Authentic Volumetric Avatars from a Phone Scan

CHEN CAO, TOMAS SIMON, JIN KYU KIM, GABE SCHWARTZ, MICHAEL ZOLLHOEFER, SHUNSUKE SAITO, STEPHEN LOMBARDI, SHIH-EN WEI, DANIELLE BELKO, SHOOU-I YU, YASER SHEIKH, and JASON SARAGIH, Reality Labs, USA

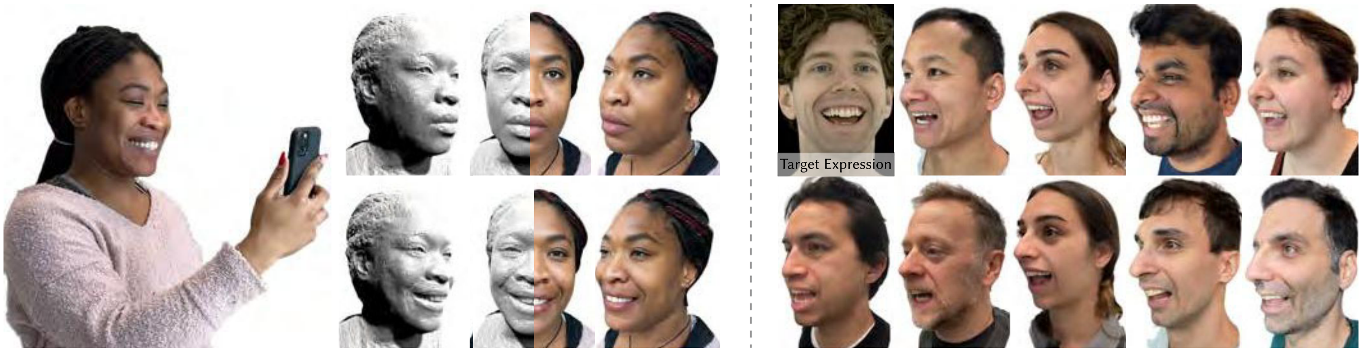


Fig. 1. We present a novel approach to create volumetric avatars using only a short phone capture as input. The resulting avatars produce high-fidelity renderings from novel viewpoints in realtime (left), and can generate novel animations using a common latent space of expressions (right).

Creating photorealistic avatars of existing people currently requires extensive person-specific data capture, which is usually only accessible to the VFX industry and not the general public. Our work aims to address this drawback by relying only on a short mobile phone capture to obtain a drivable 3D head avatar that matches a person's likeness faithfully. In contrast to existing approaches, our architecture avoids the complex task of directly modeling the entire manifold of human appearance, aiming instead to generate an avatar model that can be *specialized* to novel identities using only small amounts of data. The model dispenses with low-dimensional latent spaces that are commonly employed for hallucinating novel identities, and instead, uses a conditional representation that can extract person-specific information at multiple scales from a high resolution registered neutral phone scan. We achieve high quality results through the use of a novel universal avatar prior that has been trained on high resolution multi-view video captures of facial performances of hundreds of human subjects. By fine-tuning the model using inverse rendering we achieve increased realism and personalize its range of motion. The output of our approach is not only a high-fidelity 3D head avatar that matches the person's facial shape and appearance, but one that can also be driven using a jointly discovered shared global expression space with disentangled controls for gaze direction. Via a series of experiments we demonstrate that our avatars are faithful representations of the subject's likeness. Compared to other state-of-the-art methods for lightweight avatar creation, our approach exhibits superior visual quality and animateability.

Authors' address: Chen Cao, zju.caochen@gmail.com; Tomas Simon, tomas.simon@oculus.com; Jin Kyu Kim, jinkyuk@fb.com; Gabe Schwartz, gbschwartz@fb.com; Michael Zollhoefer, zollhoefer@fb.com; Shunsuke Saito, shunsukesaito@fb.com; Stephen Lombardi, stephen.lombardi@oculus.com; Shih-En Wei, swei@fb.com; Danielle Belko, dbelle3@gmail.com; Shoou-I Yu, shoou-i.yu@fb.com; Yaser Sheikh, yasers@fb.com; Jason Saragih, jason.saragih@oculus.com, Reality Labs, 131 15th Street, Pittsburgh, Pennsylvania, USA.

© 2022 Copyright held by the owner/author(s).
0730-0301/2022/7-ART163
https://doi.org/10.1145/3528223.3530143

CCS Concepts: • **Computing methodologies** → **Animation**.

Additional Key Words and Phrases: 3D Avatar Creation, Neural Rendering

ACM Reference Format:
Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shunsuke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason Saragih. 2022. Authentic Volumetric Avatars from a Phone Scan. *ACM Trans. Graph.* 41, 4, Article 163 (July 2022), 19 pages. https://doi.org/10.1145/3528223.3530143

1 INTRODUCTION

More than any other attribute, a person's face is their most important marker of self-identification, what Kundera referred to as "the serial number of a human specimen" [Kundera 1999]. Being the primary social display, evolutionary pressures have made people very sensitive to faces [Sheehan and Nachman 2014], especially familiar ones This presents a significant challenge for digital avatar creation, as even small deviations from a person's real facial appearance, structure or motion can result in an uncanny effect [Mori et al. 2012], greatly diminishing the avatar's utility for facilitating communication and perceived authenticity. Overcoming this difficulty traditionally relies on extensive person-specific data captures as well as artist-driven manual processing that is costly and time consuming. Automating the avatar creation process, with lightweight data capture, low latency, and acceptable quality, is thus, highly desirable, and is the subject of our work.

The core challenge of automatic avatar creation from limited data lies in the trade-off between prior and evidence. A prior is required to complement the limited information about a person's appearance, geometry, and motion that can be acquired in a lightweight way (e.g., using a cellphone camera). However, despite significant progress in recent years [Blanz and Vetter 1999; Booth et al. 2016; Karras et al. 2021b, 2020], learning the manifold of human faces at high resolution remains challenging. Modeling the long tail of the distribution, necessary for capturing personal idiosyncrasies like specific freckles, tattoos, or scars, likely requires models with much higher dimensional latent spaces, and consequently, much more data than what is currently used to train such models. Modern approaches [Chan et al. 2021; Karras et al. 2021a,b, 2020] are capable of hallucinating plausible non-existing faces, but fail to generate representations of real people at a fidelity that makes them recognizable as themselves. Recent approach achieve good inverse reconstruction by optimizing outside of the latent space (e.g. $\mathcal{W}+$ space in [Wu et al. 2021]), where there are no guarantees about the model behavior, resulting in strong artifacts in their image translation results.

In this work, we break the trade-off between prior and evidence by dispensing with the ability to hallucinate non-existing people, and instead, specialize our representation for adaptation using easily acquired cellphone data of real people. Our approach comprises three main elements; 1) a universal prior in the form of a hypernetwork that is trained on a high quality corpus of multiview video of hundreds of identities, 2) a registration technique for conditioning the model on a phone scan of the user's neutral expression, and 3) an inverse rendering-based technique to fine-tune the personalized model on additional expressive data.

Our prior's architecture is based on the observation that long tail aspects of facial appearance and structure lie in details that are best extracted directly from conditioning data of a person, instead of reconstructed from low-dimensional identity embeddings. In line with prior work [Blanz and Vetter 1999; Gross et al. 2005], we find that the performance of low-dimensional embeddings plateaus quickly, failing to capture person-specific idiosyncrasies. Instead, we find that augmenting existing approaches (e.g., [Lombardi et al. 2021]) with person-specific multi-scale 'untied' bias maps can faithfully reconstruct the high level of detail specific to a person. These bias maps can be generated from unwrapped texture and geometry of a user's neutral scan using a U-Net-style network. In this way, our model is a kind of hypernetwork that takes in data of a user's neutral face and produces parameters for a personalized decoder in the form of bias maps. Together, our universal prior and adaptation strategy enable the creation of highly realistic avatars instantly from even a single neutral scan, and can produce a model that spans a person's expressive range with additional frontal cellphone captures of only a few expressions.

Our approach improves upon the state of the art in avatar generation from cellphone captures [Grassal et al. 2021; Ichim et al. 2015; Luo et al. 2021; Nagano et al. 2018] without significantly increasing requirements on the user end. Whereas existing methods can produce plausible hallucinations of people, our approach produces avatars that look and move like a *specific* person. Furthermore, our model inherits the speed, resolution, and rendering quality of an existing person-specific model [Lombardi et al. 2021], since it employs a similar architecture and rendering machinery. Thus, it is well suited for interactive framerate-demanding applications such as VR. This opens up the possibility for ubiquitous photorealistic telepresence in VR that has thus far been hindered by the heavy requirements for avatar creation, or the low quality of avatars produced by lightweight captures.

The technical contributions of our work are:

- A system for producing a lifelike avatar of a person, with unprecedented appearance, structure and motion quality compared to existing approaches.
- A novel hypernetwork architecture that can produce high quality expressive avatars of a person given their neutral texture and geometry that preserves person-specific details. The resulting avatar has a consistent expression latent space with disentangled controls for viewpoint, expression, and gaze direction. The model is robust against real-world variations in the conditioning signal, including variations due to lighting, sensor noise, and limited resolution.
- An inverse-rendering strategy that specializes the avatar's expression space to the user given additional frontal cellphone captures, while ensuring viewpoints generalizability and preserving the latent space's semantics.

The remainder of this paper is structured as follows. We begin in §2 with an overview of related work. Our method is then described in §3, covering the model's architecture, dataset, and finetuning strategy. Experiments ablating our model and comparisons against existing approaches are presented in §4. We discuss limitations and future work in §5 and conclude in §6.

2 RELATED WORK

Our approach is related to several research domains in computer graphics and vision. We summarize the most related domains, such as face reconstruction, parametric 3D models, neural rendering, and avatar creation in the following. For a detailed discussion, we refer to the corresponding survey papers [Egger et al. 2020; Tewari et al. 2020, 2021; Zollhöfer et al. 2018].

Classical 3D/4D Face Reconstruction. The reconstruction of high-fidelity static and dynamic models of the human head based on photometric measurements has a long standing history in computer graphics and vision. Complex multi-view camera setups are required to obtain detailed face geometry via triangulation, motion via template tracking, and face appearance via the extraction of textures. Some recent multi-view systems employ additional active illumination, e.g., in the form of projected light patterns to help with the reconstruction of featureless regions [Ghosh et al. 2011; Klaudiny and Hilton 2012; Ma et al. 2007]. Obtaining high quality results requires multi-view capture systems [Beeler et al. 2011; Bickel et al. 2007; Bradley et al. 2010; Furukawa and Ponce 2009; Fyffe et al. 2014; Huang et al. 2004; Pighin and Lewis 2006; Zhang et al. 2004] that are expensive to build, notoriously challenging to operate, and require the participants to travel to the capture studio. While classical face reconstruction techniques have enabled the creation of the first photo-realistic actors [Alexander et al. 2013, 2010; Borshukov and Lewis 2003; Seymour et al. 2017], they do not scale to the general public and are thus not directly applicable to the creation of photo-realistic avatars for commodity applications.

Parametric Face Models. Given a large corpus of high-fidelity face reconstructions, a low-dimensional prior of facial geometry and appearance can be learned to better enable reconstruction and tracking based on light-weight sensing configurations, i.e., from monocular captures with a phone. The seminal work on 3D Morphable Models

(3DMMs) [Blanz and Vetter 1999] employs principal component analysis to extract a low-dimensional facial shape and appearance space from a set of high-quality scans. Extensions build dedicated models of higher fidelity for the complete head [Ploumpis et al. 2020], the eye region [Wood et al. 2016], learn 3DMMs from significantly more data [Booth et al. 2016], or from internet photo collections [Kemelmacher-Shlizerman 2013]. While the commonly used linear expression basis [Garrido et al. 2016; Thies et al. 2016] are independent of identity, multi-linear models learn an identity-dependent facial expression space [Cao et al. 2014; Vlasic et al. 2005]. Light-weight reconstruction [Garrido et al. 2016; Romdhani and Vetter 2005] and tracking approaches [Cao et al. 2015, 2016; Thies et al. 2016] employ the learned shape and appearance space as a prior to better constrain the ill-posed optimization problems they are tackling. The blessing and curse of these approaches is the low-dimensional global prior. While it provides ample regularization and enables to overcome under-constrained reconstruction and tracking problems in the wild, it prevents modeling of person-specific idiosyncrasies, such as wrinkle-level detail. In contrast, our UNet-based latent encoding incorporates multi-resolution reasoning into the prior, achieving high-fidelity reconstruction of avatars. Another drawback of mesh-based 3DMMs is their incompleteness, since they only model the facial skin region and thus disregard eyes, hair and the mouth interior. Our approach, on the other hand, employs a volumetric representation that enables learning of complete heads via image-based supervision. A recent work similarly employs an implicit surface representation based on coordinate-based neural networks to jointly model the face as well as hair, but its fidelity is far from photo-realistic [Yenamandra et al. 2021]. Generative models learned from a facial image collection also demonstrate the ability to synthesize photo-realistic faces of non-existing people in 2D [Karras et al. 2021a,b, 2020] or 3D [Chan et al. 2020] with low-dimensional latent codes. However, it remains difficult for such models to represent real people with authentic expressions of the person.

2D Neural Rendering of Human Heads. Reconstructing explicit high-fidelity geometry and appearance for the entire human head remains non-trivial even using state-of-the-art multi-view systems. The promise of neural rendering is to learn the synthesis of realistic imagery in an end-to-end manner from the captured data, while conditioning the output on the required control signals, such as expression or view point. Early neural rendering techniques operated in the 2D image domain and employed image-to-image translation networks to learn a mapping from a rendered conditioning image to photo-realistic output. For example, Deep Video Portraits [Kim et al. 2018] renders a deep feature buffer of conditioning information and employs a U-Net [Ronneberger et al. 2015] as image-to-image translation network. Deferred Neural Rendering [Thies et al. 2019] jointly learns a neural texture map that is attached to the face rig and is rendered to create the deep feature buffer. While these approaches are able to synthesize highly realistic images, they do not generalize well to novel view points since the 2D network struggles to learn about the underlying 3D transformations. Thus, the applicability of such approaches for photo-realistic avatars that can be rendered from any view point is limited.

3D Neural Rendering of Human Heads. More recent neural rendering techniques explicitly incorporate the underlying 3D structure for better generalization in terms of view point variation. For example, Deep Appearance Models [Lombardi et al. 2018] employ a coarse 3D triangle mesh in combination with view-dependent texture mapping. The texture is regressed by a neural network conditioned on view point and expression latent codes to account for view- and expression-dependent variation as well as to compensate for the imperfect proxy geometry. Pixel Codec Avatars (PiCA) [Ma et al. 2021] demonstrate that such models can be rendered efficiently, even on mobile hardware platforms by leveraging efficient per-pixel processing. If accurate surface geometry can be obtained, mesh-based techniques produce impressive results, but they often struggle in regions where tracking or 3D reconstruction is difficult such as for hair or the inner mouth. Neural Volumes [Lombardi et al. 2019] tackle this challenge by regressing a dense grid of appearance and opacity values that are composited using volume rendering. Mixture of Volumetric Primitives (MVP) [Lombardi et al. 2021] tackle the cubic memory complexity of grid-based approaches, such as Neural Volumes, based on a sparse 3D data structure. While the discussed approaches achieve impressive results, they require multiple hours of exhaustive data captures per user based on expensive and inaccessible multi-camera setups. In contrast, our approach regresses a drivable avatars of comparable fidelity from data captured by a phone scan, thus for the first time commoditizing the creation of photo-realistic avatars.

Light-weight Avatar Generation. There already exist several approaches, even commercial, for generating 3D avatars from light-weight sensing configurations. These approaches can be categorized based on the used input modality, such as multi-view images captured with a single camera [Cao et al. 2016; Ichim et al. 2015], a single monocular image [Hu et al. 2017; Lattas et al. 2020, 2021; Luo et al. 2021; Nagano et al. 2018; Yamaguchi et al. 2018], video [Grassal et al. 2021], depth images [Thies et al. 2018], or a high-quality neutral mesh and texture map [Li et al. 2020]. Mesh-based avatars can be reconstructed based on several color images captured by a quick phone scan [Ichim et al. 2015] and multi-view stereo. Image-based dynamic avatars [Cao et al. 2016] combine coarse face shape, hair, and neck geometry with dynamic texture mapping. Monocular reconstruction approaches primarily focus on facial appearance modeling, often leveraging learned (deep) priors. However, hair modeling is either ignored [Lattas et al. 2020, 2021; Yamaguchi et al. 2018] or based on asset retrieval from a database [Hu et al. 2017; Luo et al. 2021; Nagano et al. 2018], failing to model the user's hair style faithfully. Commercial solutions of similar technology already exist, such as AvatarSDK [1] or Pinscreen [2]. Learning a person-specific avatar is also possible using depth sensors [Thies et al. 2018] or monocular video [Grassal et al. 2021] with the help of generic 3DMM priors. While such person specific approaches can synthesize plausible facial expressions of a person, it is difficult to model the entire span of authentic facial expressions with limited capture time. Other approaches employ a high quality scan and a texture map as input to generate a dynamic facial rig [Li et al. 2020], but

[1] https://avatarsdk.com/
[2] https://www.pinscreen.com/

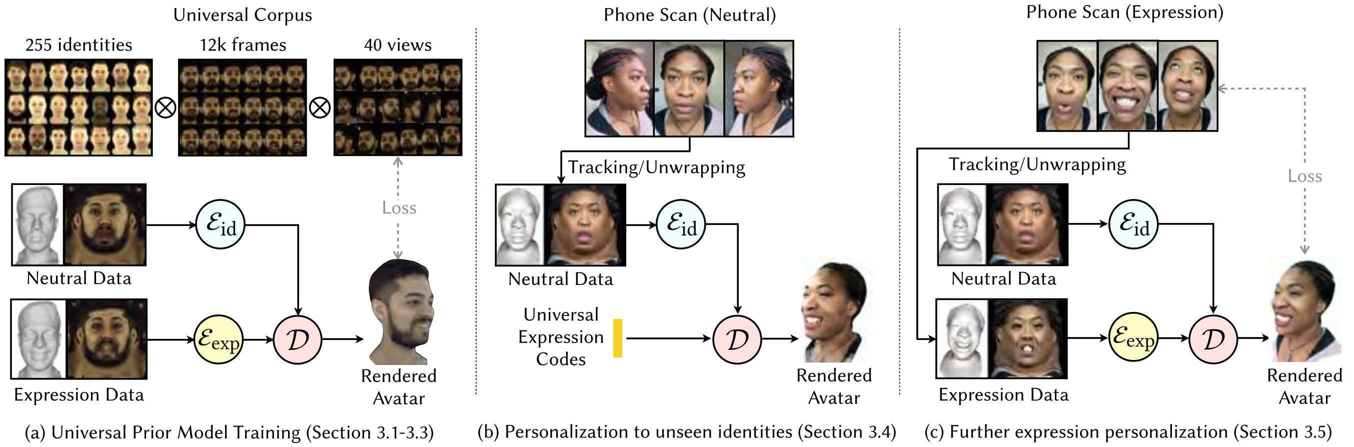


Fig. 2. Method overview. (a) We employ a large corpus of multi-view facial performances to train a cross-identity hypernetwork that can generate volumetric avatar representations. (b) The representation can be specialized to unseen individuals by conditioning on a lightweight capture of that person's neutral expression. (c) We can optionally refine the model using unstructured captures of an individual's appearance using inverse rendering.

require a multi-view capture system to obtain this conditioning data. The head avatars generated by all discussed techniques are neither fully photo-realistic nor authentic in their range of motion. In contrast, our approach enables photo-realistic modeling of the complete head and faithfully synthesizes facial expressions of a person from a phone scan that is only a few minutes long.

3 METHOD

An overview of our approach is shown in Fig. 2. We build upon the mixture of volumetric primitives (MVP) avatar representation of Lombardi et al. [2021]. However, instead of training person-specific avatars from extensive captures of each individual, our architecture trains a cross-identity hypernetwork as a prior for this representation (Fig. 2 (a)) that can be specialized to specific individuals by conditioning on a lightweight capture of that person's neutral expression (Fig. 2 (b)). The architecture for this prior and its training regime are detailed in §3.1, the dataset used to build it in §3.2, the training regime in §3.3, and the design and acquisition of the personalization data used for conditioning in §3.4. Finally, to account for person-specific details that are difficult to model using a cross-identity prior, we can optionally refine the model (Fig. 2 (c)) using unstructured captures of an individual via the inverse rendering method described in §3.6.

3.1 Universal Prior Model (UPM)

Our universal prior model (UPM) is a hypernetwork [Ha et al. 2017a] that generates parameters for a person-specific MVP-based avatar that can be animated following prior works [Wei et al. 2019]. A key observation is that person-specific avatars achieve a high degree of likeness to the target identity largely from the use of 'untied' bias maps in their architecture [Lombardi et al. 2018, 2021]. The simplest form of this is the base texture and geometry used in classical avatar representations that capture static details such as freckles, moles, wrinkles and even tattoos and small accessories like ear- and nose-rings. Thus, the ability to generate bias maps for real unseen identities is a necessary attribute of our hypernetwork. While recent advances in generative face modeling have been shown to plausibly hallucinate detailed appearance of non-existing people [Karras et al. 2021b, 2020], they can fail to span the detailed appearance of a particular unseen real person [Abdal et al. 2019, 2020], possibly stemming from the low-dimensional latent spaces they employ. The result is a similar-looking, but recognizably different, identity. Since our goal is to generate avatars of real people, we dispense with the ability to hallucinate avatars for non-existing people, and instead extract person-specific bias maps from conditioning data of real people. Thus, we call our hypernetwork an *identity encoder*, $\mathcal{E}_{\text{id}}$, and the person-specific avatar it generates, a *decoder*, $\mathcal{D}$.

An illustration of our construction is shown in Figure 3. To enable the extraction of person-specific details, $\mathcal{E}_{\text{id}}$ takes conditioning information in the form of a neutral texture map, $\mathbf{T}_{\text{neu}}$, and a neutral geometry image (an *xyz*-position map), $\mathbf{G}_{\text{neu}}$, and produces bias maps for each level of $\mathcal{D}$ via a set of skip connections, similar to a U-Net architecture [Ronneberger et al. 2015]. We refer to §3.4 for a detailed description of how the conditioning information is acquired. The model is trained to reconstruct a multiview dataset of multiple identities with multiple expressions each. During training, the expression codes, $\mathbf{e}$, are generated using an *expression encoder*, $\mathcal{E}_{\text{exp}}$, that takes, for a particular expression frame, view-averaged texture and geometry images, $\mathbf{T}_{\text{exp}}$ and $\mathbf{G}_{\text{exp}}$, as input. In summary, our universal prior model can be written as:

$$\mathbf{e} = \mathcal{E}_{\text{exp}}(\Delta\mathbf{T}_{\text{exp}}, \Delta\mathbf{G}_{\text{exp}}; \Phi_{\text{exp}}), \quad (1)$$

$$\Theta_{\text{id}} = \mathcal{E}_{\text{id}}(\mathbf{T}_{\text{neu}}, \mathbf{G}_{\text{neu}}; \Phi_{\text{id}}), \quad (2)$$

$$\mathcal{M} = \mathcal{D}(\mathbf{e}, \mathbf{v}, \mathbf{g}; \Theta_{\text{id}}, \Phi_{\text{dec}}) \ . \quad (3)$$

where $\Delta\mathbf{T}_{\text{exp}} = \mathbf{T}_{\text{exp}} - \mathbf{T}_{\text{neu}}$, $\Delta\mathbf{G}_{\text{exp}} = \mathbf{G}_{\text{exp}} - \mathbf{G}_{\text{neu}}$, and $\mathcal{M}$ is the output volumetric primitives for ray-marching, and Φ_{exp} and Φ_{id} are trainable parameters for the expression and identity encoders, respectively. The decoder is also conditioned on view- and gaze-direction vectors, $\mathbf{v}$ and $\mathbf{g}$, used for rendering, to allow explicit control over gaze and view-dependent appearance changes. The decoder parameters are comprised of two parts: 1) trainable network weights, Φ_{dec}, that model identity independent information that

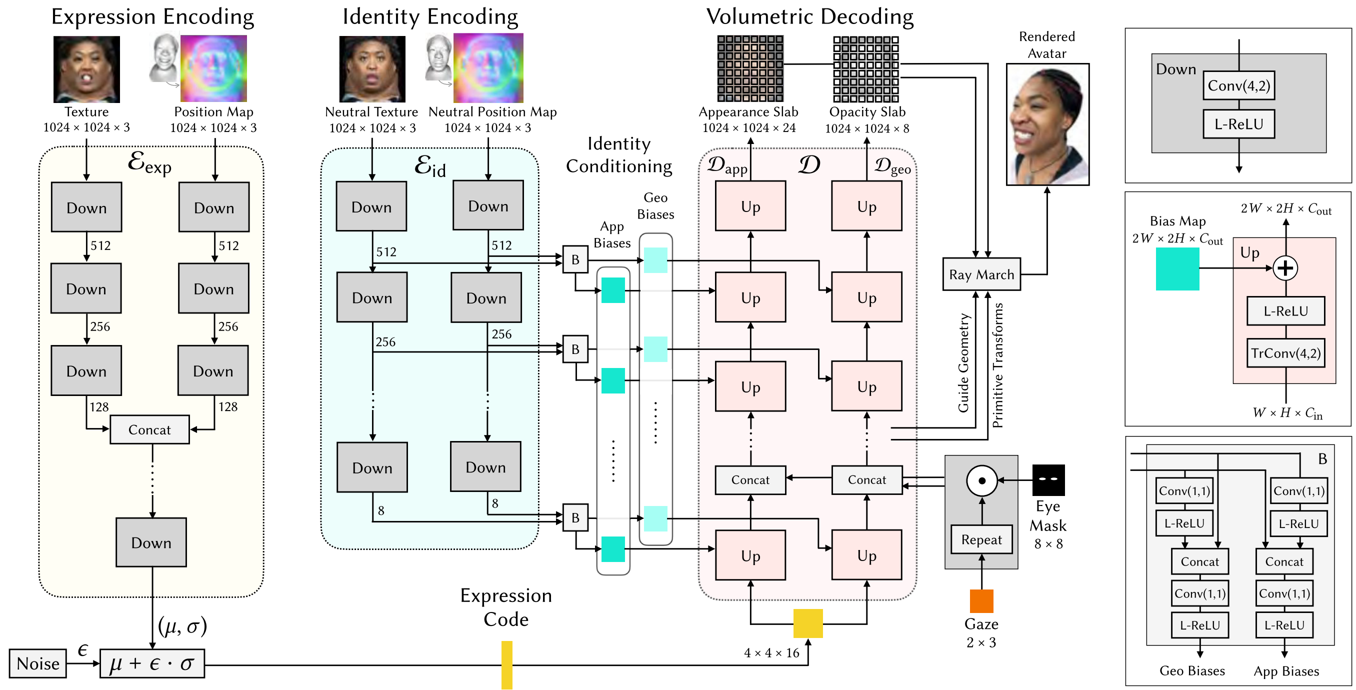


Fig. 3. Architecture diagram of our Universal Prior Model. An identity hypernetwork produces bias maps at multiple scales for an MVP-based decoder for each new identity. The bias maps have to be computed only once given a user's neutral texture and geometry. Disentangled gaze controls are input to the network by concatenation at the 8×8 feature level after appropriate masking. 'Conv(x,y)' denotes convolution with kernel width x and stride y, 'TrConv' denotes transposed convolution, and 'L-ReLU' denotes a LeakyReLU activation.

is shared across different identities, and 2) bias maps, Θ_{id}, which are regressed by the identity encoder and capture person specific information.

The novelties of our hypernetwork design include: 1) We encode person-specific details in the form of multi-scale bias maps, these bias maps are identified as a key source of identity for face modeling, enabling a compute-once-use-often setting for live facial animation. Prior methods, e.g. StyleGAN [Karras et al. 2021b] entangles the architecture for expression and identity, making it difficult to save computation for animation purposes. 2) Our proposed hypernetwork is the first effective U-net architecture from 2D conditioning data to a volumetric slabs, which can be ray marched to generated photorealistic avatar.

3.1.1 *Architecture Details.*

In the following, we provide a more detailed description of the encoder and decoder in our hypernetwork.

Upsampling with Bias Maps. The basic building block of our decoder is a convolutional upsampling layer with bias maps, i.e., one bias per output activation. Let C_{in} an C_{out} be the number of input and output channels of our upsampling layer, and let W and H be the width and height of the input activations. Thus, the input to the layer is a feature tensor of size $(W \times H \times C_{\text{in}})$, which is upsampled to dimension $(2W \times 2H \times C_{\text{out}})$. The upsampling is implemented by a transpose convolution layer (no bias, 4×4 kernel, stride 2) and is followed by an addition with a bias map of dimension $(2W \times 2H \times C_{\text{out}})$ produced by $\mathcal{E}_{\text{id}}$. The result is the output features of our layer.

Decoder. Our decoder, $\mathcal{D}$, closely resembles the architecture described in Lombardi et al. [2021], comprising two deconvolutional networks, $\mathcal{D}_{\text{geo}}$ and $\mathcal{D}_{\text{app}}$, that produce opacity $(1024 \times 1024 \times 8)$ and appearance $(1024 \times 1024 \times 24)$ *slabs*, as well as sparse guide geometry and transformations that are used to place the volumetric primitives in world space for ray-marching. We refer the reader to Lombardi et al. [2021] for further details. We make two modifications to this architecture to enable the generation of avatars of different subjects with consistent properties. First, we use a fully convolutional expression latent space [Ma et al. 2021], $\mathbf{e} \in \mathbb{R}^{4 \times 4 \times 16}$, to spatially localize the effects of each latent dimension. This promotes semantic consistency of the expression latent space across identities, which is important for some downstream tasks such as expression transfer. Second, we explicitly disentangle gaze from the expression latent space by replicating encodings of (2×3) gaze direction into an (8×8)-grid, masking these tensors to zero-out unrelated spatial regions, and conditioning the decoder by concatenating with its features at the (8×8) layer before continuing to decode to higher resolutions. Similar to view-disentanglement in prior methods, which aims to enable explicit control of view-dependent factors based on a viewer's vantage point in the scene, our construction enables explicit estimates of gaze to be directly used to control the avatar, which is well suited to VR applications where these estimates may already be present to support other functions (e.g., varifocal adjustments [Akşit et al. 2017] and foveated rendering [Patney et al. 2016]). An avatar representation that explicitly disentangles gaze controls from the rest of facial motion will be able to leverage those in-built eye tracking systems more directly.

Identity Encoder. The identity encoder, $\mathcal{E}_{\text{id}}$, uses strided convolutions to extract person-specific information from the conditioning data in the form of (1024×1024) texture- and position-maps of a subject's neutral expression. First, these inputs are processed separately using (1×1)-convolution to increase the feature channels to 8, followed by eight strided convolution layers with LeakyReLU activations [Maas et al. 2013], increasing the channel size each time[3]. At each resolution level, the intermediate features of the geometry and texture branches are concatenated and further processed using (1×1)-convolutions to produce the bias map for that level of the decoder, $\mathcal{D}$. When considering the pair $\mathcal{E}_{\text{id}}$ and $\mathcal{D}$ together, the architecture resembles a U-Net [Ronneberger et al. 2015], which provides a short-cut for transferring high resolution detail from the conditioning data directly to the decoded output, without passing through a low dimensional embedding space, allowing it to more easily reproduce intricate person-specific detail.

Although the U-Net architecture promotes the preservation of detail, the identity conditioning information may be insufficient to fully describe a person's idiosyncrasies. One example is how a person's teeth are not predictable from a closed-mouth neutral expression alone. In §3.6, we show that to better capture a person's likeness, our model can be fine-tuned on additional expressive data of the target subject.

Expression Encoder. The expression encoder, $\mathcal{E}_{\text{exp}}$, extracts expression latent codes, $\mathbf{e}$, for each sample in the training set. For this we employ a fully convolutional variational network that takes view-averaged expressive texture and position-maps as input. We follow the input format of the original MVP avatar representation, where the view-averaged input removes the view-dependent effects, forcing the decoder to make use of the view-conditioning at the bottleneck to enable explicit control. Since we use a $(4 \times 4 \times 16)$ latent code, we produce the mean and variances at that resolution instead of downsampling and reshaping to a vectorized latent space. To further promote the formation of a semantically consistent expression latent space, we subtract the neutral texture and position-map from their expressive counterparts before inputting them into the network. We will show in §4 that this simple scheme avoids identity information leaking into the expression latent space without the need for additional adversarial terms employed in other works [Lample et al. 2017; Schwartz et al. 2020].

3.2 Dataset

In this section, we describe the generation of the tracked meshes that are later on employed for supervision of the UPM training. There are three key components: the capture dome, the capture script, and the tracking pipeline.

Capture Dome. To capture synchronized multiview videos of a facial performance, we built a multiple video-camera capture dome as shown in Figure 4 (left). The dome has 40 color and 50 monochrome cameras that are placed on a spherical structure with a 1.2 meter radius. The cameras are pointed towards the center of the spherical structure where the participant's head is situated. Figure 4 (right) shows an example of captured multi-view images. We captured at a resolution of 4096×2668 with a shutter speed of 2.222 ms at 90 frames per second. 350 point light sources are evenly distributed across the structure to uniformly illuminate the participant. To compute the intrinsic and extrinsic camera parameters of each camera, we use a 3D calibration target [Ha et al. 2017b] mounted on a robot arm to perform automatic geometric camera calibration.

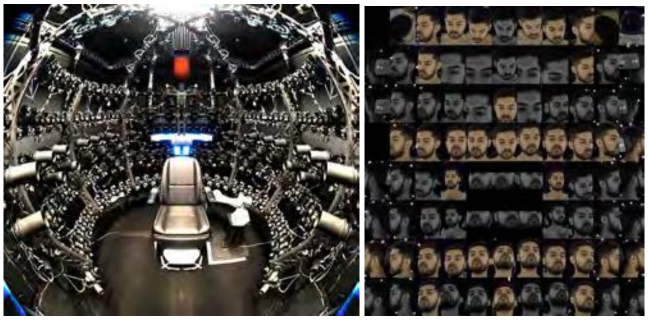


Fig. 4. Left: Image of capture dome. Right: Example captured images.

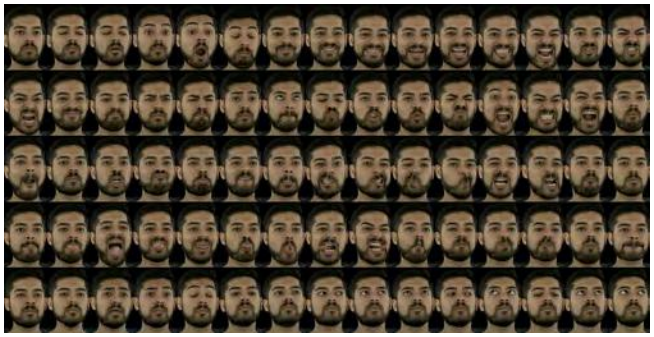


Fig. 5. Snapshot of various expressions captured during a performance.

Capture Script. The goal of the capture script is to systematically guide the participant through a wide range of facial expressions in the shortest amount of time. The participants are asked to go through the following exercises: 1) mimic 65 distinct facial expressions, 2) perform a free-form facial range-of-motion segment, 3) look in 25 different directions to represent various gaze angles, and 4) read 50 phonetically balanced sentences. Examples of captured expressions are shown in Figure 5. In total, 255 participants were captured using this capture script, and an average of 12k subsampled frames were recorded per participant to be used in the subsequent tracking stage. This leads to a total of 3.1 million frames to be processed.

Tracking Pipeline. In order to efficiently generate tracked meshes for over 3.1M frames, we implemented a highly scalable two phase approach, similar to that of Laine et al. [2017]. Our approach can process each frame independently and thus fully in parallel. In the first phase, we train a high-coverage landmark detector [Newell et al. 2016] that produces a set of 320 landmarks that are uniformly distributed across the face, covering both salient features (such as eye corners) as well as more uniform regions (such as the cheeks

[3] In this work we use a channel progression of (8, 16, 32, 64, 64, 128, 128, 256, 256) for all encoders.

and forehead). We leveraged two sources to generate training data for the high-coverage landmark detector: 1) for 30 participants, we first ran dense tracking [Wu et al. 2018] on ∼ 6k frames to cover a variety of expressions followed by sampling landmark locations from the dense tracking results, and 2) for all 255 participants, we ran non-rigid Iterative-Closest-Point-based face mesh fitting similar to [Bradley et al. 2010] on 65 expressions, and sampled landmark locations from the fitted meshes. The first source of data provides good expression coverage, but only on a limited set of identities, hence we added the second source to expand identity coverage. In the second phase, we run the high-coverage landmark detector on multiple views of each frame. The detected landmarks are then used to initialize a Principal Component Analysis (PCA) model-based tracking method [Tena et al. 2011; Wu et al. 2016] to produce the final tracked mesh.

3.3 Training and Losses

The UPM parameters, $\Phi = [\Phi_{\text{exp}}, \Phi_{\text{id}}, \Phi_{\text{dec}}]$, are optimized using:

$$\Phi^* = \operatorname*{argmin}_{\Phi} \sum_{i\in\mathcal{N}_{\mathcal{I}}} \sum_{f\in\mathcal{N}_{\mathcal{F}_i}} \sum_{c\in\mathcal{N}_C} \mathcal{L}_{\text{total}}(\Phi; \mathcal{I}_f^{i,c}) , \tag{4}$$

over $\mathcal{N}_{\mathcal{I}}$ different identities, $\mathcal{N}_{\mathcal{F}_i}$ frames and $\mathcal{N}_C$ different camera views from the dataset described in §3.2. We abuse notation slightly and use $\mathcal{I}_f^{i,c}$ to denote both the ground truth camera image as well as the set of training data associated with this frame f, namely: the tracked geometry and corresponding geometry image $\mathbf{G}_{\text{exp}}$, the view-averaged texture $\mathbf{T}_{\text{exp}}$, camera calibration, tracked gaze direction $\mathbf{g}$, and a segmentation image (described below). Our loss function comprises three main components:

$$\mathcal{L}_{\text{total}}(\Phi; \mathcal{I}_f^{i,c}) = \mathcal{L}_{\text{rec}}(\Phi; \mathcal{I}_f^{i,c}) + \mathcal{L}_{\text{mvp}}(\Phi; \mathcal{I}_f^{i,c}) + \mathcal{L}_{\text{seg}}(\Phi) . \tag{5}$$

Here, $\mathcal{L}_{\text{mvp}}$ are the losses introduced by Lombardi et al. [2021] (except their photometric ℓ_2-loss), whereas $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{seg}}$ are additions specific to our use case, which we elaborate on below. We optimize Equation (4) using stochastic gradient descent with ADAM [Kingma and Ba 2014] using a learning rate of $1e{-}3$ and all other parameters set to their default values.

Reconstruction Losses. The purpose of the reconstruction loss is to ensure that the synthesized images match ground truth. It can be split into three different parts:

$$\mathcal{L}_{\text{rec}}(\Phi; \mathcal{I}_p) = \mathcal{L}_{\text{pho}}(\Phi; \mathcal{I}_f^{i,c}) + \mathcal{L}_{\text{vgg}}(\Phi; \mathcal{I}_f^{i,c}) + \mathcal{L}_{\text{gan}}(\Phi; \mathcal{I}_f^{i,c}) . \tag{6}$$

The pixel-wise photometric reconstruction loss, $\mathcal{L}_{\text{pho}}$, compares the synthesized images with the ground truth, pixel-by-pixel:

$$\mathcal{L}_{pho}(\Phi; \mathcal{I}_p) = \lambda_{\text{pho}} \frac{1}{\mathcal{N}_{\mathcal{P}}} \sum_{p\in\mathcal{P}} \left\| \mathcal{I}_f^{i,c}(p) - \tilde{\mathcal{I}}_f^{i,c}(p) \right\|_1 . \tag{7}$$

Here, $\mathcal{P}$ is a random sample of pixels and we set the weight of this term to $\lambda_{\text{pho}} = 1$. We employ the ℓ_1-norm for sharper reconstruction results. We also estimate per-camera background images and color transformations for each identity, and sample pixels over the entire image. The next energy term, $\mathcal{L}_{\text{vgg}}$, is the VGG-loss of Johnson et al. [2016] which uses a VGG network [Simonyan and Zisserman 2015]. It penalizes the difference between the low-level VGG feature maps of the synthesized and ground truth images. In particular, it is more sensitive to low-level perceptual features, such as edges, and thus leads to sharper reconstruction results. We set the weight of this term to $\lambda_{\text{vgg}} = 1$. The final reconstruction loss is an adversarial loss, $\mathcal{L}_{\text{gan}}$, based on a patch-based discriminator [Isola et al. 2017] for sharper reconstruction results and reduced hole-artifacts that can occur in MVP representations. We set the weight of this term to $\lambda_{\text{gan}} = 0.1$.

Unlike $\mathcal{L}_{\text{pho}}$, the other two losses use a spatial receptive field to compute their values via convolutional architectures. As such, each pixel can not be independently evaluated of all others. Since memory limitations prohibit training on full resolution images, instead of randomly sampling each pixel of $\mathcal{P}$ independently, we randomly sample scaled and translated patches of (384×250) in resolution. We employ antialiased sampling on the full resolution images to generate the ground truth patches, but sample rays corresponding to pixels in those patches from our model in the usual way during ray-marching to reduce computation. We find this step to be necessary for the $\mathcal{L}_{\text{vgg}}$ and $\mathcal{L}_{\text{gan}}$ losses to effectively capture detail while avoiding overfitting to features at a specific scale.

Segmentation Losses. To promote better coverage of the subject in the scene, we employ a loss penalizing the difference between a pre-computed foreground-background segmentation mask and the integrated opacity field of the rendered avatar along pixel rays:

$$\mathcal{L}_{\text{seg}}(\Phi; \mathcal{I}_p) = \lambda_{\text{seg}} \frac{1}{\mathcal{N}_{\mathcal{P}}} \sum_{p\in\mathcal{P}} \left\| O_f^{i,c}(p) - \mathcal{S}_f^{i,c}(p) \right\|_1 , \tag{8}$$

where $\mathcal{S}$ are the segmentation maps and O is the integrated opacity computed during ray marching. We do not observe *hazy* reconstructions commonly observed in static or person-specific volumetric models [Lombardi et al. 2019], however, without this loss, we observe the model sometimes misses parts that are not well modeled by the guide geometry, such as a protruding tongue or hair structure that was not reconstructed accurately. Setting $\lambda_{seg} = 0.1$ initially and linearly reducing it to $\lambda_{seg} = 0.01$ effectively overcomes these limitations.

3.4 Conditioning Data Acquisition

To reconstruct a photo-realistic avatar for a user, we first acquire the conditioning data that is required by the UPM. To allow for broad adoption, our approach relies on an easily accessible device for capture and a simple script that a user can follow by themselves. For the device, we use an iPhone 12, which incorporates a depth sensor that can be used to extract better geometry of the user's face. For the capture script, the user is asked to maintain a fixed neutral expression while moving the phone around the user's head, left to right, then up and down, to acquire a complete capture of the entire head, including hair. We found that when performing this task with non-neutral expressions, maintaining a static expression was challenging for untrained participants. So, we capture additional expressions with a frontal camera only, without the need to maintain a static expression.

With the cellphone data acquired, it is processed as summarized in Fig. 6. First, the RGB-D camera of the iPhone 12 is employed to scan the user's neutral face from different perspectives (Fig. 6 (a)). For each captured image, we run a detector similar to [Newell

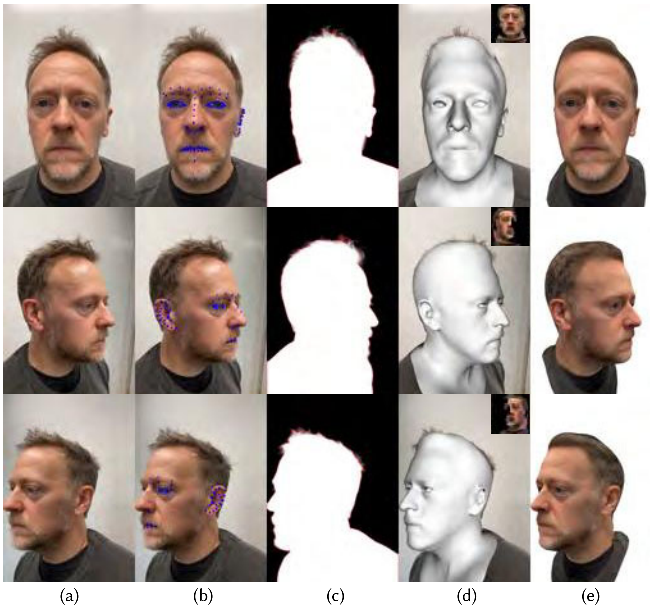


Fig. 6. Conditioning data acquisition. From left to right: (a) input image, (b) detected face landmarks, (c) portrait segmentation and traced silhouette, (d) reconstructed mesh and unwrapped texture, (e) rendered 3D face with aggregated texture.

et al. 2016] to obtain a set of landmarks (Fig. 6 (b)). In addition, we run portrait segmentation [Lin et al. 2021] to obtain segmentation masks, and trace the silhouette (Fig. 6 (c)). Using a neutral face PCA model with 150 dimensions built from our dataset in §3.2, we register a face mesh with fixed topology to the observations by solving a non-linear optimization problem. To this end, we optimize for the PCA coefficients, **a**, as well as the rigid head rotation, $\mathbf{r}_i$, and translation, $\mathbf{t}_i$, for each frame, $\mathcal{I}_i$, by minimizing a combination of a landmark, segmentation, depth, and coefficient regularization loss:

$$\mathcal{L}_{\text{cda}}(\mathbf{a}, \mathbf{r}_i, \mathbf{t}_i) = \lambda_{\text{ld}}\mathcal{L}_{\text{ld}} + \lambda_{\text{sih}}\mathcal{L}_{\text{sih}} + \lambda_{\text{d}}\mathcal{L}_{\text{d}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} \, . \tag{9}$$

Here, the landmark loss, $\mathcal{L}_{\text{ld}}$, is defined by the ℓ_1-distance between the detected 2D landmarks and the projected corresponding 3D landmark locations of the corresponding mesh vertices. For the segmentation silhouette loss, $\mathcal{L}_{\text{sih}}$, we measure the ℓ_1-distance in screen space between the vertices at the silhouette of the projected mesh and their closest points on the boundary of the portrait segmentation (see red points in Fig. 6 (c)). To compute the depth loss $\mathcal{L}_{\text{d}}$, we trace rays from each vertex in the normal and inverse normal direction, which we then intersect with triangle meshes generated from the depth maps. We define the ℓ_1-distance between the mesh vertices and the intersections as the depth loss. Finally, we regularize the PCA coefficients using Tikhonov regularization as $\mathcal{L}_{\text{reg}}$. We set $\lambda_{\text{ld}} = 5.0$, $\lambda_{\text{sih}} = 0.5$, $\lambda_{\text{d}} = 1.0$ and $\lambda_{\text{reg}} = 0.01$, and keep them fixed for all identities.

Due to the coarseness of the employed PCA model, the initial fit is only a rough approximation of the actual shape of the user's face. To further improve reconstruction quality, we employ free-form mesh deformation with Laplacian regularization [Huang et al. 2006] to minimize the aforementioned losses. This process produces a reconstructed face mesh that aligns with the input image well, see Fig. 6 (c). We use this mesh to unwrap the texture from each image (see inset of Fig. 6 (c)) and aggregate them to obtain the complete face texture. The textures are aggregated [Cao et al. 2016] by weighted averaging, where the weight of each texel is a function of the viewing angle, surface normal, and visibility. The final rendered meshes with the aggregated textures are shown in Fig. 6 (e).

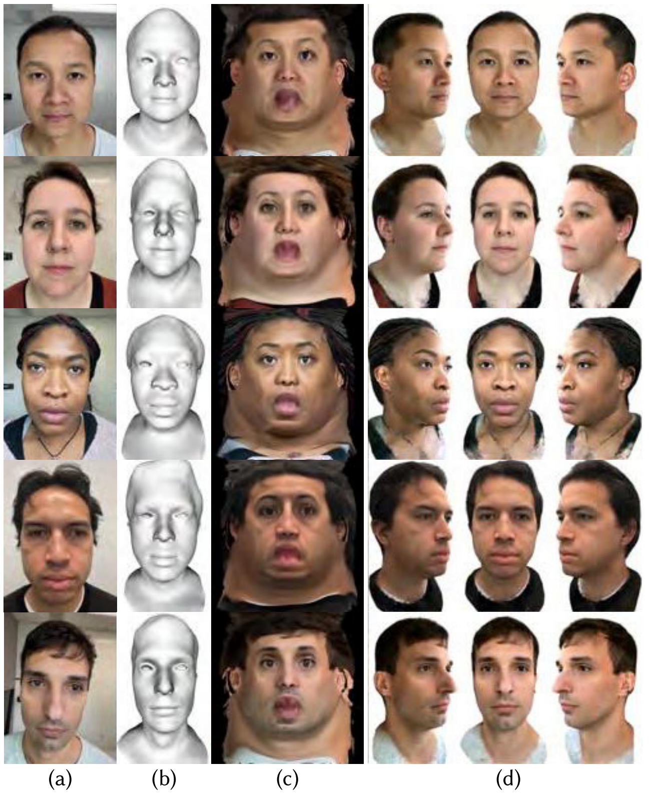


Fig. 7. Personalized decoders. From left to right: (a) example input image, (b) reconstructed mesh, (c) aggregated texture, (d) rendered avatar.

3.5 Personalized Decoder Generation

After acquiring the personalized mesh (Fig. 7 (b)), we transform it to the neutral geometry image, $\mathbf{G}_{\text{neu}}$. Together with the neutral texture $\mathbf{T}_{\text{neu}}$ (Fig. 7 (c)), both are taken as the conditioning data that is fed into the UPM to create a personalized decoder (Fig. 7 (d)). However, there exists a domain gap between the data used to train the UPM and the data acquired from the cellphone. First, the lighting environment used to build our training corpus in §3.2 is static and uniformly lit, whereas natural illumination conditions tend to exhibit more variations. Secondly, the cellphone capture data only covers the frontal half hemisphere of the head due to physical limitations.

We bridge the domain gap between cellphone and capture studio data in two steps. First, we apply the neutral face fitting algorithm in §3.4 to the capture studio data, where handheld camera motion is substituted by a discrete selection of cameras following a similar trajectory. The UPM is then trained with neutral conditioning data generated from this process, while keeping the high quality mesh tracking described in 3.2 for supervising the guide mesh and per-frame headpose in §3.3. This process significantly improves the quality of our generated avatars, as the UPM learns to inpaint regions that tend to be unobserved when following the cellphone capture script.

To account for the lighting and color transforms between the cellphone and studio data, we apply texture normalization, where we exhaustively search over our dataset of 255 identities, estimating optimal per-channel gains to match each, and pick the one with the minimal error. This normalized texture, together with the personalized mesh, are fed into the identity encoder, $\mathcal{E}_{\text{id}}$, to generate person-specific bias maps, which together with the decoder, $\mathcal{D}$, constitute the personalized decoder. Fig. 7 (d) shows the resulting avatars.

3.6 Finetuning a Personalized Decoder

Given a set of frames with arbitrary facial expression, we run an RGB-D based 3D face tracker [Weise et al. 2011]. We then unwrap the texture from the image, normalize it using the same strategy as for the neutral texture, and fill-in unobserved parts with the neutral texture. The tracked 3D face mesh and texture is used as the expression data input to the expression encoder, $\mathcal{E}_{\text{exp}}$, which along with the bias maps and $\mathcal{D}$, can be used to generate volumetric primitives that can be ray-marched to produce an image.

While the personalized decoder generates reasonable likeness with a hallucinated expression span, it often misses transient detail, such as wrinkles that are not apparent while the user's face is in a neutral expression. To build a more *authentic* avatar, we leverage data of the 65 facial expressions described in §3.2, which we capture using the cellphone from a frontal view. This capture takes 3.5 minutes on average and none of our participants experienced any difficulty in following the script.

With these expression frames, $\{\mathcal{I}_f\}$, we perform an analysis-by-synthesis to finetune the network parameters of the personalized avatar by minimizing:

$$\mathcal{L}_{\text{ref}}(\Phi;\mathcal{I}_f) = \mathcal{L}_{\text{rec}}(\Phi;\mathcal{I}_f) + \mathcal{L}_{\text{hole}}(\Phi;\mathcal{I}_f) + \mathcal{L}_{\text{seg}}(\Phi;\mathcal{I}_f), \quad (10)$$

where $\mathcal{L}_{\text{rec}}$ is the reconstruction loss in Eq. 6 and $\mathcal{L}_{\text{seg}}$ is the segmentation loss in Eq. 8 applied to all expression frames. The hole loss is defined as:

$$\mathcal{L}_{\text{hole}}(\Phi;\mathcal{I}_f) = \lambda_{\text{hole}}\left\|\max(\mathcal{T}_f - \mathcal{O}_f, 0) \cdot \mathcal{T}_f\right\|_1, \quad (11)$$

where $\mathcal{T}_f$ is a rendered mask that covers the face region and $\mathcal{O}_f$ is the integrated opacity computed during ray marching. $\mathcal{L}_{\text{hole}}$ penalizes holes that can emerge during finetuning as a result of the MVP surface primitives separating from each other. To ensure generalization to expressions not in the captured data, we also evaluate this loss on samples from the training corpus with a proportion of 1%. Fig. 16 shows the effects of the different error terms. We set $\lambda_{\text{pho}} = 1$, $\lambda_{\text{VGG}} = 3$, $\lambda_{\text{GAN}} = 0.1$, $\lambda_{\text{seg}} = 0.1$, and $\lambda_{\text{hole}} = 100$ and fix them for all finetuning experiments.

4 EXPERIMENTS

In this section, we present our experiments to evaluate the UPM and its use in our cellphone-based avatar generation pipeline. We ablate a number of design choices, presenting both qualitative and quantitative results. We also compare our method against several state-of-the-art approaches.

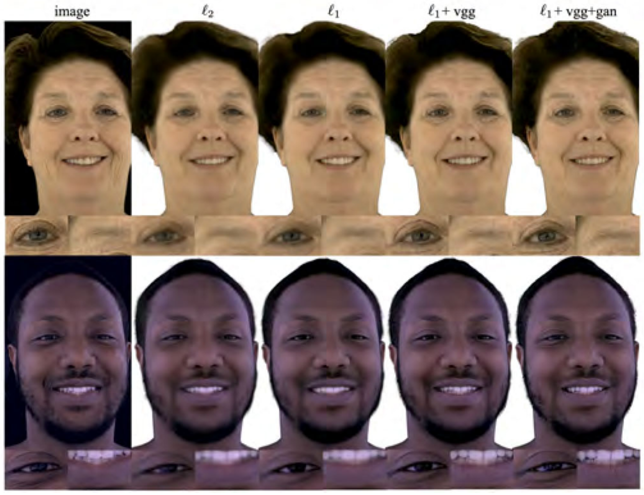


Fig. 8. The design of the reconstruction loss has a significant effect on the amount of detail that is reconstructed by the decoder. As can be seen, our complete loss function produces the highest fidelity avatars.

4.1 Universal Prior Model Experiments

We use the dataset described in §3.2 to build a UPM. To evaluate the different aspects of our construction, we group our experiments into four parts: 1) design of the loss function, 2) disentanglement and consistency of the expression latent space, 3) regularization of the identity encoder and fine-tuning, as well as 4) the effects of training corpus size on the model's performance.

4.1.1 Loss Function. Our person-specific decoder is based on Lombardi et al. [2021], which demonstrated good reconstruction quality after 500k training iterations (5 days on a single NVIDIA Tesla V100), when using the ℓ_2 reconstruction loss. Since our training corpus is much larger than the person-specific data used in that work, we found that even increasing the number of iterations to 800k (approx. 1 week processing time) did not achieve an acceptable level of detail. We found that using an ℓ_1-loss, a VGG-loss [Johnson et al. 2016], and a Patch-GAN loss [Isola et al. 2017] each add an additional level of detail to the reconstructions, enabling the model to achieve good qualitative results without greatly extending training time beyond that of the original method [4] .

Some qualitative comparisons of results after 800k iterations of training with a batch size of 384 × 250 rays are shown in Figure 8. As other work has previously found [Zhao et al. 2016], the ℓ_2-loss produces blurry images with limited detail. Switching to the ℓ_1 loss produces a small improvement, but detail is still lacking. Adding a VGG-loss significantly adds detail, but introduces additional high-frequency artifacts. Finally, the patch-GAN loss increases detail further, while removing the artifacts introduced by the VGG-loss.

[4] To reach a level of person-specific detail comparable to Lombardi et al. [2021] using our corpus, presuming each subject needs to be visited a comparable number of epochs during optimization, and presuming capacity is not the limiting factor, training with an ℓ_2-loss would take 5 days × 235 identities = 3.2 years on a single V100 GPU.

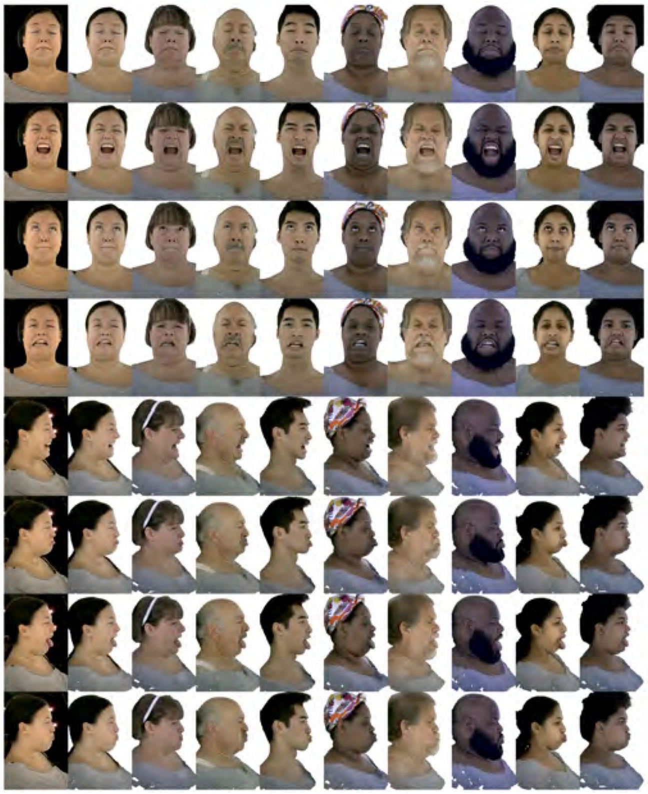


Fig. 9. Visualization of the expression consistent latent space discovered by the UPM. Left-most column: images of the source identity. Second-from-left column: UPM reconstructions. Other columns: retargeting results by decoding with our model based on different identity conditioning data.

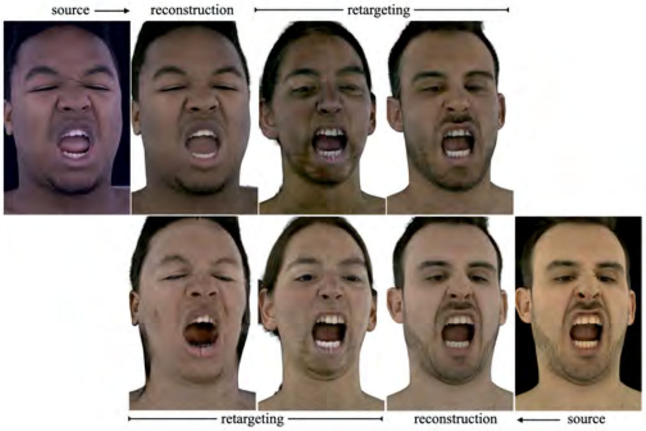


Fig. 10. Expression retargeting without neutral-subtracted input into $\mathcal{E}_{\text{exp}}$ results in an entangled expression space that fails to generalize. Top row from left to right (similarly bottom row from right to left): image, reconstruction, retargeting results.

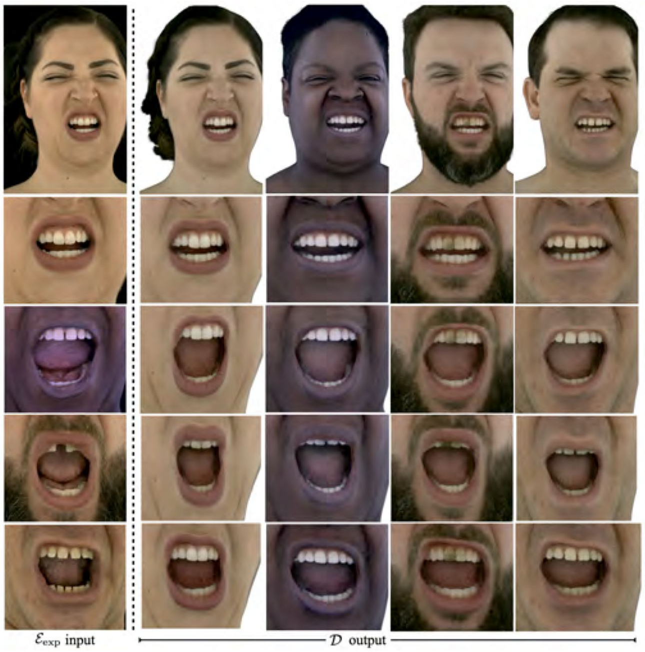


Fig. 11. Our expression latent space is identity invariant, even with respect to information that is not directly observable in the input to $\mathcal{E}_{id}$. Given expression encodings from multiple identities, the decoded results always exhibit consistent teeth.

4.1.2 Expression Space Consistency. An important element of an avatar generation system is the consistency of the controls they expose for downstream tasks. Our construction achieves this through the combination of bias maps for personalization, a fully convolutional 4 × 4 × 16 expression latent space, and neutral-differencing the input to the expression encoder. We demonstrate this in Figure 9, where we re-target one training subject's expression to another by inputting the source identity's expression data into $\mathcal{E}_{\text{exp}}$ and the target identity's neutral conditioning data into $\mathcal{E}_{\text{id}}$. The model retains the semantics of the expression space even across identities with significantly different facial shape and appearance, despite the fact that we did not explicitly define expression correspondences during training.

In Figure 10 we show the result of training without the neutral-difference input to $\mathcal{E}_{\text{exp}}$. Although the source image was reconstructed reasonably well, identity information leaks into the expression latent space, which manifests as a perturbation to the decoded avatar's identity and other visual artifacts.

Since the identity conditioning data we use to specialize our decoders are only comprised of the neutral texture and coarse geometry, some information regarding a person's expression span are not directly observed, such as expression-dependent wrinkles and the internals of the mouth. On the other hand, the input to $\mathcal{E}_{\text{exp}}$ contains identity-specific information that is not expression specific, such as the shape of teeth. A natural question to ask is whether our construction ends up encoding this information in the expression latent space. Figure 11 shows the results of expression retargeting between different identities in our training set. If information about

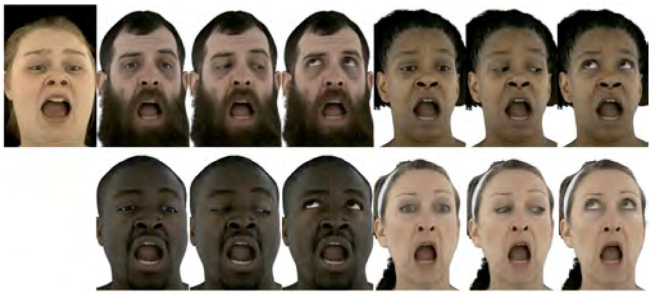


Fig. 12. Examples of explicit gaze control via a disentangled representation.

a subject's teeth are encoded into the expression latent space, we can expect the retargeted avatars to exhibit the source target's teeth. To the contrary, we find that the source identity's overall expression is successfully transferred to the target identities, but the decoded teeth remain those of each target's identity. This implies that $\mathcal{E}_{id}$ learns to correlate neutral facial appearance and geometry with teeth. However, since this correlation is weak, the model can only do so approximately, it fails to completely capture unusual appearances such as missing teeth. One possible solution is to enrich the conditioning information set with additional expressions, however, since our representation already disentangles expression from identity, we choose to instead rely on the fine-tuning strategy described in §3.6, which has more flexibility in leveraging the expressions that are available at test time, instead of requiring the set to be predefined a-priori.

Finally, the consistency of cross-identity expression transfer is difficult to evaluate empirically, due to challenges in defining a suitable metric (i.e. how different is one person's expression from another person's). However, for those attributes with a physical meaning, such as gaze direction, our model can disentangle their effects from the rest of the expression space, enabling their direct control from external sensors (e.g. eye tracking) without disturbing the rest of the expression. Some examples of this are shown in Figure 12, where expression retargeting is performed as above, but the gaze direction is modified.

4.1.3 *Identity Latent Spaces.*

One of our key design choices is to dispense with the ability to hallucinate *non-humans* that requires a latent space of identities. However, since our model relies on finetuning to capture the *last-mile* of a person's idiosyncracies, one might expect that catastrophic forgetting [French 1994] is more likely, where the model's weights specialize to reconstruct the finetuning data, but fail to generalize to other expressions and viewpoints of the same person. The regularization provided by an identity latent space avoids this phenomena since fintuning only involves traversing a latent space of identities instead of weights of the neural network [Abdal et al. 2020].

To evaluate the sensitivity of our approach to this, we excluded 22 identities and trained models with and without an identity latent space on data from the remaining identities. To make the comparisons meaningful, we keep all components of the model, the loss, and training regime the same across all experiments, and only add a fully-convolutional Gaussian VAE [Pu et al. 2016] to the input of $\mathcal{E}_{\text{id}}$; effectively a generative model over the identity conditioning. We experiment with compressing the conditioning data to a fully convolutional latent code at different spatial resolutions, before decoding back up to 1024 × 1024 for input into $\mathcal{E}_{\text{id}}$. We add a KL-divergence loss on the posterior distribution with weight $\beta = 0.01$, but do not add any reconstruction error on the conditioning data, freeing up $\mathcal{E}_{\text{id}}$'s input to take on richer representations than the original color and position maps.

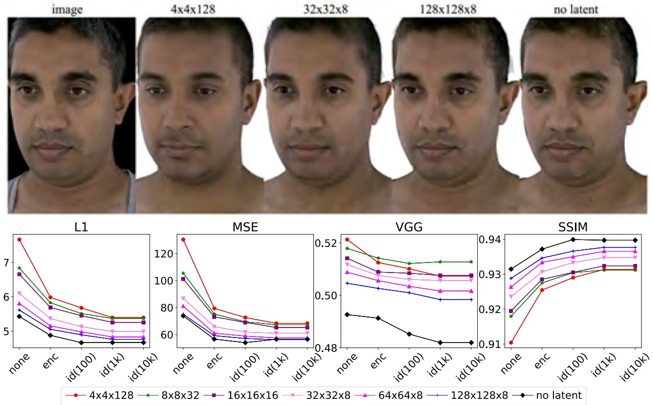


Fig. 13. Effects of finetuning models with and without an identity latent space of different spatial resolutions. Although the absence of a latent space leads to slight overfitting when finetuned for too long (>1k iterations) on a neutral frame, it is superior to the models using a latent space which fail to reconstruct fine detail. Top row: Reconstructions of an unseen identity using models with increasing latent space resolution. Bottom row: average reconstruction errors on range-of-motion sequences of 22 unseen subjects. On the x-axis, *none* refers to results without finetuning, *enc* finetunes $\mathcal{E}_{\text{exp}}$, and *id(x)* finetunes the identity latent codes (or $\mathcal{E}_{\text{id}}$ for models without a latent space) for x iterations.

Figure 13 shows results of our experiments for different settings of the latent space, ranging from 4 × 4 up to 128 × 128 in spatial resolution. To remove the effects of generalization errors from $\mathcal{E}_{\text{exp}}$, we further finetune it on the test set, while keeping all elements of $\mathcal{E}_{\text{id}}$ and $\mathcal{D}$ fixed[5]. From there, we further finetune $\mathcal{E}_{\text{id}}$ on the neutral images of the test subjects, where all weights of $\mathcal{E}_{\text{id}}$ are optimized for our standard model, whereas only the identity latent codes are optimized otherwise.

Our experiments show that as the identity latent space increases in spatial resolution, so does reconstruction performance, but the standard model without a latent space performs the best. Although it does exhibit slight overfitting, with small degradation in ℓ_1, MSE and SSIM metrics when finetuned for more than 1000 iterations, it still outperforms all other models. In a sense, increasing the spatial resolution of the identity latent space enables more flexibility to model unseen variations due to each latent code's localized spatial footprint on the output, at the expense of producing worse samples where long range dependencies are not modeled [Bagautdinov et al. 2018]. Comparing the qualitative results of the different models in

[5] The purpose of the universal prior model is to generate personalized decoders. Downstream tasks will not use $\mathcal{E}_{\text{exp}}$, which is specific to our multi-view system. So long as the expression latent space is disentangled from identity information, how an expression code is arrived at during test time has no bearing on the quality of the decoder itself. At the limit, directly optimizing the expression code for each test image has been used to evaluate the quality of generative face models [Blanz and Vetter 1999].

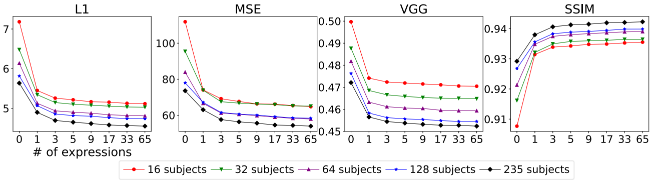


Fig. 14. Performance of UPM models trained with different numbers of subjects in the corpus (16–235) for varying amounts of available finetuning expressions (1–65). '0' denotes no finetuning.

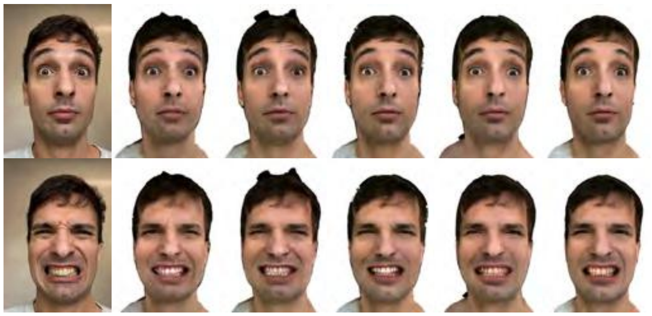
Figure 13, the ones with an identity latent space lack detail and specificity, especially for those with lower resolution latent spaces, despite being fairly high dimensional (i.e. $4 \times 4 \times 128 = 2048$), and produce avatars with a similar, but recognizably different, identity. We notice the VGG scores trend higher on these results since VGG scores are sensitive to the similarity in details between the avatar and image. The proposed model, on the other hand, captures fine details like the mole on the neck and stubble, and achieves smaller VGG scores. On the downside, without the identity latent space, our model does not support identity interpolation and relies on conditioning data of a specific individual to generate an avatar.

4.1.4 Corpus Size. Personalized avatar creation requires person-specific data, but the acquisition of such data is a major friction point for general use. The role of a UPM is to reduce the amount of person-specific data required to generate a personalized avatar. The quality of the UPM plays a critical role here, with the amount of data used to train the prior model being a key factor. To evaluate the interplay between the amount of training and person-specific data, we trained a UPM using increasing amounts of data; 16, 32, 64, 128 and 230 subjects. On the remaining subjects, we finetuned each model's $\mathcal{E}_{\text{exp}}$ and $\mathcal{E}_{\text{id}}$ on increasing amounts of expression data; 1, 3, 5, 9, 17, 33 and 65 expression frames, with five cameras each. To avoid biasing our results based on the selection of expressions that are incrementally added to the finetuning set, we ran the experiment five times, randomizing the selection each time, but keeping it consistent across the models we tested. After finetuning for 1k iterations, the models were evaluated on a held-out range-of-motion sequence for the remaining subjects from five held-out cameras. The results of this experiment are shown in Figure 14.

Our results show a clear trend where increasing the amount of training identities improves results. Similarly, additional finetuning data also leads to better results. However, improvements exhibit diminishing returns as a function of additional data. For example, the model with 16 training identities and the full finetuning set of 65 expressions is no better than the model with 230 training identities that was finetuned only on a single neutral frame. The trade-off between what can be acquired as part of the training set versus what can be acquired at the user's end will be application specific, however, our experiments suggest that improved performance may continue beyond a corpus size of 230 even when employing 65 finetuning expressions.

4.2 Finetuning Personalized Models

With the pre-trained UPM, we can generate user's personalized avatar based on the cellphone data. On a machine with 4 GPUs, cost time without finetuning is 10 minutes (mostly consumed by neutral face reconstruction, avatar generation is a single forward pass of the network), and with finetuning is about 6 hours.



(a) Input image (b) 16 ids (c) 32 ids (d) 64 ids (e) 128 ids (f) 235 ids

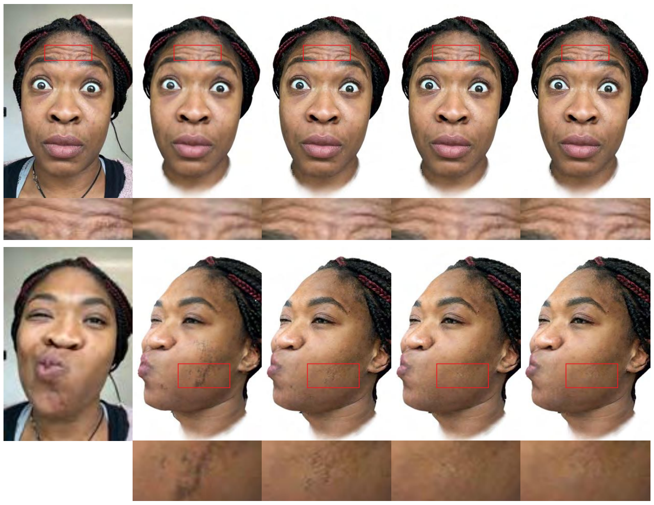
Fig. 15. Personalized avatar generated using cellphone conditioning data with a UPM that is trained with an increasing number of identities.

In the following of this section, we perform an ablation study on different components of the personalized model refinement step when using cellphone data captured in indoor environments with natural illumination using the iPhone 12's frontal camera.

4.2.1 Corpus Size. We first evaluate the effect the number of training identities has on building a personalized avatar using the models and procedures described in § 4.1.4 above. However, this time we use the reconstructed neutral face geometry and texture from a cellphone scan of subjects in natural indoor environments acquired as described in §3.4. Fig. 15 shows the reconstruction results using a universal prior model trained with different numbers of identities. As in §4.1.4, avatar generation from captures in real world conditions also benefit from larger corpus sizes, even as it is comprised entirely of a single lighting condition with uniform illumination.

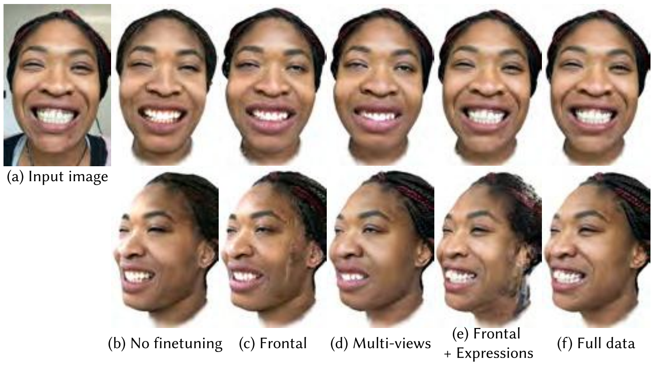
4.2.2 Finetuning Loss Functions. Fig. 16 shows reconstruction results for a cellphone personalized avatar finetuned with different losses. As in UPM training (see §4.1.1), employing only an ℓ_1-norm as the photometric loss in Equations (10) and (6) results in blurry reconstructions, as shown in Fig. 16 (b). Incorporating the VGG-loss helps to enhance sharpness of the resulting image (Fig. 16 (c)). However, it also introduces some hole-like artifacts, where the finetuned model's primitives are perturbed in a way that separates them, which results in rays missing the surface during raymaching. (Fig. 16 (b)(c) bottom row). Adding the hole-loss in Equation 11 significantly reduces such artifacts (Fig. 16 (d)). Finally, adding the GAN-loss only slightly improves quality of the result (Fig. 16 (e)) and is unable to completely remove some of the other artifacts that VGG introduces, possibly due to the limited number of iterations the discriminator is trained for during finetuning compared to during UPM training.

4.2.3 Finetuning Data Size. To build a personalized avatar, we capture two sets of user data using a cellphone: 1) a multi-view scan of the user's neutral face used to condition the universal prior model, and 2) frontal views of 65 facial expressions. We follow the same procedure described in §4.1.3 to generate different subsets for finetuning and evaluate the finetuned models on a held-out range-of-motion sequence. The results are summarized in Figure 17 and Table 1. Without any finetuning, the avatar does not reconstruct the user's



(a) Input image (b) ℓ_1 loss (c) +VGG (d) +Hole (e) +GAN

Fig. 16. Ablation study on losses used in finetuning. From left to right: (a) input image, (b) ℓ_1 reconstruction loss only, (c) + VGG loss, (d) + hole loss, (e) + GAN loss. The generated avatar is blurry with only the ℓ_1 reconstruction loss, while adding the VGG loss will improves detail, but introduces some artifacts. The hole-loss removes the artifacts and the GAN loss adds additional fine-scale detail.



(a) Input image

(b) No finetuning (c) Frontal (d) Multi-views (e) Frontal + Expressions (f) Full data

Fig. 17. Ablation study on data used for finetuning. From left to right: (a) input image, (b) no finetuning, (c) frontal neutral only, (d) multi-view neutral only, (e) frontal neutral + expressions, (f) full data. Top row: rendered in the original view, bottom row: rendered from a side view.

Table 1. Ablation study on data used for finetuning.

Neutral	Expr.	ℓ_1 (↓)	MSE(↓)	SSIM (↑)	LPIPS (↓)	VGG (↓)
None	None	14.55	137.79	0.9398	0.1226	0.2309
Frontal	None	15.40	173.31	0.9208	0.1290	0.2526
All	None	13.47	142.72	0.9359	0.1104	0.2340
Frontal	All	9.55	78.21	0.9435	0.1011	0.2304
All	2	12.26	124.44	0.9361	0.1100	0.2369
All	4	11.46	111.61	0.9395	0.1061	0.2329
All	8	10.56	96.86	0.9435	0.1019	0.2284
All	16	10.01	88.42	0.9459	0.0991	0.2254
All	32	9.72	82.82	0.9467	0.0983	0.2249
All	All	**9.18**	**74.33**	**0.9477**	**0.0966**	**0.2238**

Table 2. Ablation study on finetuning different parts of the model.

Components	ℓ_1 (↓)	MSE(↓)	SSIM (↑)	LPIPS (↓)	VGG (↓)
$\mathcal{E}_{\text{exp}}$	13.48	123.84	0.9401	0.1231	0.2329
$\mathcal{E}_{\text{id}}$	10.33	88.98	0.9485	0.1125	0.2236
$\mathcal{E}_{\text{id}} + \mathcal{E}_{\text{exp}}$	9.70	82.65	**0.9504**	0.1081	**0.2221**
$\mathcal{D}$	9.29	76.57	0.9471	0.0974	0.2244
$\mathcal{E}_{\text{id}} + \mathcal{D}$	9.27	76.59	0.9472	0.0975	0.2254
Full method	**9.18**	**74.33**	0.9477	**0.0966**	0.2238

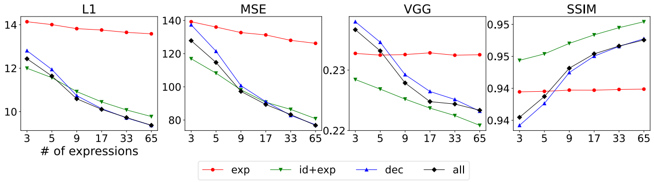


Fig. 18. The effects of finetuning dataset size on performance when finetuning different parts of the model.

expression correctly (Fig. 17 (b)), resulting in large reconstruction errors (Table 1 None/None). Finetuning on neutral frontal images alone results in overfitting, where performance degrades on the held-out set. For example, when only frontal views of the neutral scan data are used, obvious artifacts at the sides of head which are not observed (Fig. 17 (c)) are produced, resulting in an even larger reconstruction error than the model without finetuning (Table 1 Frontal/None). Using all frames of the neutral multi-view scan helps to reduce overfitting, however the model still can not accurately reconstruct expressions faithfully (Fig. 17 (d)). Finetuning on the full expression set without the multiview-neutral frames can effectively reduce the reconstruction error (Table 17 Frontal/All), however since the evaluation set comprises only frontal camera views, it does not capture generalization errors to non-frontal views. In (Fig 17 (e)), we can see that significant artifacts on the side of the head are present when finetuning using frontal data alone. Finally, when finetuning using the complete set of expression and multiview data, the personalized avatar produces accurate expression reconstructions without any artifacts when rendered in non-frontal views (Fig 17 (f)). Similar to the results in §4.1.4, Table 1 also shows a trend of improving performance as the finetuning set of expressions increases.

4.2.4 Finetuning different parts of the model. There are three major parts in our model: the decoder, $\mathcal{D}$, the identity encoder, $\mathcal{E}_{\text{id}}$ and the expression encoder, $\mathcal{E}_{\text{exp}}$. We finetune these different parts of the model on all frames of the neutral scan and the expression data, and then calculate the errors on a held-out dataset, shown in Table 2. Finetuning all parts ($\mathcal{D}$, $\mathcal{E}_{\text{id}}$ and $\mathcal{E}_{\text{exp}}$) gives the lowest ℓ_1-error, mean square error (MSE) and LPIPS metric [Zhang et al. 2018]. Finetuning only the encoder ($\mathcal{E}_{\text{id}} + \mathcal{E}_{\text{exp}}$) will achieve the best SSIM score and lowest VGG score.

To further evaluate the effects of refining different components of the model. We use the same strategy in evaluating the finetuning data size. We perform 5-fold cross-validation on different sets of expression data, finetune the different components of the model on each, and calculate the loss on the held-out set, see Fig. 18. We can

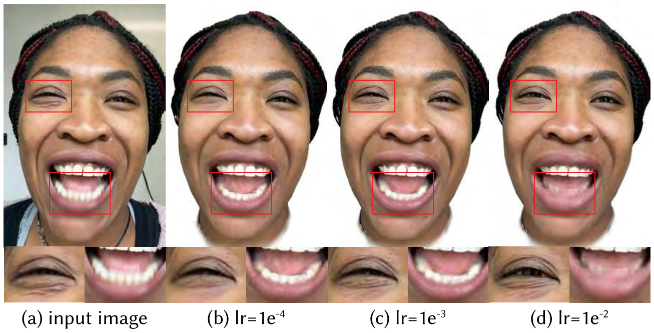


Fig. 19. The effect of learning rate on finetuning. A learning that is too small does not capture sufficient detail. If the learning rate is too high, the model looses its expression-consistent latent space. A good learning rate selection results in both high accuracy and good generalization.

see that when the finetuning dataset is small (Expression number < 10), refining $\mathcal{D}$ results in overfitting, with larger ℓ_1-error and MSE. However, with more finetuning data (Expression number > 10), refining the decoder together with the encoders help the model achieve the best results.

4.2.5 Learning rate. Another key factor that affects the quality of the avatar is the learning rate. Our universal prior model is trained on multiview data of 235 identities, while we finetune this model on a single identity with much reduced expression and viewpoints. To keep the expression space consistent and preserve view-dependent properties, we find that it is crucial to select the learning rate during finetuning carefully. Fig. 19 shows the comparison of different learning rates during the refinement of the personalized avatar. When the learning rate is too small $1e^{-4}$, we fail to recover sufficient facial detail, such as the wrinkles around the eyelids. If the learning rate is too large ($1e^{-2}$), the model easily overfits and performance degrades on held-out data. In our experiments, we found that a learning rate of $1e^{-3}$ produces detailed reconstructions while also generalizing to new expressions.

4.3 Qualitative Results

We show examples of finetuned personalized avatars in Fig. 25. Our finetuning process improves both the avatar's appearance (Fig. 25 (b)) and geometry (Fig. 25 (c)) when compared with the input images (Fig. 25 (a)). Although we only finetune on frontal view expression images, the view-dependent property of face expression is well preserved, which allows us to render the avatar from different viewpoints (Fig. 25 (d)).

Our avatars share the same global expression space. Fig. 1 right and Fig. 26 show some retargeting examples. Here, we choose one identity from our dataset (1st column in Fig. 26), pass the tracked mesh and texture into the expression encoder, to obtain the expression code, and feed it into the decoder of each personalized avatar. These results show that the expression of the source identity is transferred to the different avatars, while details such as teeth and wrinkles are preserved. In 3rd and 4th columns of Fig. 26 we show the same identity captured at different times in different environments. The recovered avatar's identity is consistent between the two captures. In building our training dataset, we designed the capture scripts to span the range of facial expressions as much as possible. Our model has satisfying results on most expressions but can exhibit artifacts, for some rare or extreme expressions (7th row of Fig. 26). The identification of these expressions and their incorporation into a more complete capture script we leave for future work.

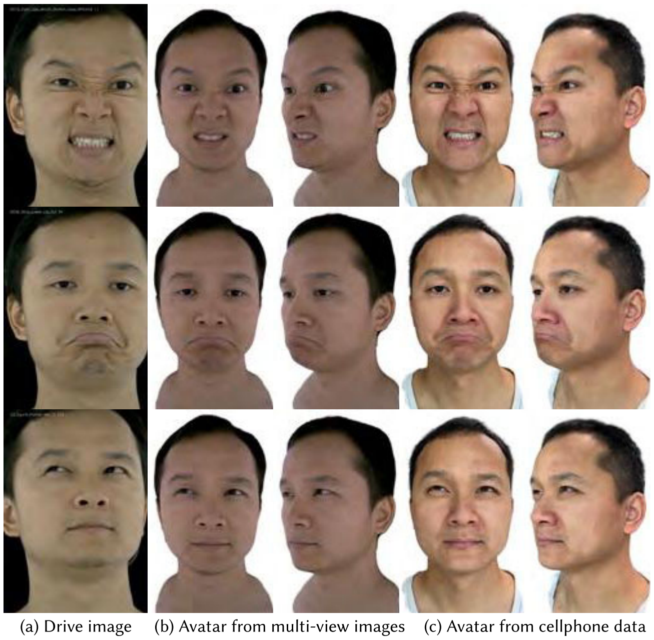

(a) Drive image
(b) Avatar from multi-view images
(c) Avatar from cellphone data

Fig. 20. Comparison of an avatar that has been produced from multi-view images and one that has been produced from cellphone data.

We also compare our generated personalized avatar with one from our training dataset from the same person, see Fig. 20. Compared to the avatar generated from the multi-view capture system, the personalized avatar generated from cellphone data has comparable quality, showing similar amounts of facial details in different expressions.

4.4 Comparisons

We compare our method against other state-of-the-art avatar creation methods. First, we compare with the stylized avatar creation method in [Luo et al. 2021]. It takes a single face image as input to a GAN-based framework that generates a normalized 3D avatar, see Fig. 21 (b). While this method generates a high-quality normalized face avatar, our method produce an authentic representation with higher realism.

We also compare our method to paGAN [Nagano et al. 2018], which takes a single face image as input and modifies the image by synthesizing different facial expressions, including the eyes and the mouth interior. Fig. 22 (b) shows their synthesized smile based on the input in (Fig. 22 (a)). We note that this method only produces the face region with a modified expression, and blends it with the original image. As a comparison, in Fig. 22 we show our results generating the same expression, based on the personalized avatar before finetuning in Fig. 22 (c), and after finetuning in Fig. 22 (d).

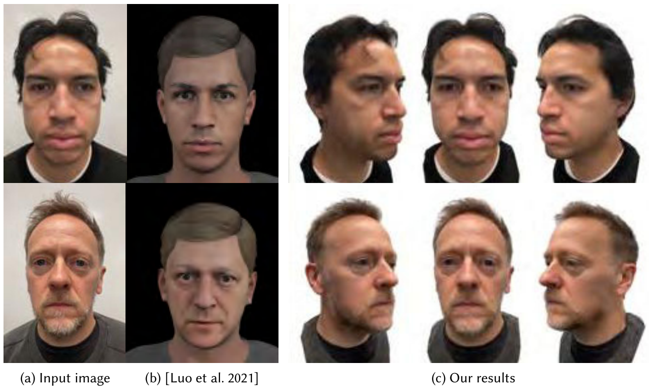


Fig. 21. Comparison of our approach to the state-of-the-art stylized avatar creation method of Luo et al. [2021]. Our approach produces photo-realistic avatars, while theirs produces a stylized representation.

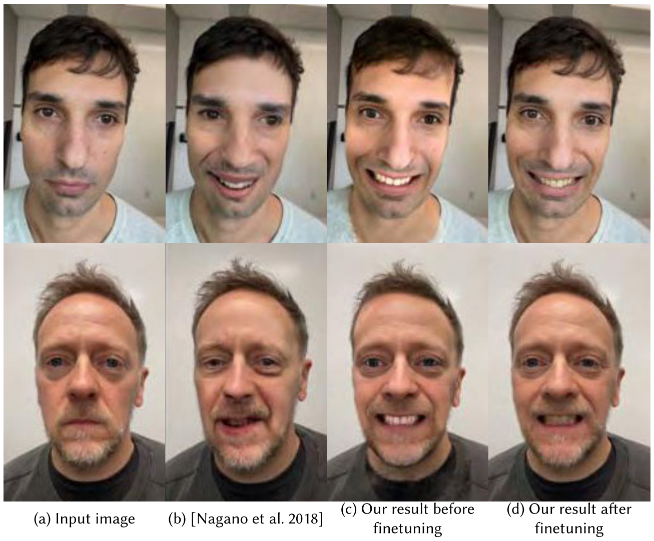

(a) Input image
(b) [Nagano et al. 2018]
(c) Our result before finetuning
(d) Our result after finetuning

Fig. 22. Comparison of our approach to paGAN [Nagano et al. 2018]. Our approach produces more realistic avatars. Note, our avatar is a full 3D model, while paGAN is restricted to the face region, which is overlaid on top of the input image.

Although paGAN produces plausible facial images, our results better preserve the user's likeness with semantically more consistent expressions.

Finally we compare our method with the RGB video-based avatar creation approaches of Grassal et al. [2021] and Gafni et al. [2021]. Given a user's monocular RGB video as input, Grassal et al. [2021] explicitly models face geometry and appearance of the user. Gafni et al. [2021] builds dynamic neural radiance fields for modeling the appearance and dynamics of a human face.

We take the RGB images from our captured data as input to both methods to train the models, and apply them to held-out data of the same user. Fig. 23 (b)(c) shows the results. Using the same frames in our method produces the results in Fig 23 (d). Our method is advantaged here, as it uses RGB-D instead of only the color images. Nonetheless, our approach produces significantly better results, with more fine-scale detail, especially in dynamic areas like the mouth.

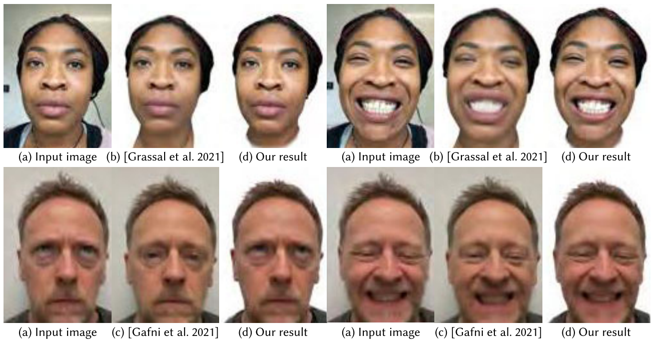


Fig. 23. Comparison to the approach of Grassal et al. [2021] and Gafni et al. [2021]. Our approach produces higher-fidelity results that exhibit more fine-scale detail.

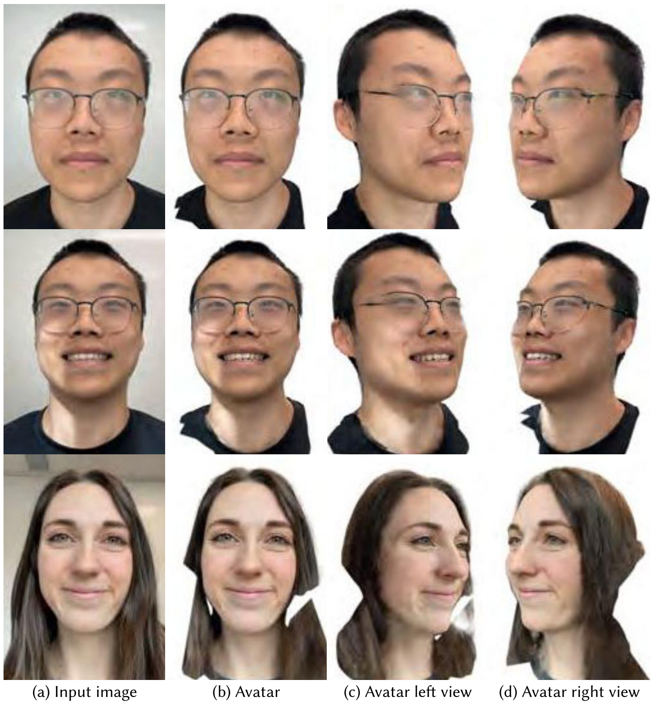


Fig. 24. Limitations of our approach. Our approach does not explicitly model glasses or long hair, results in artifacts of these parts in the generated avatar. Due to the limited lighting conditions in our training dataset, our approach cannot model challenging lighting well.

5 LIMITATIONS

While we have demonstrated state-of-the-art results for the generation of photo-realistic avatars from a cellphone scan, our approach is still subject to a few limitations that can be addressed in follow-up work.

Our proposed approach requires hours of processing time to finetune our model on additional expression data in order to create

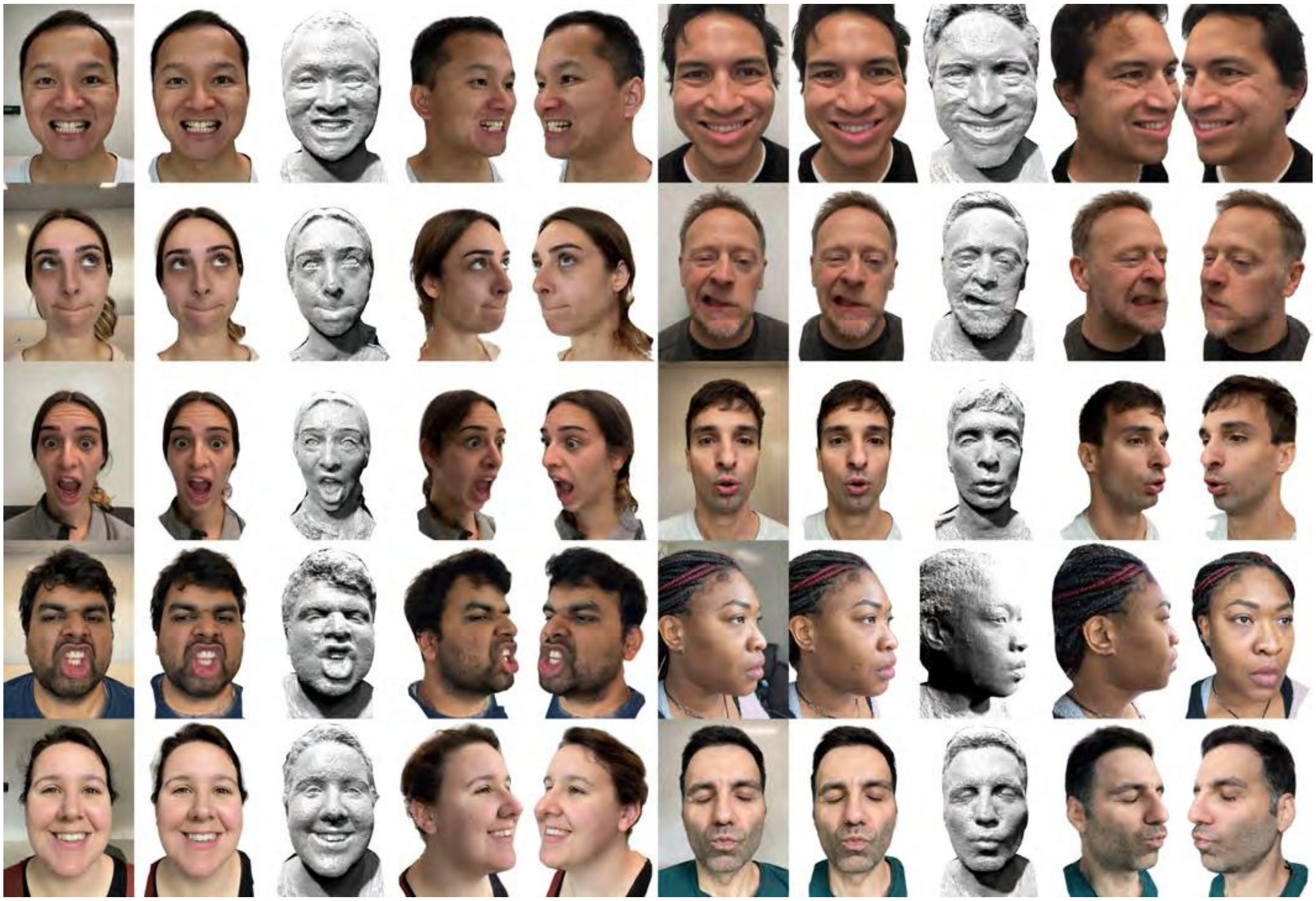


Fig. 25. Examples of refined personalized avatars. From left to right: input image, avatar, avatar depth by ray marching, avatar 3/4 left view and 3/4 right view.

a fully authentic representation of a user. It also exhibits overfitting behavior when the amount of finetuning data is deficient. One way to address this is to train a UPM that anticipates the finetuning process later. Some recent approaches have used meta-learning to do this [Zakharov et al. 2019], and employing similar strategies may reduce the number of finetuning iterations required as well as resolve the overfitting problem when finetuning data is sparse.

Another limitation of our approach stems from the domain gap between the data used to build our UPM and the real world settings. More specifically, our dataset is limited in terms of illumination and clothing variations (Fig. 24 2nd row), which can be seen in the pre-finetuned avatars we generate, where the teeth can appear unnaturally bright and the standard garment the subjects in our corpus wear is 'baked in'. This domain gap makes finetuning more challenging, possibly leading to increased data requirements and the overfitting issues described above. Collecting a corpus with more variation in terms of illumination and clothing would help alleviate this problem.

Finally, our avatars still lack completeness, without the ability to create full bodies, hands or even challenging hair styles that are not easily captured and reconstructed (Fig. 24). We cannot handle the glasses either (Fig. 24 1st and 2nd rows). Developing an easy-to-follow script to capture conditioning data for these elements is a necessary first step. With the inclusion of loose clothing and long hair, secondary dynamics and interpenetration become challenges that our proposed approach does not handle.

6 CONCLUSION

We have presented a system for generating an authentic and photorealistic avatar of a person from a short self-captured cellphone scan. Our volumetric avatars achieve an unprecedented level of realism compared to existing works and the individual's likeness is preserved throughout their avatar's range-of-motion. We performed experiments to characterize the effects of different components of our model, and also identified opportunities for further improvements. Our approach is designed for ease of use to encourage the proliferation of authentic avatar creation. In the future, such technology will given everyone the ability to create a digital version of their "serial number of a human specimen" [Kundera 1999].

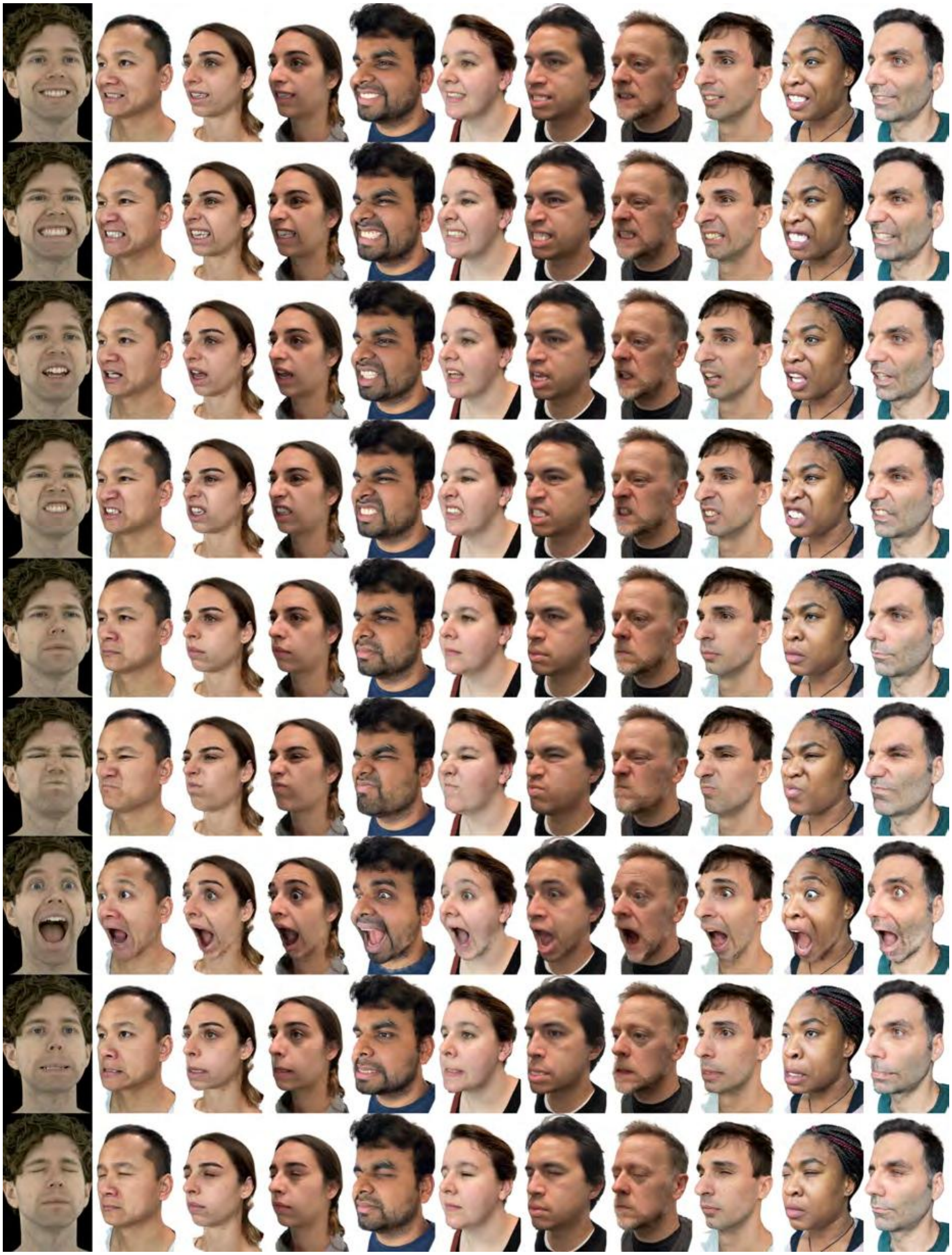


Fig. 26. Driving personalized avatars using expressions from an identity in our dataset (left column).

REFERENCES

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020. Image2StyleGAN++: How to Edit the Embedded Images?. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/CVPR42600.2020.00832 KAUST Repository Item: Exported on 2020-10-01 Acknowledged KAUST grant number(s): CRG2018-3730 Acknowledgements: This work was supported by the KAUST Office of Sponsored Research (OSR) under Award No. OSR-CRG2018-3730.

Kaan Akşit, Ward Lopes, Jonghyun Kim, Peter Shirley, and David Luebke. 2017. Near-Eye Varifocal Augmented Reality Display Using See-through Screens. *ACM Trans. Graph.* 36, 6, Article 189 (nov 2017), 13 pages. https://doi.org/10.1145/3130800.3130892

Oleg Alexander, Graham Fyffe, Jay Busch, Xueming Yu, Ryosuke Ichikari, Andrew Jones, Paul Debevec, Jorge Jimenez, Etienne Danvoye, Bernardo Antionazzi, Mike Eheler, Zybnek Kysela, and Javier von der Pahlen. 2013. Digital Ira: Creating a Real-time Photoreal Digital Actor. In *ACM SIGGRAPH 2013 Posters (SIGGRAPH '13)*. ACM, New York, NY, USA, 1:1–1:1.

Oleg Alexander, Mike Rogers, William Lambeth, Jen-Yuan Chiang, Wan-Chun Ma, Chuan-Chang Wang, and Paul Debevec. 2010. The Digital Emily Project: Achieving a Photorealistic Digital Actor. *IEEE Computer Graphics and Applications* 30, 4 (2010), 20–31.

Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. 2018. Modeling Facial Geometry Using Compositional VAEs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3877–3886. https://doi.org/10.1109/CVPR.2018.00408

Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W. Sumner, and Markus Gross. 2011. High-quality Passive Facial Performance Capture Using Anchor Frames. *ACM Trans. Graph.* 30, 4 (2011), 75:1–75:10.

Bernd Bickel, Mario Botsch, Roland Angst, Wojciech Matusik, Miguel Otaduy, Hanspeter Pfister, and Markus Gross. 2007. Multi-scale Capture of Facial Geometry and Motion. *ACM Trans. Graph.* 26, 3 (2007), 33:1–33:10.

Volker Blanz and Thomas Vetter. 1999. A Morphable Model for the Synthesis of 3D Faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '99)*. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 187–194.

J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway. 2016. A 3D Morphable Model learnt from 10,000 faces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

George Borshukov and J. P. Lewis. 2003. Realistic Human Face Rendering for "The Matrix Reloaded". In *ACM SIGGRAPH 2003 Sketches & Applications (SIGGRAPH '03)*. ACM, New York, NY, USA, 16:1–16:1.

Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer. 2010. High Resolution Passive Facial Performance Capture. *ACM Trans. Graph.* 29, 4 (2010), 41:1–41:10.

Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. 2015. Real-Time High-Fidelity Facial Performance Capture. *ACM Trans. Graph.* 34, 4, Article 46 (jul 2015), 9 pages. https://doi.org/10.1145/2766943

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2014. FaceWarehouse: A 3D Facial Expression Database for Visual Computing. *IEEE Transactions on Visualization and Computer Graphics* 20, 3 (mar 2014), 413–425. https://doi.org/10.1109/TVCG.2013.249

Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. 2016. Real-Time Facial Animation with Image-Based Dynamic Avatars. *ACM Trans. Graph.* 35, 4, Article 126 (jul 2016), 12 pages. https://doi.org/10.1145/2897824.2925873

Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2020. pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis. In *arXiv*.

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, Tero Karras, and Gordon Wetzstein. 2021. Efficient Geometry-aware 3D Generative Adversarial Networks. *CoRR* abs/2112.07945 (2021). arXiv:2112.07945 https://arxiv.org/abs/2112.07945

Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 2020. 3D Morphable Face Models—Past, Present, and Future. *ACM Trans. Graph.* 39, 5, Article 157 (jun 2020), 38 pages. https://doi.org/10.1145/3395208

Robert M. French. 1994. Catastrophic Forgetting in Connectionist Networks: Causes, Consequences and Solutions. In *Trends in Cognitive Sciences*. 128–135.

Yasutaka Furukawa and Jean Ponce. 2009. Dense 3D motion capture for human faces. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09)*. IEEE Computer Society, 1674–1681.

Graham Fyffe, Andrew Jones, Oleg Alexander, Ryosuke Ichikari, and Paul Debevec. 2014. Driving High-Resolution Facial Scans with Video Performance Capture. *ACM Trans. Graph.* 34, 1 (2014), 8:1–8:14.

Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2021. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8649–8658.

Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. 2016. Reconstruction of Personalized 3D Face Rigs from Monocular Video. *ACM Trans. Graph.* 35, 3, Article 28 (may 2016), 15 pages. https://doi.org/10.1145/2890493

Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. 2011. Multiview Face Capture Using Polarized Spherical Gradient Illumination. *ACM Trans. Graph.* 30, 6 (2011), 129:1–129:10.

Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2021. Neural Head Avatars from Monocular RGB Videos. *arXiv preprint arXiv:2112.01554* (2021).

Ralph Gross, Iain Matthews, and Simon Baker. 2005. Generic vs. Person Specific Active Appearance Models. *Image Vision Comput.* 23, 12 (nov 2005), 1080–1093. https://doi.org/10.1016/j.imavis.2005.07.009

David Ha, Andrew Dai, and Quoc V. Le. 2017a. HyperNetworks. https://openreview.net/pdf?id=rkpACe1lx

Hyowon Ha, Michal Perdoch, Hatem Alismail, In So Kweon, and Yaser Sheikh. 2017b. Deltille grids for geometric camera calibration. In *Proceedings of the IEEE International Conference on Computer Vision*. 5344–5352.

Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. 2017. Avatar digitization from a single image for real-time rendering. *ACM Trans. Graph.* 36, 6 (2017), 195:1–195:14.

Jin Huang, Xiaohan Shi, Xinguo Liu, Kun Zhou, Li-Yi Wei, Shang-Hua Teng, Hujun Bao, Baining Guo, and Heung-Yeung Shum. 2006. Subspace gradient domain mesh deformation. In *ACM SIGGRAPH 2006 Papers*. 1126–1134.

Xiaolei Huang, Song Zhang, Yang Wang, Dimitris N. Metaxas, and Dimitris Samaras. 2004. A Hierarchical Framework For High Resolution Facial Expression Tracking. In *Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops '04)*. IEEE Computer Society, Washington, DC, USA, 22.

Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. 2015. Dynamic 3D Avatar Creation from Hand-held Video Input. *ACM Trans. Graph.* 34, 4 (2015), 45:1–45:14.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. *CVPR* (2017).

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021a. Alias-Free Generative Adversarial Networks. *CoRR* abs/2106.12423 (2021). arXiv:2106.12423 https://arxiv.org/abs/2106.12423

Tero Karras, Samuli Laine, and Timo Aila. 2021b. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 12 (2021), 4217–4228. https://doi.org/10.1109/TPAMI.2020.2970919

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proc. CVPR*.

Ira Kemelmacher-Shlizerman. 2013. Internet based morphable model. In *Proceedings of the IEEE international conference on computer vision*. 3256–3263.

Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. 2018. Deep Video Portraits. *ACM Trans. Graph.* 37, 4, Article 163 (jul 2018), 14 pages. https://doi.org/10.1145/3197517.3201283

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. http://arxiv.org/abs/1412.6980 cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Martin Klaudiny and Adrian Hilton. 2012. High-Detail 3D Capture and Non-sequential Alignment of Facial Performance. In *Proceedings of the 2nd International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission (3DIMPVT '12)*. IEEE Computer Society, 17–24.

M. Kundera. 1999. *Immortality*. HarperCollins.

Samuli Laine, Tero Karras, Timo Aila, Antti Herva, Shunsuke Saito, Ronald Yu, Hao Li, and Jaakko Lehtinen. 2017. Production-level Facial Performance Capture Using Deep Convolutional Neural Networks. In *Proceedings of the ACM SIGGRAPH / Eurographics Symposium on Computer Animation*. Article 10, 10 pages.

Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Fader Networks: Manipulating Images by Sliding Attributes. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 5969–5978.

Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. 2020. AvatarMe: Realistically Renderable 3D Facial Reconstruction "In-the-Wild". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos P Zafeiriou. 2021. AvatarMe++: Facial Shape and BRDF Inference with Photorealistic Rendering-Aware GANs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

Jiaman Li, Zhengfei Kuang, Yajie Zhao, Mingming He, Karl Bladin, and Hao Li. 2020. Dynamic Facial Asset and Rig Generation from a Single Scan. *ACM Trans. Graph.* 39, 6, Article 215 (nov 2020), 18 pages. https://doi.org/10.1145/3414685.3417817

Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. 2021. Robust High-Resolution Video Matting with Temporal Guidance. *arXiv preprint arXiv:2108.11515* (2021).

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep Appearance Models for Face Rendering. *ACM Trans. Graph.* 37, 4, Article 68 (July 2018), 13 pages.

Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural Volumes: Learning Dynamic Renderable Volumes from Images. *ACM Trans. Graph.* 38, 4, Article 65 (jul 2019), 14 pages. https://doi.org/10.1145/3306346.3323020

Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. 2021. Mixture of Volumetric Primitives for Efficient Neural Rendering. *ACM Trans. Graph.* 40, 4, Article 59 (jul 2021), 13 pages. https://doi.org/10.1145/3450626.3459863

Huiwen Luo, Koki Nagano, Han-Wei Kung, Qingguo Xu, Zejian Wang, Lingyu Wei, Liwen Hu, and Hao Li. 2021. Normalized Avatar Synthesis Using StyleGAN and Perceptual Refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 11662–11672.

Shugao Ma, Tomas Simon, Jason M. Saragih, Dawei Wang, Yuecheng Li, Fernando De la Torre, and Yaser Sheikh. 2021. Pixel Codec Avatars. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021.* Computer Vision Foundation / IEEE, 64–73. https://openaccess.thecvf.com/content/CVPR2021/html/Ma_Pixel_Codec_Avatars_CVPR_2021_paper.html

Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, and Paul Debevec. 2007. Rapid Acquisition of Specular and Diffuse Normal Maps from Polarized Spherical Gradient Illumination. In *Proceedings of the 18th Eurographics Conference on Rendering Techniques (EGSR '07).* Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 183–194.

Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing.*

Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics Automation Magazine* 19, 2 (2012), 98–100. https://doi.org/10.1109/MRA.2012.2192811

Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. 2018. PaGAN: Real-Time Avatars Using Dynamic Textures. *ACM Trans. Graph.* 37, 6, Article 258 (dec 2018), 12 pages. https://doi.org/10.1145/3272127.3275075

Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. In *European conference on computer vision.* Springer, 483–499.

Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. 2016. Towards Foveated Rendering for Gaze-Tracked Virtual Reality. *ACM Trans. Graph.* 35, 6, Article 179 (nov 2016), 12 pages. https://doi.org/10.1145/2980179.2980246

F. Pighin and J.P. Lewis. 2006. Performance-Driven Facial Animation. In *ACM SIGGRAPH Courses.*

Stylianos Ploumpis, Evangelos Ververas, Eimear O'Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick Pears, William Smith, Baris Gecer, and Stefanos P Zafeiriou. 2020. Towards a complete 3D morphable model of the human head. *IEEE transactions on pattern analysis and machine intelligence* (2020).

Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. 2016. Variational autoencoder for deep learning of images, labels and captions. *Advances in neural information processing systems* 29 (2016), 2352–2360.

Sami Romdhani and Thomas Vetter. 2005. Estimating 3D Shape and Texture Using Pixel Intensity, Edges, Specular Highlights, Texture Constraints and a Prior. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02 (CVPR '05).* IEEE Computer Society, USA, 986–993. https://doi.org/10.1109/CVPR.2005.145

O. Ronneberger, P.Fischer, and T. Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI) (LNCS, Vol. 9351).* Springer, 234–241. http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a (available on arXiv:1505.04597 [cs.CV]).

Gabriel Schwartz, Shih-En Wei, Te-Li Wang, Stephen Lombardi, Tomas Simon, Jason Saragih, and Yaser Sheikh. 2020. The Eyes Have It: An Integrated Eye and Face Model for Photorealistic Facial Animation. *ACM Trans. Graph.* 39, 4, Article 91 (jul 2020), 15 pages. https://doi.org/10.1145/3386569.3392493

Mike Seymour, Chris Evans, and Kim Libreri. 2017. Meet Mike: Epic Avatars. In *ACM SIGGRAPH 2017 VR Village* (Los Angeles, California) *(SIGGRAPH '17).* ACM, New York, NY, USA, Article 12, 2 pages.

Michael Sheehan and Michael Nachman. 2014. Morphological and population genomic evidence that human faces have evolved to signal individual identity. *Nature communications* 5 (09 2014), 4800. https://doi.org/10.1038/ncomms5800

Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings,* Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1409.1556

J Rafael Tena, Fernando De la Torre, and Iain Matthews. 2011. Interactive region-based linear 3d face models. In *ACM SIGGRAPH 2011 papers.* 1–10.

A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Nießner, R. Pandey, S. Fanello, G. Wetzstein, J.-Y. Zhu, C. Theobalt, M. Agrawala, E. Shechtman, D. B Goldman, and M. Zollhöfer. 2020. State of the Art on Neural Rendering. *Computer Graphics Forum (EG STAR 2020)* (2020).

Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Yifan Wang, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, Tomas Simon, Christian Theobalt, Matthias Niessner, Jonathan T. Barron, Gordon Wetzstein, Michael Zollhoefer, and Vladislav Golyanik. 2021. Advances in Neural Rendering. arXiv:2111.05849 [cs.GR]

Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred Neural Rendering: Image Synthesis Using Neural Textures. *ACM Trans. Graph.* 38, 4, Article 66 (jul 2019), 12 pages. https://doi.org/10.1145/3306346.3323035

J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. 2016. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE.*

Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Niessner. 2018. <i>Headon</i>: Real-Time Reenactment of Human Portrait Videos. *ACM Trans. Graph.* 37, 4, Article 164 (jul 2018), 13 pages. https://doi.org/10.1145/3197517.3201350

Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. 2005. Face Transfer with Multilinear Models. In *ACM SIGGRAPH 2005 Papers* (Los Angeles, California) *(SIGGRAPH '05).* Association for Computing Machinery, New York, NY, USA, 426–433. https://doi.org/10.1145/1186822.1073209

Shih-En Wei, Jason Saragih, Tomas Simon, Adam W. Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. 2019. VR Facial Animation via Multiview Image Translation. *ACM Trans. Graph.* 38, 4, Article 67 (jul 2019), 16 pages. https://doi.org/10.1145/3306346.3323030

Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. 2011. Realtime performance-based facial animation. *ACM transactions on graphics (TOG)* 30, 4 (2011), 1–10.

E. Wood, T. Baltrusaitis, L. P. Morency, P. Robinson, and A. Bulling. 2016. A 3D morphable eye region model for gaze estimation. In *ECCV.*

Chenglei Wu, Derek Bradley, Markus Gross, and Thabo Beeler. 2016. An anatomically-constrained local deformation model for monocular face capture. *ACM transactions on graphics (TOG)* 35, 4 (2016), 1–12.

Chenglei Wu, Takaaki Shiratori, and Yaser Sheikh. 2018. Deep incremental learning for efficient high-fidelity face tracking. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–12.

Zongze Wu, Yotam Nitzan, Eli Shechtman, and Dani Lischinski. 2021. StyleAlign: Analysis and Applications of Aligned StyleGAN Models. *arXiv preprint arXiv:2110.11323* (2021).

Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. 2018. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–14.

T Yenamandra, A Tewari, F Bernard, HP Seidel, M Elgharib, D Cremers, and C Theobalt. 2021. i3DMM: Deep Implicit 3D Morphable Model of Human Heads. In *Proceedings of the IEEE / CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky. 2019. Few-shot adversarial learning of realistic neural talking head models. In *IEEE/CVF International Conference on Computer Vision.* 9459–9468.

Li Zhang, Noah Snavely, Brian Curless, and Steven M. Seitz. 2004. Spacetime Faces: High Resolution Capture for Modeling and Animation. *ACM Trans. Graph.* 23, 3 (2004), 548–558.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 586–595.

Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. 2016. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging* 3, 1 (2016), 47–57.

Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. 2018. State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications. *Computer Graphics Forum* (2018). https://doi.org/10.1111/cgf.13382