
Generative Neural Articulated Radiance Fields

Alexander W. Bergman*
Stanford University
awb@stanford.edu

Petr Kellnhofer*
TU Delft
p.kellnhofer@tudelft.nl

Wang Yifan*
Stanford University
yifan.wang@stanford.edu

Eric R. Chan*
Stanford University
erchan@stanford.edu

David B. Lindell
University of Toronto
Vector Institute
lindell@cs.toronto.edu

Gordon Wetzstein
Stanford University
gordonwz@stanford.edu

computationalimaging.org/publications/gnarf/

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 GNARE, 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

*Equal contribution

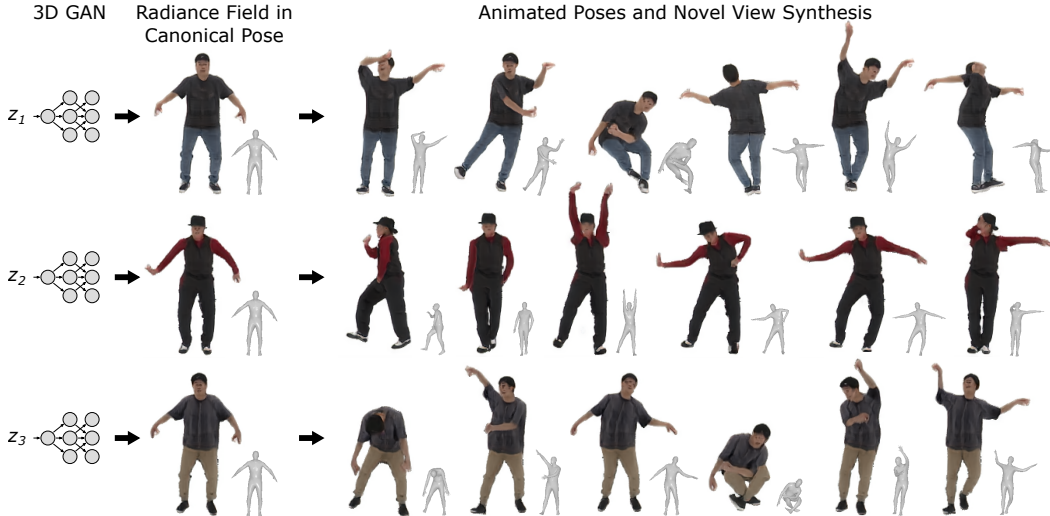


Figure 1: Our method, GNARF, maps a latent space to radiance fields representing human identities. These generated humans can then be animated and rendered from novel views. Qualitative results for our method trained on the AIST++ dataset [23] are shown here.

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 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 [37–41]. 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 [42], have emerged as a powerful tool that enable differentiable representations of 3D scenes [43–56] and learning view-dependent neural radiance fields [57–83]. While initial proposals have focused on static scenarios,

recent work has demonstrated successful representations of dynamic scenes [84–91]. Articulated neural radiance fields further extend these approaches by providing editability for neural representations of human heads [92–97] and bodies [9–21, 98, 99] or animals [100], 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. [101]. 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, 102–104], 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 [105, 106], dense [107–112] or sparse [113] 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 [114, 115]. 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 [116] is related to our approach in that they generate 3D heads conditioned on various attributes. Both HeadNeRF and GNARF condition on identity and facial expression independently, with the former also conditioning on illumination and albedo. 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. [117] 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 work of Noguchi et al. [118] 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

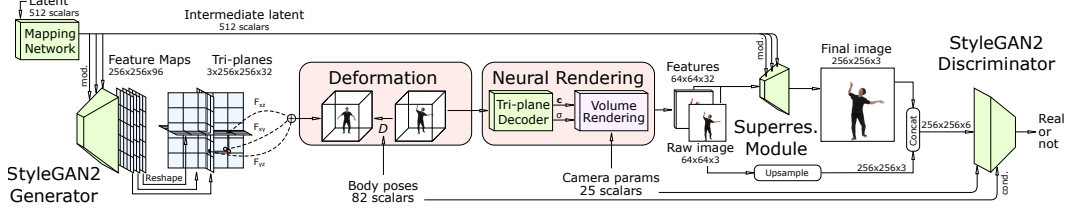


Figure 2: Illustration of the GNARF pipeline, including the StyleGAN2 generator, the tri-plane feature representation, feature volume deformation, neural volume rendering, image super-resolution as well as camera view and body-pose conditioned dual discrimination. The resolution of intermediate data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

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

$$(\mathbf{c}(\mathbf{x}), \sigma(\mathbf{x})) = \text{MLP}((F_{xy} \circ D)(\mathbf{x}) + (F_{yz} \circ D)(\mathbf{x}) + (F_{xz} \circ D)(\mathbf{x})). \quad (1)$$

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 transferred from videos of other people. Therefore, the SF method is generally sufficient to apply to different body parts and it can be intuitively edited in a number of ways, resulting in accurate volume deformation for our class of volumetric models.

The SF approach assigns each 3D coordinate \mathbf{x} to its nearest triangle $t_{\mathbf{x}}^D = [\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2] \in \mathbb{R}^{3 \times 3}$ 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 $t_{\mathbf{x}}^C$ and its normal $\mathbf{n}_{t_{\mathbf{x}}^C}$. The deformed coordinate can then be computed as

$$D(\mathbf{x}) = t_{\mathbf{x}}^C \cdot [u, v, w]^T + \left\langle \mathbf{x} - t_{\mathbf{x}}^D \cdot [u, v, w]^T, \mathbf{n}_{t_{\mathbf{x}}^D} \right\rangle \mathbf{n}_{t_{\mathbf{x}}^C}, \quad (2)$$

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 [62, 119]. For this purpose, the aggregated feature $F(\mathbf{r})$ of a ray \mathbf{r} is computed by integrating the volumetric features \mathbf{f} and density σ as

$$\mathbf{F}(\mathbf{r}) = \int_{t_n}^{t_f} T(t) \sigma(\mathbf{r}(t)) \mathbf{f}(\mathbf{r}(t)) dt, \quad T(t) = \exp\left(-\int_{t_n}^{t_f} \sigma(\mathbf{r}(s)) ds\right), \quad (3)$$

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 [119].

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. Thus, the generator depends only on the latent code controlling identity, which is input into a StyleGAN2 generator. This architectural choice takes advantage of the state-of-the-art 2D generative model architectures by using them to generate the tri-plane 3D representation. The discriminator having access to the camera and body poses ensure 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, source code, and pre-trained models can be found in the supplement or on our website.

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 [120]) and faces (FFHQ) [104]. Training details and hyper-parameters are discussed in the supplement.

AIST++ is a large dataset consisting of 10.1M images capturing 30 performers in dance motion. Each frame is annotated with a ground truth camera and fitted SMPL body model. SURREAL contains 6M images of synthetic humans created using SMPL body models 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, redistribution, and adaptation for non-commercial 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}^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}^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.

	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.

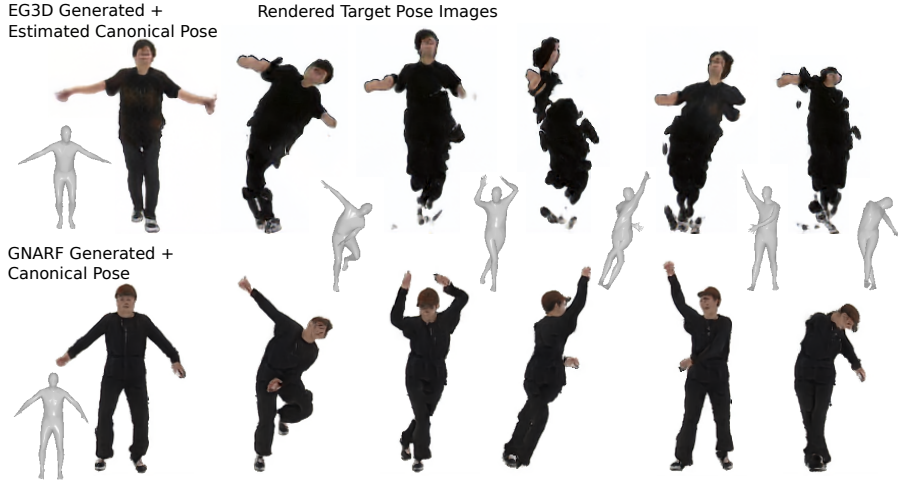


Figure 3: Qualitative comparison of generated target poses using our model vs. warping a pre-trained EG3D model on the AIST++ dataset.

To speed up MVC and SF computation, we decimate the source and deformed SMPL mesh using Quadric Error Metric Decimation [121] in the Open3D library [122] 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.

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 [120] 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

	AIST++ @256 ²		SURREAL @128 ²		
	FID (50k) ↓	PCKh@0.5 ↑	FID (10k) ↓	FID (50k) ↓	PCKh@0.5 ↑
ENARF-VAE [118]*	—	—	63.0	—	—
ENARF-GAN [118]*	—	—	21.3	—	<u>0.966</u>
EG3D (no warping)	<u>8.3</u>	—	<u>14.3</u>	<u>13.3</u>	—
EG3D (+ pose est. & re-warping)	66.5	0.855	163.9	162.2	0.348
GNARF	7.9	0.980	4.7	5.7	0.999

Table 2: Quantitative evaluation of our GAN trained on SURREAL [120] and AIST++ [23]. Our method only considers foreground images (backgrounds masked). *Metrics provided by authors of [118] after standardizing the evaluation protocol, which may differ from initial values in their original report.

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 [118] 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 [123] trained on MPII [124] publicly available on the MMPose Project [125] 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 [126] to remove backgrounds to stabilize the GAN training. To speed up training, we use GAN transfer learning [127]. 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.



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

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

	FID (500) ↓	FID (50k) ↓	AED ↓	APD ↓	ID-Consistency ↑
EG3D (+3DMM est. & re-warping)	<u>22.9</u>	<u>11.6</u>	0.29	<u>0.028</u>	0.81
PIRenderer [129]	64.4	—	0.28	0.040	0.70
3D GAN inversion [130]	31.2	—	0.36	0.039	0.73
GNARF	17.9	6.6	0.23	0.025	<u>0.80</u>

Table 3: Quantitative comparison on FFHQ dataset [104].

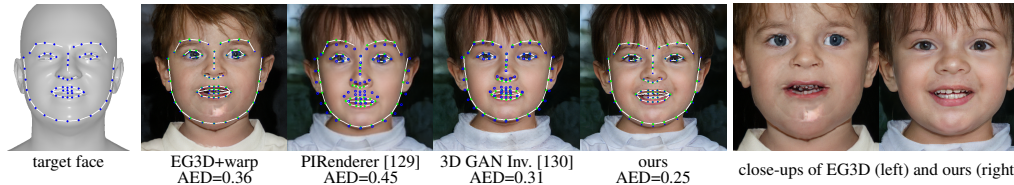


Figure 5: Deformation fidelity on FFHQ dataset. We plot the target and detected facial landmarks in blue and green, respectively, to visualize the expression fidelity. PIRender [129] and 3D GAN inversion [130] both show large discrepancy to the target face, while the baseline EG3D has strong artifacts from warping (right). Our generated results show high visual quality and they align accurately with the driving deformation.

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.

In Tab. 2, we see that our method trained with deformation produces improved FID scores as the version trained without SF feature volume deformation (5th vs 3rd row). Attempting to deform generated radiance fields results in an immense degradation in quality, shown by both the FID and PCKh@0.5 metrics (4th row).

Additional qualitative results and evaluations are included on the website.

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 [128] 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 [129] and 3D GAN inversion [130]. Several metrics are used by our quantitative evaluation. FID (500) follows the evaluation protocol of Lin et al. [130], 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 [130], we evaluate the fidelity of the animation with the Average Expression Distance (AED) and the Average



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.

Pose Distance (APD) computed using DECA estimation, as well as the identity consistency based on a face recognition model [131].

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.

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. [132]). 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 [117]. 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 [133], 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. Currently, the background is not modeled in the generative framework and requires pre-processing of the dataset to separate foreground from background. 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 [132] 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. In general, using the surface of a parametric mesh to guide an volume may not be the optimal choice for more complex volumetric scenes. 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.

Acknowledgments and Disclosure of Funding

Alexander W. Bergman was supported by a Stanford Graduate Fellowship. Gordon Wetzstein was supported by NSF Award 1839974, Samsung, Stanford HAI, and a PECASE from the ARO. We thank Connor Lin for helping standardize comparisons to [130]. We thank Atsuhiko Noguchi and the authors of [118] for helping in standardization of the comparisons of our methods.

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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [2] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *IEEE Conference on Computer Vision and Pattern Recognition (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 *IEEE Conference on Computer Vision and Pattern Recognition (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 *IEEE Conference on Computer Vision and Pattern Recognition (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 *IEEE Conference on Computer Vision and Pattern Recognition (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 *IEEE International Conference on Computer Vision (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 *Advances in Neural Information Processing Systems (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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [13] Atsuhiko Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.
- [14] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [15] Haotong Lin, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields with learned depth-guided sampling. *arXiv preprint arXiv:2112.01517*, 2021.
- [16] Tao Hu, Tao Yu, Zerong Zheng, He Zhang, Yebin Liu, and Matthias Zwicker. Hvtr: Hybrid volumetric-textural rendering for human avatars. *arXiv preprint arXiv:2112.10203*, 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 Transactions on Graphics (SIGGRAPH ASIA)*, 2021.
- [18] Tianhan Xu, Yasuhiro Fujita, and Eiichi Matsumoto. Surface-aligned neural radiance fields for controllable 3d human synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [21] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *IEEE Conference on Computer Vision and Pattern Recognition (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 *IEEE International Conference on Computer Vision (ICCV)*, 2021.
- [24] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of SIGGRAPH*, 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 Transactions on Graphics (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 Transactions on 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 Transactions on Graphics (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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [29] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 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)*, 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 *IEEE International Conference on Computer Vision (ICCV)*, 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [34] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [35] Tao Ju, Scott Schaefer, and Joe Warren. Mean value coordinates for closed triangular meshes. In *Proceedings of SIGGRAPH*. 2005.
- [36] Pushkar Joshi, Mark Meyer, Tony DeRose, Brian Green, and Tom Sanocki. Harmonic coordinates for character articulation. *ACM Transactions on Graphics (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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 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), 2020.

- [39] Tomas Jakab, Richard Tucker, Ameesh Makadia, Jiajun Wu, Noah Snavely, and Angjoo Kanazawa. Keypointdeformer: Unsupervised 3D keypoint discovery for shape control. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 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)*, 2021.
- [41] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Neural articulated shape approximation. In *European Conference on Computer Vision (ECCV)*, 2020.
- [42] 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 *Computer Graphics Forum*, volume 41, pages 641–676, 2022.
- [43] 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.
- [44] 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.
- [45] 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.
- [46] 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.
- [47] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [48] 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.
- [49] 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.
- [50] 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.
- [51] Thomas Davies, Derek Nowrouzezahrai, and Alec Jacobson. Overfit neural networks as a compact shape representation. *arXiv preprint arXiv:2009.09808*, 2020.
- [52] 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.
- [53] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision (ECCV)*, 2020.
- [54] 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.
- [55] 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.
- [56] 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.
- [57] 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.
- [58] 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.
- [59] 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.

- [60] Shichen Liu, Shunsuke Saito, Weikai Chen, and Hao Li. Learning to infer implicit surfaces without 3D supervision. *arXiv preprint arXiv:1911.00767*, 2019.
- [61] 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.
- [62] 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.
- [63] 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.
- [64] 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.
- [65] 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.
- [66] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020.
- [67] 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.
- [68] 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.
- [69] 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.
- [70] 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.
- [71] 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.
- [72] 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.
- [73] 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.
- [74] 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.
- [75] 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.
- [76] 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.
- [77] 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.
- [78] 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.
- [79] 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.
- [80] 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.

- [81] Matthew Tancik, Vincent Casser, Xinchun Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [82] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (SIGGRAPH)*, 41(4):102:1–102:15, 2022.
- [83] 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.
- [84] 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.
- [85] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [86] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [87] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *IEEE International Conference on Computer Vision (ICCV)*, 2021.
- [88] 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 Transactions on Graphics (SIGGRAPH ASIA)*, 40(6), 2021.
- [89] 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 *IEEE International Conference on Computer Vision (ICCV)*, 2021.
- [90] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [91] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (ToG)*, 40(4), 2021.
- [92] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [93] Ziyang Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhofer. Learning compositional radiance fields of dynamic human heads. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [94] 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 *IEEE International Conference on Computer Vision (ICCV)*, 2021.
- [95] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. IM Avatar: Implicit morphable head avatars from videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [96] Kacper Kania, Kwang Moo Yi, Marek Kowalski, Tomasz Trzcíński, and Andrea Tagliasacchi. CoNeRF: Controllable Neural Radiance Fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [97] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. *arXiv preprint arXiv:2112.02308*, 2021.
- [98] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision (ECCV)*, 2022.
- [99] Markus Knoche, István Sárándi, and Bastian Leibe. Reposing humans by warping 3D features. In *CVPR Workshop on Towards Human-Centric Image/Video Synthesis*, 2020.
- [100] Gengshan Yang, Minh Vo, Neverova Natalia, Deva Ramanan, Vedaldi Andrea, and Joo Hanbyul. Banmo: Building animatable 3D neural models from many casual videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

- [101] 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 *Computer Graphics Forum*, volume 41, pages 703–735, 2022.
- [102] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2016.
- [103] 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.
- [104] 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.
- [105] Attila Szabó, Givi Meishvili, and Paolo Favaro. Unsupervised generative 3D shape learning from natural images. *arXiv preprint arXiv:1910.00287*, 2019.
- [106] 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.
- [107] Matheus Gadelha, Subhansu Maji, and Rui Wang. 3D shape induction from 2D views of multiple objects. In *International Conference on 3D Vision*, 2017.
- [108] 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.
- [109] 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.
- [110] 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.
- [111] 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.
- [112] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3D-aware image synthesis via learning structural and textural representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [113] 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.
- [114] 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.
- [115] 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.
- [116] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [117] Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. Stylepeople: A generative model of fullbody human avatars, 2021.
- [118] Atsuhiko Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. *arXiv preprint arXiv:2204.08839*, 2022.
- [119] N. Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 1995.
- [120] 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.
- [121] Michael Garland and Paul S Heckbert. Surface simplification using quadric error metrics. In *Proceedings of SIGGRAPH*, 1997.
- [122] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018.
- [123] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

- [124] 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)*, 2014.
- [125] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. <https://github.com/open-mmlab/mmpose>, 2020.
- [126] 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.
- [127] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *European Conference on Computer Vision (ECCV)*, 2018.
- [128] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (SIGGRAPH)*, 40(4):1–13, 2021.
- [129] 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)*, 2021.
- [130] Connor Z. Lin, David B. Lindell, Eric R. Chan, and Gordon Wetzstein. 3D GAN inversion for controllable portrait image animation. *arXiv preprint arXiv:2203.13441*, 2022.
- [131] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [132] Anna Frühstück, Krishna Kumar Singh, Eli Shechtman, Niloy J Mitra, Peter Wonka, and Jingwan Lu. Insetgan for full-body image generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [133] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

Generative Neural Articulated Radiance Fields

–Supplementary Document–

Alexander W. Bergman*
Stanford University
awb@stanford.edu

Petr Kellnhofer*
TU Delft
p.kellnhofer@tudelft.nl

Wang Yifan*
Stanford University
yifan.wang@stanford.edu

Eric R. Chan*
Stanford University
erchan@stanford.edu

David B. Lindell
University of Toronto
Vector Institute
lindell@cs.toronto.edu

Gordon Wetzstein
Stanford University
gordonwz@stanford.edu

computationalimaging.org/publications/gnarf/

Contents

1	Implementation Details	18
1.1	Generator and tri-plane representation	18
1.2	Deformation and volume rendering	18
1.3	Discriminator	19
1.4	Training	19
2	Additional Results and Evaluation Details	19
2.1	Single-scene overfitting.	19
2.2	Human body generation and animation	21
2.3	Human face generation and editing	21

*Equal contribution

1 Implementation Details

Applications of our method to human bodies and faces share the same framework, but differ in a few implementation details. For clarity, in the following section we describe the implementation details for GNARF applied to human bodies and outline the differences for faces in Sec. 2.3.

1.1 Generator and tri-plane representation

We use the generator architecture from EG3D [1], which is built on top of the public StyleGAN2 [2] architecture located at <https://github.com/NVlabs/stylegan3> (this StyleGAN3 repository contains backward compatibility for StyleGAN2). The generator is composed of four components: a mapping network, a convolutional backbone, an MLP decoder, and a convolutional super-resolution module.

The generator is conditioned on a 512-dimensional Gaussian noise input using a two-layer mapping network of 512 hidden units. We do not condition the generator on either camera pose or body pose. The mapping network produces a 512-dimensional latent code. This latent code modulates the layers of a StyleGAN2-based convolutional backbone, which produces a 96-channel 256×256 feature image. This is reshaped into three axis-aligned tri-planes, each of shape $256 \times 256 \times 32$. This architecture is trained from scratch rather than using a pre-trained StyleGAN2 network.

The MLP decoder which operates on top of sampled plane features consists of a single hidden layer of 64 units. The decoder maps the 32-dimensional sampled plane feature to a 33-channel feature consisting of a scalar density and 32-dimensional feature. These are integrated per the volume rendering equation (Eq. 3 in the main paper) to obtain a $64 \times 64 \times 32$ feature image, where 64 is the spatial resolution and 32 is the number of channels.

As in EG3D, a separate super-resolution module (implemented as CNN) up-samples and converts the feature images to the final RGB output. The final resolution of the output is 128^2 for SURREAL and 256^2 for AIST++ respectively. As in EG3D, this module is implemented with two StyleGAN2 convolutional blocks, with channel depth of 128 and 256 respectively.

1.2 Deformation and volume rendering

The SF deformation is performed using a simplified version of the SMPL mesh. As described in the main text, this simplified version is obtained using Quadratic Error Metric Decimation [3] in the Open3D library [4] to reduce SMPL from 6890 vertices and 13,776 faces to 690 vertices and 1376 faces.

We perform volume rendering in the canonical space with 64 uniformly-spaced samples plus 64 additional samples based on importance sampling [5] per ray. Additionally, we sample rays only inside an expanded version of the simplified SMPL mesh. The mesh is expanded using a growth offset parameter g , which controls the new position of a vertex v with vertex normal n , by moving v to $\hat{v} = v + gn$. We use $g = 0.05$ during training.

As described in the main paper, deformation methods such as Mean Value Coordinates (MVC) [6] and skinning introduce large artifacts when the template mesh (accidentally) intersects itself or comes very close to doing so. This happens often in practice, even with the unmodified SMPL mesh, due to imperfect SMPL parameter estimation or a subject actually touching somewhere on their body.

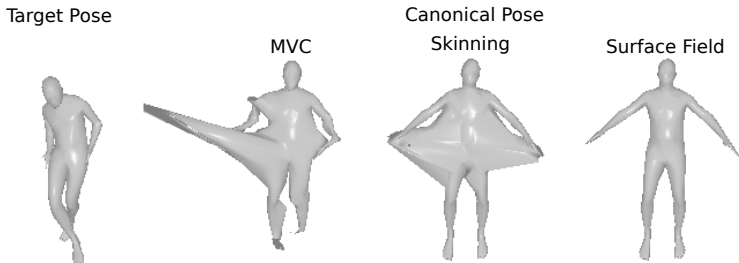


Figure 1: Deforming the vertices of the target pose on the left into the canonical pose results in artifacts for the MVC and skinning deformation method, while the surface field deformation method results in a perfect deformation by construction.

This is shown in Fig. 1: the hands coming close to the body result in large deformation artifacts in the canonical space for MVC and skinning. In the case of MVC, this is mainly attributed to MVC being non-local - a point has non-zero weight w.r.t. *all* the vertices in the driving mesh. This is worsened by the grid approximation mentioned in the main paper (this is necessary to enable feasible training time using MVC): the grid approximation to MVC does not guarantee that surface points remain on-surface after deformation (unlike the full MVC computation) since the rapid change in MVC weights near the surface cannot be sufficiently captured in feasible grid resolution. Similarly, in the skinning method, when parts of the driving mesh get close together, individual vertices may become closer to a different bone than the one which they are a part of. This causes them to deform to incorrect places in the canonical space. These artifacts are significantly mitigated when using the proposed surface field (SF) deformation, as SF is very local (in contrast to MVC) and the lookup of the closest triangle is less error-prone than of the closest bone.

1.3 Discriminator

Dual discrimination. Similarly to EG3D, we use dual discrimination to ensure consistency between the raw neural rendering and the final super-resolved output. As in EG3D, we concatenate a resized copy of the raw neural rendering to the super-resolved input to form a 6-channel discriminator input tensor. This raw neural rendering consists of the first three channels of the rendered feature image.

Discriminator pose conditioning. As in EG3D, we condition the discriminator on the camera pose via a mapping network that modulates the layers of the discriminator. Unlike EG3D, we additionally condition the discriminator on the expected body pose / facial expression by concatenating the body/facial pose parameters (SMPL or FLAME parameters) to the camera parameters as input to the mapping network.

By conditioning on body / facial poses, we give the discriminator the ability to ensure that the applied deformation matches the specified pose. Empirically, we found that corrupting the poses with 1 standard deviation of Gaussian noise before passing them as input to the discriminator aided training convergence. We hypothesize that in the absence of noise, the discriminator was able to overfit on the specific poses and cameras of the ground truth dataset, which destabilized training.

Note that unlike EG3D, which conditions the generator on the camera parameters for training with FFHQ (*modeling pose-correlated attributes*) we do not condition the generator with camera pose. We also do not condition the generator with body pose or facial expression. This is done in order to ensure that the generator learns to generate a body/face in the canonical space which is robust to custom deformations, rather than one which is specific for a warping or camera viewpoint.

1.4 Training

Many of our training hyperparameters are adopted from those of EG3D and StyleGAN2: generator learning rate (0.0025), discriminator learning rate (0.002), batch size (32), blurring images (GT and generated) over the first 200K iterations, and R1 regularization [7]. As recommended, the gamma parameter of R1 regularization is tuned according to the dataset: FFHQ: $\gamma = 1$; SURREAL: $\gamma = 1$; AIST $\gamma = 4$.

We use 8 Tesla V100 GPUs, training each model for roughly 2 days.

2 Additional Results and Evaluation Details

2.1 Single-scene overfitting.

Data pre-processing. The single-scene overfitting experiment is conducted using the sequence *gBR_sBM_cAll_d04_mBR0_ch01* from the AIST++ dataset [8]. We extract all 719 frames sampled at 60 Hz from each of the 8 available cameras and we crop the human body using the same procedure as in the GAN training. We skip the camera number 4 because the annotation in the dataset does not match the video. We hold out the cameras number 2 and 7 for testing and we train our models using 30 frames uniformly sampled from the remaining six cameras.

Model. We use the same triplane representation as in our GAN experiments with a few modifications to avoid overfitting to the sparse training data. First, we limit the capacity of the decoder network by reducing the latent space size from 64 to 32 and by reducing the resolution of the triplanes from 256 to 128. Next, we increase the number of samples per ray to 128 and disable the second stage of importance sampling. Finally, unlike in the GAN setup, we do not remove background from the training images. Instead, we train a representation of the entire scene as is common in other overfitting papers [9]. To avoid mixing of the static background and dynamic foreground in the neural representation and to allow for efficient sampling of both regions with different depth ranges, we include a separate identical triplane representation for the background. We use the accumulated optical density from the foreground network to alpha-blend the foreground image over the background image. No ground-truth foreground masks are used during the training.

We utilize the same model for deformation function D implemented using mesh skinning, Surface Field and MVC. For skinning and Surface Field, we compute the transformations on-the-fly. However, this is not feasible for the relatively slow MVC computation. Therefore, we precompute the MVC transformations for $16 \times 16 \times 16$ points uniformly sampled withing a bounding cube of the human body for all training poses, and we sample them during training using a trilinear interpolation.

We train our model for 500 000 steps with Adam optimizer [10] and step size of 0.002 and we use L2 loss to supervise the training at 128×128 resolution with batch size of 3 images on Nvidia RTX3090 graphical processing unit.

Evaluation. We rely on well known image metrics to compare performance of individual warping methods. Since the goal is to evaluate efficacy in compensating human body motion and not capacity for learning the background scenery, we use foreground masks for computing the image metrics. To this goal, we compute human body masks using a pre-trained image segmentation model [11]. Then, we filter out the background pixels for the PSNR metric. For the structural metrics of SSIM and LPIPS [12], we set the background pixels to zero in both the predicted and ground-truth images.

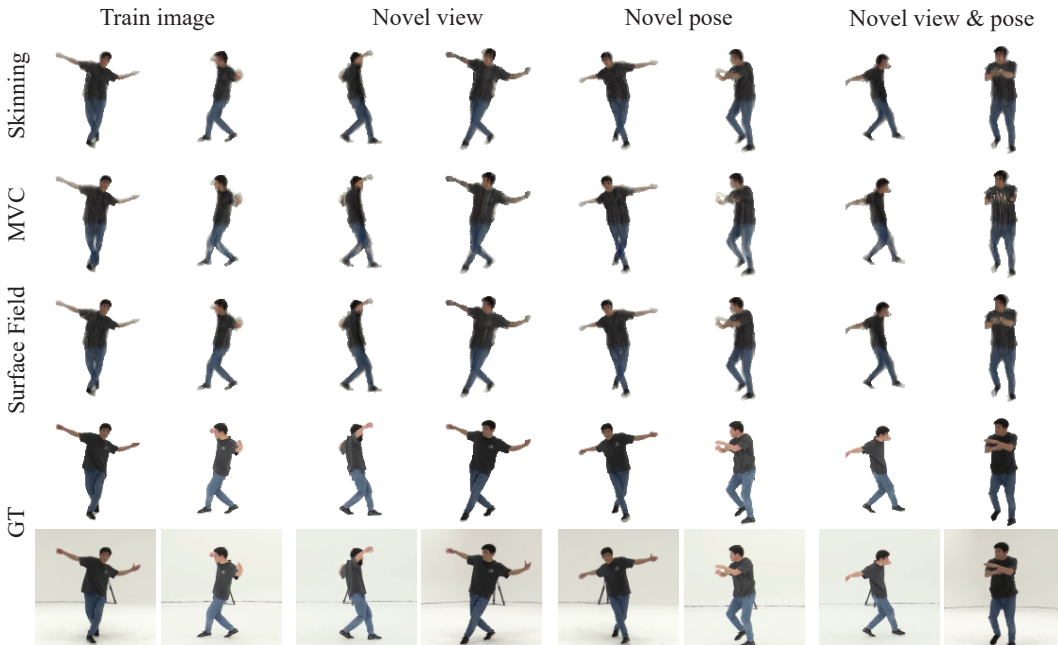


Figure 2: Qualitative comparison of results from our single-scene overfitting experiment. The figure presents the original training views and poses, interpolation of novel views, interpolation of novel poses and interpolation of novel poses under novel views. All images except the last row are presented with the same foreground masks as used for the metric evaluation.

2.2 Human body generation and animation

SURREAL data pre-processing. For training, we use the official SURREAL training split, and extract the first frame from each video. We square-crop each frame from head-to-toe using the ground truth segmentation mask, and resize the image to 128×128 . The backgrounds are set to be black. We use the provided camera poses and intrinsics from the SURREAL dataset for training. Images for which the SMPL mesh scale is not consistent with the image size are filtered, leaving 35,332 total images for training.

The SURREAL dataset provides ground truth SMPL parameters for each frame, which we use. We compute the mean SMPL parameters across the frames in the dataset in order to assign the canonical pose, such that it is close to each of the target SMPL poses.

AIST++ data pre-processing. For training, we extract 30 frames uniformly sampled from each video in the AIST++ dataset. We then filter out frames whose effective camera distance to a normalized SMPL model is above a threshold or the projected human bounding box is partially outside of the image as a form of heuristic detecting poorly estimated SMPL parameters. We square-crop these frames at 600×600 resolution centered at the pelvis joint of the SMPL mesh, and resize each image to 256×256 . Since ground truth masks are not provided for this dataset, we run an off-the-shelf segmentation model [11] to remove backgrounds. This stabilizes the GAN training, as the GAN no longer has to attempt to model a 3D-consistent background.

The AIST++ dataset also provides ground truth camera and body pose parameters. We move the translation of the SMPL mesh into the camera extrinsics, and simulate the scaling of the SMPL mesh by either moving the camera further back or closer. We additionally rescale all meshes and camera parameters such that the mean distance from camera to mesh is 1.7. Similarly to SURREAL, we compute the mean SMPL parameters across the selected training frames to assign the canonical pose.

Evaluation. We use the FID metric [13] for evaluation of the quality of generated images. This metric compares the distribution of intermediate features extracted from an inception network run on both generated and ground truth images. FID (10k) refers to the evaluation consistent with [14], which compares the distribution of 10,000 generated images with 10,000 randomly sampled ground truth images. FID (50k) compares the distribution of 50,000 generated images with the entire training dataset, giving a better estimate of the true distance. For EG3D, FID is run on the generated outputs with no warping. For the EG3D + warping baseline, the generated results from EG3D (not in the canonical pose) are warped by estimating the pose of the generated body and using this as the canonical pose. The FID is then applied to images generated in this fashion. For GNARF, we simply apply the FID to generated images with our deformation method.

In order to measure deformation accuracy for our method and the EG3D + warping method, we use the PCKh@0.5 metric. The use of this metric was inspired by the concurrent work [14]. After correspondence with the authors of this work, we have standardized our evaluation of this metric in order to compare reported values in the main paper table. Details of the evaluation method have a large effect on the magnitude of the PCKh@0.5 numbers reported, but describe the correct trend within a consistent evaluation standardization. For both the AIST++ and SURREAL dataset, we compute the GT keypoints by running a off-the-shelf body keypoint estimator [15] trained on MPII [16] (publicly available on the MMPose Project [17]) on each GT image. We then generate an image and deform the generated result with the body-pose parameters corresponding to this GT image, and render the generated result from the same camera position. The keypoint estimator is then run on this generated image, and the keypoints detected are compared in 2D. We discard keypoints that the keypoint estimator is not confident on in order to factor out estimator error, and only compare confident keypoints. We determine the head size (interocular distance) using the detected keypoints from the GT image.

2.3 Human face generation and editing

Here, we outline the implementation differences in the human body application.

Deformation. We use the FLAME head model to drive the deformation. The original template FLAME [18] template model has 5023 vertices and 9976 faces. This mesh contains 3 unconnected

parts, modeling the base face and the two eyeballs respectively. Since there is no suitable method to accurately extract the FLAME parameters related to eye movements from training images, we remove the eyeball parts. In original template, the neck and mouth are modeled with holes. We find having holes slightly degrades the deformation quality for points around the hole area, likely because worse point-to-triangle correspondence. Furthermore, we find the small triangles can lead to numeric instabilities when e.g. computing the barycentric coordinates for deformation, therefore we decimate the mesh as described in the main paper which makes the triangle sizes more uniform and also speeds up the SF deformation. The resulting mesh template, shown in Fig. 3 contains 1, 252 vertices and 2, 500 triangles.

The default FLAME head model is in a different scale and origin as the head generated by the pretrained EG3D model. To ensure meaningful warping in the transfer learning, we rescale by 2.6 and fix the root joint at $[-0.0013, -0.1344, -0.0390]^T$, which we determined by visually aligning the heads from FLAME and a pretrained EG3D.

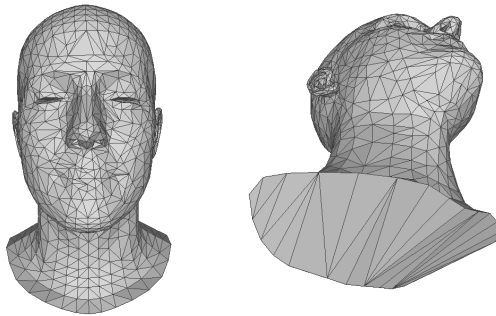


Figure 3: Processed FLAME model for training GNARF on FFHQ.

Generator pose conditioning. Unlike for the body, we use camera pose conditioning for the generator as proposed by EG3D. EG3D found that by using swapping regularization, camera pose conditioning will not negatively impact 3D consistency but rather improve generation quality. We found this true in our experiments for faces.

However, same as the body experiment, we do not provide the generator with any information related to the deformation (i.e. expression, shape and jaw rotation). As explained before, this is crucial for generating consistent canonical faces.

Volume Rendering. Unlike for bodies, the final resolution of the output is 512×512 , consistent with the original EG3D.

Dataset preprocessing. Our data preparation is based on the original EG3D [1]. For each training image, we fix the camera intrinsics and estimate the camera extrinsics assuming that the head is inside a unit-length bounding box, front-facing and at a fixed position. Additionally, we use DECA [19] to estimate the expression, pose (only the jaw rotation, since we assume that the head is front-facing) and the shape parameters of the FLAME models. Note that, originally, DECA also outputs the camera parameters. However it uses an orthographic camera model, which is not directly transferable to the camera used in the pretrained EG3D. We therefore use the camera pose estimation from EG3D’s data preprocessing procedure.

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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [2] 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.
- [3] Michael Garland and Paul S Heckbert. Surface simplification using quadric error metrics. In *Proceedings of SIGGRAPH*, 1997.
- [4] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018.

- [5] 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.
- [6] Tao Ju, Scott Schaefer, and Joe Warren. Mean value coordinates for closed triangular meshes. In *Proceedings of SIGGRAPH*. 2005.
- [7] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *International Conference on Machine Learning (ICML)*, 2018.
- [8] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. AI choreographer: Music conditioned 3D dance generation with aist++. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.
- [9] Kai Zhang, Gernot Riegler, Noah Snively, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020.
- [10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- [11] 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.
- [12] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [14] Atsuhiko Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. *arXiv preprint arXiv:2204.08839*, 2022.
- [15] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [16] 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)*, 2014.
- [17] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. <https://github.com/open-mmlab/mmpose>, 2020.
- [18] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics (SIGGRAPH ASIA)*, 36(6):194:1–194:17, 2017.
- [19] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (SIGGRAPH)*, 40(4):1–13, 2021.

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]