Morphable Diffusion:
3D-Consistent Diffusion for Single-image Avatar Creation

Xiyi Chen[1]   Marko Mihajlovic[1]   Shaofei Wang[1,2,3]   Sergey Prokudin[1,4]   Siyu Tang[1]

ETH Zürich[1]; University of Tübingen[2]; Tübingen AI Center[3];
ROCS, University Hospital Balgrist, University of Zürich[4]
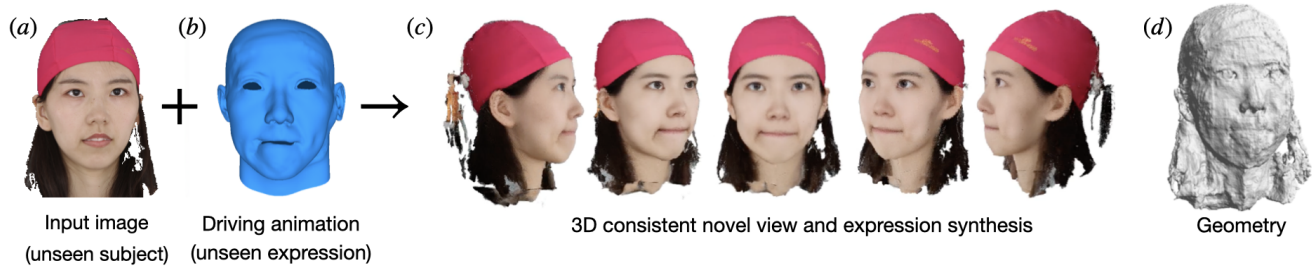
https://xiyichen.github.io/morphablediffusion/

Figure 1. **Morphable diffusion**. We introduce a morphable diffusion model to enable consistent controllable novel view synthesis of humans from a single image. Given a single input image *(a)* and a morphable mesh model with a target facial expression *(b)* our method directly generates 3D consistent and photo-realistic images from novel viewpoints *(c)*. Using the generated multi-view consistent images, we can reconstruct a coarse 3D model *(d)* using off-the-shelf neural surface reconstruction methods such as [80].

Abstract

Recent advances in generative diffusion models have enabled the previously unfeasible capability of generating 3D assets from a single input image or a text prompt. In this work, we aim to enhance the quality and functionality of these models for the task of creating controllable, photorealistic human avatars. We achieve this by integrating a 3D morphable model into the state-of-the-art multi-view-consistent diffusion approach. We demonstrate that accurate conditioning of a generative pipeline on the articulated 3D model enhances the baseline model performance on the task of novel view synthesis from a single image. More importantly, this integration facilitates a seamless and accurate incorporation of facial expression and body pose control into the generation process. To the best of our knowledge, our proposed framework is the first diffusion model to enable the creation of fully 3D-consistent, animatable, and photorealistic human avatars from a single image of an unseen subject; extensive quantitative and qualitative evaluations demonstrate the advantages of our approach over existing state-of-the-art avatar creation models on both novel view and novel expression synthesis tasks. The code for our project is publicly available.

1. Introduction

The field of photorealistic controllable human avatar generation has been subject to several technological leaps in the recent decade. The introduction of large 3D scan collections has facilitated the construction of expressive, articulated models of 3D human bodies [47, 55], faces [5, 39, 56], and hands [65]. From the outset, one of the primary applications of these models was to reconstruct a 3D avatar from highly under-constrained setups, such as monocular video or a single image [5, 18, 37, 55].

While allowing for rich semantic information to be inferred, these 3D morphable models were limited in the level of photorealism due to their focus on minimally clothed bodies and face regions, as well as their reliance on the standard mesh-based computer graphics pipelines for rendering.

Recently, the task of generating *photorealistic* avatars [57] gained significant attention from the research community due to its potential to revolutionize our ways of digital communication. Here, combining novel neural rendering techniques [50, 86] with articulated human models allowed for a new level of generated image quality. However, the best-performing models here still require a significant visual input, such as calibrated multi-view images [21, 78] or monocular video sequences of the subject [61, 82, 95, 99].

Concurrently, the field of generative modeling emerged with its ability to create highly photorealistic assets by learning a complex distribution of real-world image data [33, 64]. Here, a substantial body of research has been dedicated to better controlling the generative process of such models in the case of human-related imagery [17, 29, 93]. This is regularly done by conditioning the pipelines on 2D keypoints [8], 3D deformable models [73], or text [48]. Another emerging topic is building geometry-aware generative models that allow view-consistent image synthesis and 3D asset extraction [9, 45, 59].

The method proposed in this work can be considered the next logical step in the evolution and unification of the three aforementioned branches of research on photorealistic controllable virtual humans. We begin with investigating the performance of a state-of-the-art view-consistent diffusion model [44, 45] on the task of novel view synthesis of human heads and full bodies. However, we show that a simple finetuning of the baseline on a limited dataset (e.g. [90, 92]) leads to sub-optimal results, with the model failing to preserve the identity or the facial expression / body pose.

We follow in the footsteps of the view-consistent diffusion model and propose a novel finetuning strategy and architecture that enhances the reconstruction quality and allows animation. Our key idea is to leverage a well-studied statistical model of human shapes [39, 47] to introduce human prior and guide the reconstruction process. More importantly, a controllable 3D model also allows us to perform a more challenging task of photorealistic animation of a human head from a single image under novel expression (Figure 1).

A similar type of diffusion process guidance has been investigated by several prior works [16, 93]. However, as we will show later, the models conditioned purely on 2D rasterizations of meshes or projected keypoints struggle to achieve truly 3D-consistent novel view and expression synthesis results.

To alleviate this problem, we introduce the conditioning of the diffusion process on a 3D morphable model that performs an uplifting of the noisy image features and associates them with the corresponding mesh vertices *in 3D space* (Figure 2A). We also introduce a shuffled training scheme for the diffusion model: during training, the model is trained to predict a view-consistent image set of novel facial expressions, given a single head image with a different expression as a reference and an articulated 3D model as a driving signal. The resulting combination of a powerful diffusion network and 3D model conditioning allows for the first-time building of a highly photorealistic animatable head model of an unseen subject from a single input image with an unseen facial expression as driving signal.

To summarize, our contributions are as follows: (a) We analyze the applicability of state-of-the-art multi-view con-

sistent diffusion models for the task of human avatar creation and propose a superior pipeline that consistently improves the quality of generated images across most metrics, thanks to the efficient conditioning of the generative process on a deformable 3D model; (b) We further propose a more efficient training scheme to enable the generation of new facial expressions for an unseen subject from a single image.

2. Related work

3D morphable models: Traditional methods for constructing shapes have been predominantly centered around parameterized 3D mesh representations, particularly in the context of human faces [5, 39, 56], bodies [47, 55, 89], and hands [41, 65]. These data-driven models are typically created from high-quality 3D scans obtained in expensive studio environments and they have laid the foundation for reconstructing humans in the wild from as few as a single input image [18, 31, 37]. However, these mesh-based representations are limited in quality for representing thin geometric details (*e.g.* hair) and are not suitable for truly photo-realistic novel view synthesis.

Photorealistic avatars: Neural radiance fields (NeRFs) [50] have recently enabled rapid development of modeling and reconstructing photorealistic scenes. Several recent works have successfully combined neural rendering with articulated human models [46, 57, 86] to enable high-quality photorealistic avatar reconstruction. However, they still require a substantial visual input such as a monocular video [30, 61, 82, 95, 99] or several calibrated multi-view images or videos [11, 21, 32, 49, 78].

Avatars from a single image: Single-image human avatar creation has been an active research area in the past few years. The seminal work PIFu [67] has proposed to directly model the human body shape as an implicit representation conditioned on pixel-aligned features. Several follow-ups [1, 22, 23, 28, 68, 87, 88] have focused on improving robustness and reconstruction quality; others have focused on reconstructing human faces [6, 12, 35, 63, 74]. However, these methods lack controllability and cannot be directly animated or used to synthesize novel views for diverse novel expressions or poses of an unseen subject.

Controllable face avatars: Controllable face avatars allow rigging of appearance, identity, pose, or expression from a single image [26, 40, 97]. However, these methods only demonstrate the manipulation ability under frontal views and produce incomplete or blurry renderings under extreme head poses, as we will demonstrate in the following sections. Several other works achieve the rigging from multiple views or monocular videos of the input subject [95, 96, 99]. Closer to our work is DiffusionRig [16] that leverages the
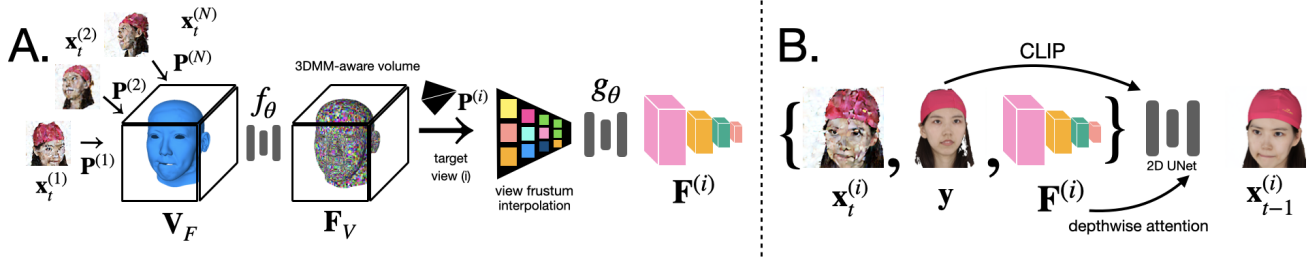
Figure 2. **Morphable diffusion step.** This figure gives an overview of a single denoising step of the proposed 3D morphable diffusion pipeline. Our morphable denoiser takes as input a single image $\mathbf{y}$ and the underlying human model $\mathcal{M}$ and generates N novel views from pre-defined viewpoints. Given the noisy images of N fixed views $x_t^{(1:N)}$ obtained from the previous iteration, camera projection matrices $P^{(1:N)} = (K^{(1:N)}, R^{(1:N)}, T^{(1:N)})$, and the target articulated 3D mesh model, we construct A) a morphable noise volume by attaching the 2D noise features onto mesh vertices that are processed by a SparseConvNet f_θ to output a 3DMM-aware feature volume $\mathbf{F}_V$, which is further interpolated to the frustum $\mathbf{F}^{(i)}$ of a target view (i) that we wish to synthesize. B) The noisy target image $\mathbf{x}_t^{(i)}$, the input image $\mathbf{y}$, and the target feature frustum are then processed by a pre-trained 2D UNet akin to [45] to predict the denoised image in the next iteration $\mathbf{x}_{t-1}^{(i)}$.

photorealism of the diffusion models [64] with the explicit controllability achieved via conditioning of the denoising process on 2D rasterizations of a morphable model. However, this work struggles to achieve a multi-view consistent novel view synthesis, as we will demonstrate in the experiment section later.

Generative models: As reconstruction from a single view is inherently a challenging ill-posed task, many works have attempted to formulate the problem as a generative task instead. A rich set of related methods leverage generative adversarial networks [33] or powerful diffusion models [64] to learn complex data distributions from real-world imagery and to directly synthesize novel images without modeling an underlying 3D model. In the context of human avatars [17, 29, 93], prior works explore different input conditions to control the generative process such as keypoints [8], deformable models [73], or just textual queries [7, 38, 48]. However, achieving a high degree of photorealism in these models comes at the expense of 3D consistency.

3D-consistent generative models: Several attempts have been made towards constructing a 3D consistent generative model [2, 4, 9, 20, 42, 51–53, 69, 70, 72, 79, 83, 84]. On the other side, the recent work Zero-1-to-3 [44] learns to control the camera perspective in large-scale diffusion models and enable consistent generation of novel views. Sync-Dreamer [45] further improves the multi-view consistency through a 3D-aware feature attention module. However, the novel view synthesis from a single image of this model cannot be explicitly controlled, *e.g.* by a human expression or pose, which is essential for avatar creation. In this work, we build on the state-of-the-art multi-view consistent diffusion framework [45] and enhance the reconstruction quality while enabling additional control for generating animatable photorealistic avatars.

Finetuning diffusion models: The main limitation of diffusion models is the long and expensive training. Therefore, many recent works [15, 27, 43, 45] focus instead on finetuning large pre-trained models. Here, a popular approach [93] is to inject the control parameters as an additional signal when finetuning. However, adjusting this framework operating in 2D space to finetune multi-view consistent diffusion models [45] which contains a 3D conditioning module is not straightforward. In the following, we will examine how to effectively control the diffusion models for photorealistic avatar creation.

3. Morphable diffusion model

Given an input image $\mathbf{y}$ of a human, our goal is to synthesize N multi-view consistent novel views from predefined viewing angles and allow for their animation via a morphable model $\mathcal{M}$. Our method builds on the recent multi-view consistent diffusion model [45] while further improving the reconstruction quality and allowing for explicit manipulation of the synthesized images. We start by reviewing the 3D-consistent diffusion model (Sec. 3.1) and later detail our morphable diffusion model (Sec. 3.2).

3.1. Preliminaries: multi-view diffusion

Multi-view diffusion [45] consistently learns the joint distribution of novel views at N fixed viewpoints $\{\mathbf{x}_0^{(i)}\}_{i=1}^N$ given an input condition $\mathbf{y}$: $p_\theta(\mathbf{x}_0^1, \ldots, \mathbf{x}_0^N, |\mathbf{y})$; $\mathbf{y}$ is omitted for brevity in the following. Analogously to the diffusion model [71], it defines the *forward process*

$$q(\mathbf{x}_{1:T}^{(1:N)}|\mathbf{x}_0^{(1:N)}) = \prod_{t=1}^{T}\prod_{n=1}^{N} q(\mathbf{x}_t^{(n)}|\mathbf{x}_{t-1}^{(n)}), \quad (1)$$

$$q(\mathbf{x}_t^{(n)}|\mathbf{x}_{t-1}^{(n)}) = \mathcal{N}(\mathbf{x}_t^{(n)}; \sqrt{1-\beta_t}\mathbf{x}_{t-1}^{(n)}, \beta_t\mathbf{I}), \quad (2)$$

and the *reverse process*

$$p_\theta(\mathbf{x}_{0:T}^{(1:N)}) = p(\mathbf{x}_T^{(1:N)}) \prod_{t=1}^{T} \prod_{n=1}^{N} p_\theta(\mathbf{x}_{t-1}^{(n)}|\mathbf{x}_t^{(1:N)}), \quad (3)$$

$$p_\theta(\mathbf{x}_{t-1}^{(n)}|\mathbf{x}_t^{(1:N)}) = \mathcal{N}(\mathbf{x}_{t-1}^{(n)}; \mu_\theta^{(n)}(\mathbf{x}_t^{(1:N)}, t), \sigma_t^2 \mathbf{I}), \quad (4)$$

where $\mu_\theta^{(n)}$ is trainable, while β_t and σ_t are fixed time-dependent coefficients.

To learn the joint distribution by minimizing the negative log-likelihood, multi-view diffusion [45] follows the DDPM's [25] parameterization:

$$\mu_\theta^{(n)}(\mathbf{x}_t^{(1:N)}, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t^{(n)} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta^{(n)}(\mathbf{x}_t^{(1:N)}, t) \right),$$
$$(5)$$

where ϵ_θ is a trainable *noise predictor* (in practice, parameterized by a UNet [66]) and other constants $(\alpha_t, \bar{\alpha}_t)$ are derived from β_t. Then, the training is performed by minimizing the loss

$$\ell = \mathbb{E}_{t,\mathbf{x}_0^{(1:N)},n,\epsilon^{(1:N)}} \left[\|\epsilon^{(n)} - \epsilon_\theta^{(n)}(\mathbf{x}_t^{(1:N)}, t)\|_2 \right], \quad (6)$$

where $\epsilon^{(1:N)} \sim \mathcal{N}(0, \mathbf{I})$ is the sampled Gaussian noise.

3.2. Morphable multi-view diffusion model

To increase the reconstruction fidelity and extend the baseline model to animatable avatar, we introduce a morphable diffusion model for animatable high-fidelity multi-view consistent novel view synthesis of humans from a single input image.

Our model takes as input a single image $\mathbf{y}$ and an underlying morphable model $\mathcal{M}$ (*e.g.* SMPL [47], FLAME [39], or bilinear [90] model) that maps a low-dimension identity $\beta \in \mathbb{R}^I$ and expression $\Theta \in \mathbb{R}^E$ parameters to a mesh containing n_v vertices:

$$\mathcal{M} : \beta, \Theta \mapsto \mathbb{R}^{n_v \times 3}. \quad (7)$$

Given the image $\mathbf{y}$ and mesh $\mathcal{M}$, we synthesize N images $\mathbf{x}_0^{(1:N)}$ of the same resolution through an iterative diffusion process. These novel views are generated from fixed viewpoints with the relative camera translations $T^{(1:N)} \in \mathbb{R}^3$, rotations $R^{(1:N)} \in \mathbb{R}^{3\times3}$ and their respective calibration parameters $K^{(1:N)} \in \mathbb{R}^{3\times3}$. An overview of a single diffusion step is illustrated in Fig. 2.

We leverage the mesh of the underlying morphable model to unproject and interpolate the d-dimensional target image noise features $x_0^{(1:N)}$ onto n_v mesh vertices in the world space. These pixel-aligned noise features for all target views are then fused via channel-wise mean pooling to make the vertex features invariant of the order of target

views. The output vertex features $\mathbf{V}_F \in \mathbb{R}^{n_v \times d}$ are then processed via a sparse 3D ConvNet [19, 57] f_θ and trilinearly interpolated to create a 3D morphable-model-aware feature volume:

$$\mathbf{F}_V = f_\theta(\mathbf{V}_F) \in \mathbb{R}^{x \times y \times z \times f_V}, \quad (8)$$

where x, y, z, and f_V denote the dimensionality and number of channels of the interpolated 3DMM-aware feature volume $\mathbf{F}_V$.

Given a target view i, we uniformly sample d_F points on every ray inside the view frustum defined by fixed near and far planes. This formulates a 3D grid $\mathbf{F}^{(i)} \in \mathbb{R}^{d_F \times h_F \times w_F \times f_V}$ inside the view frustum. We then assign the trilinearly interpolated features from $\mathbf{F}_V$ to each point in $\mathbf{F}^{(i)}$, which is then processed via a convolutional network to extract a feature volume $\mathbf{F}_j^{(i)}$ at each layer j of the L-layer denoising UNet. This interpolation and the neural network operation are jointly denoted as g_θ and output a view frustum volume with L grids:

$$\mathbf{F}_{1:L}^{(i)} = g_\theta(\mathbf{F}^{(i)}) = g_\theta(\mathbf{F}_V | R^{(i)}, T^{(i)}, K^{(i)}). \quad (9)$$

Lastly, the target $x_t^{(i)}$ and the CLIP encoding [62] of the input image $\mathbf{y}$ are propagated through a pre-trained UNet that is conditioned on $\mathbf{F}_{1:L}^{(i)}$ through depth-wise attentions as in SyncDreamer. Specifically, a cross-attention [75] layer is applied on every intermediate feature map f_j of the current view i in the UNet:

$$\text{Attention}(Q_j, K_j, V_j) = \text{softmax}(\frac{Q_j K_j^T}{\sqrt{d}}) \cdot V_j, \quad (10)$$

with $Q_j = W_{Q,j} \cdot \varphi_j, K_j = W_{K,j} \cdot \mathbf{F}_j^{(i)}, V_j = W_{V,j} \cdot \mathbf{F}_j^{(i)}$, where $\varphi_j \in \mathbb{R}^{N \times d_\epsilon^j \times h_j \times w_j}$ is the intermediate representation of f_j cross-attentioned with the CLIP feature of the input image, $\mathbf{F}_j^{(i)} \in \mathbb{R}^{N \times d_r \times d_j \times h_j \times w_j}$ is the feature volume downsampled from $\mathbf{F}^{(i)}$ corresponding to the j-th layer of the UNet. $W_{Q,j} \in \mathbb{R}^{d \times d_\epsilon^j}, W_{K,j} \in \mathbb{R}^{d \times d_r}, W_{V,j} \in \mathbb{R}^{d \times d_r}$ are learnable projection matrices [64]. The attention computation and feature aggregation is only performed along the depth dimension d_j.

4. Experiments

We demonstrate the effectiveness of our method on the novel view synthesis (Sec. 4.1) and the animation from a single image (Sec. 4.2). For ablation studies on design choices and training regimes, and discussion about the effect of the topology and expressiveness of the input meshes, please refer to the supplementary material.

Training: To train our model, we exploit diffusion models Zero-1-to-3 [44] and SyncDreamer [45] that have been
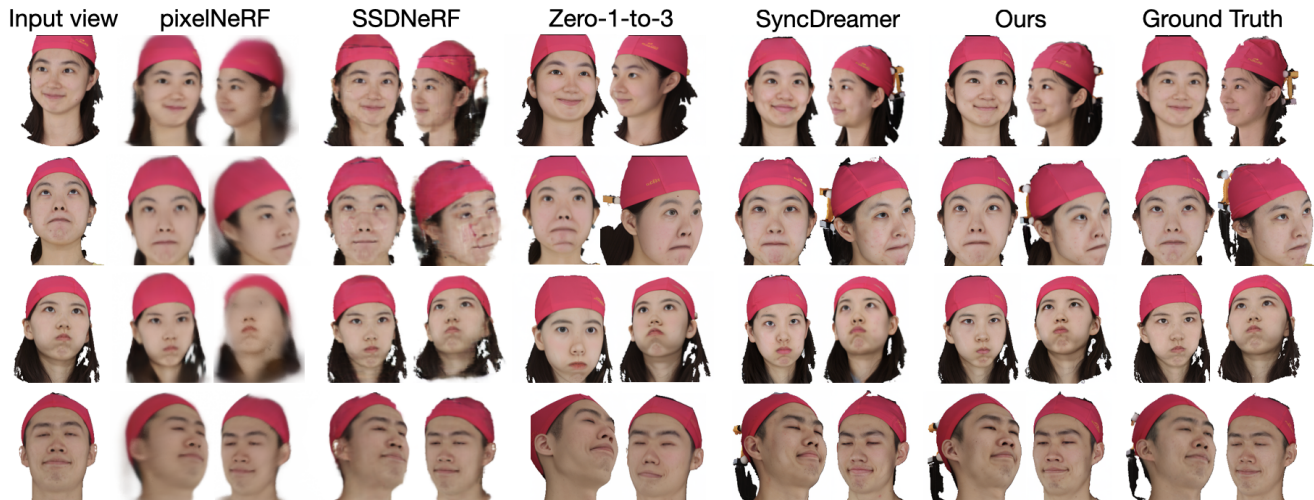
Figure 3. **Single-view reconstruction of human faces.** In addition to the single input view, our method also takes as input a mesh of the facial expression corresponding to the input image. Our method produces more plausible and realistic novel views compared to state-of-the-art methods. While providing multi-view consistency, PixelNeRF [91] and SSDNeRF [10] produce overly blurry results. Zero-1-to-3 [44] generates images of good quality which however fail to preserve multi-view consistency and do not align with the ground truth target views. SyncDreamer [45] produces multi-view consistent images with relatively accurate facial expressions that however lose the resemblance. For more details and discussion, please see section 4.1.

trained on a large dataset such as Objaverse [14] (800k objects). To further allow better generalization to any user-input images, we remove the input view parameter embeddings by setting all the corresponding values to zeros. Additionally, we drop the assumption that the input and the first target image have the same azimuth.

Figure 4. **Single-view reconstruction of human bodies.** Our method is the only one that reconstructs the correct body poses. The relatively low resolution of all methods, however, limits the amount of details in the generated images.

Experimental setup: To evaluate our model, we use a dataset of human faces (FaceScape [90]) and bodies (THuman 2.0 [92]). For the novel view synthesis task of the human faces, we train our model on 323 subjects, each having 20 unique facial expressions, and evaluate its performance on the remaining 36 subjects, including the 4 subjects who agree to be shown in publications. The results of these 4 subjects are displayed in our qualitative evaluation figures. Since the capture devices are visible from the back views, we only include views that have azimuth in the range of

[-90, 90] degrees for our training and evaluation. During testing, we use input images that have absolute azimuth and elevation both less than 15 degrees, since using more extreme facial poses does not provide a sufficient amount of information for any meaningful reconstruction. We follow DINER [60] and apply cropping, rescaling, and color calibration. FaceScape provides coarse meshes of the fitted bilinear model for each frame, which we use to build the voxel grids for SparseConvNet (Sec. 3.2).

For the novel view synthesis of full bodies, we evaluate our method on the THuman 2.0 dataset [92], which is a synthetic dataset that consists of 526 meshes with ground truth SMPL-X [55] fittings. We follow [45] and render 16 input and 16 target views (with the same azimuth) for each mesh in a spherical camera path with an elevation of 30 degrees. The SMPL-X vertices for each mesh are used to build voxel grids and query the SparseConvNet. We use the first 473 meshes for training and the last 53 meshes for testing.

Metrics: We compare SSIM [81], LPIPS [94], Frechet Inception Distance (FID) [24], Percentage of correct Keypoints (PCK) [3], and face re-identification accuracy (Re-ID) [58] on the generated images of our method and the baselines. We use an off-the-shelf facial keypoints regressor [13] with HRNetV2 backbone [76] to predict 68 facial keypoints and normalize them using the intercanthal distance between two eyes. Due to the slightly inaccurate facial expressions in the ground truth bilinear meshes, we also run the keypoint regressor on ground truth images and use the predicted keypoints as pseudo ground truth, instead of us-
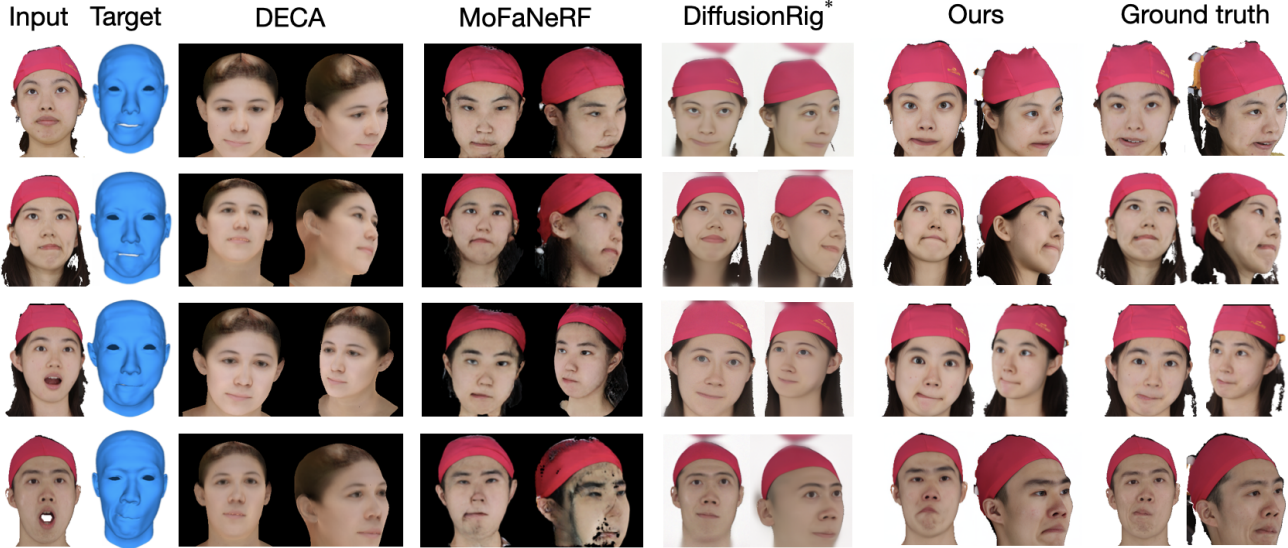
Figure 5. **Novel facial expression synthesis.** Qualitative comparison with DECA[18], MoFaNeRF [97], and DiffusionRig [16] on novel facial expression synthesis. DiffusionRig is denoted with * since it requires per-subject finetuning with additional images. Our morphable diffusion model is the only one that successfully synthesizes novel views for a novel facial expression while retaining high fidelity. For more details and discussion, please see section 4.2.

ing the projected 3D keypoints from the bilinear meshes. We then label the keypoints predicted from the generated face images that are within 0.2 pixels away from the pseudo ground truth as correct ones. Note that we only evaluate PCK for camera poses that have both absolute azimuth less than 30 degrees and absolute elevation less than 15 degrees, in order to avoid potential failure cases of the keypoint regressor on larger angles where some keypoints are invisible. We also follow [16] to report face re-identification accuracy, where we extract 128-dimensional face descriptor vectors from the face images using the dlib library [36]. If two vectors between ground truth and the generated image have an Euclidean distance of less than 0.6, we classify the generated image to be re-identified as the same person. In addition, we reconstruct meshes by training NeuS2 [80] on the generated images. We show more results on mesh reconstruction in the supplementary.

4.1. Novel view synthesis

Baselines: We adopt pixelNeRF [91], Zero-1-to-3 [44], SyncDreamer [45], and SSD-NeRF [10] as baselines for the novel view synthesis task. For fair comparisons, we finetune the baselines or train them from scratch on our training data. Note that although SyncDreamer proposes not to finetune the UNet along with its conditioning module, we find it beneficial to do so when we transfer the domain of the SyncDreamer model pretrained on general objects to human face / bodies. Therefore, the UNet is also finetuned in the baseline SyncDreamer models that we report in this paper. We

Method	LPIPS ↓	SSIM ↑	FID ↓	PCK@0.2 ↑	Re-ID ↑
pixelNeRF [91]	0.2200	0.7898	92.61	72.34	97.46
Zero-1-to-3 [44]	0.4248	0.5656	10.97	4.69	96.77
SSD-NeRF [10]	0.2225	0.7225	34.88	92.65	98.74
SyncDreamer [45]	0.1854	0.7732	**6.05**	94.07	99.60
Ours	**0.1653**	**0.8064**	6.73	**95.80**	**99.86**

Table 1. **Novel view synthesis of human faces.** Quantitative comparison on the FaceScape [90] dataset demonstrates that our method produces more realistic face images with more accurate facial expressions and better resemblance.

Method	LPIPS ↓	SSIM ↑	FID ↓
pixelNeRF [91]	0.1432	0.8759	104.42
Zero-1-to-3 [44]	0.1163	0.8764	49.94
SyncDreamer [45]	0.0960	0.8826	41.33
Ours	**0.0625**	**0.9181**	**30.25**

Table 2. **Novel view synthesis of full human bodies.** Quantitative comparison for full-body novel view synthesis on the THuman 2.0 [92] dataset. Our method demonstrates a considerable improvement over all the baselines across all metrics.

refer to the supplementary for the advantages of finetuning the UNet and further details on training the baseline methods, as well as an additional qualitative evaluation of novel view synthesis of faces with EG3D [9].

Novel view synthesis of faces: Fig. 3 and Tab. 1 show the qualitative and quantitative results of our method versus the baselines for novel view synthesis on the FaceScape test subjects. PixelNeRF and SSD-NeRF both preserve the re-

semblance, however, produce blurry results. SyncDreamer generates views with good multi-view consistency and reconstructs relatively accurate facial expressions, but sometimes fails to preserve the resemblance. Zero-1-to-3 preserves resemblance to some extent, but the generated results are slightly distorted and are misaligned with the target views, which causes inferior results on the structural and keypoint metrics. This is caused by the fact that the method assumes a uniform camera intrinsic matrix for all data; this way, the varying intrinsic camera information in the considered dataset is never taken into account by the model. Our method produces the best scores on most of the metrics while preserving good resemblance, which is attributed to the effective conditioning on the morphable model.

Novel view synthesis of full bodies: Fig. 4 and Tab. 2 demonstrates more significant improvement of full body reconstructions using our method compared to the facial data, indicating the necessity of a human prior in a diffusion pipeline. However, the generated full body images for all methods lack details due to the relatively low resolutions.

4.2. Novel facial expression synthesis

We further evaluate our method for multi-view consistent novel facial expression synthesis. We hold out the "jaw_right" expression in the FaceScape dataset from training and train our model on the remaining 19 expressions.

Baselines: Using the same metrics as for novel view synthesis, we compare our method with DiffusionRig [16], DECA [18], and MoFaNeRF [97]. Note that DiffusionRig requires per-subject finetuning with additional images and is thus denoted with the * symbol. For this baseline, we randomly choose 20 views from the 20 expressions and finetune the second stage for 5k steps for each subject. For a fair comparison to DECA, we fit FLAME [39] parameters to the ground truth meshes of the test expression in FaceScape's bilinear topology, instead of using the predicted ones. We then simply render the predicted albedo maps onto the ground truth meshes and use them to render albedo and lambertian images for 2D conditioning in DiffusionRig, Although the pre-trained model of MoFaNeRF has already been trained on all 20 facial expressions and some of our test subjects in the FaceScape dataset, we still treat the input image in a random facial expression as a novel subject and finetune the pre-trained model for 2k steps (as suggested by the authors).

Disentanglement of reconstruction and animation: We propose an efficient training scheme that disentangles the reconstruction (guided by the input image) and the animation (guided by the underlying morphable model). At every training step, given an input image of a subject, we randomly sample a novel facial expression from the dataset that does not align with the expression of the input im-

	LPIPS ↓	SSIM ↑	FID ↓	PCK@0.2 ↑	Re-ID ↑
DiffusionRig* [16]	0.2534	0.7438	42.93	89.74	97.67
DECA [18]	0.3393	0.6904	182.25	91.06	25.23
MoFaNeRF [97]	0.2877	0.6956	32.92	79.81	88.97
Ours	**0.1693**	**0.8026**	**14.34**	**95.46**	**99.89**

Table 3. **Novel expression synthesis.** Quantitative evaluation for novel expression synthesis on the held-out "jaw_right" expression. When evaluating against ground truth images with a white background, we convert the background pixels to white for all baselines using the ground truth masks. Our method outperforms the baselines on all metrics.

age. This cross-expression training, besides allowing explicit control of generated images, further improves the training efficiency as it serves as a data augmentation procedure and reduces potential overfitting on small datasets. At inference time, we could either use the expression that corresponds to the input image to enable novel view synthesis or use a novel expression to animate the reconstruction. We refer the readers to the ablation studies section in the supplementary for more discussion about the effectiveness of this proposed training regime.

Fig. 5 and Tab. 3 show the qualitative and quantitative results of our method versus the baselines on the novel expression synthesis task. DiffusionRig is unable to control eye poses since DECA that it depends on for mesh reconstruction during the first stage training does not model eye poses. It also generates slightly distorted faces in extreme head poses. DECA renders meshes with the predicted albedo maps, which is unable to produce photorealistic facial details. Expression rigging with MoFaNeRF loses resemblance and produces artifacts when the input facial expressions are not neutral. Our method, on the other hand, is able to generate realistic facial images under novel expressions from various views while retaining high visual quality, with input image in any facial expression.

5. Limitations and future work

While our morphable model exhibits promising capabilities, it is essential to acknowledge its inherent limitations that could impact its widespread applicability. As shown in Fig. 6, one significant constraint arises from the current inability to generalize effectively across various ethnicities and hair types, primarily stemming from the constraints of the FaceScape dataset, which predominantly features Asian subjects wearing a distinctive red cap. This limited diversity poses a challenge in training the model to handle the rich variability present in real-world scenarios, where individuals showcase diverse ethnicities and hair textures, lengths, and styles. We believe that this limitation can be alleviated with the advent of more diverse head datasets, such as RenderMe-360 [54].

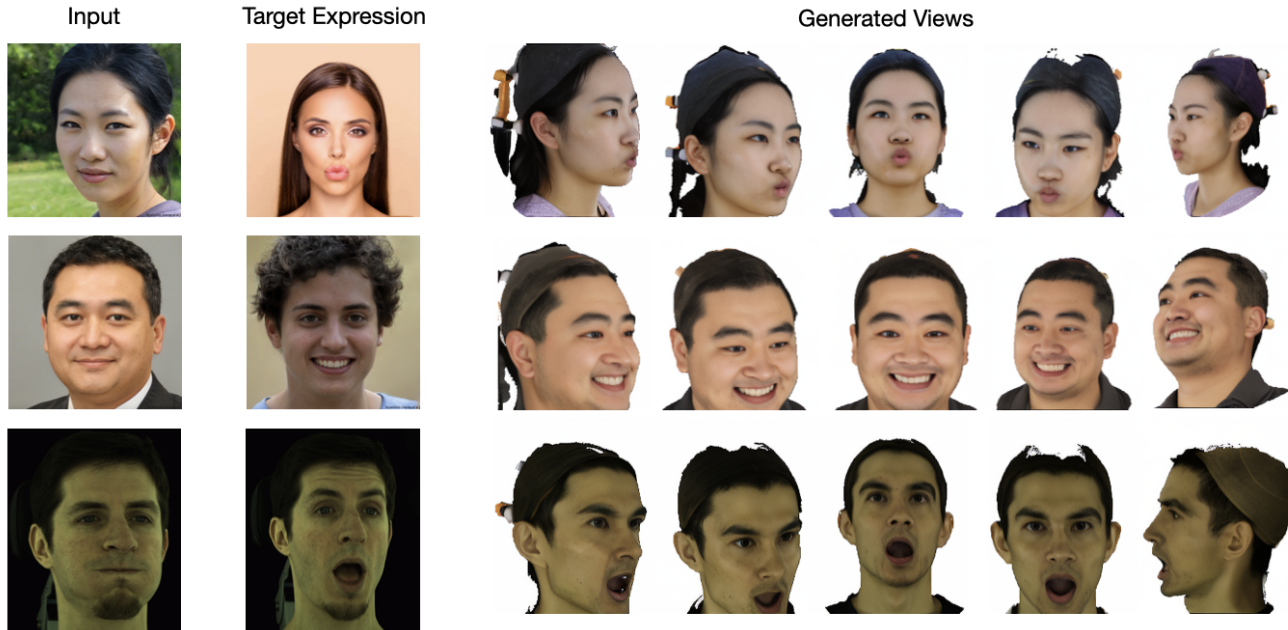Additionally, we find that our pipeline does not gener-

Figure 6. **Generalization abilities.** We show the results of our novel facial expression synthesis model on two Asian subjects generated by StyleGAN2 [34] and one subject from the Multiface dataset [85]. Our model is capable of maintaining 3D consistency of the generated imagery and preserving target facial expression. However, it struggles with the hairstyle reconstruction and out-of-distribution ethnicities due to the strong bias of the training data towards hair caps and Asian subjects. We use MICA [98] to optimize for the shape and expression parameters of the FLAME model from the input and target expression images, respectively. All results in this figure are generated using the model trained with FLAME fittings. For more details about models trained with bilinear / FLAME topologies, please refer to the supplementary. The target expression image in the first row is courtesy of © Deagreez / iStock.

alize well to out-of-distribution camera parameters that are significantly different from the ones used during training. Therefore, our current implementation relies on an external NeRF-based method, such as [77, 80], for comprehensive 3D facial reconstruction and free-path novel view synthesis. While this approach has been proven effective, it introduces a dependency on an external system, potentially impacting the model's standalone usability and flexibility. Improving the generalizability to any camera views and exploring ways to integrate a more self-contained 3D reconstruction process within our model could be a valuable avenue for future research.

Finally, the resolution of our model is limited at 256×256 as we inherit this limitation from the components proposed in the previous works [44, 45]. Rendering full bodies under such a low resolution provides a limited amount of details. Future works could involve training a full-body specific model from scratch at a higher resolution or integrating a super-resolution module into the pipeline.

6. Conclusion

Our work presents a method for avatar creation through the introduction of a novel morphable diffusion model. By seamlessly integrating a 3D morphable model into a multi-

view consistent diffusion framework, we have successfully incorporated a powerful human prior. This serves a dual purpose: firstly, it enhances the finetuning process for generative diffusion models, allowing for effective adaptation from extensive datasets to more limited, human-specific datasets. Secondly, it empowers the direct manipulation of synthesized novel views, offering a level of control and realism that surpasses current methodologies. Our qualitative and quantitative evaluation demonstrates the superior quality achieved by our model in comparison to state-of-the-art methods. We hope that our method will positively contribute to accelerating the field of photorealistic digitization of humans and promote follow-up research to address the current limitations.

Acknowledgements

References

[1] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *CVPR*, 2022. 2

[2] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *CVPR*, 2023. 3

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 5

[4] Alexander Bergman, Petr Kellnhofer, Wang Yifan, Eric Chan, David Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. In *NeurIPS*, 2022. 3

[5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 2023. 1, 2

[6] Egor Burkov, Ruslan Rakhimov, Aleksandr Safin, Evgeny Burnaev, and Victor Lempitsky. Multi-neus: 3d head portraits from single image with neural implicit functions. *IEEE Access*, 2023. 2

[7] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In *CVPR*, 2024. 3

[8] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *ICCV*, 2019. 2, 3

[9] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022. 2, 3, 6

[10] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *ICCV*, 2023. 5, 6

[11] Jianchuan Chen, Wentao Yi, Liqian Ma, Xu Jia, and Huchuan Lu. Gm-nerf: Learning generalizable model-based neural radiance fields from multi-view images. In *CVPR*, 2023. 2

[12] Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. GPAvatar: Generalizable and precise head avatar from image(s). In *ICLR*, 2024. 2

[13] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 5

[14] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023. 5

[15] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *NeurIPS*, 2023. 3

[16] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *CVPR*, 2023. 2, 6, 7

[17] Zijian Dong, Xu Chen, Jinlong Yang, Michael J Black, Otmar Hilliges, and Andreas Geiger. AG3D: Learning to Generate 3D Avatars from 2D Image Collections. In *ICCV*, 2023. 2, 3

[18] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Trans. Graph.*, 2021. 1, 2, 6, 7

[19] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018. 4

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

[21] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Trans. Graph*, 2021. 1, 2

[22] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *NeurIPS*, 2020. 2

[23] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *ICCV*, 2021. 2

[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 5

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 4

[26] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *CVPR*, 2022. 2

[27] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 3

[28] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *CVPR*, 2020. 2

[29] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2, 3

[30] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *CVPR*, 2023. 2

[31] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2

[32] Kacper Kania, Kwang Moo Yi, Marek Kowalski, Tomasz Trzciński, and Andrea Tagliasacchi. Conerf: Controllable neural radiance fields. In *CVPR*, 2022. 2

[33] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 3

[34] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 8

[35] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *ECCV*, 2022. 2

[36] Davis E. King. Dlib-ml: A machine learning toolkit. *JMLR*, 2009. 6

[37] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 1, 2

[38] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *NeurIPS*, 2023. 3

[39] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *SIGGRAPH Asia*, 2017. 1, 2, 4, 7

[40] Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot neural head avatar. *NeurIPS*, 2023. 2

[41] Yuwei Li, Longwen Zhang, Zesong Qiu, Yingwenqi Jiang, Nianyi Li, Yuexin Ma, Yuyao Zhang, Lan Xu, and Jingyi Yu. Nimble: a non-rigid hand model with bones and muscles. *ACM Trans. Graph.*, 2022. 2

[42] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *CVPR*, 2020. 3

[43] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. *arXiv preprint arXiv:2311.07885*, 2023. 3

[44] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023. 2, 3, 4, 5, 6, 8

[45] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Learning to generate multiview-consistent images from a single-view image. In *ICLR*, 2024. 2, 3, 4, 5, 6, 8

[46] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Trans. Graph.*, 2021. 2

[47] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 2015. 1, 2, 4

[48] Mohit Mendiratta, Xingang Pan, Mohamed Elgharib, Kartik Teotia, Mallikarjun B R, Ayush Tewari, Vladislav Golyanik, Adam Kortylewski, and Christian Theobalt. Avatarstudio: Text-driven editing of 3d dynamic human head avatars. *ACM Trans. Graph.*, 2023. 2, 3

[49] Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. Keypointnerf: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In *ECCV*, 2022. 2

[50] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2

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

[52] Thu H Nguyen-Phuoc, Christian Richardt, Long Mai, Yongliang Yang, and Niloy Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. In *NeurIPS*, 2020.

[53] 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. 3

[54] Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, and Kwan-Yee Lin. Renderme-360: Large digital asset library and benchmark towards high-fidelity head avatars". *NeurIPS*, 2023. 7

[55] 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*, 2019. 1, 2, 5

[56] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *IEEE international conference on advanced video and signal based surveillance*, 2009. 1, 2

[57] 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. 1, 2, 4

[58] P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. The feret evaluation methodology for face-recognition algorithms. *PAMI*, 2000. 5

[59] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 2

[60] Malte Prinzler, Otmar Hilliges, and Justus Thies. Diner: Depth-aware image-based neural radiance fields. In *CVPR*, 2023. 5

[61] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *CVPR*, 2024. 1, 2

[62] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4

[63] Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. Lolnerf: Learn from one look. In *CVPR*, 2022. 2

[64] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 4

[65] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *SIGGRAPH Asia*, 2017. 1, 2

[66] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4

[67] 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 *ICCV*, 2019. 2

[68] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020. 2

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

[70] Zifan Shi, Sida Peng, Yinghao Xu, Andreas Geiger, Yiyi Liao, and Yujun Shen. Deep generative models on 3d representations: A survey. *arXiv preprint arXiv:2210.15663*, 2022. 3

[71] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3

[72] Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. Next3d: Generative neural texture rasterization for 3d-aware head avatars. In *CVPR*, 2023. 3

[73] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *CVPR*, 2020. 2, 3

[74] Alex Trevithick, Matthew Chan, Michael Stengel, Eric Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. Real-time radiance fields for single-image portrait view synthesis. *ACM Trans. Graph.*, 2023. 2

[75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[76] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *PAMI*, 2021. 5

[77] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 2021. 8

[78] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *ECCV*, 2022. 1, 2

[79] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *CVPR*, 2023. 3

[80] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *ICCV*, 2023. 1, 6, 8

[81] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004. 5

[82] 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. 1, 2

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

[84] Yiqian Wu, Hao Xu, Xiangjun Tang, Hongbo Fu, and Xiaogang Jin. 3dportraitgan: Learning one-quarter headshot 3d gans from a single-view portrait dataset with diverse body poses. *arXiv preprint arXiv:2307.14770*, 2023. 3

[85] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Xuhua Huang, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. Multiface: A dataset for neural face rendering. *arXiv preprint arXiv:2207.11243*, 2022. 8

[86] 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*, 2022. 1, 2

[87] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *CVPR*, 2022. 2

[88] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *CVPR*, 2023. 2

[89] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *CVPR*, 2020. 2

[90] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *CVPR*, 2020. 2, 4, 5, 6

[91] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 5, 6

[92] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *CVPR*, 2021. 2, 5, 6

[93] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3

[94] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[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 *CVPR*, 2022. 1, 2

[96] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *CVPR*, 2023. 2

[97] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. In *ECCV*, 2022. 2, 6, 7

[98] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *ECCV*, 2022. 8

[99] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *CVPR*, 2023. 1, 2