DPHMs: Diffusion Parametric Head Models for Depth-based Tracking

Jiapeng Tang[1] Angela Dai[1] Yinyu Nie[1] Lev Markhasin[2] Justus Thies[3] Matthias Nießner[1]

[1] Technical University of Munich [2] Sony Semiconductor Solutions Europe
[3] Technical University of Darmstadt

https://tangjiapeng.github.io/projects/DPHMs

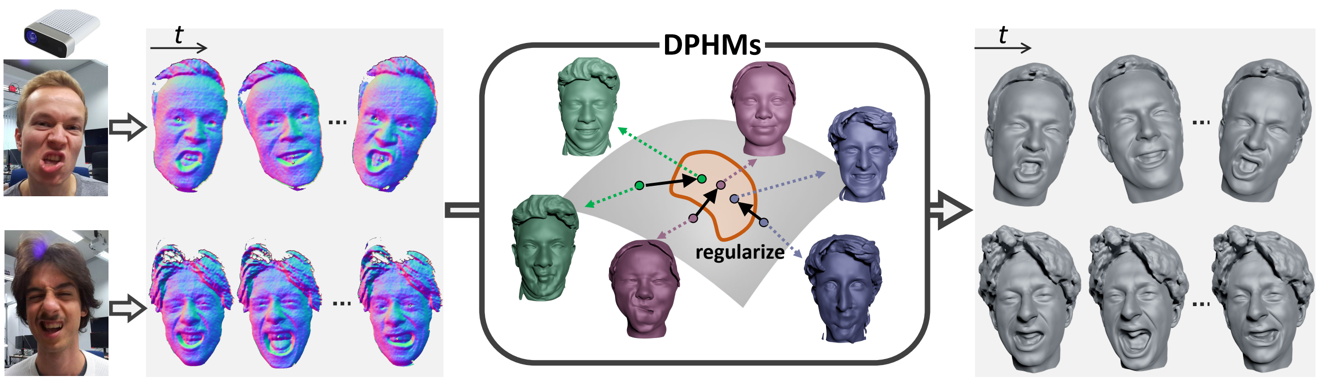

Figure 1. We present DPHMs, a diffusion parametric head model that is used for robust head reconstruction and expression tracking from monocular depth sequences. Leveraging the DPHM diffusion prior, we effectively constrain the identity and expression codes on the underlying latent manifold when fitting to noisy and partial observations of commodity depth sensors.

Abstract

We introduce Diffusion Parametric Head Models (DPHMs), a generative model that enables robust volumetric head reconstruction and tracking from monocular depth sequences. While recent volumetric head models, such as NPHMs, can now excel in representing high-fidelity head geometries, tracking and reconstructing heads from real-world single-view depth sequences remains very challenging, as the fitting to partial and noisy observations is underconstrained. To tackle these challenges, we propose a latent diffusion-based prior to regularize volumetric head reconstruction and tracking. This prior-based regularizer effectively constrains the identity and expression codes to lie on the underlying latent manifold which represents plausible head shapes. To evaluate the effectiveness of the diffusion-based prior, we collect a dataset of monocular Kinect sequences consisting of various complex facial expression motions and rapid transitions. We compare our method to state-of-the-art tracking methods and demonstrate improved head identity reconstruction as well as robust expression tracking.

1. Introduction

The fascination with 3D models of human heads spans across millennia, evolving from sculptural artistry to the realm of computer graphics. Creating a digital twin for each person is poised to revolutionize entertainment and communication, potentially transforming applications in video calls, augmented reality, virtual reality, animated movies, and gaming. For this revolution, we have to use affordable hardware that is accessible to everyone like webcams or Kinect-like depth sensors that are built into smartphones and laptops. Especially to capture the metrical size of a human, the depth sensors play an important role, as actual distances are measured. However, with a single depth sensor, only a part of the person is visible and the visible part contains sensor and surface-dependent noise. This leads to a generally under-constraint reconstruction problem which needs to be addressed with data priors [5–8, 10, 11, 40, 51, 53].

The most common data priors for face reconstruction are so-called 3D morphable models (3DMM) [5, 40] which capture facial shape and expression variations using Principal Component Analysis. However, these 3DMMs use a fixed template mesh which restricts their capacity to model the full landscape of head identity geometries and expressions (*e.g.*, diverse hairstyles, wrinkles, intricate faces, etc.). Recently, Neural Parametric Head Models (NPHMs) [24] have overcome these limitations by modeling full-head avatars with a broad spectrum of hair geometries and intricate non-linear facial deformations through an over-parameterized coordinate MLP-based neural field. This over-parametrization, however, is a key limitation when NPHMs are used for an underconstrained reconstruction task. Noisy or sparse input data leads to severe overfitting of NPHMs, with highly unrealistic head reconstructions.

To this end, we introduce DPHMs, the first diffusion generative model designed to generate clean and diverse 3D heads from noisy NPHM latent representations. Our key idea is to couple NPHMs with denoising diffusion models that can produce high-fidelity and diverse samples by navigating in the latent space. We learn identity and expression parametric diffusion through iterative transitions between noisy and clean latent representations, over-parametrized by NPHMs. We leverage the noise estimation during diffusion of the DPHM model to represent the gradient of the identity and expression latent distributions, enabling effective regularization of the identity and expression while fitting to real sequences for commodity depth sensors. As illustrated in Fig. 1, when fitting NPHMs to noisy and partial data, latent vectors might fall outside the underlying latent surface manifold, generating implausible head geometries. Using our DPHM prior, those latents can be regularized towards latent vectors that are on the surface manifold, generating high-quality head geometries.

We evaluate our proposed method on a new challenging benchmark that contains various extreme facial expression motions with rapid transitions captured with a monocular Kinect Azure sensor. Extensive experiments and comparisons against recent state-of-the-art head reconstruction methods demonstrate that our approach can reconstruct more accurate head geometries and expressions and achieve more robust and coherent facial expression tracking.

Our contributions can be summarized as follows:

- We propose the first diffusion generative model that creates clean and diverse 3D heads by explicitly learning the distributions of identity and expression latent defined in neural parametric head models.
- We design novel regularization terms based on diffusion parametric head models, effectively constraining the latent optimization when fitting sparse and noisy observations from monocular depth sequences.
- We collect a dataset of monocular Kinect scan sequences with various challenging facial expression motions for evaluation benchmark.

2. Related Work

3D morphable face and head models. The conception of 3D morphable face and head models can be dated back to the pioneering work of Blanz and Vetter [5]. They introduced the concept of a 3D Morphable Face Model (3DMM) based on a dataset of 200 3D face scans and used Principal Component Analysis (PCA) to represent facial shape and texture variations compactly. To enhance expressiveness, Cootes et al. introduced the Active Appearance Model (AAM)[17]. Some subsequent models incorporated more captured data [7, 8, 11, 51, 53]. Advanced facial models were designed to go beyond linear spaces. These include multi-linear models [6, 10], non-linear models [76], and the FLAME model [40], which seamlessly integrates linear shape spaces with articulated head components. Recently, researchers have explored integrating Signed Distance Functions (SDFs) and deformation fields for human faces [24, 25, 84, 89], bodies [48, 49, 70], and animals [38, 71]. While these neural parametric models excel in producing high-fidelity geometries and estimating complex non-linear deformations, they often struggle to generate reasonable samples from random noise. In contrast, our approach enhances neural parametric models with diffusion models, effectively mapping random noise latent vectors onto the desired surface manifold.

Head reconstruction and tracking. Building upon the data priors of 3D morphable models, many works [19, 20, 23, 26, 28, 59, 62, 73, 75, 77, 93] tried to reconstruct 3D faces from monocular images or videos. Our work is more closely related to those endeavors focusing on the 3D face/-head reconstruction from scans [9, 42, 88]. For a comprehensive overview of early face tracking from scan sequences, we recommend referring to [52]. An early attempt at tracking using commodity depth sensors can be found in [79], where pre-recorded animation priors were applied. Li [39] proposed a real-time method for Kinect tracking. Thies et al. integrated RGBD face tracking with facial reenactment in [74]. More recently, several works [2, 13, 14, 21, 41, 55, 82, 90, 91] have attempted to reconstruct human heads using coordinate-MLP representations. A concurrent work of MonoNPHM [25] extends NPHMs for head tracking from monocular RGB videos. IMAvatar [90] and PointAvatar [91] utilize coordinate-MLP to build personalized canonical head geometries, which are tracked through skinning weight fields generalized from the FLAME model. However, these methods struggle to reconstruct high-quality head avatars with intricate details due to a lack of effective geometry priors. In contrast, we learn high-quality priors from high-resolution full-head scans.

Diffusion models. Denoising Diffusion Models (DDMs) [30, 63–66] have shown unprecedented generation diversity and realism in various data domains, including images [4, 22, 30, 31, 35, 43, 45, 47, 56, 58], videos [16, 29, 34, 83], and shapes [12, 32, 44, 72, 85–87, 92] To synthesize high-dimensional data effectively, some methods compress physical data into a high-dimensional latent space [1, 12, 56, 72, 85, 87] and learn the distribution of latent features. Our approach adopts a similar strategy, parametrizing high-resolution head avatar geometry into two separate latent spaces (identity and expression) and leveraging latent diffusion models to learn their data distributions. Some recent works leverage the learned priors from diffusion models to guide the NeRF optimization [46], such as DreamFusion [54], Diffusion-NeRF [81], and Single Stage NeRF [15]. A concurrent work of FaceTalk [3] uses diffusion models to synthesize temporally consistent 3D motion sequences of human heads based on NPHMs [24]. In this work, we design diffusion prior regularizers to effectively constrain the identity and expression codes to lie on the underlying latent manifold that represents plausible head shapes, producing robust head reconstruction tracking when fitting to noisy and partial scans of commodity depth sensors.

3. Approach

Given a monocular sequence of depth maps $\mathcal{I}$, we aim to reconstruct a series of full-head avatars $\mathcal{O}$, see Fig. 2. To accurately reconstruct various facial expressions with topological variations, we choose to predict a continuous signed distance field for each frame using a modified NPHM model, which is parametrized by $\mathbf{z}^{id}$ and $\mathbf{z}_i^{ex}$. However, the latent optimization to real-world single-view depth sequences is extremely challenging, as the fitting to partial and noisy observations is underconstrained. To overcome these challenges, we introduce Diffusion Parametric Head Models (DPHMs), the first diffusion generative model tailored for generating clean and diverse 3D heads from noisy latent representations. It is used as a prior to regularize the NPHM identity and expression codes which significantly improves head tracking robustness from noisy and partial scans.

3.1. Diffusion Parametric Head Model

Our Diffusion Parametric Head Model (DPHM) is designed to learn a diffusion generative model to enable robust head tracking and reconstruction. We build a disentangled latent space of shape and expression, inspired by Neural Parametric Head Models (NPHMs) [24]. NPHMs represent head geometries using an SDF decoder in canonical space and capture facial expressions through forward deformations. However, this approach maintains the same mesh connectivity in the canonical identities, limiting the ability to change topologies during expression tracking. To address this limitation, we replace forward deformations with backward deformations, which learn the deformation fields from arbitrary expressions to canonical space. This enables

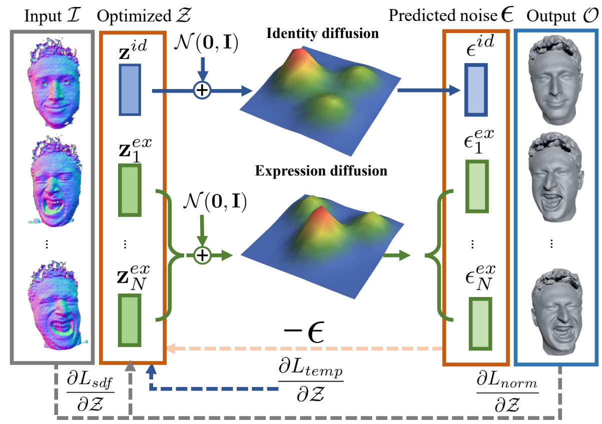

Figure 2. **DPHMs for depth-based tracking.** Given a sequence of depth maps $\mathcal{I}$ of N frames, our objective is to reconstruct a full-head avatar $\mathcal{O}$ including its expression transitions. To achieve this, we optimize the parametric latent $\mathcal{Z} = \{\mathbf{z}^{id}, \mathbf{z}_1^{ex}, ..., \mathbf{z}_N^{ex}\}$ of NPHM that can be decoded into continuous signed distance fields $\mathcal{O}$ by identity and expression decoders. To align with the observations, we calculate data terms L_{sdf} and L_{norm} between $\mathcal{I}$ and $\mathcal{O}$. However, high-level noise still makes navigating the latent optimization extremely challenging. At the core of our method is an effective latent regularization using diffusion priors; we add Gaussian noises to $\mathcal{Z}$ and then pass them into identity and expression diffusion models to predict perturbed noise ϵ for updating $\mathcal{Z}$. The diffusion regularizer guides $\mathbf{z}^{id}$ and $\mathbf{z}_i^{ex}$ towards the individual manifold of their distributions via ϵ^{id} and ϵ^{ex}, ensuring plausible head geometry reconstruction and robust tracking. To enhance temporal coherence, L_{temp} penalizes inconsistency between $\mathbf{z}_i^{ex}$ of nearby frames.

the reconstruction of a continuous signed distance field for each expression space by wrapping points from the expression to canonical space and querying SDF values. Please refer to the supplementary material for more details about the revised NPHM model based on backward deformations.

3.1.1 NPHM Identity Space

Following NPHMs [24], we represent the signed distance field of canonical identity by an ensemble of several smaller local MLP-based networks individually responsible for local regions centered at 39 pre-defined anchors of human heads. Concretely, we define $K = 2K_{\text{sym}} + K_{\text{usym}}$ facial anchors, denoted as $\mathbf{a} \in \mathbb{R}^{K\times 3}$, which are estimated by a small MLP_{anc} from the global latent $\mathbf{z}_{\text{glo}}^{id}$. K_{sym} anchors are on the left face. They are mirrored to the other K_{sym} ones on the right face. K_{usym} anchors are in the middle of the face, shared by both the left and right faces. For k_{th} local region of facial anchor, we represent its local geometry via a local latent vector $\mathbf{z}_{\mathbf{k}}^{\mathbf{id}}$, along with the global latent vector $\mathbf{z}_{\text{glo}}^{\mathbf{id}}$ as well as an SDF decoder MLP_{θ_k} parametrized by learnable weights θ_k:

$$f_k(\mathbf{p}, \mathbf{z}_{\mathbf{k}}^{\mathbf{id}}, \mathbf{z}_{\mathbf{glo}}^{\mathbf{id}}) = \text{MLP}_{\theta_k}([\mathbf{p} - \mathbf{a}_k, \mathbf{z}_{\mathbf{k}}^{\mathbf{id}}, \mathbf{z}^{\mathbf{glo}}]). \quad (1)$$

Finally, we can composite local fields into a global field:

$$F_{id}(\mathbf{p}) = \sum_{k=1}^{K} \omega_k(\mathbf{p}, \mathbf{a_k}) f_k(\mathbf{p}, \mathbf{z_k^{id}}, \mathbf{z_{glo}^{id}}). \tag{2}$$

The blending weights are calculated by a Gaussian kernel based on the Euclidean distance between the query point $\mathbf{p}$ and $\mathbf{a_k}$.

3.1.2 Backward-Deformation Expression Space

In contrast to NPHMs [24], which define their expression space through forward deformations, limiting topology to that of the canonical shape, we model the expression space through globally conditioned backward deformation fields learned by a MLP_ψ [3, 24]. We use a latent expression vector $\mathbf{z}^{ex}$ to explain the geometry variations caused by expression transitions. Since such a deformation is also closely related to the identity geometry, the deformation decoder also receives the identity code $\mathbf{z}^{id} = \mathbf{z}^{id}_{glo} \oplus \mathbf{z}^{id}_1 \cdots \oplus \mathbf{z}^{id}_K$ as an additional condition. To enforce the deformation network to learn a disentangled expression representation independent of the identity latent, we impose a constraint that neutral expressions for the canonical space are close to zeros:

$$F_{ex}(\mathbf{p}) = \mathrm{MLP}_\phi([\mathbf{p}, \mathbf{z}^{id}, \mathbf{z}^{ex}]). \tag{3}$$

We jointly train the identity and expression networks in an auto-decoder fashion. Once finished, we can obtain a set of identity latents $\mathcal{Z}^{id} = \{\mathbf{z}^{id}_j\}_j$ for canonical geometries and a set of expression latents $\mathcal{Z}^{ex} = \{\mathbf{z}^{ex}_{j,l}\}_{j,l}$, where $\mathbf{z}^{ex}_{j,l}$ means the l_{th} expression of j_{th} subject.

3.1.3 Neural Parametric Diffusion

In this section, we will explain how DPHMs are learned and their connection with score functions. To learn the identity and expression latent diffusion, we consider them as flattened 1D latent vectors. The identity and expression diffusion models are treated analogously; the only difference is the number of input and output channels in the denoising network. In the following, we will use the expression latent diffusion to explain the details. Given an expression code $\mathbf{x}_0$ sampled from $\mathcal{Z}^{ex}$ over-parametrized from the training dataset. The forward diffusion process progressively adds Gaussian noise to $\mathbf{x}_0$, obtaining a series of corrupted versions $\mathbf{x}_1, ..., \mathbf{x}_T$, according to a linearly increased noise variance schedule $\beta_1, ..., \beta_T$ along the step t ($\beta_1 < ... < \beta_T$). The diffusion step at time t is defined as:

$$\mathbf{x}_t := \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \beta_t \epsilon_{t-1}, \tag{4}$$

where $\epsilon_{t-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. With the re-parametrization trick, we obtain:

$$\mathbf{x}_t := \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{5}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\alpha_t := 1 - \beta_t$, $\bar{\alpha}_t := \prod_{r=1}^{t} \alpha_s$. The reverse diffusion process is tasked to remove the noise gradually. In each denoising step t, we have:

$$p_\phi(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\phi(\mathbf{x}_t, t), \tilde{\beta}_t(\mathbf{I})), \tag{6}$$

where $\tilde{\beta}_t := (1 - \overline{\alpha}_{t-1})\beta_t/(1 - \overline{\alpha}_t)$. Instead of directly predicting $\mu_\phi(\mathbf{x}_t, t)$, we opt to predict the noise $\epsilon_\phi(\mathbf{x}_t, t)$ using the denoiser ϵ_ϕ and the noise data $\mathbf{x}_t$. Then, $\mu_\phi(\mathbf{x}_t)$ can be re-parametrized by subtracting the predicted noise:

$$\mu_\phi(\mathbf{x}_t, t) := \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\phi(\mathbf{x}_t, t)). \tag{7}$$

By training the denoiser on various noise levels and time steps, we obtain the empirical risk expectation as the loss:

$$\mathbb{E}_{\mathbf{x}_0, \epsilon, t}[\frac{\beta_t}{2\alpha_t(1-\overline{\alpha}_t)}\|\epsilon - \epsilon_\phi(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2]. \tag{8}$$

As depicted in [78], a DDM noise estimator has a connection to score matching [33, 64–69] and is proportional to the score function:

$$\epsilon_\phi(\mathbf{x}_t, t) \propto -\nabla_{\mathbf{x}} \log p(\mathbf{x}). \tag{9}$$

Therefore, moving in the opposite direction of the noise predicted by the neural network is akin to approaching the modes of the data distribution. This principle can be harnessed for generating samples that closely resemble the data distribution through the Langevin dynamics [63, 80].

3.2. DPHM for Monocular Depth-based Tracking

Given a depth map sequence $\mathcal{I} = \{[\mathbf{P}_i, \mathbf{N}_i]\}_{i=1}^N$ of N frames, where $\mathbf{P}_i, \mathbf{N}_i$ are back-projected points and normals of i_{th} frame, our goal is to discover the identity vector $\mathbf{z}^{id}$ along with N expressive latent vectors $\mathbf{z}^{ex}_{1:N}$ that can be interpreted into full-head avatars $\mathcal{O}$ with hairs and accurate expression transitions through the pre-trained identity and deformation decoders in Sec. 3.1.1 and 3.1.2. According to Bayes's theorem, it is equivalent to maximizing the *posterior* probability of identity and expression latent $\mathcal{Z} = \{\mathbf{z}^{id}, \mathbf{z}^{ex}_{1:N}\}$. This can be further formulated as:

$$p(\mathcal{Z}|\mathcal{I}) = \frac{p(\mathcal{I}|\mathcal{Z})p(\mathcal{Z})}{p(\mathcal{I})} \tag{10}$$

In practice, we can maximize the log-posterior as follows:

$$\begin{aligned} \log p(\mathcal{Z}|\mathcal{I}) &= \log p(\mathcal{I}|\mathcal{Z}) + \log p(\mathcal{Z}) - \log p(\mathcal{I}) \\ &\geq \log p(\mathcal{I}|\mathcal{Z}) + \log p(\mathbf{z}^{id}) + \log(\mathbf{z}^{ex}_{1:N}). \end{aligned} \tag{11}$$

As $\mathcal{I}$ is neither independent on $\mathbf{z}^{id}$ nor $\mathbf{z}^{ex}_{1:N}$, we can drop $\log p(\mathcal{I})$ that is a normalizing constant. Based on the assumption that identity and expression are disjoint, we can consider $\mathbf{z}^{id}$ and $\mathbf{z}^{ex}_{1:N}$ are independent. We can update $\mathcal{Z}$ with stochastic gradient descent:

$$\begin{aligned} \nabla_{\mathcal{Z}} \log p(\mathcal{Z}|\mathcal{I}) = \nabla_{\mathcal{Z}} \log p(\mathcal{I}|\mathcal{Z}) + \nabla_{\mathbf{z}^{id}} \log p(\mathbf{z}^{id}) \\ + \nabla_{\mathbf{z}^{ex}_{1:N}} \log p(\mathbf{z}^{ex}_{1:N}), \end{aligned} \tag{12}$$

with the first term being the gradient of the log-likelihood $\log p(\mathcal{I}|\mathcal{Z})$, the other two terms are the gradient of the log-prior $\log p(\mathbf{z}^{id})$ and $\log p(\mathbf{z}^{ex}_{1:N})$ respectively. $\nabla_{\mathbf{z}^{id}} \log p(\mathbf{z}^{id})$ is exactly the score function of identity diffusion of DPHMs, while the log-prior of all expression latents within a sequence $\nabla_{\mathbf{z}^{ex}_{1:N}} \log p(\mathbf{z}^{ex}_{1:N}) \propto$

$\sum_{i=1}^{N} \log p(\mathbf{z}_i^{ex})$ which is the sum of score function of our expression latents for all observed frames.

By plugging Eq. 9 into 12, we can use the learned prior of our DPHMs over $\mathbf{z}^{id}, \mathbf{z}_i^{ex}$, and the temporal smoothness constrain between $\mathbf{z}_i^{ex}$ and $\mathbf{z}_{i-1}^{ex}$ to approximate the dependencies of near-by frames. For the first term of $\log p(\mathcal{I}|\mathcal{Z})$, we use the data terms by considering SDF prediction and normal in-consistency errors of observed points from $\mathcal{I}$. Thus, the gradient with respect to $\mathcal{Z}$ while minimizing the following loss function L is obtained by:

$$\nabla L = \nabla L_{sdf}(\mathcal{I}) + \lambda_{norm} \nabla L_{norm}(\mathcal{I}) - \lambda_{id} \epsilon^{id}(\hat{\mathbf{z}}^{id}) - \lambda_{ex} \sum_{i=1}^{N} \epsilon^{ex}(\hat{\mathbf{z}}_i^{ex}) + \lambda_{temp} \nabla L_{temp}(\mathbf{z}_{1:N}^{ex}). \tag{13}$$

note that $\hat{\mathbf{z}}$ is the corrupted version of $\mathbf{z}$ by adding Gaussian noise at diffusion step t defined in Eq. 5. $L_{sdf}(\mathcal{I})$ enforces a constraint that requires observed points from the i_{th} frame $\mathbf{P}_i$, when transformed back into the canonical space to get $\mathbf{P}_i^{id}$, to be precisely positioned on the zero iso-surface of the identity geometry:

$$L_{sdf}(\mathcal{I}) = \sum_{i=1}^{N} |F_{id}(\mathbf{P}_i^{id}, \mathbf{z}^{id}) - 0| \qquad \mathbf{P}_i^{id} = F_{ex}(\mathbf{P}_i, \mathbf{z}^{id}, \mathbf{z}_i^{ex}). \tag{14}$$

To further attain the facial geometry details, the normal in-consistency is penalized:

$$L_{norm}(\mathcal{I}) = \sum_{i=1}^{N} < \mathbf{N}_i^{ex}, \mathbf{N}_i >, \tag{15}$$

where $\mathbf{N}_i^{ex}$ is the gradient of the SDF values $F_{id}(\mathbf{P}_i^{id}, \mathbf{z}^{id})$ with respect to $\mathbf{P}_i$, based on the theorem in IGR [27]. The temporal smoothness term for the expressions is defined as:

$$L_{temp}(\mathbf{z}_{1:N}^{ex}) = \sum_{i=2}^{N} \|\mathbf{z}_i^{ex} - \mathbf{z}_{i-1}^{ex}\|_2^2. \tag{16}$$

4. DPHM-Kinect Dataset

To assess head tracking performance, researchers commonly use datasets with rendered depth map sequences from talking face mesh sequences, such as VOCA [18]. However, these sequences often capture limited lip movements during speech without complex, intricate facial expressions. In our work, we establish a challenging benchmark for head tracking and reconstruction. We construct a benchmark containing 130 single-view depth scan sequences, capturing diverse facial expressions in motion sequences, including rapid transitions. Specifically, we use a Microsoft Kinect Azure RGB-D camera to record data. The dataset includes eighteen men and eight women from different skin tones and ethnicities. For each participant, we recorded five sequences, where they quickly switched their expressions and could have global rotations, including 'smile and laugh,' 'eyeblinks,' 'fast-talking,' 'random facial expressions,' and 'mouth movements.' An example of the collected RGBD sequences is depicted in Fig. 3.

Figure 3. An example of captured DPHM-Kinect sequences with complex facial expressions and fast transitions.

5. Experiments

Dataset To learn high-quality head geometry priors, we use 4,760 high-fidelity head scans with varying expressions of 239 identities from NPHM as the training dataset. To evaluate the performance of head reconstruction and tracking from monocular scans, we additionally use the reconstructed single-view depth sequences from a multi-view video dataset [37] by COLMAP [60, 61]. We use twenty head motion sequences from 10 different identities.

Baselines We compare against state-of-the-art head tracking methods: FLAME [40], ImFace [89], ImAvatar [90], and NPHM [24]. FLAME is the most recent and advanced template-based PCA model. ImFace [89] is one of the pioneering works to integrate neural signed distance and deformation fields for face reconstruction. We also include ImFace trained on the NPHM dataset notated as ImFace*. ImAvatar [90] is an optimization method that utilizes a coordinate-MLP to reconstruct the personalized canonical head geometry and animates head motions through skinning weight fields, building upon FLAME. We also include depth supervision in the optimization of ImAvatar. To isolate the effects of our proposed parametric diffusion, we compare against the improved version of NPHM based on backward-deformation expression space.

Evaluation Metrics To evaluate the accuracy of the head reconstruction and tracking, we calculate the ℓ_2 distance error between the observed scans and reconstructed meshes. Concretely, for each point in the captured scan, we find the closest point in the reconstructed surface mesh and then calculate ℓ_2 distance (mm). Based on the nearest neighborhood search correspondence, we calculate the normal consistency (NC) through cosine similarity. Also, we provide Recall scores (RC) with thresholds of $\tau = 1.5mm$

Figure 4. Head Tracking on the DPHM-Kinect dataset. Note that RGB images are only used for reference not used by all the methods except ImAvatar. Compared to state-of-the-art methods, our approach achieves more accurate identity reconstruction with detailed hair geometries while tracking more plausible expressions, even during extreme mouth movements.

and $2\tau = 3mm$, demonstrating the percentage of observed points well approximated under a defined threshold.

Implementations Our DPHM is trained on a single RTX A6000. Initially, we train the identity and expression decoders on the NPHM dataset to obtain over-parametrized latent representations. This training involves using a learning rate of 0.001 with a batch size of 32 for 6,000 epochs. The learning rate is decayed by a factor of 2 every 500 epochs. Subsequently, we train the identity and expression parametric diffusion models, employing UNet-1D [57] as the denoising network based on DDPM [30]. These diffusion models are individually trained using a batch size of 32 and a learning rate of $8\mathrm{e}{-5}$ for 200,000 iterations.

The test-time optimization of head tracking involves three phases. We firstly optimize $\mathbf{z}^{id}$ and $\mathbf{z}_1^{ex}$ for the first frame. The optimization is performed for 200 iterations, with a learning rate of 0.01 and a decay of 0.1 every 50 iterations. Subsequently, we fix $\mathbf{z}^{id}$ and incrementally optimize $\mathbf{z}_{2:N}^{ex}$ for subsequent frames. The optimization uses a learning rate of 0.001 for 50 iterations, with a decay of 0.1 after 30 iterations. Finally, we jointly fine-tune $\mathbf{z}^{id}$ and $\mathbf{z}_{1:N}^{ex}$ with a learning rate of 0.0001 for 20 iterations. The hyperparameters in Equation 13 are set to $\lambda_{norm} = 0.025$, $\lambda_{id} = 0.25$, $\lambda_{ex} = 0.25$, $\lambda_{temp} = 0.5$. At each iteration, the noisy latent $\hat{\mathbf{z}}$ is generated by Eq. 5 with $t \in [0.4, 0.6]$. Please refer to the supplementary material for more details.

5.1. Comparison against state of the art

Head Tracking on DPHM-Kinect dataset. The qualitative comparison of our method with state-of-the-art reconstruction methods on the collected Kinect data is presented in Fig. 4. FLAME struggles to model hair geometries and has a limited capacity for intricate facial expressions related to mouth and eye movements. ImFace focuses only on the front face region. ImAvatar always produces over-smooth results due to a lack of high-quality geometry priors. In contrast, NPHMs and our DPHMs successfully reconstruct fine-grained hair geometries by utilizing volumetric SDFs to model complete head geometries, accommodating different hairstyles. However, NPHMs often fail to reconstruct plausible expressions in challenging scenarios with partial and noisy scans. On the other hand, our DPHMs effectively regulate latent optimization, resulting in plausible reconstructions and accurate tracking, even in complicated expressions shown in the second and third rows. Quantitative comparisons in Table 1 show that our approach consistently

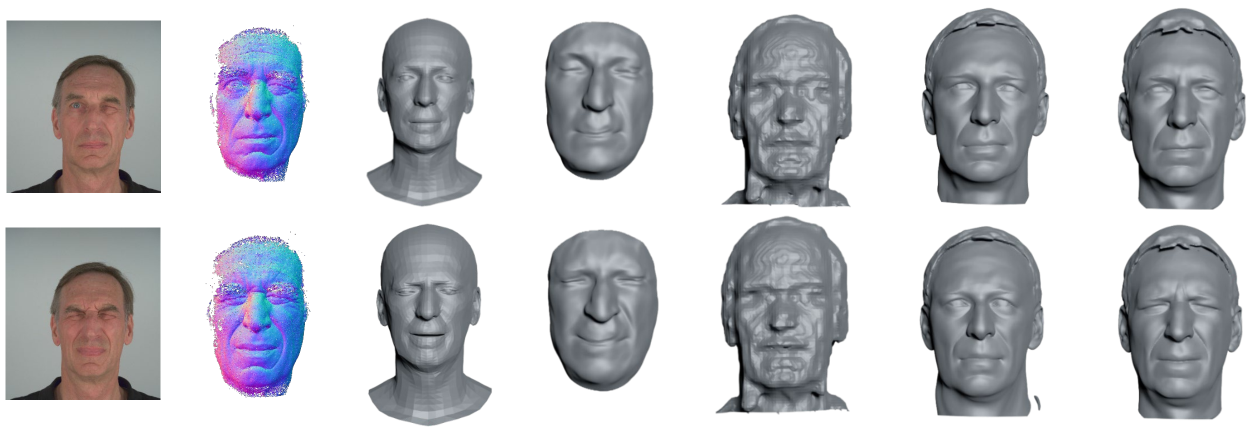

Figure 5. Head Reconstruction and Tracking on the single-view depth sequences of NerSemble [37]. Note that RGB images are only used for reference and not used by all methods except ImAvatar. Compared to state-of-the-art methods, our approach demonstrates the ability to reconstruct realistic head avatars with hairs and accurately capture intricate facial expressions such as eyelid movements.

outperforms all state-of-the-art methods, illustrating more accurate and robust tracking.

Method	FLAME	ImFace	ImFace*	IMAvatar	NPHM	Ours
$\ell_2 \downarrow$	5.251	11.21	11.00	5.581	1.579	**1.465**
NC $\uparrow$	83.88	84.82	84.00	70.23	85.97	**86.80**
RC $\uparrow$	15.80	57.70	58.74	11.30	65.95	**70.79**
RC2 $\uparrow$	53.11	71.88	72.99	40.75	88.33	**90.98**

Table 1. Quantitative comparison on the DPHM-Kinect dataset.

Head Tracking on NerSemble dataset Fig. 5 depicts the comparisons on reconstructed single-view depth sequences from NerSemble [37]. Again, we showcase the superiority of the diffusion prior-based regularization through improved numerical results across all metrics in Table 2. This is evident in the more accurate tracking of our results, particularly in terms of eye movements.

5.2. Ablation Studies

We conduct detailed ablation studies to verify the effectiveness of each design in our approach (see Table 3 and Fig. 6).

What is the effect of backward deformation networks? An alternative approach to modeling facial expressions is to use forward deformation networks, as employed in the original NPHMs, which warp points from canonical space to expression space. Since all frames share the same mesh topology as in canonical space, it has a limited capacity to track complicated expressions. They always fail to reconstruct facial expressions with an open mouth.

What is the effect of the diffusion regularizer? We also explore using other regularizers to constrain the latent optimization. In line with V-Poser [50], we trained a variational autoencoder (VAE) [36] using the over-parametrized latent and attempted to constrain the latent optimization using the VAE prior. However, we observed that the VAE priors cannot ensure plausible head identity reconstruction and fail to track correct expressions. This highlights the superiority of our diffusion prior-based regularizer.

Method	FLAME	ImFace	ImFace*	IMAvatar	NPHM	Ours
ℓ_2	3.286	1.560	1.557	3.337	1.201	**1.103**
NC	84.00	87.24	87.44	77.79	87.76	**88.73**
RC	21.06	80.57	80.60	14.97	84.73	**89.96**
RC2	60.70	92.69	93.16	54.26	96.63	**97.20**

Table 2. Quantitative comparison on dynamic point cloud sequences reconstructed from multi-view video dataset [37].

Method	forward	VAE prior	w/o. exp. diff.	w/o. iden. diff	Ours
$\ell_2 \downarrow$	1.367	2.929	1.283	1.159	**1.146**
NC $\uparrow$	80.58	82.82	90.60	91.13	**91.32**
RC $\uparrow$	80.47	70.72	75.58	80.61	**81.33**
RC2 $\uparrow$	93.20	68.85	94.23	95.93	**96.08**

Table 3. Ablation studies on the DPHM-Kinect dataset.

What is the effect of the expression diffusion regularizer? We replace the expression diffusion regularizer with the simple one used in NPHM, which constrains the expression around the neutral state. Due to the weaker constraints, it always falls outside of the underlying surface manifold, leading to incorrect expression reconstruction. The obvious improvement in terms of numerical results can also verify the effective constraint of expression diffusion regularizer.

What is the effect of the identity diffusion regularizer? We replace the identity diffusion regularizer with the simpler one used in NPHM, which constrains the identity geometry to be close to the mean head in the training set. However, we observed that this simplified regularizer could not accurately reconstruct facial geometry details, particularly resulting in high errors in the nose part.

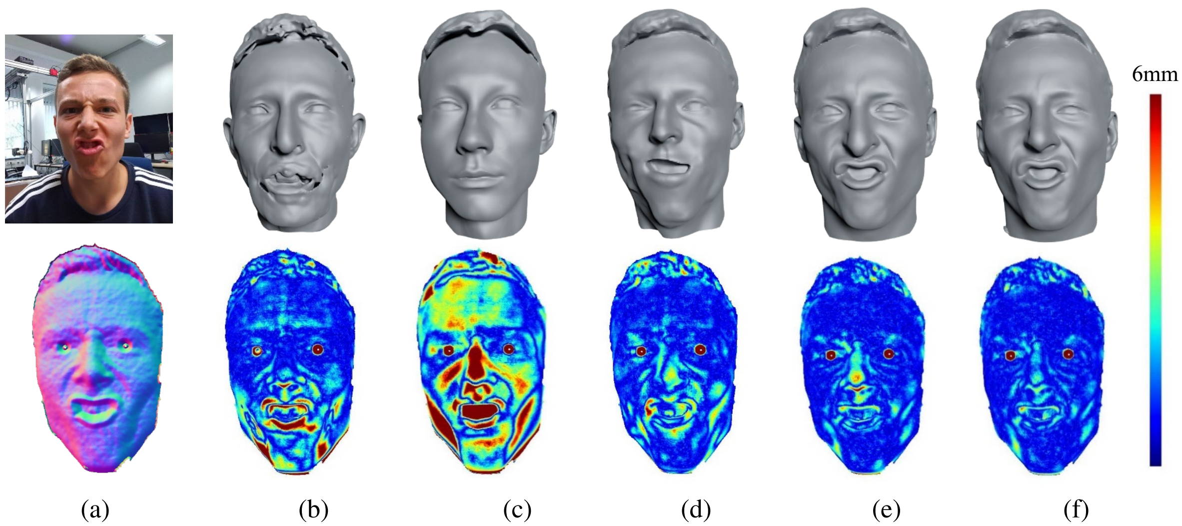

Figure 6. **Ablation Studies** (a) RGB reference & Input Scans; (b) Ours with forward deformations; (c) Ours with VAE priors; (d) Ours without expression diffusion; (e) Ours without identity diffusion; (f) Ours. Note that RGB images are only used for reference not used by all the methods except ImAvatar. We visualize the scan2mesh distance error map at the bottom. Our final model captures complicated expressions with lower identity reconstruction errors.

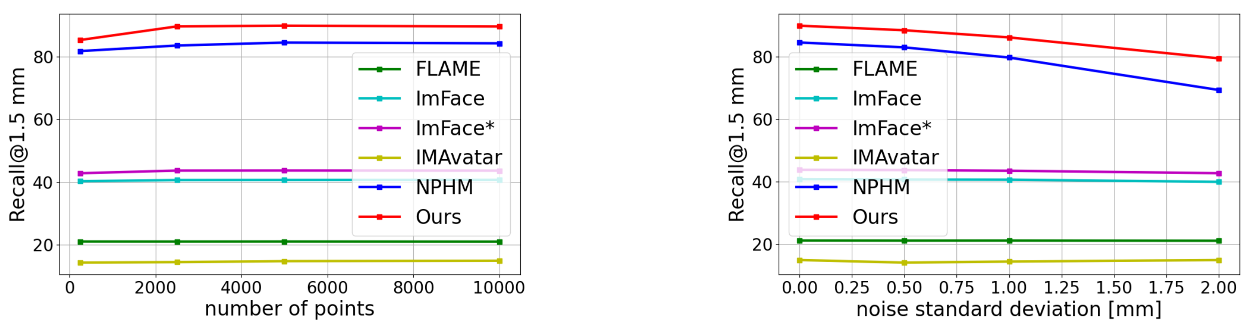

Figure 7. Robustness analysis with varying sparsity levels and additive Gaussian noise intensities. DPHMs demonstrate superior robustness compared to NPHMs across diverse imperfect observations

5.3. Robustness Analysis

To assess the robustness of our diffusion prior regularizations across various imperfect observations, we apply DPHMs to downsampled scans from NerSemble at sparsity levels of 10,000, 5,000, 2,500, and 250 points. We also evaluate varying intensities of additive Gaussian noise (standard deviations of 0mm, 0.5mm, 1mm, and 2mm) to sparse point clouds of size 5,000. Fig. 7 illustrates the Recall@$1.5mm$ curve under different sparsity and noise levels. Our method consistently outperforms all state-of-the-arts, demonstrating the superior robustness of our method.

5.4. Limitations

Despite achieving superior results, our approach currently has a limitation of slower inference due to the test-time optimization of neural parametric models. In the future, we will focus on real-time head-tracking solutions.

6. Conclusion

We introduced Diffusion Parametric Head Models (DPHMs), the first diffusion generative model enabling robust head reconstruction and tracking from real-world single-view depth sequences. Leveraging the diffusion priors of DPHMs, we designed a novel regularizer that effectively constrains the identity and expression codes on the underlying latent manifolds. Extensive experiments and comparisons against state-of-the-art head reconstruction methods on a new challenging benchmark demonstrate that our method can reconstruct more accurate head geometries and achieve more robust and coherent expression tracking.

Acknowledgement. This work was supported by the ERC Starting Grant Scan2CAD (804724), the German Research Foundation (DFG) Research Unit "Learning and Simulation in Visual Computing, the ERC Starting Grant SpatialSem (101076253), as well as Sony Semiconductor Solutions. We thank Simon Giebenhain for close discussions and constructive suggestions during the project.

References

[1] Stable diffusion. *https://github.com/Stability-AI/stablediffusion*, 2022. 3

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

[3] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. Facetalk: Audio-driven motion diffusion for neural parametric head models, 2023. 3, 4

[4] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 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*, pages 157–164. 2023. 1, 2

[6] Timo Bolkart and Stefanie Wuhrer. A groupwise multilinear correspondence optimization for 3d faces. In *Proceedings of the IEEE international conference on computer vision*, pages 3604–3612, 2015. 2

[7] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5543–5552, 2016. 2

[8] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3d face morphable models" in-the-wild". In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 48–57, 2017. 1, 2

[9] Popa T. Bradley D., Heidrich W. and Sheffer A. High resolution passive facial performance capture. *ACM Trans. Graph.*, 29, 2010. 2

[10] Alan Brunton, Timo Bolkart, and Stefanie Wuhrer. Multilinear wavelets: A statistical shape space for human faces. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 297–312. Springer, 2014. 1, 2

[11] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013. 1, 2

[12] Wei Cao, Chang Luo, Biao Zhang, Matthias Nießner, and Jiapeng Tang. Motion2vecsets: 4d latent vector set diffusion for non-rigid shape reconstruction and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[13] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *arXiv*, 2020. 2

[14] 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 *arXiv*, 2021. 2

[15] 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. 3

[16] Hong Chen, Xin Wang, Guanning Zeng, Yipeng Zhang, Yuwei Zhou, Feilin Han, and Wenwu Zhu. Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning, 2023. 3

[17] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. In *Computer Vision—ECCV'98: 5th European Conference on Computer Vision Freiburg, Germany, June 2–6, 1998 Proceedings, Volume II 5*, pages 484–498. Springer, 1998. 2

[18] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10101–10111, 2019. 5

[19] Radek Danecek, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20311–20322, 2022. 2

[20] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Computer Vision and Pattern Recognition Workshops*, 2019. 2

[21] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[22] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3

[23] 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 (ToG), Proc. SIGGRAPH*, 40(4):88:1–88:13, 2021. 2

[24] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Learning neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21003–21012, 2023. 2, 3, 4, 5

[25] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Mononphm: Dynamic head reconstruction from monocular videos. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[26] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022. 2

[27] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 5

[28] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[29] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. 2022. 3

[30] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3, 6

[31] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47–1, 2022. 3

[32] Ka-Hei Hui, Ruihui Li, Jingyu Hu, and Chi-Wing Fu. Neural wavelet-domain diffusion for 3d shape generation. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 3

[33] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005. 4

[34] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 3

[35] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 3

[36] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 7

[37] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *arXiv preprint arXiv:2305.03027*, 2023. 5, 7

[38] Jiahui Lei and Kostas Daniilidis. Cadex: Learning canonical deformation coordinate space for dynamic surface representation via neural homeomorphism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6624–6634, 2022. 2

[39] Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. Realtime facial animation with on-the-fly correctives. *ACM Transactions on Graphics*, 32(4):1–10, 2013. 2

[40] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 1, 2, 5

[41] Connor Z. Lin, Koki Nagano, Jan Kautz, Eric R. Chan, Umar Iqbal, Leonidas Guibas, Gordon Wetzstein, and Sameh Khamis. Single-shot implicit morphable faces with consistent texture parameterization. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 2

[42] I-Chen Lin and Ming Ouhyoung. Mirror mocap: Automatic and efficient capture of dense 3d facial motion parameters from video. *The Visual Computer 21*, 21:355–372, 2005. 2

[43] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 3

[44] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021. 3

[45] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 3

[46] 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. *Communications of the ACM*, 65(1):99–106, 2021. 3

[47] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[48] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12695–12705, 2021. 2

[49] Pablo Palafox, Nikolaos Sarafianos, Tony Tung, and Angela Dai. Spams: Structured implicit parametric models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12851–12860, 2022. 2

[50] 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 *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 7

[51] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 1, 2

[52] F. Pighin and J. P. Lewis. Performance-driven facial animation. In *Siggraph*, 2006. 2

[53] Stylianos Ploumpis, Haoyang Wang, Nick Pears, William AP Smith, and Stefanos Zafeiriou. Combining 3d morphable models: A large scale face-and-head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10934–10943, 2019. 1, 2

[54] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 3

[55] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-

Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5620–5629, 2021. 2

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3

[57] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 6

[58] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[59] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[60] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[61] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 5

[62] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. *arXiv preprint arXiv:2007.12494*, 2020. 2

[63] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2, 4

[64] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4

[65] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

[66] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020. 2

[67] Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pages 574–584. PMLR, 2020.

[68] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[69] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34: 1415–1428, 2021. 4

[70] Jiapeng Tang, Dan Xu, Kui Jia, and Lei Zhang. Learning parallel dense correspondence from spatio-temporal descriptors for efficient and robust 4d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6022–6031, 2021. 2

[71] Jiapeng Tang, Lev Markhasin, Bi Wang, Justus Thies, and Matthias Nießner. Neural shape deformation priors. In *Advances in Neural Information Processing Systems*, 2022. 2

[72] Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[73] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 1274–1283, 2017. 2

[74] Justus Thies, Michael Zollhoefer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM Transactions on Graphics*, 34(6):1–14, 2015. 2

[75] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 2

[76] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7346–7355, 2018. 2

[77] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5163–5172, 2017. 2

[78] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. 4

[79] Li H. Weise T., Bouaziz S. and Pauly M. Realtime performance-based facial animation. In *Siggraph*, 2011. 2

[80] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011. 4

[81] Jamie Wynn and Daniyar Turmukhambetov. Diffusionerf: Regularizing neural radiance fields with denoising diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4180–4189, 2023. 3

[82] Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. In *Proceedings of*

the IEEE/CVF International Conference on Computer Vision (ICCV), pages 2195–2205, 2023. 2

[83] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Make-your-video: Customized video generation using textual and structural guidance. *arXiv preprint arXiv:2306.00943*, 2023. 3

[84] Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. i3dmm: Deep implicit 3d morphable model of human heads. In *CVPR*, 2021. 2

[85] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint arXiv:2210.06978*, 2022. 3

[86] Biao Zhang and Peter Wonka. Functional diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[87] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–16, 2023. 3

[88] S. Zhang and P. Huang. High-resolution, real-time 3d shape acquisition. In *CVPR Workshop*, 2005. 2

[89] Mingwu Zheng, Hongyu Yang, Di Huang, and Liming Chen. Imface: A nonlinear 3d morphable face model with implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20343–20352, 2022. 2, 5

[90] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 5

[91] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[92] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021. 3

[93] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision*, 2022. 2