turn reducing the task of StyleGANv2 generator $f_G(\cdot)$ to synthesizing a single RGB image C and the corresponding per-plane alpha maps $\boldsymbol{\alpha}_i$, given by

$$\boldsymbol{M} \triangleq \{C, \{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_L\}\} = f_G(z, \{d_1, \cdots, d_L\}), \quad (1)$$

where z denotes the latent vector input to $f_G(\cdot)$. For generating the alpha maps at a resolution $r \in \{4, 8, \cdots\}$, a single convolutional layer $f^r_{ToAlpha}$ is first used to generate the $\hat{\boldsymbol{\alpha}}^r_i$ from intermediate feature representations $F^r_{\boldsymbol{\alpha}_i}$, given by

$$\hat{\boldsymbol{\alpha}}^r_i = f^r_{ToAlpha}(F^r_{\boldsymbol{\alpha}_i}), \quad (2)$$

$$F^r_{\boldsymbol{\alpha}_i} = \frac{F^r - \mu(F^r)}{\sigma(F^r)} + f_{Emb}(d_i, e), \quad (3)$$

where $\sigma(F^r)$, $\mu(F^r) \in \mathbb{R}^{dim_r}$ denote the standard deviation and mean of the feature $F^r \in \mathbb{R}^{r \times r \times dim_r}$. The plane specific embedding $f_{Emb}(d_i, e)$ is computed using the style embedding e and depth d_i of plane i, similar to StyleGANv2. Note that $F^r_{\boldsymbol{\alpha}_i}$ is specific to each plane, while $f^r_{ToAlpha}$ is shared across all planes. Finally, the alpha maps $\boldsymbol{\alpha}^r_i$ at different resolutions r are accumulated through an upsampling operation that is consistent with the StyleGANv2 design. With this formulation, GMPI introduces a branch to generate alpha maps conditioned on the plane depths d_i by utilizing the intermediate feature representations and conditions the discriminator on the camera poses to make the 2D StyleGANv2 3D-aware [65].

Our baseline framework extends the existing 2D StyleGANv2 to make it 3D-aware using implict and explicit representations. Moreover, by first generating multiplane images at canonical view and then warping them to target poses, it avoids a volumetric rendering leading to an efficient training and inference. However, the baseline model does not effectively render object images at extrapolated views, likely due to fewer multiplanes employed to overcome memory issues during training. Furthermore, since the baseline optimizes by warping from canonical view to a single random target pose, it leads to inconsistent artifacts across multiple views (see Fig. 2). Next, we present our approach that collectively address the above issues without any significant change in training and inference time.

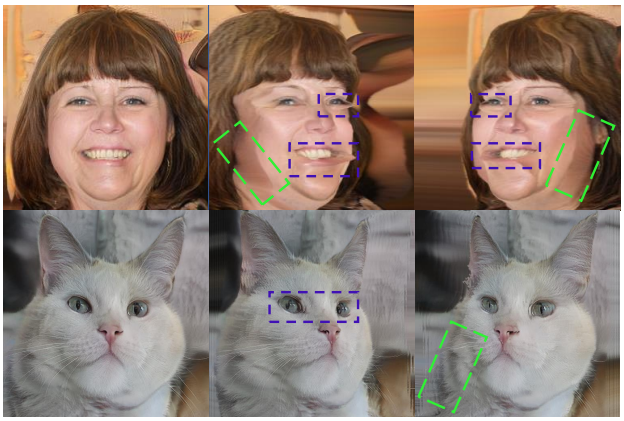

Figure 2. Example generated images using the baseline approach depicting the frontal (canonical) and target views. Here, the baseline model is trained on FFHQ (rows 1) and AFHQv2-Cats (row 2). While being effective in synthesizing frontal views (col. 1), the baseline struggles when generating the target views and introduces artifacts during the rendering (col. 2 and 3). In these cases, the generated images depict repeated textures (highlighted as blue box) due to the same RGB content in each plane and layered artifacts (highlighted as green box) due to fewer planes employed during training. For more examples, see supplementary material.

3. Proposed Approach

Overall Architecture: Fig. 3 presents the overview of our proposed GMNR framework. Within our GMNR, the RGB$\boldsymbol{\alpha}$ generator adapts a conventional 2D generator by integrating an $\boldsymbol{\alpha}$-branch with StyleGANv2, yielding a set of fronto-parallel alpha maps $\{\boldsymbol{\alpha}_i\}^L_{i=1}$, as in the baseline. The focus of our design is the introduction of an $\boldsymbol{\alpha}$-*guided view-dependent representation* ($\boldsymbol{\alpha}$-VdR) that learns image-specific view-dependent information and modifies the RGB values at the different planes according to the view-dependent directions, which is crucial for rendering images with diminished artifacts in target poses. The $\boldsymbol{\alpha}$-VdR module learns a view-dependent pixel representation using a linear combination of coefficients obtained from two MLPs by efficiently sampling pixel positions through an $\boldsymbol{\alpha}$-guided pixel sampling technique. This enables the $\boldsymbol{\alpha}$-VdR module to learn modeling the image-specific view-dependent 3D characteristics. Moreover, we employ a *view-consistency loss* to enforce photometric consistency across different views of the rendered images. Consequently, 3D-aware view-consistent images of high-resolution at target poses are synthesized by the RGBα generator together with the $\boldsymbol{\alpha}$-VdR module and renderer at inference.

3.1. α-guided View-dependent Representation

As discussed earlier, the baseline model renders 3D images using an MPI representation without explicitly utilizing the view-dependent information during training, which is desired for 3D-aware view-consistent image genera-

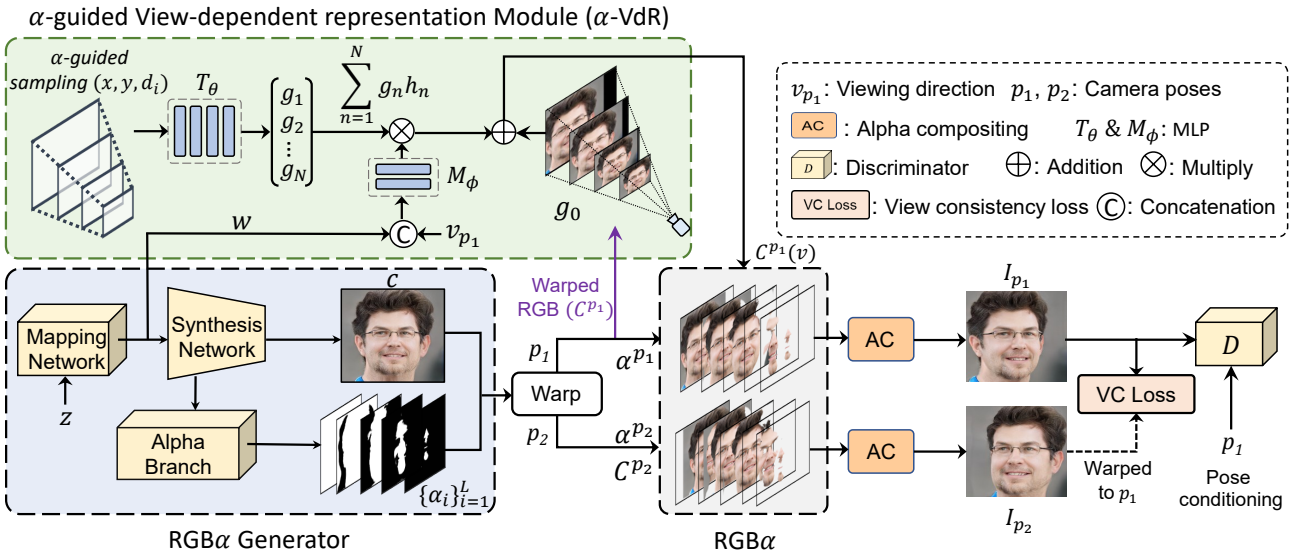

Figure 3. Overall architecture of our GMNR for generating 3D-aware and view-consistent images at high-resolution. Our GMNR takes a latent vector $z \in \mathbb{R}^{d_z}$ and outputs an RGB image at a target view. GMNR comprises an RGBα generator, an $\boldsymbol{\alpha}$-guided view-dependent representation ($\boldsymbol{\alpha}$-VdR) module, a differentiable renderer and a pose-conditioned discriminator. The RGBα generator synthesizes the RGB image and the alpha maps $\{\boldsymbol{\alpha}_i\}_{i=1}^{L}$ corresponding to the canonical pose. The generated RGB image that is warped to a target pose p_1 is then input along with the style-code $\boldsymbol{w}$ to the $\boldsymbol{\alpha}$-VdR module (Sec. 3.1). The $\boldsymbol{\alpha}$-VdR module learns a view-dependent pixel representation using a linear combination of coefficients (image-agnostic $\{g_n\}_{n=1}^{N}$ and image-specific $\{h_n(\boldsymbol{w})\}_{n=1}^{N}$) computed from the $\boldsymbol{\alpha}$-guided sampling positions (x, y, d_i), viewing direction v_{p_1} and style-code $\boldsymbol{w}$ using two MLP networks $T_\theta(\cdot)$ and $M_\phi(\cdot)$, respectively. Here, the α-guided pixel sampling aids in efficiently sampling the pixel positions for computing the view-dependent representation. As a result, $\boldsymbol{\alpha}$-VdR module learns to model the image-specific view-dependent 3D characteristics. Moreover, a view-consistency loss $\mathcal{L}_{vc}$ (Sec. 3.2) is employed for enhancing the photometric consistency across different views of the rendered images. Consequently, the RGBα generator along with the $\boldsymbol{\alpha}$-VdR module and renderer synthesize 3D-aware view-consistent images at target poses during inference.

tion. To learn view-dependent information, we introduce an $\boldsymbol{\alpha}$-guided view-dependent representation module ($\boldsymbol{\alpha}$-VdR) that comprises two separate MLP networks $T_\theta(\cdot)$ and $M_\phi(\cdot)$. We consider a pixel to be a discrete sample of a radiance function $R(q, v)$ with $q, v \in \mathbb{R}^3$ as the pixel coordinate and the target viewing direction. Motivated by [31, 32], we note that $R(q, v)$ can be approximated by a sum of products $g_n \cdot h_n$. Here, g_n and h_n are computed using two MLP networks $T_\theta(\cdot)$ and $M_\phi(\cdot)$ with inputs q and v, respectively.

Given the pixel location $q = (x, y, d_i)$ for a plane depth d_i ($i \in \{1, \cdots, L\}$), the MLP $T_\theta(\cdot)$ predicts the *image-agnostic* position coefficients $\{g_1^q, \cdots, g_N^q\}$. Similarly, to enable *image-specific* view-dependent modeling, the normalized viewing direction $v = (v_x, v_y, v_z)$ is utilized along with the style-code $\boldsymbol{w}$ generated by the mapping network of the StyleGANv2 in the RGBα generator. To this end, v and $\boldsymbol{w}$ are concatenated and input to the MLP $M_\phi(\cdot)$, which outputs image-specific viewing direction coefficients $\{h_1^v(\boldsymbol{w}), \cdots, h_N^v(\boldsymbol{w})\}$. The color representation $s^q(v)$ of a pixel q for a target viewing direction v is computed using

$$s^q(v) = g_0^q + \sum_{n=1}^{N} g_n^q \cdot h_n^v(\boldsymbol{w}). \tag{4}$$

Note that g_0 is computed by performing a homography warping operation on the RGB image generated at canonical pose to the target camera pose p_1.

$\boldsymbol{\alpha}$-guided Pixel Sampling: Sampling all the pixels for color representation learning (Eq. 4) to generate *high-resolution* images is cost-prohibitive, since it is infeasible to compute the view-dependent representation $s^q(v)$ of all the pixels q in an MPI. To alleviate this issue, we introduce a new sampling technique with the aid of the generated alpha maps $\{\boldsymbol{\alpha}_i\}_{i=1}^{L}$, which reduces the volume of points that are required for sampling while retaining the fine details for coefficient learning. To this end, we compute a weight matrix $\boldsymbol{A}_i \in \mathbb{R}^{H \times H}$ corresponding to the plane i, given by

$$\boldsymbol{A}_i = \boldsymbol{\alpha}_i^{p_1} \cdot \prod_{j=1}^{i-1} (1 - \boldsymbol{\alpha}_j^{p_1}), \tag{5}$$

where $\boldsymbol{\alpha}_i^{p_1}$ denotes $\boldsymbol{\alpha}_i$ warped to target pose p_1. A pixel location (x, y, d_i) is a candidate for sampling if the corresponding weight is greater than 0, *i.e.*, if $\boldsymbol{A}_i(x, y) > 0$. Note that Eq. 5 ensures that a pixel location (x, y, d_i) is not considered during sampling if either $\boldsymbol{\alpha}_i^{p_1}(x, y) = 0$ or if $\boldsymbol{\alpha}_j^{p_1}(x, y) = 1$ for any $j < i$. Furthermore, among the candidate locations (x, y, d_i), a *per-plane* random sampling is performed to select only a certain percentage of the candidate locations in a plane i. This balanced sampling strat-

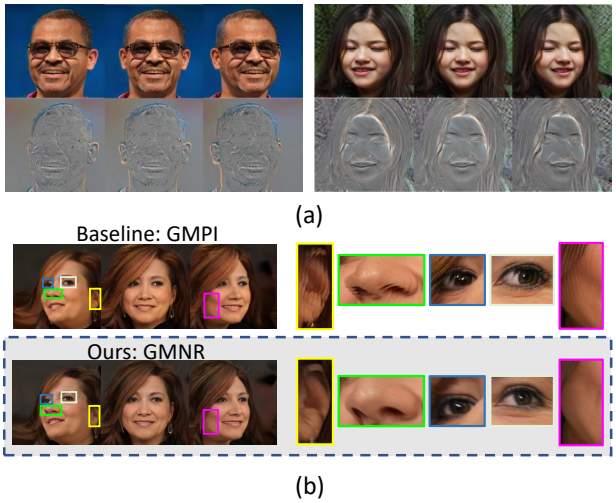

Figure 4. (a) Visualization of view-dependent information integrated by $\boldsymbol{\alpha}$-VdR module (first row) at various poses of two example generated images (second row) (b) Qualitative comparison between baseline GMPI and our GMNR. Repeated texture artifacts at extrapolated views are seen in the baseline due to novel regions being rendered as duplicates of visible areas. These repeated texture artifacts are reduced at extrapolated views of GMNR due to the modification of RGB values at different planes by our α-VdR module, *e.g.*, ear, nose, eyes in the zoomed-in crops.

egy guided by alpha maps ensures that every plane i is adequately represented during the color representation computation (Eq. 4), thereby leading to effective learning of view-dependent information. This technique of per-plane $\boldsymbol{\alpha}$-guided sampling mitigates the issue of under-sampling of valid pixels in nearby planes, which can arise if sampling is performed by considering pixels together across all planes. As such, our $\boldsymbol{\alpha}$-VdR module outputs a view-dependent RGB image $C^{p_1}(v)$ at a target pose p_1 and incorporates 3D-awareness during image generation. Fig. 4(a) shows the view-dependent information integrated by $\boldsymbol{\alpha}$-VdR module when rendering objects at different views.

As discussed earlier, the standard view-independent MPI used in GMPI utilizes the same RGB (generated) image for all the planes resulting in repeated textures at extrapolated views due to novel regions being rendered as duplicates of visible areas (see nose and ear crops of baseline in Fig. 4(b)). To address this issue, the proposed α-guided view-dependent representation (α-VdR) module in our GMNR modifies the RGB values at different planes according to the viewing direction (Eq. 4), thereby reducing the artifacts (bottom row in Fig 4(b)) without requiring additional multi-planes. Furthermore, Fig. 5 shows the impact of $\boldsymbol{\alpha}$-VdR module on generated examples from FFHQ and AFHQv2-Cats. Compared to the baseline, our approach with $\boldsymbol{\alpha}$-VdR module synthesizes high-resolution images at target views with diminished artifacts.

Baseline **Our Approach**

Figure 5. Synthesized rendered RGB image at a target view (left), its corresponding canonical view (center) along with the mesh (right) for the baseline and our GMNR, respectively. Compared to the baseline, our GMNR better renders the target objects at large non-frontal views due to learning the view-dependent information through the $\boldsymbol{\alpha}$-VdR module. Stretched eyes can be observed in the case of baseline generated image (row 1, col. 1), while such an artifact is mitigated by our approach (row 1, col. 4).

3.2. View-consistency Loss

To achieve a better photometric consistency across multiple views, we employ an image-level optimization loss. Let I_{p_1} and I_{p_2} denote the images rendered at target poses p_1 and p_2. Note that only I_{p_1} is generated by integrating the view-dependent information from $\boldsymbol{\alpha}$-VdR module, while I_{p_2} is rendered directly from the RGBα generator output, as shown in Fig. 3. We first warp I_{p_2} to target pose p_1 to obtain the warped image I_ψ. The warping is performed by utilizing the accumulated depth of the alpha maps. Afterwards, we employ the image-level optimization loss that enhances the view-consistency between the primary image I_{p_1} and the warped image I_ψ in order to satisfy the geometry requirements between views. Similar to the image reconstruction problem [20, 67], we formulate the image-level optimization loss as a combination of SSIM [58] and ℓ_1 [64], given by

$$\mathcal{L}_{vc} = \frac{\delta}{2}(1 - SSIM(I_{p_1}, I_\psi)) + (1-\delta)||I_{p_1} - I_\psi||_1. \quad (6)$$

3.3. Training and Inference

Training: Our GMNR framework is trained using a pose conditioned discriminator D, which compares the fake images generated by the RGBα generator and real training images I_{gt}. The overall loss formulation, utilizing a non-saturating GAN loss with $R1$ penalties [37] and the view-

consistency loss $\mathcal{L}_{vc}$ is given by

$$\mathcal{L} = \mathbb{E}_{I_{p_t},p_t}[f(\log Q(I_{p_t},p_t))] + \mathbb{E}_{I_{gt},p_t}[f(\log Q(I_{gt},p_t))] + \eta|\nabla_{I_{gt}} \log P(y=\text{real}|I_{gt},p_t)|^2] + \lambda\mathcal{L}_{vc}, \quad (7)$$

where $Q(I,p_t) = P(y = \text{real}|I,p_t)$ denotes the probability that image I from a camera pose p_t is real, $f(x) = -\log(1+\exp(-x))$ and $\eta = 10.0$.

Inference: During inference, we use the RGBα generator to synthesize RGB image C and alpha maps $\boldsymbol{\alpha}_i$. The $\boldsymbol{\alpha}$-VdR module takes the semantic information $\boldsymbol{w}$ generated by the StyleGANv2 along with target viewing direction v_t as inputs, and computes an image-specific view-dependent representation $C^{p_t}(v_t) \in \mathbb{R}^{H\times W\times 3\times L}$ (based on Eq. 4). This $C^{p_t}(v_t)$ together with $\boldsymbol{\alpha}_i$ are used to obtain multiplane images, which are then input to the MPI renderer. The renderer warps them to the target pose p_t followed by alpha composition for combining the planes to obtain the desired 3D-aware and view-consistent image I_{p_t}. Note that while the image-specific coefficients $\{h_n(\boldsymbol{w})\}$ in Eq. 4 are computed once for an image, the image-agnostic coefficients $\{g_n^q\}$ are computed only once, leading to minimal computational overhead during inference.

4. Experiments

Datasets: The proposed GMNR is evaluated on three datasets: FFHQ, AFHQv2 and MetFaces. The **FFHQ** [29] dataset comprises $70,000$ high-quality images of real people's faces at 1024×1024 resolution captured from various angles. The **AFHQv2-Cats** [11, 27] dataset has $5,065$ images (512×512 size) of cat faces at various angles. The **MetFaces** [26] dataset consists of $1,336$ high-quality face images extracted from the Metropolitan Museum of Art's collection. While an off-the-shelf pose estimator [14] is employed to compute a face's pose required for the pose conditioning for FFHQ and MetFaces, a cat face landmark predictor [3] along with OpenCV [1] perspective-n-point technique are used to compute the pose for AFHQv2-Cats images. Furthermore, horizontal flips are used as augmentation for AFHQv2-Cats and MetFaces.

Evaluation Metrics: In this work, five metrics are employed for quantitatively comparing the generated image quality, as in [65]. The Frechet Inception Distance (FID) [25] and Kernel Inception Distance (KID) [5] are computed between $50K$ generated images rendered at different random poses and (a) $50K$ real images for FFHQ; (b) $5,065$ real images with flip augmentation for AFHQv2-Cats. The multi-view facial identity consistency (ID) is computed by first generating $1,024$ MPI representations and then employing the mean Arcface [12] cosine similarity score between pairs of rendered views at random poses for the same face. The depth accuracy (Depth) is measured as the MSE between the rendered depth and the pseudo ground-truth depth obtained from a pre-trained face reconstruction model [14] on the face mask area. Similarly, the 3D pose accuracy (Pose) is computed by comparing the pose input used for rendering and the yaw, pitch and roll predicted by [14] for the rendered image.

4.1. Implementation Details

Within our GMNR, the MLP $T_\theta(\cdot)$ comprises 4 fully-connected (FC) layers with hidden size of 384, while $M_\phi(\cdot)$ has 3 FC layers with hidden size 64. Here, Leaky-ReLU activation is used in both MLPs. We add sinusoidal positional encodings to the pixel location and the viewing direction inputs, as in [40]. While the batch size is set to 32 for AFHQv2-Cats, it equals 64, 32 and 16 for FFHQ256, FFHQ512 and FFHQ1024, respectively. We set δ in L_{vc} loss to 0.85 and λ to 0.5. For $\boldsymbol{\alpha}$-guided pixel sampling, 6% of valid pixels are sampled from each plane for FFHQ (256^2, 512^2 sizes) and AFHQv2-Cats, while it is 4% for Metfaces and FFHQ (1024×1024). The learning rate for training our GMNR is set to 2×10^{-3}. As in [65], we use 32 planes during training and 96 for inference in all experiments. Near and far depth of the MPI are set as $0.95/1.12$ (FFHQ and Metfaces), $2.55/2.8$ (AFHQv2-Cats). Depth normalization is performed as in [65]. Our model is trained using 8 Tesla V100 GPUs using PyTorch-1.9 [45]. In our framework, we refer to the L^{th} plane as the background plane of the MPI representation. To color the background plane, we use the leftmost and rightmost 5% pixels of the synthesized image C as the left and right boundaries, respectively. The RGB values of the remaining pixels in the background plane are linearly interpolated between the left and right boundaries. For generating the mesh cubes, we use the marching cube algorithm [37] implemented in PyMCubes [2] and utilize a smoothing function to have the better visualization.

4.2. Experimental Results

4.2.1 Baseline Comparison

We first present a quantitative and qualitative comparison of our GMNR approach with the baseline GMPI on both FFHQ and AFHQv2-Cats datasets. Tab. 1 shows the comparison in terms of FID, KID, ID, Depth, Pose metrics and training time. As in the baseline GMPI [65], the comparison is presented at three different resolutions: 256×256, 512×512 and 1024×1024. Compared to the baseline, our GMNR achieves consistent improvement in performance on all metrics, without any significant degradation in the training time and inference speed. In the case of 512^2 resolution, the baseline obtains FID scores of 8.29 and 7.79 on FFHQ and AFHQv2-Cats datasets, respectively. In comparison, our GMNR achieves favorable performance with FID scores of 6.81 and 6.01, respectively. Similarly, GMNR obtains improved performance by reducing the Depth er-

Table 1. Comparisons between the baseline and our GMNR on FFHQ and AFHQv2-Cats. We also report the training time when utilizing 8 Tesla V100 GPUs. Our GMNR achieves consistent improvement in performance on all metrics and different resolutions, compared to the baseline. Furthermore, this improvement in performance over the baseline is achieved without any significant degradation in the training time and the inference speed of the model.

	Method	FFHQ							AFHQv2-Cats	
		$FID\downarrow$	$KID\downarrow$	$ID\uparrow$	$Depth\downarrow$	$Pose\downarrow$	Train Time$\downarrow$	Infer Speed$\uparrow$	$FID\downarrow$	$KID\downarrow$
256^2	Baseline	11.4	0.738	0.700	0.53	0.0040	3h	328 FPS	n/a	n/a
	Ours: GMNR	9.20	0.720	0.730	0.39	0.0032	3h 33m	313 FPS	n/a	n/a
512^2	Baseline	8.29	0.454	0.740	0.46	0.0060	5h	83.5 FPS	7.79	0.474
	Ours: GMNR	6.81	0.370	0.760	0.40	0.0052	5h 42m	78.9 FPS	6.01	0.450
1024^2	Baseline	7.50	0.407	0.750	0.54	0.0070	11h	19.4 FPS	n/a	n/a
	Ours: GMNR	6.58	0.351	0.769	0.43	0.0064	12h	17.6 FPS	n/a	n/a

Table 2. Effect of progressively integrating our proposed contributions into the baseline on FFHQ dataset with 512^2 resolution. The introduction of the proposed $\boldsymbol{\alpha}$-VdR (Sec. 3.1) into the baseline results in a consistent improvement in performance on all metrics. The results are further improved when integrating the proposed view-consistency loss $\mathcal{L}_{vc}$ (Sec. 3.2).

Method	$FID\downarrow$	$KID\downarrow$	$ID\uparrow$	$Depth\downarrow$	$Pose\downarrow$
Baseline	8.29	0.454	0.740	0.457	0.0060
Baseline + $\boldsymbol{\alpha}$-VdR	7.01	0.381	0.751	0.412	0.0054
Baseline + $\boldsymbol{\alpha}$-VdR + $\mathcal{L}_{vc}$	6.81	0.370	0.760	0.400	0.0052

ror from 0.46 to 0.40 and KID from 0.454 to 0.370 on the FFHQ dataset, compared to the baseline.

4.2.2 Ablation Study

Tab. 2 shows the impact of progressively introducing each of our contributions into the baseline on the FFHQ dataset at 512^2 resolution. When integrating the proposed $\boldsymbol{\alpha}$-VdR (Sec. 3.1) into the baseline framework, we observe a consistent improvement in results highlighting the importance of learning view-dependent information during training to render images with diminished artifacts in target poses. Notably, the FID scores reduce from 8.29 to 7.01, and the KID scores from 0.454 to 0.370. The results are further improved by the introduction of the view-consistency loss $\mathcal{L}_{vc}$ (Sec. 3.2), leading to a consistent gain on all metrics. Our final approach (row 3) that generates 3D-aware view-consistent images achieves an absolute improvement of 1.48 in terms of FID score over the baseline.

We further conduct an experiment to ablate the pixel sampling rate in our proposed plane-specific sampling within the $\boldsymbol{\alpha}$-VdR of our GMNR. Here, we ablate the rate from 1% to 6%, since 6% is the maximum rate that can be accommodated during our GMNR training under the same batch size setting as the baseline. We observe the results to consistently improve when increasing the sampling rate (1%: 8.32, 3%: 7.53, 6%: 6.81 in terms of FID score). As a next step, we also compare our plane-specific sampling with random sampling across planes at the optimal sampling rate (6%). The random sampling scheme obtains FID and KID scores of 7.68 and 0.39. In comparison, our plane-specific sampling-based GMNR improves the results, achieving FID and KID scores of 6.81 and 0.37, respectively.

Similarly, to see the effect of λ, we set it to a high value (λ=20) for $\mathcal{L}_{vc}$ which improves the ID score to 0.732 at the cost of FID 9.70 for 256×256 resolution, leading to sub-optimal generation. Furthermore, we observe that setting $p_2 = 0$ throughout the experiment for the view-consistency loss leads to the sub-optimal FID of 9.63 for 256×256 resolution.

4.2.3 Comparison with Existing Approaches

Here, we present the comparison of our GMNR with existing works published in literature. Tab. 3 presents the comparison on FFHQ and AFHQv2-Cats datasets. The results are reported with 256×256, 512×512 and 1024×1024 images. Both the GMPI and our GMNR approach utilize 96 planes during the test time and do not employ any truncation tricks [6, 28]. Among existing works, the recent GMPI obtains comparable performance to other methods in literature with faster training, when evaluating on 256^2 resolution images. For instance, the training time of GMPI for 256^2 resolution is 3 hours (using 8 Tesla V100 GPUs), excluding the StyleGANv2 pre-training time [65]. Compared to the GMPI training duration (including both StyleGANv2 pre-training and GMPI training), other existing methods including EG3D, GRAM, and StyleNeRF require a longer training time. Moreover, GMPI demonstrates the ability to generate high-resolution images of 1024×1024 resolutions, where most other existing works struggle to generate. When comparing with GMPI approach, our GMNR achieves consistently improved performance on all metrics at different resolutions, without any significant degradation in training time as well as operating at a comparable inference speed. For the high-resolution of 1024×1024, our GMNR achieves FID score of 6.58 on the FFHQ dataset, performing favorably against existing published works in the literature while operating at an inference speed of 17.6 FPS on a single Tesla V100. Fig. 6 presents a compari-

Table 3. Comparisons with existing approaches. Here, both GMPI and our approach use 96 planes during inference and without applying any truncation tricks [6, 28]. Further, the KID score is reported in KID×100. In case the corresponding work does not report the results, we denote it as '-'. We report the results of GIRAFFE, pi-GAN and LiftedGAN from the EG3D paper. Results reported on the entire AFHQv2 dataset instead of cats only are denoted by '*'. For 256×256 and 1024×1024 resolutions, no results are reported on AFHQv2-Cats for both GMPI and our GMNR since the corresponding pre-trained StyleGANv2 checkpoints are unavailable. Compared to the recent GMPI, our GMNR achieves consistent improvement in performance, while running at comparable inference speed (FPS). Particularly, GMNR better generates high-resolution (1024×1024) images where most existing works struggle to operate demonstrating its flexibility. Further, GMNR achieves consistent improvement on all metrics and significantly reduces the FID from 7.50 to 6.58 on FFHQ, compared to GMPI.

	Method	Infer Speed↑	FFHQ					AFHQv2-Cats	
			$FID\downarrow$	$KID\downarrow$	$ID\uparrow$	$Depth\downarrow$	$Pose\downarrow$	$FID\downarrow$	$KID\downarrow$
256^2	GIRAFFE [42]	250	31.5	1.992	0.64	0.94	0.0890	16.1	2.723
	pi-GAN 128^2 [9]	1.63	29.9	3.573	0.67	0.44	0.0210	16.0	1.492
	LiftedGAN [49]	25	29.8	-	0.58	0.40	0.0230	-	-
	GRAM [13]	180	29.8	1.160	-	-	-	-	-
	StyleSDF [43]	-	11.5	0.265	-	-	-	12.8*	0.447*
	StyleNeRF [23]	16	8.00	0.370	-	-	-	14.0*	0.350*
	CIPS-3D [66]	-	6.97	0.287	-	-	-	-	-
	EG3D [8]	36	4.80	0.149	0.76	0.31	0.0050	3.88	0.091
	GMPI [65]	328	11.4	0.738	0.70	0.53	0.0040	n/a	n/a
	Ours: GMNR	313	9.20	0.720	0.73	0.39	0.0032	n/a	n/a
512^2	EG3D [8]	35	4.70	0.132	0.77	0.39	0.0050	2.77	0.041
	StyleNeRF [23]	14	7.80	0.220	-	-	-	13.2*	0.360*
	GMPI [65]	83.5	8.29	0.454	0.74	0.46	0.0060	7.79	0.474
	Ours: GMNR	78.9	6.81	0.370	0.76	0.40	0.0052	6.01	0.450
1024^2	CIPS-3D [66]	-	12.3	0.774	-	-	-	-	-
	StyleNeRF [23]	11	8.10	0.240	-	-	-	-	-
	GMPI [65]	19.4	7.50	0.407	0.75	0.54	0.0070	n/a	n/a
	Ours: GMNR	17.6	6.58	0.351	0.76	0.43	0.0064	n/a	n/a

son of StyleNeRF[23] and GMPI[65] with GMNR in terms of high-resolution generated image quality at various target views. We show six sets of images that are generated at high-resolution (1024×1024) for each method. For each set, we show a generated face at a canonical view in the center as well as the four non-frontal views at different angles. Additional results are presented in supplementary.

5. Related Work

Several works have explored rendering 2D multi-view images [18, 53, 61, 62, 63] as neural representations of 3D scenes that are differentiable and 3D-aware[4, 7, 10, 16, 22, 38, 39, 55]. Based on the scene geometry used, representations can be implicit or explicit. Explicit representations, such as voxel grid and multiplane images have been employed to render novel views in [52, 35, 68, 53, 57, 19, 33] for their speed. However, these representations often encounter memory overheads, making it challenging to scale to high-resolutions. Differently, approaches [38, 40, 50, 56] employing implicit 3D representation, such as neural radiance fields [40] (NeRF) are memory efficient and can handle complex scenes. However, these approaches struggle with slow rendering limiting the resolution of rendered images. Few works [15, 34, 36, 46, 8] have also explored integrating the merits of these two representations.

In the context of 3D-aware 2D GAN-based image generation, earlier works adopted 3D representations such as, voxel [17, 24, 41, 59, 69] and mesh [54] for 2D image synthesis. The voxel-based methods are difficult to train on higher-resolution images due to memory requirements of voxel grids and computational overhead of 3D convolution. While [42] partially alleviates this issue via low-resolution rendering followed by 2D upsampling, it struggles to synthesize view-consistent images due to the lack of 3D inductive biases. Furthermore, few works [9, 47, 44] utilize neural radiance fields (NeRF) [40] to generate 3D-aware images, constrained by the slow querying and inefficient GAN training. Recent NeRF-based works such as, StyleSDF[43], StyleNeRF[23], CIPS3D[66], GRAM[13], VolumeGAN[60] and EG3D[8] attempt to generate high-resolution images. However, these NeRF-based approaches still require significant training time to achieve convergence. In contrast, the recent GMPI method [65] adopts an MPI-based representation and obtains fast training and rendering speed. However, it struggles to accurately render object shapes at extrapolated views, leading to inconsistent artifacts cross multiple views. Our GMNR addresses these

Figure 6. Qualitative comparison of StyleNeRF [23], GMPI [65] and GMNR on FFHQ. For each method, six sets of example images generated at high-resolution (1024×1024) are shown. Each set comprises a generated face at canonical view in the center along with four non-frontal views at various angles. StyleNeRF-generated images exhibit inconsistent geometry across views, *e.g.*, hair style variation (rows 1, 2, 4 and 5), pupil distortion (row 4), varying artifacts on hat (row 6). On the other hand, GMPI generates near-frontal views of images reasonably well (rows 2 and 4), while artifacts occur at large non-frontal views (rows 1, 3 and 5). Compared to both methods, we observe our GMNR to generate images with enhanced view-consistency and diminished artifacts even at large non-frontal views.

issues by learning a view-dependent representation to generate 3D-aware high-resolution images without degrading the training efficiency and inference speed.

6. Conclusion

We introduced an approach, named GMNR, that focuses at efficiently generating 3D-aware high-resolution images that are view-consistent across multiple camera poses. To this end, the proposed GMNR introduces a novel $\boldsymbol{\alpha}$-VdR module that computes the view-dependent representation in an efficient manner by learning viewing direction and position coefficients. Additionally, we employ a view-consistency loss that aims at improving the photometric consistency across multiple views. Qualitative and quantitative experiments on three datasets demonstrate the merits of our contributions, leading to favorable performance in terms of image generation quality and computational efficiency, compared to existing works.

References

[1] Opencv. https://opencv.org/. 6

[2] Pymcubes. https://github.com/pmneila/PyMCubes. 6

[3] Cat hipsterizer. https://github.com/kairess/cat_hipsterizer, accessed: 2022-11-07. 6

[4] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *CVPR*, 2020. 8

[5] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 6

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 7, 8

[7] Rohan Chabra, Jan E Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *ECCV*, 2020. 8

[8] Eric Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, S. Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022. 1, 8

[9] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, 2021. 8

[10] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019. 8

[11] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 2, 6

[12] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 6

[13] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *CVPR*, 2022. 1, 8

[14] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPR Workshops*, 2019. 6

[15] Terrance Devries, Miguel Ángel Bautista, Nitish Srivastava, Graham W. Taylor, and Joshua M. Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *ICCV*, 2021. 8

[16] SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 2018. 8

[17] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. *3DV*, 2017. 8

[18] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *ICCV*, 2021. 8

[19] Sushobhan Ghosh, Zhaoyang Lv, Nathan Matsuda, Lei Xiao, Andrew Berkovich, and Oliver Cossairt. Liveview: Dynamic target-centered mpi for view synthesis. *arXiv preprint arXiv:2107.05113*, 2021. 8

[20] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. 5

[21] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1

[22] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 8

[23] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. In *ICLR*, 2022. 1, 8, 9

[24] Philipp Henzler, Niloy Jyoti Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *ICCV*, 2019. 8

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 6

[26] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020. 2, 6

[27] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 1, 6

[28] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 7, 8

[29] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 6

[30] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 1, 3

[31] Jan Kautz et al. Hardware rendering with bidirectional reflectances. 1999. 4

[32] Jan Kautz and Michael D McCool. Interactive rendering with arbitrary brdfs using separable approximations. In *EGWR*, 1999. 4

[33] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *ICCV*, 2021. 8

[34] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *arXiv preprint arXiv:2007.11571*, 2020. 8

[35] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. 1, 8

[36] Julien N. P. Martel, David B. Lindell, Connor Z. Lin, Eric Chan, Marco Monteiro, and Gordon Wetzstein. Acorn: Adaptive coordinate networks for neural scene representation. *arXiv preprint arXiv:2105.02788*, 2021. 8

[37] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *ICML*, 2018. 5, 6

[38] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 8

[39] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *ICCV*, 2019. 8

[40] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 6, 8

[41] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *ICCV*, 2019. 8

[42] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 1, 8

[43] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *CVPR*, 2022. 1, 8

[44] Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. In *NeurIPS*, 2021. 8

[45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *NeurIPS*, 2019. 6

[46] Songyou Peng, Michael Niemeyer, Lars M. Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. *arXiv preprint arXiv:2003.04618*, 2020. 8

[47] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *NeurIPS*, 2020. 1, 8

[48] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. In *NeurIPS*, 2022. 1

[49] Yichun Shi, Divyansh Aggarwal, and Anil K. Jain. Lifting 2d stylegan for 3d-aware face generation. In *CVPR*, 2021. 8

[50] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *arXiv preprint arXiv:2006.09661*, 2020. 8

[51] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. In *CVPR*, 2019. 1

[52] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *CVPR*, 2019. 8

[53] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *CVPR*, 2021. 8

[54] Attila Szabó, Givi Meishvili, and Paolo Favaro. Unsupervised generative 3d shape learning from natural images. *arXiv preprint arXiv:1910.00287*, 2019. 8

[55] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *CVPR*, 2021. 8

[56] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *arXiv preprint arXiv:2006.10739*, 2020. 8

[57] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *CVPR*, 2020. 8

[58] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004. 5

[59] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Joshua B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NeurIPS*, 2016. 8

[60] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *CVPR*, 2022. 8

[61] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *NeurIPS*, 2020. 8

[62] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *ICCV*, 2021. 8

[63] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 8

[64] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE TIP*, 2016. 5

[65] Xiaoming Zhao, Fangchang Ma, David Güera, Zhile Ren, Alexander G Schwing, and Alex Colburn. Generative multiplane images: Making a 2d gan 3d-aware. In *ECCV*, 2022. 1, 2, 3, 6, 7, 8, 9

[66] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788*, 2021. 8

[67] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 5

[68] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 8

[69] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Joshua B. Tenenbaum, and Bill Freeman. Visual object networks: Image generation with disentangled 3d representations. In *NeurIPS*, 2018. 8