Generative Neural Articulated Radiance Fields

Anonymous Author(s)
Affiliation
Address
email

Abstract

Unsupervised learning of 3D-aware generative adversarial networks (GANs) using only collections of single-view 2D photographs has very recently made much progress. These 3D GANs, however, have not been demonstrated for human bodies and the generated radiance fields of existing frameworks are not directly editable, limiting their applicability in downstream tasks. We propose a solution to these challenges by developing a 3D GAN framework that learns to generate radiance fields of human bodies or faces in a canonical pose and warp them using an explicit deformation field into a desired body pose or facial expression. Using our framework, we demonstrate the first high-quality radiance field generation results for human bodies. Moreover, we show that our deformation-aware training procedure significantly improves the quality of generated bodies or faces when editing their poses or facial expressions compared to a 3D GAN that is not trained with explicit deformations.

1 Introduction

Unsupervised learning of 3D-aware generative adversarial networks (GANs) using large-scale datasets of unstructured single-view images is an emerging research area. Such 3D GANs have recently been demonstrated to enable photorealistic and multi-view consistent generation of radiance fields representing human faces [1–7]. These approaches, however, have not been shown to work with human bodies, partly because learning the body pose distribution is much more challenging given the significantly higher diversity in articulations compared to facial expressions.

Yet, generative 3D models of photorealistic humans have significant utility in a wide range of applications, including visual effects, computer vision, and virtual or augmented reality. In these scenarios, it is critical that the generated people are editable to support interactive applications, which is not necessarily the case for existing 3D GANs. While variations of linear blend skinning [8] have been adopted to articulate radiance fields for single-scene scenarios [9–21], it is unclear how to efficiently apply such deformation methods to generative models.

With our work, dubbed generative neural articulated radiance fields or GNARF, we propose solutions to both of these challenges. Firstly, we demonstrate generation of high-quality 3D (i.e., multi-view consistent and geometry aware) human bodies using a GAN that is trained in an unsupervised manner on datasets containing single-view images. To this end, we adopt the recently proposed tri-plane feature representation [1], which is extremely efficient for training and rendering radiance fields, while being compatible with conventional 2D CNN–based generators, such as StyleGAN [22]. While this framework has been successfully demonstrated for faces in prior work, we are the first to adapt it to generating radiance fields of full human bodies. Secondly, we tackle the editability of the generated radiance fields by introducing an explicit radiance field deformation step as part of our GAN training procedure. This step ensures that the generator synthesizes radiance fields of people in a canonical body pose, which is then explicitly warped according to the body pose distribution of the training data. We show that this new approach generates high-quality, editable, multi-view-consistent human

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

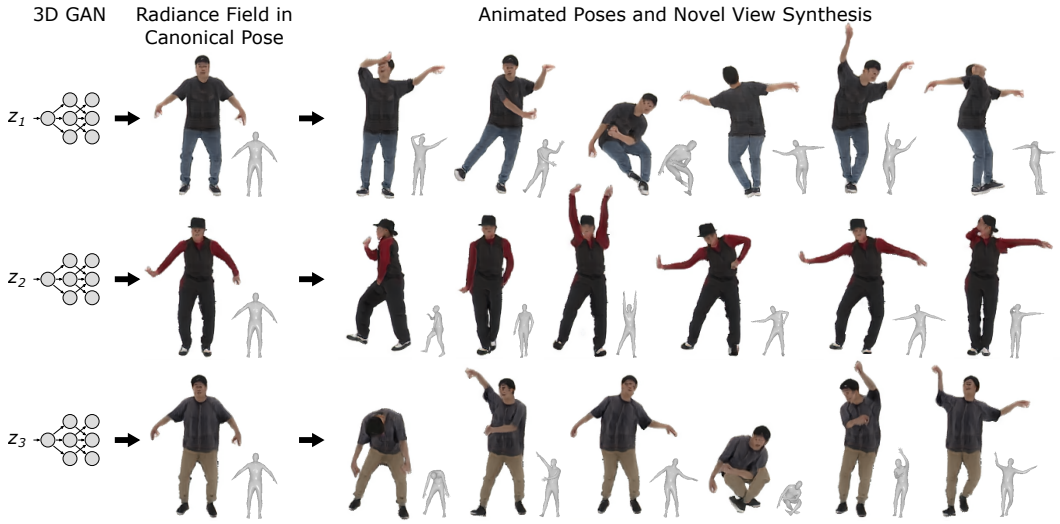

Figure 1: Our method, GNARF, maps a latent space to radiance fields representing human identities. These generated humans can then be animated and rendered from novel views. Qualitative results for our method trained on the AIST++ dataset [23] are shown here.

bodies and that our approach can also be applied to editing faces, increasing the controllability of existing generative models for this task (see Fig. 1).

To summarize, the contributions of our approach are:

- We present a 3D-aware GAN framework for the generation of editable radiance fields of human bodies. To our knowledge, this is the first approach of its kind.
- Our framework introduces an efficient neural representation for articulated objects, including bodies and heads, that combines the recently proposed tri-plane feature volume representation with an explicit feature volume deformation that is guided by a template shape.
- We demonstrate high-quality results for unconditional generation and animation of human bodies using the SURREAL and AIST++ datasets and faces using the FFHQ dataset.

2 Related Work

Articulated 3D Representations. Parametric shape templates are one of the most common types of articulated 3D representations adopted in recent neural scene representation and rendering approaches. These templates, including faces [24, 25], bodies [26], hands [27], or a combination of these parts [28], and even animals [29], have been widely utilized for pose estimation and reconstruction, e.g. [30–34]. Volume deformation is also commonly used in computer graphics, for example using mean value coordinates (MVC) [35] or biharmonic coordinates [36] for shape deformation and editing [37–40]. GNARF combines parametric template shapes, such as FLAME [25] for heads and SMPL [26] for bodies, with an intuitive surface-driven volume deformation approach that is both computationally efficient and qualitatively comparable to or better than both skinning and MVC-based deformation. This provides intuitive editing control for articulated radiance field deformation.

Neural Radiance Fields. Coordinate networks, also known as neural fields [41], have emerged as a powerful tool that enable differentiable representations of 3D scenes [42–55] and learning view-dependent neural radiance fields [56–82]. While initial proposals have focused on static scenarios, recent work has demonstrated successful representations of dynamic scenes [83–90]. Articulated neural radiance fields further extend these approaches by providing editability for neural representations of human heads [91–96] and bodies [9–21] or animals [97], often by deforming the underlying radiance fields using traditional 3D morphable models or skeleton-based parameterizations, or alternatively conditioning the radiance field decoder with pose-related parameters. A more detailed survey of static, dynamic, and articulated neural radiance fields can be found in the recent state-of-the-art report by Tewari et al. [98]. Note that all of these techniques are supervised with scene-specific multi-image data and focus on representing, i.e., "overfitting", a single scene. Therefore, it is not

easily possible to train these models using unstructured 2D image data and then apply them to generate and edit new and unseen objects or humans.

Generative 3D-aware Radiance Fields. Building on the success of 2D image-based GANs [22, 99–101], recent efforts have focused on training 3D-aware multi-view consistent GANs from collections of single-view 2D images in an unsupervised manner. Achieving this challenging goal requires a combination of a neural scene representation and differentiable rendering algorithm. Recent work in this domain builds on representations using meshes [102, 103], dense [104–109] or sparse [110] voxel grids, multiple planes [2], fully implicit networks [3–7], or a combination of low-resolution voxel grids combined with 2D CNN-based image upsampling layers [111, 112]. Our 3D GAN architecture is most closely related to the recent work by Chan et al. [1], which uses an efficient tri-plane-based volume representation combined with neural volume rendering. We extend this work by including an explicit deformation field in our GAN architecture to model diverse articulations, which allows the generator to synthesize radiance fields of human bodies or heads in a canonical pose while being supervised by 2D image collections that contain arbitrary pose distributions. Explicitly disentangling radiance field generation and deformation enables us to drastically improve the quality of generated human bodies and faces when their poses or facial expressions are edited.

HeadNerf [113] is related to our approach in that they generate 3D heads conditioned on various attributes, e.g. facial expressions. However, to disentangle individual attributes they need to acquire training images of the same person performing various expressions in different lighting conditions. In contrast, and similar to other 3D GANs, our approach only requires single-view images of different people and can therefore work with readily available 2D image collections. The recent work by Grigorev et al. [114] also generates human bodies. However, their work proposes a 2D GAN that generates textures which are used in combination with a standard articulated template mesh whereas we aim at generating and editing radiance fields using a 3D GAN. The concurrent and unpublished work of Noguchi et al. [115] is closest to ours as the method also includes a radiance field deformation step in a tri-plane-based 3D GAN. Our evaluation shows superior generation quality and, more importantly, while their approach ties the network architecture to the specific choice of skeleton, our surface-driven volume deformation is agnostic to the particular choice of template and can be used with human bodies, faces, or other object types.

3 Generative Articulated Neural Radiance Fields

GNARF is a novel general framework to train 3D-aware GANs for deformable objects that have a parametric template mesh, e.g. human bodies and faces. It builds on the efficient tri-plane feature representation [1] for the generated neural radiance field, but additionally applies an explicit deformation which alleviates the requirement for the generator to learn a complicated distribution of articulations. As a result, the generator automatically learns to generate radiance fields of objects in the canonical pose, which are then warped explicitly to produce target body poses and facial expressions in a fully controllable and interpretable manner.

3.1 Modeling Articulated Radiance Fields

We first discuss our approach to modeling and rendering articulated radiance fields, before describing how this is integrated into the 3D GAN in Sec. 3.2.

Scene representation. To represent an object, we leverage the recently proposed tri-plane feature representation [1]. This representation uses three axis-aligned 2D feature planes, each with resolution $N \times N \times C$, where N and C denote the spatial resolution and number of channels. The feature of any 3D point $\mathbf{x} \in \mathbb{R}^3$ is queried by projecting $\mathbf{x}$ onto the planes, retrieving three feature vectors via bilinear interpolation, and aggregating the vectors by summation, i.e., $F(\mathbf{x}) = F_{xy}(\mathbf{x}) + F_{yz}(\mathbf{x}) + F_{xz}(\mathbf{x})$ where $F_{ij} : \mathbb{R}^3 \mapsto \mathbb{R}^C$ is a function mapping 3D coordinates to features on the ij plane via projection and interpolation.

Articulated deformation. We use the deformation function $D : \mathbb{R}^3 \mapsto \mathbb{R}^3$ (detailed later) to warp a coordinate $\mathbf{x}$ from the target (deformed) space into the canonical space. Using a small multilayer perceptron $\text{MLP} : \mathbb{R}^C \mapsto \mathbb{R}^4$, we convert the deformed 3D feature volume into a neural field of

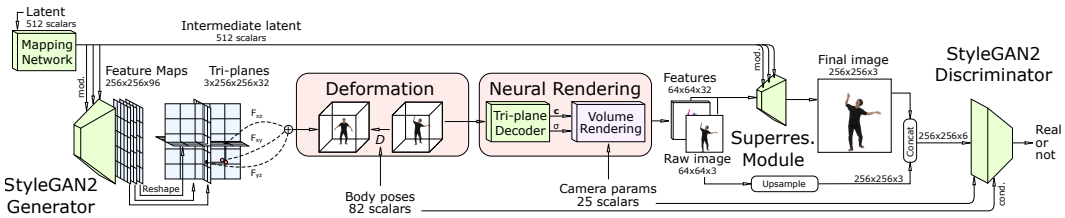

Figure 2: Illustration of the GNARF pipeline, including the StyleGAN2 generator, the tri-plane feature representation, feature volume deformation, neural volume rendering, image super-resolution as well as camera view and body-pose conditioned dual discrimination. The resolution of intermediate data and the final image is indicated for experiments with the AIST++ dataset.

spatially varying RGB colors $\mathbf{c}$ and volumetric densities σ as

$$\left(\mathbf{c}\left(\mathbf{x}\right), \sigma\left(\mathbf{x}\right)\right) = \mathrm{MLP}\left(\left(F_{xy} \circ D\right)\left(\mathbf{x}\right) + \left(F_{yz} \circ D\right)\left(\mathbf{x}\right) + \left(F_{xz} \circ D\right)\left(\mathbf{x}\right)\right). \tag{1}$$

There are many possible choices for how to specify the deformation field in an intuitive manner. For example, linear blend skinning can be used to deform the entire volume "rigged" by a skeleton (see e.g. [8]). While skinning is popular for human body articulations, it cannot explain subtle deformation due to varying facial expressions. Another option is to use the object-specific template mesh as a cage and apply cage-based deformation for the entire volume using mean value coordinates (MVC) [35]. However, the high computational cost of evaluating MVCs on the full-resolution grid (see Tab. 1) is prohibitive for GAN training and, more critically, this approach generally leads to severe artifacts when the template mesh (accidentally) includes self-intersections (see supplement).

To alleviate these problems, we use an intuitive surface-driven deformation method, which we label as the Surface Field (SF) method. This method only requires canonical and target template meshes with correspondences, which are readily available for faces [25] and bodies [26]. These template shapes, in turn, can be driven using skeletons, manual editing, or using keypoints or landmarks that could be detected in and transfered from videos of other people. Therefore, the SF method is general sufficient to apply to different body parts and it can be intuitively edited in a number of ways.

The SF approach assigns each 3D coordinate $\mathbf{x}$ to its nearest triangle $t^{\mathrm{D}}_{\mathbf{x}} = [\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2]$ on the target (deformed) mesh. We compute the barycentric coordinates $[u, v, w]$ of the coordinate projected onto this triangle and find the corresponding triangle on the canonical mesh $t^{\mathrm{C}}_{\mathbf{x}}$ and its normal $\mathbf{n}^{\mathrm{C}}_{t_{\mathbf{x}}}$. The deformed coordinate can then be computed as:

$$D\left(\mathbf{x}\right) = t^{\mathrm{C}}_{\mathbf{x}}\left[u, v, w\right]^{\mathsf{T}} + \left(\mathbf{x} - t^{\mathrm{D}}_{\mathbf{x}}\left[u, v, w\right]^{\mathsf{T}}\right)\mathbf{n}^{\mathrm{C}}_{t_{\mathbf{x}}}, \tag{2}$$

The SF approach is very fast to compute and mitigates artifacts from self-intersections of the template shape, thereby combining the benefits of linear blend skinning and MVC-based approaches for the task of radiance field deformation.

Rendering deformed radiance fields. We render the radiance field using (neural) volume rendering [61, 116]. For this purpose, the color $C(\mathbf{r})$ of a ray $\mathbf{r}$ is computed by integrating the volumetric color $\mathbf{c}$ and density σ as

$$\mathbf{C}(\mathbf{r}) = \int_{t_n}^{t_f} T\left(t\right)\sigma\left(\mathbf{r}(t)\right)\mathbf{c}\left(\mathbf{r}(t)\right)\mathrm{d}t, \quad T\left(t\right) = \exp\left(-\int_{t_n}^{t_f}\sigma\left(\mathbf{r}\left(s\right)\right)\mathrm{d}s\right), \tag{3}$$

where t_n and t_f indicate near and far bounds along the ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ pointing from its origin $\mathbf{o}$ into direction $\mathbf{d}$. The volume rendering equation (eq. 3) is typically approximated using numerical methods, such as the quadrature rule [116].

3.2 3D GAN Framework

An overview of our 3D pipeline is shown in Fig. 2. Several components, including the StyleGAN generator, the tri-plane representation, the volume rendering, the CNN-based image super-resolution module, and (dual) discrimination are directly adopted from the EG3D framework [1].

Instead of directly generating the radiance field with the target body pose or facial expression, however, GNARF is unique in generating the radiance field in a canonical pose and then applying the deformation field discussed in the previous section to warp the feature volume. We additionally remove the pose conditioning on the generator, and only use camera pose and body pose conditioning in the discriminator. This removes the ability for the generator to incorporate any knowledge about the final view or pose in the canonical radiance field generation, ensuring that the generated results will be robustly animatable beyond just the image rendered at training time. However, the discriminator having access to this information ensures that the GAN learns to generate warping accurate to a target pose rather than just being in the correct distribution. Finally, we adopt a radiance field rendering strategy which samples along each ray inside of an expanded template mesh. This ensures that the integration samples are taken in regions of the radiance field with the most detail and not taken in empty space, simultaneously improving the quality of the generated results and speeding up training.

Additional implementation details are included in the supplement. Source code and pre-trained models will be released to the public.

4 Experiments

We first evaluate the proposed deformation field by overfitting a single representation on a single dynamic full body scene. Then we apply this deformation method in a GAN training pipeline for both bodies (AIST++ [23] and SURREAL [117]) and faces (FFHQ) [101]. Training details and hyper-parameters are discussed in the supplement.

AIST++ is a large dataset consisting of 10.1M images capturing 30 performer in dance motion. Each frame is annotated with ground truth camera and fitted SMPL body model. SURREAL contains 6M images of synthetic humans created by SMPL body model in various poses rendered in indoor scenes. FFHQ is a large dataset of high-resolution images of human faces collected from Flickr. All images have licenses that allow free use and redistribution and adaptation for non-commencial use.

4.1 Single-scene Overfitting

We compare the proposed surface-driven deformation method, SF, with two alternative methods, MVC and skinning, in a single-scene overfitting task. MVCs require a set of weights (called the mean value coordinates) to be computed w.r.t. every vertex of the target mesh $\mathcal{M}^{\mathrm{D}}$ for every sample point. The sample point is then deformed into the canonical pose by linearly combining the vertices of the canonical mesh $\mathcal{M}^{\mathrm{C}}$ with these computed weights. In skinning, the sampling points are deformed to the canonical pose by the rigid transformation of the closest bone as measured by point to line-segment distance. We find this simplified definition of skinning effective in avoiding blending between two topologically distant body parts (e.g., hand and pelvis) if the starting pose brings them to a geometric proximity.

We select a multi-view video sequence from the AIST++ dataset [23] and optimize tri-plane features in the canonical pose using a subset of the views and frames for supervision. We then evaluate the quality of the estimated radiance field warped into these training views and poses but also into held-out test views and poses. We apply several modifications to the tri-plane architecture to reduce overfitting; details regarding these changes as well as the selection of training and evaluation views are provided in the supplemental material.

To speed up MVC and SF computation, we decimate the source and deformed SMPL mesh using Quadric Error Metric Decimation [118] in the Open3D library [119] from the original 13,776 faces to 1,376 faces, while tracking the correspondence between the source and deformed meshes. Nonetheless computing MVC for each deformed pose is still prohibitively expensive for online training (3.7 s per example). We thus precompute the deformation for training and test body poses on a fixed 16^3 grid and retrieve the deformation for arbitrary sampling points using trilinear interpolation.

As shown in Tab. 1, our SF method outperforms the others for both training and test images. MVC performs worst, partially due to the grid approximation, which is essential in practice. The skinning method is comparable to SF in terms of image quality but it is $3\times$ slower. Moreover, skinning cannot sufficiently deform subtle facial expressions. Therefore, the SF approach is the most flexible among these deformation methods by being compatible with different human body parts while also offering computational and memory efficiency.

	Training images			Test images			Run time [ms] ↓
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	
Skinning	18.8	0.942	0.060	17.9	**0.940**	0.067	95.6
MVC [35]	18.1	0.937	0.067	17.2	0.934	0.074	0.2 (3782.1)
Surface Field	**19.0**	**0.943**	**0.058**	**18.0**	**0.940**	**0.065**	**31.6**

Table 1: Single-scene overfitting. We evaluate three deformation approaches for the task of estimating a single radiance field in a canonical body pose supervised by a video sequence showing a person from different views and in different poses. The SF approach achieves the best quality for both training and unseen test images while being the fastest. Note that the MVC method is only faster than SF when using precomputed grid at the cost of significantly lower deformation accuracy (the runtime without such approximation is reported in parentheses). The timings are measured to deform a single feature volume on an RTX3090 graphics processing unit.

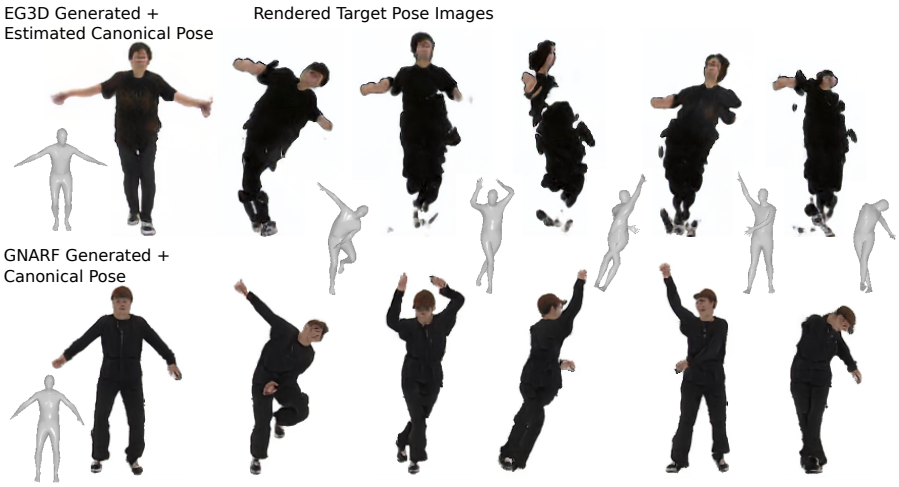

Figure 3: Qualitative comparison of generated target poses using our model vs. warping a pre-trained EG3D model on the AIST++ dataset.

4.2 Human Body Generation and Animation

We now use our SF approach as the deformation method for feature volumes generated by GNARF. Our method is trained and evaluated on the captured AIST++ [23] and the synthetic SURREAL [117] datasets. For both datasets, our method generates high-quality multi-view consistent human bodies in diverse poses that closely match the target pose.

Baselines & Evaluation. Because GNARF is the first method to learn a generative model of radiance fields representing bodies, we propose a baseline where we use the original EG3D trained without deformation to generate a feature volume (not in the canonical pose) then warp it into various target poses during inference time using the proposed SF deformation method. Without the feature volume deformation, the generator is forced to learn to model both identity and pose in its latent space. As a result, the tri-plane features no longer represent a human body in a consistent canonical pose, but rather match the distribution of poses in the dataset. The animation of generated bodies is similar to our proposed approach, except that the generated (arbitrarily posed) human body is used as the canonical pose, for which we obtain a SMPL mesh by applying the human shape reconstruction method SPIN [31]. Additionally, we included the concurrent work ENARF-GAN and its variant ENARF-VAE [115] in the evaluation whenever a fair and standardized comparison is possible.

We compare FID scores to evaluate the quality and diversity of generated images as well as the Percentage of Correct Keypoints (PCK) metric to evaluate the quality of the animation. PCK computes the percentage of 2D keypoints detected on a generated rendered image that are within an error threshold (half the size of the head in the case of PCKh@0.5) of keypoints on a ground truth image in the same pose and view. We use an off-the-shelf body keypoint estimator [120] trained on MPII [121]

	AIST++ @256^2		SURREAL @128^2		
	FID (50k) ↓	PCKh@0.5 ↑	FID (10k) ↓	FID (50k) ↓	PCKh@0.5 ↑
ENARF-VAE [115]*	—	—	63.0	—	—
ENARF-GAN [115]*	—	—	21.3	—	—
EG3D (no warping)	8.3	—	**9.2**	**8.2**	—
EG3D (+ pose est. & re-warping)	66.5	0.855	79.6	78.8	0.785
GNARF	**7.9**	**0.980**	11.9	10.9	**0.913**

Table 2: Quantitative evaluation of our GAN trained on SURREAL [117] and AIST++ [23]. Our method only considers foreground images (backgrounds masked).

publically available on the MMPose Project [122] to estimate keypoints from a corresponding ground truth and generated image.

AIST++. AIST++ is a challenging dataset as the body poses are extremely diverse. We collect 30 frames per video as our training data after filtering out frames whose camera distance is above a threshold or the human bounding box is partially outside the image. Then we extract the human body by cropping a 600×600 patch centered at the pelvis joint, and resize these frames to 256×256. Since this dataset does not provide ground truth segmentation masks, we use a pre-trained segmentation model [123] to remove backgrounds to stabilize the GAN training. To speed up training, we use GAN transfer learning [124]. Rather than initializing our network weights randomly, we begin training from a pre-trained EG3D [1] model. Fine-tuning allows for quicker convergence and saves computational resources during training.

As shown in Tab. 2, we can see that our method outperforms the naive reanimation of EG3D by a large margin (see the 4th row vs. the 5th row). Furthermore, our animated method also generates better images than the EG3D baseline that does not support animations. This is likely due to GNARF allowing the generator to focus on generating a specific identity in a canonical pose instead of learning both identity and complex pose distribution in a combined latent space. Similarly, our model enables high-quality re-animation, as demonstrated by the PCKh@0.5 metric. In Fig. 3, our method produces significantly better qualitative results than those generated by the baseline. The baseline results using re-warped EG3D are significantly degraded, since it is difficult to accurately estimate the SMPL mesh from the generated images, which we use as the canonical pose in the deformation function. Additionally, floating artifacts which exist in the radiance field outside of camera views and make no difference in the conventionally rendered images become visible after being warped can cause false occlusions. In Fig. 1, we show that our method generates bodies in a canonical pose with diverse identities. Additionally, we show that by changing the SMPL parameters, we can drive each radiance field to a desired target pose and render at an arbitrary novel view.

SURREAL. We also test our method on the SURREAL dataset. The training data is extracted from the first frame of each video in SURREAL's training split. Each frame is cropped based on the provided segmentation mask from head-to-toe, and resized to 128×128. We filter out the backgrounds and set them to be black. The SURREAL dataset provides ground truth SMPL parameters and camera intrinsics and extrinsics for each frame. Similarly to AIST++, we use transfer learning from a pre-trained EG3D model at the appropriate resolution.

Figure 4: Example of generated humans in canonical pose and in target pose using model trained on SURREAL dataset.

In Tab. 2, we see that our method trained with deformation produces similar but slightly degraded FID scores as the version trained without SF feature volume deformation (first row). A possible explanation is that the much lower diversity of poses in the SURREAL dataset and the lack of details allow the EG3D model to learn both the distribution of identity and body poses effectively in the latent space. However, attempting to deform generated radiance fields results in an immense degradation in quality, shown by both the FID and PCKh@0.5 metrics (second row).

Additional qualitative results and evaluations are included as part of the supplemental material.

	FID (500) ↓	FID (50k) ↓	AED ↓	APD ↓	ID-Consistency ↑
EG3D (+3DMM est. & re-warping)	22.9	11.6	0.29	0.028	**0.81**
PIRenderer [126]	64.4	—	0.28	0.040	0.70
3D GAN inversion [127]	31.2	—	0.36	0.039	0.73
GNARF	**17.9**	**6.6**	**0.23**	**0.025**	0.80

Table 3: Quantitative comparison on FFHQ dataset [101].

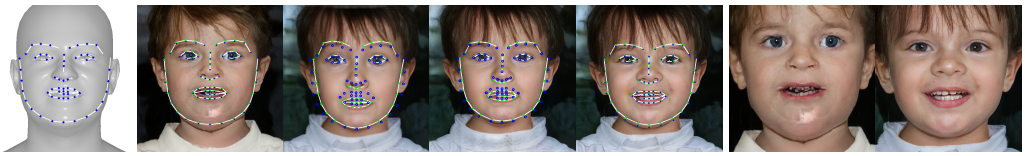

Figure 5: Deformation fidelity on FFHQ dataset. We plot the target and detected facial landmarks in blue and green, respectively, to visualize the expression fidelity. PIRender [126] and 3D GAN inversion [127] both show large discrepancy to the target face, while the baseline EG3D has strong artifacts from warping (right). Our generated results show high visual quality and they align accurately with the driving deformation.

4.3 Human Face Generation and Editing

Thanks to the surface-driven deformation method, SF, our method can directly utilize expressive parametric face models to drive subtle deformations. We use the FLAME [25] head model and apply the state-of-the-art facial reconstruction method DECA [125] to estimate the flame parameters (100D identity vectors, 6D joint position vectors and 50D expression vectors) from the training dataset. Since DECA does not account for eyeball movement, we remove the eyeballs from the original FLAME template. Additionally, we add triangles to close the holes at the neck and mouth area in the FLAME template, which improves the consistency of the SF deformation. These preprocessing steps produce a mesh with 3,741 vertices and 7,478 faces. Finally we apply the same decimation method as in the body experiments to obtain a coarse mesh (2,500 triangles), which we use during the GAN training for faster SF deformation.

To train GNARF, we start with a conventional EG3D model pre-trained using the FFHQ dataset (see [1] for details on the training details and data processing). We then fine-tune this model using the proposed framework, which includes the SF deformation module. Note that we apply a global scaling and translation to the FLAME template mesh to roughly align it to the faces generated by the pre-trained EG3D model.

For evaluation, our first baseline is the original EG3D model with re-warping at inference time, as described in Sec. 4.2. We also include two state-of-the-art facial reenactment methods, PIRenderer [126] and 3D GAN inversion [127]. Several metrics are used by our quantitative evaluation. FID (500) follows the evaluation protocol of Lin et al. [127], where the ground truth dataset consists of 500 randomly sampled identities and the test dataset is constructed by animating the ground truth using randomly sampled target poses. FID (50k) follows the protocol in EG3D, where the entire FFHQ dataset is treated as the ground truth and the test dataset includes 50k generated images using randomly sampled latent vectors, camera poses and FLAME facial parameters. Following [127], we evaluate the fidelity of the animation with the Average Expression Distance (AED) and the Average Pose Distance (APD) computed using DECA estimation, as well as the identity consistency based on a face recognition model [128].

As shown in Tab. 3, our method is superior to the baseline methods in both FIDs and in the edited fidelity, and comparable to the baseline EG3D in terms of identity consistency. The advantage of our method in editing ability is clearly demonstrated in Fig. 5: our method reproduces the target expression more accurately when compared to the face reenactment methods and mitigates the warping artifacts present in the baseline EG3D. In Fig. 6, we further demonstrate the quality of our results. Notice that even though the FLAME model does not include teeth, the SF deformation guides the neural radiance field to construct teeth consistently at the correct location.

Figure 6: Qualitative results on FFHQ dataset. We show each identity in their canonical (left) and a different (right) expression from two different camera poses. Our method shows excellent multi-view consistency.

5 Discussion

Limitations and future work. Our work is not without limitations. The level of detail in the generated bodies, for example, is relatively low. This is partly due to the limited resolution of the training data in the SURREAL and AIST++ datasets, but also due to the limited resolution that the tri-plane representation offers for any one body part, such as the face. An interesting avenue of future work could include the exploration of adaptive radiance field resolution for human bodies, allocating more resolution to salient parts (see e.g. [129]). The image quality of the generated bodies is currently on par with simpler approaches that only need to generate an RGB texture on the SMPL mesh [114]. Yet, details in faces and hair cannot be handled by a texture-generating approach and we expect the quality of generative radiance fields to surpass that of simpler alternatives with increasing and perhaps adaptive radiance field resolutions. Another bottleneck of our current framework is the challenge of being able to work with large-scale datasets showing a diversity of visible humans. In-the-wild datasets, such as MSCOCO [130], do contain this diversity but also contain a significant amount of occlusion and detailed backgrounds, requiring the generator to spend capacity on modeling the distribution of backgrounds and occlusions themselves. An interesting avenue of future work consists of explicitly modeling occlusion and background in the GAN training pipeline independently from identity and pose. Additionally, pose estimation from in-the-wild images is still not very accurate, degrading the quality of our deformation function. On the other hand, large-scale custom curated datasets such as the one used in InsetGAN [129] are not publicly available. Finally, the deformation we use as part of the 3D GAN training is limiting in several ways: it does not allow for topology changes, it does not prevent solid parts like teeth or eyeglasses from being stretched, and the deformation quality degrades for points far from the surface. More advanced deformation methods, perhaps including kinematic constraints, could be an interesting direction of future research.

Ethical considerations. GANs could be misused for generating edited imagery of real people. Such misuse of image synthesis techniques poses a societal threat, and we do not condone using our work with the intent of spreading disinformation. We also recognize a potential lack of diversity in our results, stemming from implicit biases of the datasets we process.

Conclusion. Our work takes important steps towards photorealistic 3D-aware image synthesis of articulated human bodies and faces with applications in visual effects, virtual or augmented reality, and teleconferencing among others.

References

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

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

[3] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: Generative radiance fields for 3D-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[4] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[5] 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.

[6] 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.

[7] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. Fenerf: Face editing in neural radiance fields. In *CVPR*, 2022.

[8] Alec Jacobson, Zhigang Deng, Ladislav Kavan, and JP Lewis. Skinning: Real-time shape deformation. In *ACM SIGGRAPH 2014 Courses*, 2014.

[9] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021.

[10] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021.

[11] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *NeurIPS*, 2021.

[12] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. In *NeurIPS*, 2021.

[13] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *ICCV*, 2021.

[14] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. In *NeurIPS*, 2021.

[15] Haotong Lin, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields with learned depth-guided sampling. In *arXiv*, 2021.

[16] Tao Hu, Tao Yu, Zerong Zheng, He Zhang, Yebin Liu, and Matthias Zwicker. Hvtr: Hybrid volumetric-textural rendering for human avatars. In *arXiv*, 2021.

[17] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021.

[18] Tianhan Xu, Yasuhiro Fujita, and Eiichi Matsumoto. Surface-aligned neural radiance fields for controllable 3d human synthesis. In *CVPR*, 2022.

[19] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. *arXiv preprint arXiv:2112.02789*, 2021.

[20] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, 2022.

[21] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *CVPR*, 2022.

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[23] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021.

[24] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *ACM SIGGRAPH*, page 187–194, 1999.

[25] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph. (SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017.

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015.

[27] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. Graph. (SIGGRAPH Asia)*, 36(6), November 2017.

[28] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019.

[29] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *CVPR*, pages 6365–6373, 2017.

[30] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018.

[31] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019.

[32] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019.

[33] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020.

[34] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020.

[35] Tao Ju, Scott Schaefer, and Joe Warren. Mean value coordinates for closed triangular meshes. In *ACM SIGGRAPH*, pages 561–566. 2005.

[36] Pushkar Joshi, Mark Meyer, Tony DeRose, Brian Green, and Tom Sanocki. Harmonic coordinates for character articulation. *ACM Trans. Graph. (SIGGRAPH)*, 26(3), 2007.

[37] Wang Yifan, Noam Aigerman, Vladimir G Kim, Siddhartha Chaudhuri, and Olga Sorkine-Hornung. Neural cages for detail-preserving 3d deformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 75–83, 2020.

[38] Minhyuk Sung, Zhenyu Jiang, Panos Achlioptas, Niloy J. Mitra, and Leonidas J. Guibas. Deformsyncnet: Deformation transfer via synchronized shape deformation spaces. *ACM Transactions on Graphics (ToG)*, 39(6), nov 2020. ISSN 0730-0301. doi: 10.1145/3414685.3417783. URL `https://doi.org/10.1145/3414685.3417783`.

[39] Tomas Jakab, Richard Tucker, Ameesh Makadia, Jiajun Wu, Noah Snavely, and Angjoo Kanazawa. Keypointdeformer: Unsupervised 3d keypoint discovery for shape control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12783–12792, 2021.

[40] Minghua Liu, Minhyuk Sung, Radomir Mech, and Hao Su. Deepmetahandles: Learning deformation meta-handles of 3d meshes with biharmonic coordinates. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12–21, 2021.

[41] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *arXiv*, 2021.

[42] 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.

[43] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[44] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[45] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[46] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[47] Matan Atzmon and Yaron Lipman. SAL: Sign agnostic learning of shapes from raw data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[48] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *International Conference on Machine Learning (ICML)*, 2020.

[49] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, and Thomas Funkhouser. Local implicit grid representations for 3D scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[50] Thomas Davies, Derek Nowrouzezahrai, and Alec Jacobson. Overfit neural networks as a compact shape representation. *arXiv preprint arXiv:2009.09808*, 2020.

[51] Rohan Chabra, Jan Eric Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local SDF priors for detailed 3D reconstruction. In *European Conference on Computer Vision (ECCV)*, 2020.

[52] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision (ECCV)*, 2020.

[53] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[54] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[55] Julien N.P. Martel, David B. Lindell, Connor Z. Lin, Eric R. Chan, Marco Monteiro, and Gordon Wetzstein. ACORN: Adaptive coordinate networks for neural representation. *ACM Transactions on Graphics (SIGGRAPH)*, 2021.

[56] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3D-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[57] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. DeepVoxels: Learning persistent 3D feature embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[58] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[59] Shichen Liu, Shunsuke Saito, Weikai Chen, and Hao Li. Learning to infer implicit surfaces without 3D supervision. *arXiv preprint arXiv:1911.00767*, 2019.

[60] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Transactions on Graphics (SIGGRAPH)*, 2019.

[61] 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 *European Conference on Computer Vision (ECCV)*, 2020.

[62] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the wild: Neural radiance fields for unconstrained photo collections. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[63] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[64] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

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

[66] Thomas Neff, Pascal Stadlbauer, Mathias Parger, Andreas Kurz, Joerg H. Mueller, Chakravarty R. Alla Chaitanya, Anton S. Kaplanyan, and Markus Steinberger. DONeRF: Towards Real-Time Rendering of Compact Neural Radiance Fields using Depth Oracle Networks. *Computer Graphics Forum*, 40(4), 2021.

[67] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. DIST: Rendering deep implicit signed distance function with differentiable sphere tracing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[68] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[69] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[70] David B Lindell, Julien NP Martel, and Gordon Wetzstein. AutoInt: Automatic integration for fast neural volume rendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[71] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[72] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. SDFDiff: Differentiable rendering of signed distance fields for 3D shape optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[73] Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W. Taylor, and Joshua M. Susskind. Unconstrained scene generation with locally conditioned radiance fields. *arXiv preprint arXiv:2104.00670*, 2021.

[74] Petr Kellnhofer, Lars Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. Neural lumigraph rendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[75] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. KiloNeRF: Speeding up neural radiance fields with thousands of tiny MLPs. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[76] Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[77] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. PlenOctrees for real-time rendering of neural radiance fields. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[78] Peter Hedman, Pratul P. Srinivasan, Ben Mildenhall, Jonathan T. Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[79] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. FastNeRF: High-fidelity neural rendering at 200fps. *arXiv preprint arXiv:2103.10380*, 2021.

[80] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-NeRF: Scalable large scene neural view synthesis. *arXiv*, 2022.

[81] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv:2201.05989*, January 2022.

[82] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. *Eurographics Association*, 2020.

[83] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. *arXiv preprint arXiv:2011.13961*, 2020.

[84] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, 2021.

[85] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *CVPR*, pages 9421–9431, 2021.

[86] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021.

[87] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph. (SIGGRAPH Asia)*, 40(6), 2021.

[88] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *ICCV*, 2021.

[89] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *CVPR*, 2021.

[90] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Trans. Graph.*, 40(4), 2021.

[91] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *CVPR*, 2021.

[92] Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhofer. Learning compositional radiance fields of dynamic human heads. In *CVPR*, pages 5704–5713, 2021.

[93] Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *ICCV*, 2021.

[94] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M Avatar: Implicit morphable head avatars from videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2022.

[95] Kacper Kania, Kwang Moo Yi, Marek Kowalski, Tomasz Trzciński, and Andrea Tagliasacchi. CoNeRF: Controllable Neural Radiance Fields. In *CVPR*, 2022.

[96] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. *arXiv preprint arXiv:2112.02308*, 2021.

[97] Gengshan Yang, Minh Vo, Neverova Natalia, Deva Ramanan, Vedaldi Andrea, and Joo Hanbyul. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022.

[98] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Yifan Wang, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, Tomas Simon, Christian Theobalt, Matthias Niessner, Jonathan T. Barron, Gordon Wetzstein, Michael Zollhoefer, and Vladislav Golyanik. Advances in neural rendering. In *arxiv*, 2021.

[99] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2016.

[100] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

[101] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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

[103] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3D controllable image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[104] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3D shape induction from 2D views of multiple objects. In *International Conference on 3D Vision*, 2017.

[105] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Joshua B. Tenenbaum, and William T. Freeman. Visual object networks: Image generation with disentangled 3D representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[106] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping Plato's cave: 3D shape from adversarial rendering. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[107] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. HoloGAN: Unsupervised learning of 3D representations from natural images. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[108] Thu Nguyen-Phuoc, Christian Richardt, Long Mai, Yong-Liang Yang, and Niloy Mitra. BlockGAN: Learning 3D object-aware scene representations from unlabelled images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[109] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *arXiv preprint arXiv:2112.10759*, 2021.

[110] Zekun Hao, Arun Mallya, Serge Belongie, and Ming-Yu Liu. GANcraft: Unsupervised 3D neural rendering of minecraft worlds. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[111] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing scenes as compositional generative neural feature fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[112] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. StyleNeRF: A style-based 3D-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021.

[113] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *CVPR*, 2022.

[114] Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. Stylepeople: A generative model of fullbody human avatars, 2021.

[115] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. In *arxiv*, 2022.

[116] N. Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 1995.

[117] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[118] Michael Garland and Paul S Heckbert. Surface simplification using quadric error metrics. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 209–216, 1997.

[119] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018.

[120] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[121] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014.

[122] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020.

[123] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*, 2020.

[124] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 218–234, 2018.

[125] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. volume 40, 2021. URL `https://doi.org/10.1145/3450626.3459936`.

[126] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *IEEE International Conference on Computer Vision (ICCV)*, pages 13759–13768, 2021.

[127] Connor Z. Lin, David B. Lindell, Eric R. Chan, and Gordon Wetzstein. 3d gan inversion for controllable portrait image animation, 2022.

[128] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[129] Anna Frühstück, Krishna Kumar Singh, Eli Shechtman, Niloy J. Mitra, Peter Wonka, and Jingwan Lu. Insetgan for full-body image generation, 2022.

[130] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We plan to release the code completely, but have not yet with submission. Details on the architecture, training, and data are available in the supplemental material.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See supplemental document and sections 4.2 and 4.3.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] GNARF models require a significant amount of computational resources (see supplemental document), and thus averaging training over many seeds would be computationally infeasible. However, we evaluate our trained models with a large number of generated images, demonstrating the robustness of our method.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See supplemental document.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [Yes] See Section 4.

(c) Did you include any new assets either in the supplemental material or as a URL? [No]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]