

Driving-Signal Aware Full-Body Avatars

TIMUR BAGAUTDINOV, Facebook Reality Labs, USA

CHENGLI WU, Facebook Reality Labs, USA

TOMAS SIMON, Facebook Reality Labs, USA

FABIÁN PRADA, Facebook Reality Labs, USA

TAKAAKI SHIRATORI, Facebook Reality Labs, USA

SHIH-EN WEI, Facebook Reality Labs, USA

WEIPENG XU, Facebook Reality Labs, USA

YASER SHEIKH, Facebook Reality Labs, USA

JASON SARAGIH, Facebook Reality Labs, USA



Fig. 1. We present an approach for building photorealistic full-body avatars that can be animated using input driving signals such as 3D keypoints (left) or body pose and facial animation codes (right). We build a disentangled latent space that is *driving-signal aware* to ensure that the model generalizes to novel sequences while still producing photorealistic results.

We present a learning-based method for building driving-signal aware full-body avatars. Our model is a conditional variational autoencoder that can be animated with incomplete driving signals, such as human pose and facial keypoints, and produces a high-quality representation of human geometry and view-dependent appearance. The core intuition behind our method is that better drivability and generalization can be achieved by disentangling the driving signals and remaining generative factors, which are not available during animation. To this end, we explicitly account for information deficiency in the driving signal by introducing a latent space that exclusively captures the remaining information, thus enabling the imputation of the missing factors required during full-body animation, while remaining faithful to the driving signal. We also propose a learnable localized compression for the driving signal which promotes better generalization, and helps

minimize the influence of global chance-correlations often found in real datasets. For a given driving signal, the resulting variational model produces a compact space of uncertainty for missing factors that allows for an imputation strategy best suited to a particular application. We demonstrate the efficacy of our approach on the challenging problem of full-body animation for virtual telepresence with driving signals acquired from minimal sensors placed in the environment and mounted on a VR-headset.

CCS Concepts: • Computing methodologies → Neural networks; Animation.

Additional Key Words and Phrases: full-body avatars, disentanglement

ACM Reference Format:

Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabián Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragh. 2021. Driving-Signal Aware Full-Body Avatars. *ACM Trans. Graph.* 40, 4, Article 143 (August 2021), 17 pages. <https://doi.org/10.1145/3450626.3459850>

Authors' addresses: Timur Bagautdinov, Facebook Reality Labs, Pittsburgh, USA, timurb@fb.com; Chenglei Wu, Facebook Reality Labs, Pittsburgh, USA, chenglei@fb.com; Tomas Simon, Facebook Reality Labs, Pittsburgh, USA, tsimon@fb.com; Fabián Prada, Facebook Reality Labs, USA, fabianprada@fb.com; Takaaki Shiratori, Facebook Reality Labs, Pittsburgh, USA, tshiratori@fb.com; Shih-En Wei, Facebook Reality Labs, Pittsburgh, USA, swei@fb.com; Weipeng Xu, Facebook Reality Labs, Pittsburgh, USA, xuweipeng@fb.com; Yaser Sheikh, Facebook Reality Labs, Pittsburgh, USA, yasers@fb.com; Jason Saragh, Facebook Reality Labs, USA, jsaragh@fb.com.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2021 Copyright held by the owner/author(s).
0730-0301/2021/8-ART143
<https://doi.org/10.1145/3450626.3459850>

1 INTRODUCTION

The goal of this work is to build high quality full body models of geometry and appearance that can be driven from commodity sensors placed in the environment. Building expressive and animatable virtual humans is a well studied problem in the graphics community. The creation of so called digital doubles has roots in the special effects industry [Alexander et al. 2010], and in recent years has begun to see examples of real-time uses as well such as Siren of Epic Games

and DigiDough of Digital Domain. These models are typically built using sophisticated multi-view capture systems with elaborate scripts that span variations in pose and expression. They capture a low-dimensional prior representation of human shape and appearance that allows for animation using more modest capture settings [Blanz and Vetter 1999; Loper et al. 2015; Romero et al. 2017]. Nonetheless, these models rely on brittle hand-crafted assumptions to encourage generalization to unseen poses, and overall sensor requirements for animation remains significant (e.g., a full-body motion capture suit for animation such as Xsens¹ and OptiTrack²). Since these models are learned independently from the sensor configurations used to drive the model during animation, they can simultaneously be over- and under-constrained, with limited ability to precisely match poses observed by the sensors while exhibiting unrealistic contortions in unobserved areas.

One approach to better integrate the driving signal into the model building process is to simultaneously capture using both the modeling sensors (i.e. an elaborate multi-view capture system) and the animation sensors (i.e. a small set of cameras placed in the environment). In this scenario, one can learn a model that directly regresses full shape and appearance from the information-deficient driving signals captured by the animation sensors. Although this approach significantly restricts the types of driving signals that can be used, e.g. a motion capture suit for a regularly clothed digital double, we argue that there is a more fundamental problem with this scenario. That is, there exists information asymmetry between the modeling and animation sensors which results in a one-to-many mapping problem, where multiple combinations of model states can equally likely explain the measurements. For example, a driving signal based on body joint angles does not contain complete information about clothing wrinkles and muscle contraction. Similarly, facial keypoints typically do not encode the hair, gaze or tongue motion. As a result, a model that is trained naively, without specifically taking into account such missing information, will either under-fit and produce averaged appearance, or, given enough capacity, will over-fit and learn chance correlations only present in the training set. Some existing works have addressed the issue of information asymmetry through the use of temporal models and adversarial training [Alexanderson et al. 2020; Ginosar et al. 2019; Ng et al. 2020; Yoon et al. 2020]. However, these methods tend to prescribe a specific imputation strategy that may not be appropriate in some applications. Furthermore, they operate on models post-training, making it difficult to overcome over- or under-fitting behavior that may already exist in the model.

Our work aims to address the problem of learning a model for a full-body digital avatar that is faithful to information-deficient driving signals while also providing an explicit data-driven space of plausible configurations for the missing information. To this end, we introduce a variational model that explicitly captures two types of factors of variation: *observed* factors that can be reliably estimated from the driving signals during animation, and *missing* factors, which are available only during the modeling stage. Our core strategy is to encourage better generalization by minimizing

the correlations between the observed factors, while maximizing it for the missing factors, such that during animation the model is able to produce plausible/realistic appearance and shape configurations that are fully consistent with the driving signal. We achieve the first by building a spatially-varying representation of the driving signal that localizes its effects and breaks global chance-correlations that might exist in the training set. The second is achieved by introducing a latent space for variation that is disentangled from the observed factors, forcing it to capture only the missing factors that are necessary to reconstruct the data. In particular, better disentanglement is achieved by explicitly accounting for coarse-long range effects of the driving signal that would otherwise be modeled by the latent space due to the localized conditioning used to encourage generalization. We achieve this using a coarse model of limb motion and an ambient occlusion map that helps model self-shadowing without overfitting. The resulting model can generate the space of plausible animations that agree with the information contained in the driving signal (Figure 1). Because of the explicit separation between the observed and missing factors, our approach enables the freedom to employ imputation techniques best suited to a particular application. We demonstrate the effectiveness a particularly simple approach; we assign the mean value to the missing factors for all frames during a sequence, which results in compelling animations that avoids over- or under-fitting effects observed in other approaches.

To summarize, the contributions of this work are as follows:

- A representation for a full body model that explicitly accounts for the driving signal during its construction. The model can generate a diverse space of plausible configurations that agree with the information contained in the driving signal.
- A method for achieving good generalization to novel inputs while producing high quality reconstructions by employing localized conditioning, accounting for coarse long range effects, and disentangling driving signals from the latent space.
- A demonstration of the utility of this approach on two scenarios where driving signal information is deficient: performance capture with a different attire and avatar animation for VR telepresence.

2 RELATED WORK

Our main goal is to build personalized full-body avatars that can be animated with information-deficient driving signals, while providing flexibility to practitioners to impute missing information appropriately for the application at hand.

2.1 Avatar Modeling

In the last decade, many efforts have been made for achieving expressive and animatable 3D models for human avatars, including face, hands and body. Due to the complexity in geometric deformation and appearance, data-driven methods have become popular [Blanz and Vetter 1999; Loper et al. 2015; Romero et al. 2017]. Linear models, e.g. PCA or Blendshapes, have been employed to model the muscle-activated skin deformation space, and were demonstrated to be effective for facial models [Blanz and Vetter 1999; Lau et al. 2009; Lewis et al. 2014; Vlasic et al. 2005]. For hands and body, an articulated prior is usually explicitly modeled by a kinematic chain,

¹<https://www.xsens.com/>

²<https://www.optitrack.com/>

while the surface deformation associated with the pose is obtained through skinning [Lewis et al. 2000], which generates the deformed surface by a weighted set of the influences from neighboring joints. Combining with the statistical modeling tools, a more expressive model can be developed by learning the pose-dependent corrective deformation across different identities for a body [Anguelov et al. 2005; Loper et al. 2015] and a hand [Moon et al. 2020; Romero et al. 2017]. Expressiveness can be further improved by localizing the deformation space [Osman et al. 2020; Tena et al. 2011; Wu et al. 2016]. These geometric models enable to learn a shading-free appearance model across different identities a hand [Qian et al. 2020], to automatically create a textured full-body avatar from a video [Alldieck et al. 2018; Alldieck et al. 2018], and to render avatars under new poses and camera views by using neural rendering [Prokudin et al. 2021]. Building on the above, unified avatar models, which model face, hand and body altogether, have been developed, including the Frank model [Joo et al. 2018] and the SMPL-X model [Pavlakos et al. 2019]. However, even given the limited training data, the generated results by these models are still underfit and constrained by the fidelity they can produce.

With the advent of deep neural networks, deep generative models have been successfully applied to modeling human bodies. For example, various convolutional mesh autoencoders have been proposed for building models for faces [Ranjan et al. 2018], or hands and bodies [Zhou et al. 2020b]. Compositional VAEs [Bagautdinov et al. 2018] model facial geometry with a hierarchical deep generative model, leading to a more expressive learned space of deformations and better modeling of high-frequency detail. Deep Appearance Models [Lombardi et al. 2018] learns a personalized variational model for both geometry and texture of a face, enabling a photorealistic rendering. However, extending those methods to full-body avatars is non-trivial, and, since those do not explicitly account for the missing information, they are prone to artifacts in scenarios with deficient driving signals.

Modeling clothing is another aspect closely related to our work, as we are primarily interested in animating clothed bodies. [Stoll et al. 2010] combine the cloth simulation with a body model from a multi-view capture so that the clothed body can be animated with the simulation. Recently [Joo et al. 2018] has extended the Frank model to Adam for modeling clothed surface with deformation spaces. Simulated dynamic clothes are used to model the interaction between cloth and the underneath body by explicitly factoring out the dynamic deformation [Guan et al. 2012]. CAPE [Ma et al. 2020] employs Graph-CNN to dress 3D meshes of human body from SMPL. However, it is learnt purely on geometry and only on the scans for sparse sampling of poses, and thus cannot be driven to generate photorealistic renderings with dynamic clothing.

Another line of work on animating photorealistic human rendering skips the complicated 3D geometry modeling, and rather focuses on synthesizing photorealistic images or videos by solving an image translation problem, e.g., learning the mapping function from joint heatmaps [Aberman et al. 2019], rendered skeleton [Chan et al. 2019; Esser et al. 2018; Pumarola et al. 2018; Shysheya et al. 2019; Si et al. 2018], or rendered meshes [Liu et al. 2019c,b; Sarkar et al. 2020; Thies et al. 2019; Wang et al. 2018], to real images. Although these methods often do generate plausible images, they tend

to have challenges in generalizing to different poses, due to complex articulations of the human body and dynamics of the clothing deformations. For example, methods [Chan et al. 2019; Liu et al. 2019c] do manage to reproduce overall body motion, but tend to produce severe artifacts on the hands and do not transfer facial expressions correctly. Moreover, these methods are not capable of producing renderings from arbitrary viewpoints, which is critical e.g. for telepresence applications.

Concurrent work to ours, SCANimate [Saito et al. 2021], introduces a method for building animatable full-body avatars from unregistered scans, which relies on implicit representations to learn both skinning transformations and pose-dependent geometry correctives; interestingly, that work also discovered that localized pose conditioning is critical to tackle spurious correlations. [Peng et al. 2021] introduces a method for novel-view synthesis for human-centered videos, which combines an articulated model (SMPL) with implicit appearance representation based on neural radiance fields. Although this method is capable of producing renders of arbitrary viewpoints, it does not allow for generalization across different poses, and is primarily tailored to short videos. Neural Parametric Models [Palafox et al. 2021] introduce a multi-identity parametric model for human body geometry that uses learnable implicit functions for shape and deformation modeling. Although promising, this method does not model appearance, and requires expensive optimization during inference, thus making it unsuitable for driving with incomplete signals.

2.2 Disentangled Representations

The ability of a learning algorithm to discover disentangled factors of variation in the data is considered to be a crucial property for building robust and generalizable representations [Bengio et al. 2013]. A large body of work has been focusing on building generic methods for building disentangled representations both with [Lample et al. 2017; Schwartz et al. 2020] and without [Higgins et al. 2016; Jiang et al. 2020; Zhou et al. 2020a] supervision.

β -VAE [Higgins et al. 2016] identified that a slight modification to the original VAE objective - putting a stronger weight on the prior term - can lead to automatic discovery of disentangled representations. In [Burgess et al. 2018] authors study the properties of β -VAE from the perspective of the information bottleneck method, suggesting the reason behind the emergence of disentangled representations. MINE [Belghazi et al. 2018] provides an efficient way to compute a lower bound on the mutual information between two sets of variables, which can be used as a proxy objective to encourage disentanglement.

In the context of human modeling, disentanglement has recently received a lot of attention as a way to improve model generalization. FaderNet [Lample et al. 2017] incorporates adversarial training with facial attributes for images synthesis of human faces, allowing disentanglement based on attributes. In [Schwartz et al. 2020] authors propose a generative model for facial avatars which uses several disentanglement techniques including FaderNet to encourage a model to better make use of gaze conditioning information. [Zhou et al. 2020a] introduces a generative model for human body meshes that uses a set of consistency losses as a way to separate the space of

pose- and shape-based deformations. [Jiang et al. 2020] achieve disentanglement for cross-identity shapes and poses by incorporating a deep hierarchical neural network.

In our settings, there are by design explicitly two groups of factors, corresponding to the observed and missing data, respectively. Thus our main objective is not to discover the unknown underlying factors of variation in an unsupervised way, but rather to encourage separation between the known observed factors (which are pre-defined) and the missing factors (learned space). In Section 3.5 we discuss our approach to handling missing information in more detail.

3 METHOD

Our goal is to build a data-driven model for full-body avatars that stays faithful to the driving signal, while also providing the flexibility to generate the information that might be missing from the inputs. For example, the driving signal for a human body might consist of sparse keypoint detections around the skeletal joints, which are insufficient to disambiguate different states of clothing, hair and facial expressions. Our key insight is that generalizability and controllability can be achieved through disentanglement, while explicitly taking into account the information that is missing from the driving signals.

When driving signals are sufficiently reliable, it is desirable to break their inter-dependencies as much as possible to achieve better generalization. For human bodies, this naturally leads to the notion of *spatially localized* control, such that, for example, a change in a facial expression does not have any influence on the state of the legs. At the same time, driving signals often do not contain sufficient information to fully describe free viewpoint images of the body, leading to a many-to-one mapping problem. This results in one of two cases; 1. overly smooth estimates, as the model averages over all possible unobserved states given the driving signal, or 2. overfitting to the dataset, which manifests as chance correlations between the driving signal and all unobserved factors. One approach to mitigate this problem is to add an additional latent code that spans the missing information space. However, as we will see in §4, without special care, the model can learn to partially ignore the control signal and use the latent space to capture the full information state. The resulting model, then, will fail to faithfully reconstruct the driving signal at test time, where the *true* latent code is unobserved.

In this work, we argue that, while indeed a latent space is useful to explicitly account for the missing information, it is also necessary to ensure that the latent space *only* contains information that is *not present* in the driving signal. In other words, the latent space and the driving signals should be disentangled. As we will describe in the sections that follow, the elements of our proposed architecture are explicitly designed to achieve disentanglement through multiple and complimentary means. The result is a model that exhibits good generalization, is faithful to the driving signal during animation, and achieves good reconstruction accuracy efficiently for a representation that covers the full human body.

3.1 Overview

An overview of our architecture is shown in Figure 2. It takes as input the driving signals and viewing direction, and produces registered geometry and a view-dependent texture as output using a deconvolutional architecture. Together, these outputs can be used to synthesize an image through rasterization. There are three main components of our construction that encourage generalization through disentanglement, namely: spatially localized conditioning, capturing coarse long range effects, and information disentanglement.

3.1.1 Spatially Localized Conditioning. To reduce overfitting to driving signals seen during training, we employ a location-specific low-dimensional embedding. These embeddings are used to *late-condition* the deconvolutional architecture, such that their *footprint* in the output texture and geometry have localized spatial extent. Together, the embedding and late conditioning extract the most relevant information from the driving signal at each spatial location. This reduces the tendency to learn spurious long-range correlations which might be present in the training data.

3.1.2 Coarse Long Range Effects. Although localizing the effects of driving signals can reduce overfitting, there exists some long range effects that are difficult to capture with such a representation. To capture these effects without reintroducing overfitting, we identify two major sources of long range effects for human bodies and model them explicitly: rigid limb transformations and the effects of shadowing. For rigid limb transformations we employ linear-blend skinning (LBS) [Kavan et al. 2008; Magnenat-Thalmann et al. 1989] that explicitly models rigid motion of the limbs through a composition of transformations along an articulated tree structure that spans the entire body. We thus decouple our decoder into using joint angles via LBS, and producing correctives in the form of a displacement map generated using the deconvolutional architecture. To handle long range shadowing effects, we explicitly compute an ambient occlusion map using the LBS-generated geometry, and pass it through a UNet [Ronneberger et al. 2015] to produce a gain map that is applied on the output texture. The ambient occlusion map serves as a substitute for a shadow-map since our capture space has roughly uniform illumination. Together, the LBS and shadow branches complement the localized conditioned described above, so that major long-term effects are accounted for while maintaining good generalization performance.

3.1.3 Information Disentanglement. The main motivation for introducing a latent space is to capture information not contained in the driving signal, yet necessary to fully explain image evidence. Unlike the driving signal, where avoiding overfitting is primary, variations that have no supporting evidence during animation require the strongest possible prior to ensure compelling imputation. Thus, we use a vector-space representation at the bottle-neck of our architecture for the latent space that has spatial support over the entire output. To ensure drivability, we encourage the latent codes to contain the least amount of information about the driving signal as possible by employing disentanglement strategies during training. We train all these components of our model jointly using a Conditional Variational Auto Encoder (cVAE) [Sohn et al. 2015] that has been used effectively to produce well-structured latent

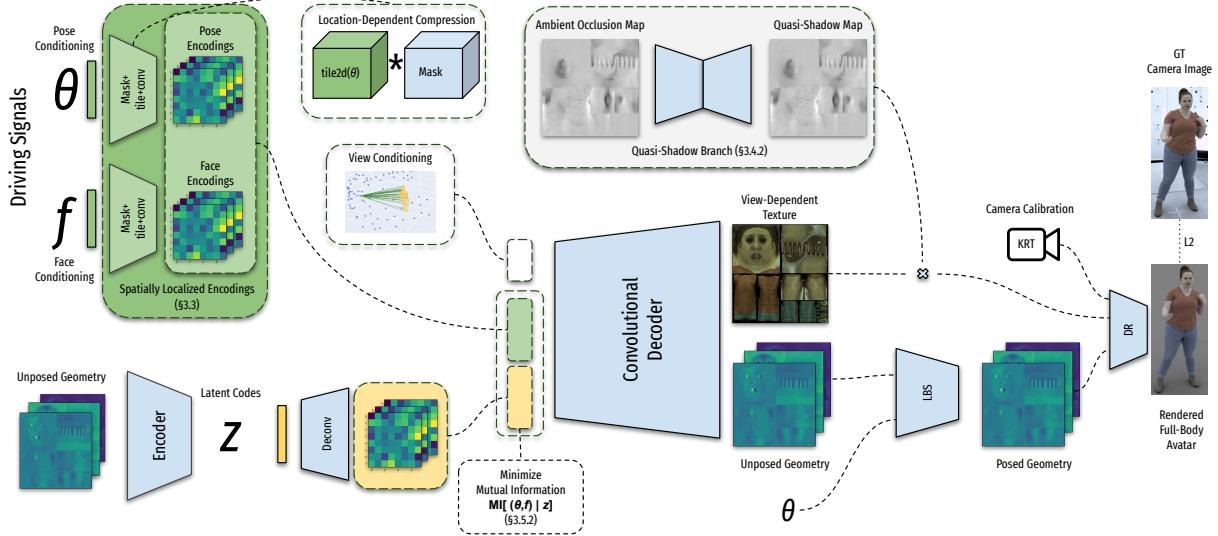


Fig. 2. General overview of the architecture. The core of our full-body model is a conditional variational auto-encoder, which takes as input driving signals and view direction, and outputs geometry and view-dependent texture. These are rendered to produce a full-body avatar. Spatially localized encoding of the driving signals helps reduce spurious correlations. Additionally, an LBS module and a quasi-shadow branch capture coarse, long-range effects. Information not present in the driving signals (e.g., clothing state) is captured by a disentangled latent code z .

spaces in human-centric datasets [Lample et al. 2017; Lombardi et al. 2018]. For supervision, we directly use multi-view images, which we achieve by applying differentiable rendering [Liu et al. 2019a] to generate synthetic images for a direct comparison with the ground truth.

3.2 Variational Autoencoder

The core of our model is a conditional variational auto-encoder (cVAE), consisting of an encoder $E(\cdot | \mathbf{W}_E)$, and a decoder $D(\cdot | \mathbf{W}_D)$, both parameterized as convolutional neural networks, with weights \mathbf{W}_E and \mathbf{W}_D , respectively. The cVAE is trained end-to-end to reconstruct images of a subject captured from a multi-view camera rig. Our system presumes the availability of a registered mesh, $G_i \in \mathbb{R}^{N_g \times 3}$, for each frame i that we acquire using LBS-based tracking with surface registration [Gall et al. 2009], followed by Laplacian deformation to 3D scans for better registration [Botsch and Sorkine 2008; Sorkine et al. 2004]. The LBS joint angles, $\theta_i \in \mathbb{R}^{100}$, as well as 3D facial keypoints, $f_i \in \mathbb{R}^{N_f}$, are the driving signals that we use to condition our model. For settings where 3D face keypoints are not available, such as in the experiment with a VR headset in §4.2.2, we use the latent codes of the personalized facial model from [Wei et al. 2019] instead. Finally, we note that the goal of our cVAE training is to produce a decoder that we can use to animate an avatar from driving signals at test time. To this end, the role of the encoder is strictly to facilitate learning and ensure a well structured latent space. It can be discarded once training is complete.

During training, the encoder takes as input geometry, G_i , that has been unposed using LBS, and produces parameters of a Gaussian distribution, \mathcal{N} , from which a latent code $z \in \mathbb{R}^{512}$ is sampled:

$$\mu_i, \sigma_i \leftarrow E(G_i | \mathbf{W}_E), z_i \sim \mathcal{N}(\mu_i, \sigma_i^2).$$

Here, G_i is first rendered to a position map in UV space before being passed to the convolutional encoder. The reparameterization trick [Kingma and Welling 2013] is used to ensure differentiability of the sampling process. Given the latent codes z_i , driving signals (θ_i, f_i) and a viewpoint $v_i \in \mathbb{R}^3$, the decoder produces reconstructed geometry \tilde{G}_i and view-dependent texture \tilde{T}_i :

$$\tilde{G}_i, \tilde{T}_i \leftarrow D(\theta_i, f_i, z_i, v_i | \mathbf{W}_D).$$

\tilde{G}_i is then passed to the LBS module to produce the final geometry, and \tilde{T}_i is multiplied by a quasi-shadow map to produce the final texture. The generation of the quasi shadow map is detailed in §3.4.2. Differentiable rasterization [Liu et al. 2019a] is used to render an image using the output geometry and texture, which is then compared with the ground truth image using an L2-error. Along with other supervision and regularization losses described in §3.6, the system is trained end-to-end, solving for the cVAE parameters \mathbf{W}_E and \mathbf{W}_D .

Once the model is trained, animation is performed by taking the driving signals (θ, f) e.g. estimated from an external sensor, imputing the latent code z (e.g. by sampling or employing a temporal model), and then synthesizing the geometry and texture by running the decoder $D(\theta_i, f_i, z_i, v_i | \mathbf{W}_D)$ followed by rasterization. The reason for separating conditioning signal between (θ, f) and z comes from the fact that, in practice, access to the *complete* conditioning signal is available only during training. This stems from differences in their capture setup: training data collection often allows for sophisticated multi-camera rigs whose data can be processed in post, whereas during animation, the capture system is typically more constrained and requires real-time processing. Our specific choice of (θ, f) in this work is informed by attributes that can be reliably inferred from

limited sensing setups in real-time, such as keypoints and skeleton joint angles from a simple stereo camera pair [Gall et al. 2009; Tan et al. 2020]. Ultimately, this means that the information contained in the driving signal is often insufficient to fully describe the output, and we need to introduce z that describes the remaining part of the signal, so as to avoid model overfitting or smoothing out the results. In Section 4, we provide a comparison with a baseline version of the model that does not use a latent space.

3.3 Spatially Localized Driving Signals

For a model to be truly drivable, it has to generalize well. That is, it should produce realistic outputs for all real combinations of the driving signals. We are focusing on building data-driven models, and thus a naive approach to achieving generalization would be to collect an exhaustive amount of data that would sufficiently cover the space of variations. Unfortunately, even for personalized full-body models with a single attire, collecting a dataset that can cover all possible combinations of body posture, facial expressions and hand gestures is intractable due to the combinatorial explosion of part variations. A common approach is to instead capture range-of-motion data, with the aim of spanning the full range of each body part with the hope that the model can learn to factorize these parts accordingly. However, if one considers highly-expressive models, such as deep ConvNets, relying on such limited data can lead to situations where the model discovers spurious correlations and thus learns to encode capture-specific dependencies that are not present in other sequences. For example, if during the capture of hand gestures the subject keeps the same facial expression throughout, there is no incentive for the model *not to* learn an association between that facial expression and the hand gestures. An example of this phenomenon occurring on real data is shown in Figure 3.

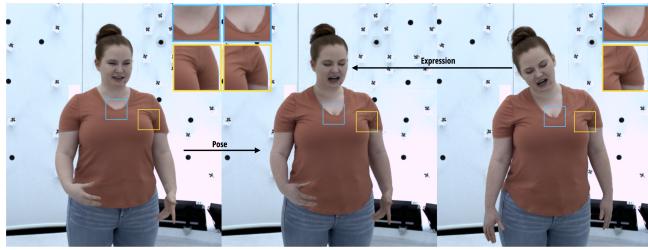


Fig. 3. Spurious Correlations. A naive model learns to associate clothing and shadows facial expression. Combining the pose from the sample on the left, and the facial expression from the image on the right produces an unrealistic sample where pose-dependent effects are fully controlled by the facial expressions (e.g. the collar and armpits, which should be independent of expression).

However, if we assume that the driving signal is sufficiently reliable, and it is not necessary to capture correlations between all of its individual components, we can build a model in a way that encourages decoupling between those individual components. For the human body this would correspond to the fact that spatially distant parts (e.g. fingers on different hands) can move completely independently from each other. Similar intuition has been applied for human geometry in STAR [Osman et al. 2020], which proposes a

localized version of pose correctives aimed at reducing spurious correlations: instead of *global* blendshapes as in SMPL [Bogo et al. 2016], the authors propose *local* blendshapes with a non-linear blending scheme. Note that this reasoning is valid primarily for geometrical deformations and appearance effects which are *local*, but does not hold for global ones, such as shadowing, and one has to model them separately, as we will discuss in §3.4.2.

In this work, we rely on the structure of our decoder network to achieve conditional spatial independence given our driving signal. Specifically, our decoder is a fully-convolutional network which takes as inputs several encoding maps, $\mathbf{e}_* \in \mathbb{R}^{N_* \times 32 \times 32}$, one for each driving source, $* \in \{\theta, f\}$, where:

$$\mathbf{e}_* = \text{proj}(\mathbf{M}_* \odot \text{tile2d}(*, 32, 32) \mid \mathbf{W}_*).$$

Here, $\text{tile2d}(\mathbf{x}, h, w)$ is an operation that repeats a given vector \mathbf{x} into a $h \times w$ feature map, and $\text{proj}(\cdot \mid \mathbf{W}_*)$ is a projection operation that applies a location-dependent compression at each point of the input feature map. It is implemented using a two-layer 1x1-convolutional network with untied biases. We also apply a binary mask $\mathbf{M}_x \in \{0, 1\}^{100 \times 32 \times 32}$, where each channel roughly defines the region of influence for each parameter, which we define using downsampled LBS skinning weights and capture the local effects of each joint angle. In principle, these masks can also be learned, but we did not observe improved results by doing so in our experiments. Figure 4 visualizes the effects of using localized driving signals, where varying one of them leads to spatially (and semantically) coherent changes in the corresponding region of the output.

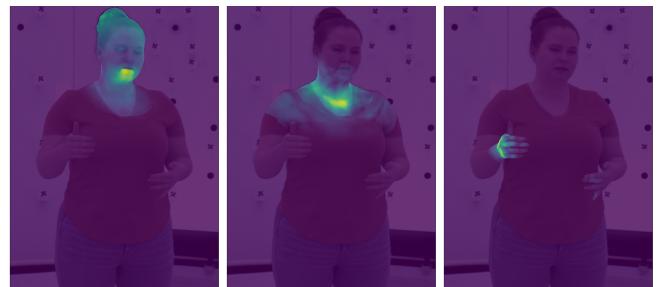


Fig. 4. Effects of localized representations. Heatmaps indicate the areas with largest changes in the output of our model when varying a single control variable. From left to right: face, neck, right wrist.

3.4 Coarse Long Range Effects

Some common driving signals used for human body animation have long range effects that couple distant locations on the body. Explicitly localizing the effects of the driving signal, as described previously in §3.3, can therefore limit the model’s ability to reconstruct real data with high accuracy. But naively reintroducing global influence can lead to over-fitting. As such, we identify two major sources of long-range effects in human body appearance, rigid limb motion and self-shadowing, and propagate these effects spatially using parametric forms that have been shown to generalize well to novel poses.

3.4.1 Rigid Limb Motion. Similar to some existing work [Anguelov et al. 2005; Loper et al. 2015; Osman et al. 2020], our geometric model composes the localized model, described in §3.3, with an explicit skeleton model based on LBS, which captures large rigid transformations of limbs:

$$\bar{\mathbf{G}} = \text{LBS}(\boldsymbol{\theta}, D_G(\boldsymbol{\theta}, \mathbf{f}, \mathbf{z}) + \bar{\mathbf{G}}_T).$$

Here, D_G is the geometry branch of the decoder, $\bar{\mathbf{G}}_T$ is the *template* mesh in canonical pose, and $\bar{\mathbf{G}}$ is the final geometry after posing³. In particular, given the template and the pose parameters, each vertex $\mathbf{x}^i \in \mathbb{R}^3$ of $\bar{\mathbf{G}}$ is computed as:

$$\mathbf{x}^i = \sum_j w_{ij} \cdot (\mathbf{R}_j \cdot \mathbf{x}_T^i + \mathbf{t}_j),$$

where w_{ij} are the pre-defined skinning weights, which describe the influence of j -th joint on the i -th vertex \mathbf{x}^i , and \mathbf{R}_j , \mathbf{t}_j are per-joint transformation parameters, which are computed from $\boldsymbol{\theta}$ via forward kinematics.

The LBS model has been demonstrated to capture large coarse motion of the human body, linking distant locations via a kinematic chain. Although many of its limitations are well known, e.g., the candy wrapper effect [Kavan et al. 2008] and inability to model more complex nonlinear effects [Jacobson and Sorkine 2011; Loper et al. 2015], these deficiencies tend to be localized and have been shown to be addressable by using localized corrective models [Osman et al. 2020]. Our approach builds on these prior findings, complementing LBS with the localized model described previously in §3.3.

3.4.2 Self-Shadowing. Whereas rigid limb motion accounts for much of the long range geometrical effects, self-shadowing accounts for the majority of long-range appearance effects. This is particularly the case during self-interaction, where limbs touch, or are in close proximity to each other. Explicitly accounting for these effects frees up the localized model in §3.3 to focus on the remaining sources of appearance variations, which tend to be more localized, such as skin wrinkles, clothing shifts, and view-dependent effects.

Physically correct estimation of self-shadowing is an extensive process [Hill et al. 2020]. It requires an accurate model of the lighting environment and geometry that may not be readily available. In our case, this would require first running the whole geometry branch before an estimate of shadowing can be used to inform the texture branch, which limits the use of parallelism for achieving fast decoding. Instead, in this work we compute a fast approximate model for depicting environment-occlusion, and then learn the shadowing effect implicitly.

Ambient occlusion (AO) captures environment visibility at each location of the body. It is a strong feature for self-shadowing in a static environment that can be computed efficiently [Miller 1994]. Additionally, for roughly uniform illumination conditions, as in our case, dependence on the global pose with respect to the environment is unnecessary. For computing an AO map, we use the LBS-posed template geometry as it is efficient to compute and does not require the full evaluation of the geometry branch to generate. Although this reduces the precision of the AO map, our goal here is to capture coarse long-range appearance effects, and the differences resulting

³We have dropped the frame index i from $\boldsymbol{\theta}_i$, \mathbf{f}_i and \mathbf{z}_i to reduce clutter.

from using the posed template geometry tend to be localized. This AO map is then passed through the shadow branch; a neural network with a UNet architecture [Ronneberger et al. 2015] that produces a quasi-shadow map. The result is multiplied with the output of the texture branch to produce the final texture. Note that the shadow branch is not supervised directly. Rather, it is used as an inductive bias in cVAE, where its supervision is implicit through a comparison of the final rendered image with the ground truth capture (see §3.2). In practice, we found that it is sufficient to produce a low-resolution quasi-shadow map (4x less compared to the texture) to obtain plausible soft shadows. An example of input and output of the shadow branch is given in Figure 5.

Owing to its physically inspired formulation, AO computation generalizes to novel geometry generated through LBS, which has already demonstrated its ability to generalize to new poses, given that reliable joint angles can be acquired from the driving signals. The UNet architecture has been shown to exhibit good generalization [Ronneberger et al. 2015], especially when its input and outputs differ mostly by local changes, as in our case, by allowing it to rely mostly on shallow skip connections⁴. As with LBS for geometry, the shadow branch frees up the texture branch to focus on local appearance effects since the majority of long range effects are already accounted for through the shadow maps.



Fig. 5. **Shadow branch inputs and outputs.** The left image is an example of a conditioning signal for the network, and the right image is the prediction of our network on a test sample. Note that even though there is no explicit supervision provided to the shadow branch, it naturally learns to capture approximate global shadowing.

3.5 Handling Information Deficiency

In this section, we describe our approach for handling the problem of information deficiency during animation that was outlined in §3.2. Our main assumption is that one can identify an *explicitly* defined set of control variables, such as $\boldsymbol{\theta}$ and \mathbf{f} , that can be reliably estimated during both training and animation. The choice for driving signals is mainly driven by the application and the sensors it affords.

⁴For non-uniform lighting environments, the generalization performance of this network may be affected since the input AO map would differ from the output shadow map more severely. One solution to this might be to perform a full shadow map calculation that the UNet would then only need to refine. We leave that investigation as future work.

A common set of signals that can be acquired for driving include the full-body pose and facial keypoints estimated from cameras placed in the scene, or, when deploying to a VR-based application, on a headset.

Only providing the model with body joints and face keypoints is, however, insufficient, as they do not contain all the information required to explain the full appearance of a body in motion, such as the specific state of clothing, hair and internals of the mouth for a given image. Due to the high dimensionality of the inputs as well as the high capacity of neural networks, directly regressing the output from these signals alone typically overfits to the data, where it learns correlations between the driving signal and outputs that are not generalizable, e.g. by remembering a certain shadow or wrinkle combination.

To address this problem, a naive solution would be to simply introduce an additional latent variable \mathbf{z} that contains the remaining information about the outputs, e.g. by encoding that latent code from the outputs. However, this approach leads to another problem; there is no guarantee that \mathbf{z} is *independent* of the driving signal. In practice, it tends to retain some information about (θ, \mathbf{f}) , which leads the decoder to ignore either of the two variables, or worse, learn to "spread" the information between the two, eliminating the ability to remain faithful to the driving signal during animation⁵. One example of such a failure mode is illustrated in Figure 6.

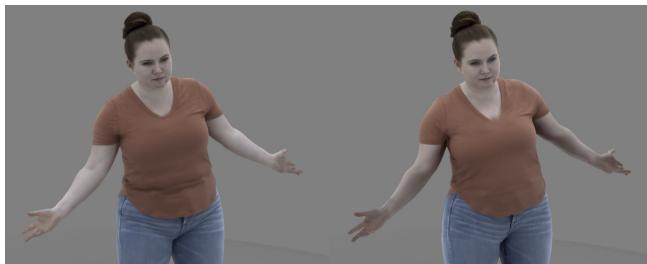


Fig. 6. The need for disentangling. A model with a latent space but no disentangling mechanism produces reasonable reconstruction (left) but ignores the pose conditioning, which leads to severe artifacts when driving (right). For example, incorrect shading on the arms, ghosting artifacts on the collar and overly smoothed clothing wrinkles.

In this work, we propose three strategies to *disentangle* the two sets of variables, which one can select between to best match the characteristics of the data at hand.

3.5.1 Variational Preferencing. cVAEs have a built-in mechanism for disentangling conditioning variables from the latent space. This stems from the propensity of a VAE with a factorized Gaussian posterior to push a portion of its latent dimensions to become uninformative in order to better minimize its KL-regularized loss. In effect, it squeezes information at the input to the encoder into a

⁵A related problem is that of posterior collapse [Aliakbarian et al. 2019], where a high capacity decoder or strong conditioning signal can lead to the latent space being ignored. This scenario is more common in temporal auto-regressive models, where the history used for conditioning can be highly predictive of the current state. We did not observe this behavior in our setting, possibly due to the conditioning signal not being strong enough or exhibiting a complex relationship with the output, where the information path from encoder to decoder is learned more easily.

lower-dimensional subspace within its latent space, the dimensionality of which can only be reduced as the KL weight increases.

In cVAE, when the conditioning variable shares some information with the input to the encoder, there is a preference to let the decoder use the *copy* of that information contained in the driving signal, so as to allow it to be squeezed out of the latent space, thereby achieving lower KL-divergence while still providing the decoder with information necessary for good reconstruction. This results in disentanglement between the latent space and conditioning variables.

By employing a cVAE architecture, our model already exhibits a built-in disentangling mechanism between the latent space and the driving signal which is used to condition the decoder. However, VAE's disentangling mechanism is not perfect. Architecture choices for the encoder and decoder, as well as optimization strategies and even the characteristics of the data can either encourage or impede disentanglement. Furthermore, increasing the KL-weight as a way of achieving better disentanglement can negatively impact reconstruction accuracy as the effective dimensionality of the latent space becomes further reduced.

3.5.2 Mutual Information Minimization. A more direct approach to *disentangle* the two sets of variables is by minimizing the mutual information (MI) between them:

$$\mathcal{I}(\mathbf{z}, \mathbf{c}) = \text{KL}(p(\mathbf{c}, \mathbf{z}) || p(\mathbf{c})p(\mathbf{z})) ,$$

i.e. a KL-divergence between the joint distribution and the factorized one, where $\mathbf{c} = (\theta, \mathbf{f})$. Since directly computing MI is intractable, we use MINE [Mattei and Frellsen 2019] - a generic data-driven approach for estimating the mutual information between random variables. It introduces a parametric approximation to MI through a "statistics network" implemented using a deep neural network:

$$m \leftarrow f(\mathbf{z}, \mathbf{c}) ,$$

where $m \in \mathbb{R}$ is a scalar that is high when the mutual information between \mathbf{z} and \mathbf{c} is high, and low otherwise. This statistics network $f(\cdot)$ is trained using the following loss function with each mini-batch:

$$\mathcal{L}_{\text{mi}} = -\left(\frac{1}{B} \sum_{b=1}^B f(\mathbf{c}_b, \mathbf{z}_b) - \log \sum_{b=1}^B \frac{1}{B} e^{f(\mathbf{c}_b, \hat{\mathbf{z}}_b)}\right) , \quad (1)$$

where B is the size of a mini-batch, and $\hat{\mathbf{z}}_b$ is a shuffled, or randomly sampled, embedding. Thus, the network is trained to maximize the difference in f -function scores between when the data is paired vs. randomized. This results in a (biased) estimate of MI, which can be used to evaluate how independent \mathbf{c} and \mathbf{z} are. The loss in Eq.1 can be used to define an adversarial term $\mathcal{L}_{\text{dis}} = -\mathcal{L}_{\text{mi}}$, which we add to the full model training objective described in §3.6. This way, our cVAE learns to reconstruct the training data while disentangling the latent space from the driving signals.

3.5.3 Perturbation Consistency. When the driving signal has a direct physical interpretation with respect to the model's output, there is an even more direct way to encourage disentanglement. For example, when the driving signal comprises a set of keypoints on the surface of the body, it may be possible to define direct correspondence with vertices on the output geometry. In this case,

disentanglement between the latent code, \mathbf{z} , and the driving signal, \mathbf{c} , can be achieved using a perturbation consistency loss:

$$\mathcal{L}_{\text{pc}} = \mathbb{E}_{p(\mathbf{z})} \left[\sum_{k \in C} \|\mathbf{c}_k - S_k D(\mathbf{z}, \mathbf{c} | \mathbf{W}_D)\|^2 \right], \quad (2)$$

where p is the prior over the VAE's latent space, C is the set of driving signal components that admit a correspondence with the output from the decoder, D , and S_k is a selection matrix that picks the corresponding elements in the output. In practice, the expected value is approximated by a discrete sampling of \mathbf{z} for each minibatch during training. If \mathbf{z} contains information about \mathbf{c} , a random sample will perturb the corresponding elements of the decoded output away from \mathbf{c} . Similarly, if \mathbf{c} and \mathbf{z} are independent, modifying \mathbf{z} will have little impact on \mathbf{c} -corresponded outputs of D , so long as D can reconstruct the data well.

Although the perturbation consistency loss is the most direct way of encouraging independence between the latent space and driving signal, it is not always possible to define direct correspondence between all elements of the driving signal and the output of the decoder. In practice, we find a good heuristic is to start by relying simply on variational preferencing, and if that is not enough, add mutual-information minimization, and finally to add perturbation consistency whenever it is available.

3.6 Training Details

We train all our models with Adam, using an initial learning rate of 1.0e-3, and batch size 8. All models are trained until convergence on 8 NVidia V100 GPU with 32GB RAM. For full body models, training takes approximately 1.5–2 days. We optimize the following composite loss:

$$\mathcal{L} = \lambda_I \mathcal{L}_I + \lambda_M \mathcal{L}_M + \lambda_{\text{lap}} \mathcal{L}_{\text{lap}} + \lambda_G \mathcal{L}_G + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}} + \lambda_{\text{dis}} \mathcal{L}_{\text{dis}}, \quad (3)$$

where \mathcal{L}_I is the inverse rendering losses for images \mathbf{I} and foreground masks \mathbf{M}

$$\mathcal{L}_I = \|\mathbf{I}_{\text{gt}} - \hat{\mathbf{I}}\|_1, \mathcal{L}_M = \|\mathbf{M}_{\text{gt}} - \hat{\mathbf{M}}\|_2^2,$$

\mathcal{L}_G is the per-vertex tracked mesh loss:

$$\mathcal{L}_G = \|\mathbf{G}_{\text{gt}} - \hat{\mathbf{G}}\|_2^2$$

\mathcal{L}_{lap} is the Laplacian loss with respect to the ground truth tracked mesh, which encourages smoothness

$$\mathcal{L}_{\text{lap}} = \|L(\mathbf{G}_{\text{gt}}) - L(\hat{\mathbf{G}})\|_2^2,$$

where $L(\cdot)$ is the mesh Laplacian operator. \mathcal{L}_{KL} is the variational term, which is a standard VAE KL-divergence penalty [Kingma and Welling 2013], and \mathcal{L}_{dis} is the disentangling loss as described in §3.5. In practice, we only use mesh supervision \mathcal{L}_G during the first 2000 iterations so as to provide a reasonable initialization for the model. After that, the inverse rendering loss on images and masks further improves the shape and correspondence, as the initial tracked mesh may exhibit artifacts due the tracking challenges of human body surface. We find that this two phase procedure leads to improved estimates of shape and correspondence, which boosts the accuracy at which our decoder can reconstruct the images.



Fig. 7. **Data Processing Pipeline.** From left to right: input image, foreground body mask, reconstructed 3d shape, detected keypoints, registered surface mesh.

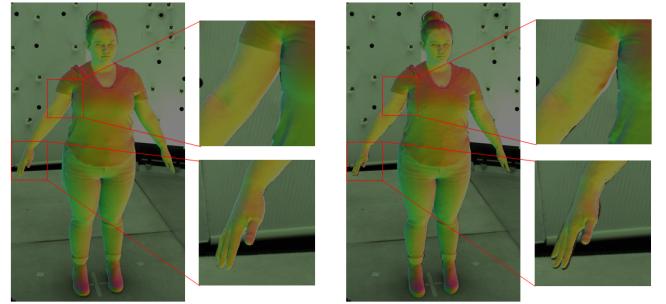


Fig. 8. **Geometry Improvement by Inverse Rendering.** Left: our result after inverse rendering; right: initial tracked mesh from data processing.

4 EXPERIMENTS

In this section, we provide an experimental evaluation of our approach for creating realistic full-body human avatars. We first shortly describe our capture setup and the data. We then provide qualitative results on multiple identities on reconstruction and driving tasks. Finally, we provide a quantitative evaluation and an ablation study on different components of our model.

4.1 Data

Our capture setup is a multi-camera dome-shaped rig with 140 synchronized cameras, each capable of producing 4096x2668 images. We collect data from three subjects, where for each subject, we capture sequences of 50–70k frames in length, which include a range of motion and natural conversation sequences. Out of those frames roughly 5000 were left out for testing purposes. For one of the subjects, we also collected two additional testing sequences: one where the individual is wearing different clothing, and another where they are wearing a VR headset. In both cases, the full state of the model built from the subject's original capture is not observable. In the first, the specific state of clothing is not transferable due to differences in attire between the two captures. In the second, parts of the face are occluded by the VR headset.

To train our full body models, we need to obtain skeletal poses as well as registered surface meshes for every frame of the multi-view captures. For this, we employ a traditional computer vision pipeline that is capable of generating reasonable estimates for both. Specifically, we follow a two-step approach: first, we run multi-view 3d reconstruction [Galliani et al. 2015], keypoint detection



Fig. 9. Qualitative Evaluation: Reconstruction Quality. We demonstrate the ability of our model to produce high-quality reconstructions given full inputs (body pose θ , facial keypoints f and latent codes z). Note that none of these frames were observed during training.



Fig. 10. **Qualitative Results: Driving.** For all identities the model is conditioned on pose θ , facial keypoints f , and an imputed latent code z . We use a naive imputation strategy with a constant latent code.

and triangulation [Tan et al. 2020] and foreground body segmentation [Kirillov et al. 2020]. Second, we perform skeletal pose estimation using a personalized body rig and LBS by matching the various features computed in the first step as well as pose priors over joint angles [Gall et al. 2009]. We further refine these estimates by running a deformable ICP algorithm on top of the tracked meshes so as to fit the 3D scan details better with surface Laplacian regularization [Sorkine et al. 2004]. One example from the various stage of the pipeline can be seen in Figure 7. Processing the entire dataset for a single subject takes 14 days with 160 GPUs (20 DGX servers with 8 GPU each). This data processing pipeline is robust enough to process most of the data, and only has challenges in poses with heavy self-occlusions, like heavy hand-cloth interactions. We discard on average roughly 10 percent of frames by filtering out frames where the tracked surface has large error compared with the scans. Notice that even on successfully registered frames, due to limited 3D scan resolution or foreground mask inaccuracies, the registered mesh may exhibit errors in shape or correspondence. We observe that these errors are greatly reduced during model training by using inverse rendering with an image reconstruction loss. One example of such improvement can be seen in Figure 8, where the decoded mesh shows greatly improved alignment with respect to the initial tracking mesh in high detail areas such as the fingers and around creases or folds.

4.2 Qualitative Results

4.2.1 Reconstruction. We first evaluate our model’s ability to generalize to new poses that are held out from the training set. For this we optimize all the model’s parameters, including the global pose, skeletal joint angles θ , face parameters f and latent codes z , to minimize the reconstruction loss over images from the multi-view cameras as well as the tracked mesh. Figure 9 shows the reconstruction results for three different identities and different combinations of poses and facial expressions, along with the ground truth capture images as a reference. The results demonstrate good reconstruction, capable of capturing subtle details of expression and pose.

4.2.2 Driving. Our main goal is to build drivable models, which produce realistically looking virtual humans given real driving signals. In this section, we thus evaluate the quality of our model on the task of *driving*, where only partial information about the full model’s state is available via the driving signals. The latent codes are not provided and have to be imputed.

Matched Setting. First we would like to evaluate how much information is really missing from a full body model that is not accounted for by skeletal joint angles and facial expressions. In Figure 10 we show driving results where only the pose θ and facial keypoints f are presumed to be available, and we simply set the latent code to zero; a maximum likelihood imputation. A video sequence of this driving can be found in the supplemental video. The first point to note is that the pose and facial expressions appear to be matched well between the avatar and the ground truth images, demonstrating the efficacy of the disentanglement scheme described in §3. The second point to notice, is that the reconstruction attained by the

driven model is of less quality than the reconstructions, with errors concentrated around clothing; areas where we expect joint and facial keypoints to have limited disambiguation capabilities. A comparison between driving and reconstruction can be found in Figure 11.

Now that we have established the extent of missing information, we would like to evaluate the effects this has on models that ignore it or model it without disentangling. In Figure 14 we provide a comparison between our method and two baselines: one without the latent space (pose+face) and one without latent-space disentanglement (pose+face+latent). Both baselines appear to exhibit a number of visual artifacts, including incorrect shadows, ghosting, and over-smoothing. For example, extraneous shadows on the pants in (pose+face) and missing shading in the armpits for (pose+face+latent) in the top row images. Another example is the unnatural wrinkles in the torso for (pose+face) and ghosting artefact in the neck area of the shirt for (pose+face+latent) in the bottom row. At a glance, static views of these models without a latent space can look acceptable, but over-fitting artifacts are significantly more noticeable in dynamic sequences. We encourage the reader to view the accompanying supplementary materials, in which we also provide additional comparison to an image-space method [Thies et al. 2019].

Unmatched Setting. In the matched setting above, information deficiency was contrived in order to evaluate its effect on animation. In practice, one could have chosen to solve for the latent code that minimizes the reconstruction error since the capture setting for the model and the performer are identical. However, a far more common scenario is when there is a domain gap between the capture setting during modeling and during driving. In this case, the model can not fully span the appearance in the driving image, and must rely instead on robust features that can be equivalently extracted in both domains to drive the model. We consider two such scenarios for one of our capture subjects.

In the first, the subject is captured in a different attire; different colored clothing, sandals instead of shoes, and hair out instead of tied in a bun. This mimics the scenario where a new performance for free-viewpoint video needs to be captured at a different time, where the actor’s appearance has changed or the old attire is not accessible. Here, the skeletal pose and facial keypoints extracted from the capture studio are used to drive the avatar. Some frames of this capture along with the animated avatar, are shown in Figure 12.

In the second, the subject is driving the model while wearing a VR headset which occludes parts of her face. There is a pair of stereo cameras in the scene as well as on the headset that extract driving signals that comprise the skeletal keypoints and facial expression codes obtained using the approach in [Wei et al. 2019]. This setting is a minimal sensing configuration for enabling VR-based telepresence for two-way interaction. Some frames demonstrating the performance of the system are shown in Figure 13.

In both the attire and headset settings, our model remains faithful to the information content present in the driving signal, while producing plausible imputations of information that is missing by simply setting the latent code in all cases to zero. Further results of these driving results can be found in the supplementary video,

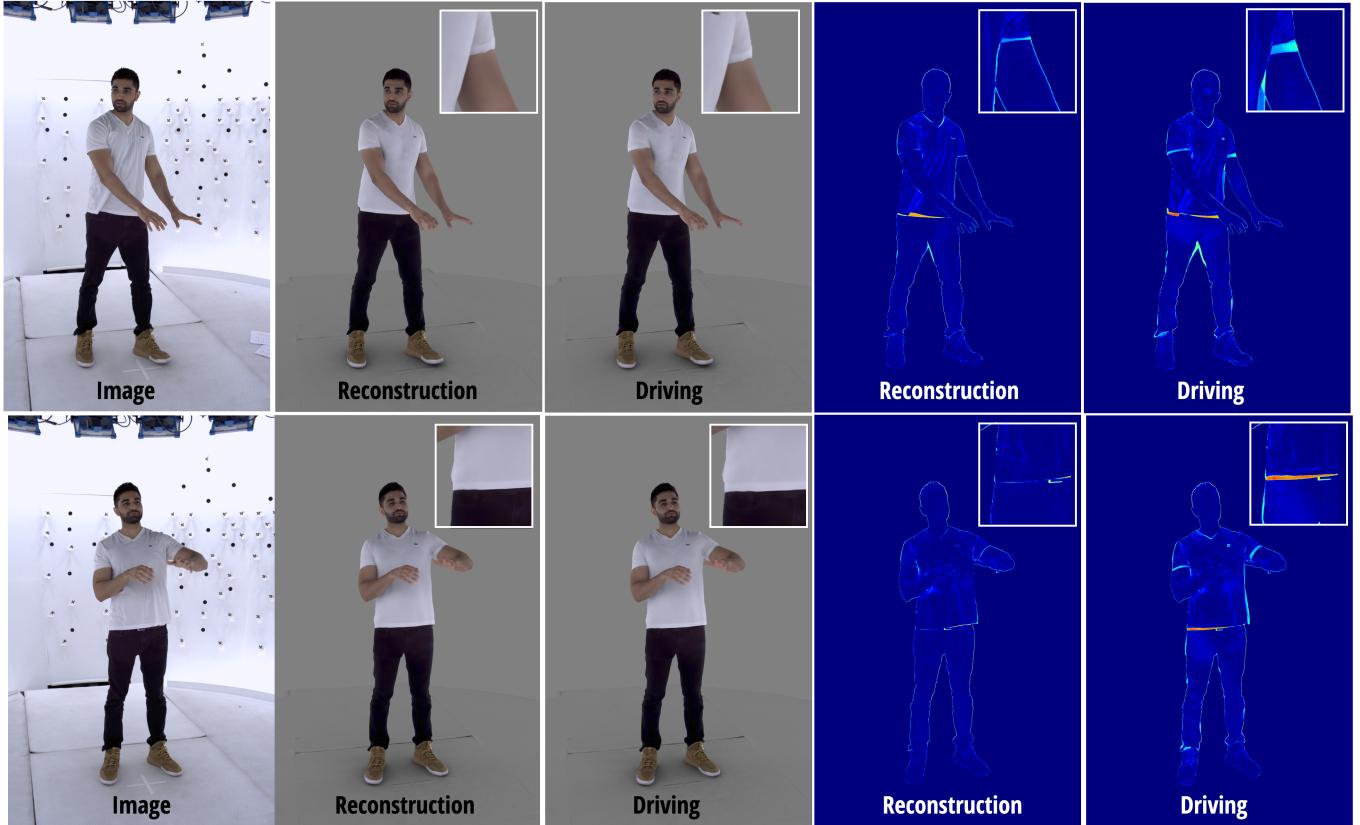


Fig. 11. **Information Deficiency.** Driving by setting the latent code to zero produces plausible results but does not match the data where the driving signal is deficient, for example to model specific clothing states.



Fig. 12. **Qualitative Evaluation: Driving in Different Clothing.** The model is conditioned on pose θ , facial keypoints f , and an imputed latent code z . We use a naive imputation strategy with constant latent code.



Fig. 13. **Qualitative Evaluation: Driving with a Headset.** The model is conditioned on pose θ , facial expression codes f , and an imputed latent code z .

where we additionally visualize results that use different methods for latent code synthesis, such as through random sampling and by using a temporal model. By disentangling the latent and driving signals, our method opens up the possibility of applying a variety of different imputation strategies that could be suited to particular applications. A deeper investigation into latent code synthesis is out of the scope of this paper, but it is an interesting direction of future work.

4.3 Quantitative Evaluation

In Table 1, we report numerical results in terms of image error with respect to capture images. Because our main goal is to evaluate how *drivable* the models are, all models are conditioned on the same set of driving signals (i.e. facial expression codes f and skeletal poses θ), and all the models that have a latent space are all conditioned on the same constant latent code $z = 0$. On the training set, the model without a latent space (i.e. pose+face) achieves the lowest error, albeit at the cost of severe overfitting, as it suffers a very significant performance drop on the test set. Among the variations of our methods, the lowest error is obtained by a version without compression mechanism, which we also attribute to overfitting. The baseline with a latent space (pose+face+latent) performs worse on both train and test - which we attribute to the fact that it does not employ any disentanglement-promoting mechanism. Similarly, out of all the versions of our approach, the model without an explicit disentanglement performs slightly worse on both.

On the test set, our method has a clear advantage over the model without latent space disentanglement (pose+face+latent). Performance of the model without the shadow branch indicates that it is useful for generalization; the model without the shadow branch is able to fit the training data well, but suffers significant degradation on the test set. Similarly, the version of our model without spatially localized embeddings has significantly lower training error, but performs significantly worse on the test set. We can also infer

that the version without disentanglement is prone to poor generalization; the model is able to fit the training data but leads to the worst performance amongst all variants of our model.

Table 1. **Quantitative Evaluation.** Image reconstruction error for driving on training data (train) and testing (unseen) samples. The baseline without a latent code (pose+face) is prone to learning chance correlations and leads to severe overfitting, the baseline with a latent code (pose+face+latent) without disentanglement mechanisms has poor drivability.

method	train	test
OURS	11.578	15.546
pose+face	7.647	17.509
pose+face+latent	14.861	19.885
OURS (no disent.)	11.926	16.364
OURS (no spat.local.)	9.424	16.461
OURS (no shadow)	10.872	16.792

5 LIMITATIONS AND FUTURE WORK

Our model relies on tracked meshes for supervision and is thus limited by the tracking quality. As tracking extremely loose clothing with topological changes is still an area of active research, it would be challenging to apply our model to those complex scenarios. Moreover, our network architecture relies heavily on the assumption of fixed topology, as it operates on UV-based mesh representation, which limits the applicable scenarios, and may be tackled by instead relying on implicit surface representations [Park et al. 2019; Remelli et al. 2020]. UV-based representations can also be prone to seam artifacts (noticeable e.g. in Figure 13), especially when combined with 2D convolutions [Groueix et al. 2018]. This issue can be addressed by using mesh convolutions [Bronstein et al. 2017; Zhou et al. 2020b], albeit at a higher computational cost.

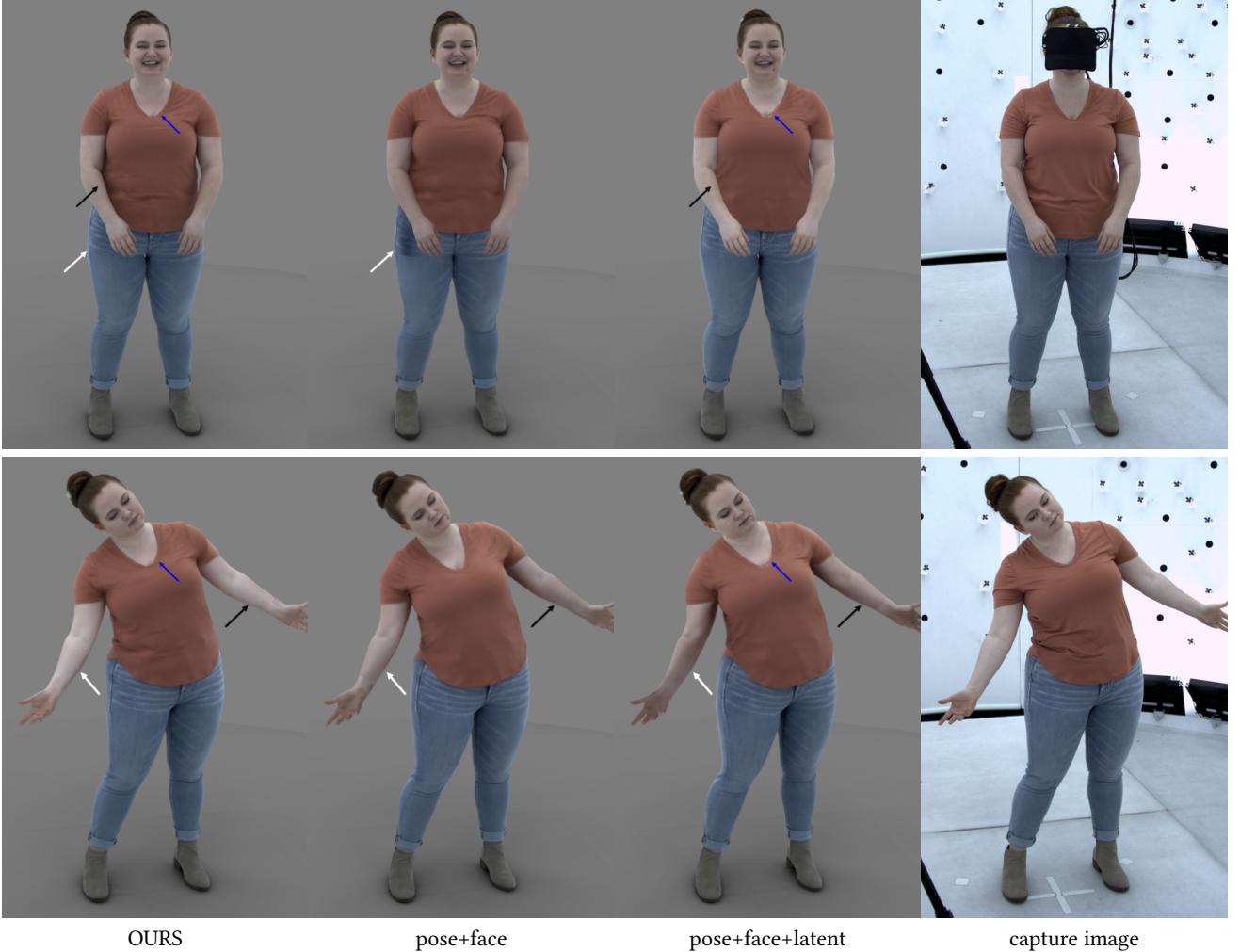


Fig. 14. Qualitative Comparison. We use a naive imputation strategy with a constant latent code. Both baselines struggle at capturing shadows (white and black arrows) and avoiding ghost artefact (blue arrows).

Although latent codes do capture clothing variations and some of the high-frequency deformations, the limited capacity of the latent space can lead to loss of details, which could be the reason that reconstructions in Figure 9 do not capture all of the clothing deformations. One potential future direction to tackle this would be to apply hierarchical generative models [Bagautdinov et al. 2018; Vahdat and Kautz 2020], which tend to have better representational power.

We employed a naive strategy for imputing missing information by setting the latent codes to a constant value for every frame, which means that the state of clothing will be fixed during animation (of course, pose-dependent deformations can still be captured). An interesting direction of future work is to investigate more advanced approaches that can be tailored to a specific application. For example, employing temporal models or style-dependent generation [Alexander et al. 2020; Ginosar et al. 2019] may enable more semantically

meaningful imputations. At the same time, methods for visualizing the space of uncertainty may be relevant for applications that rely on the authenticity of the animation, such as telepresence.

Our data collection and experimental evaluation are primarily focused on natural conversations, and thus it is not fully clear if our pipeline generalizes to particularly challenging poses which are far apart from the training distribution. In the supplementary, we provide an interactive tool for t-SNE-based visualization [Van der Maaten and Hinton 2008] of the poses in our training and testing sets.

6 CONCLUSION

We introduced a novel method for building high-quality photorealistic full-body avatars that integrates, in its construction, the specific modality of driving signal that is available during the model’s use. Information deficiency in the driving signal is accounted for by

using a latent space that is disentangled from the driving signal, enabling the generation of diverse plausible configurations that remain faithful to the information contained in the driving signal. We showcase the capabilities of the model by applying it in two example scenarios where the driving signal is information deficient, demonstrating improved generalization and fidelity compared with other approaches.

ACKNOWLEDGMENTS

We would like to thank Carsten Stoll for providing body tracking code, Georgios Pavlakos for the face fitting pipeline, Sahana Vijai for providing assets for our body rig, Breannan Smith for high-quality hand tracking, and Anuj Pahuja for the hand fitting pipeline and help on VR demo.

REFERENCES

- Kfir Aberman, Mingyi Shi, Jing Liao, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. 2019. Deep video-based performance cloning. In *Computer Graphics Forum*, Vol. 38. Wiley Online Library, 219–233.
- O. Alexander, M. Rogers, W. Lambeth, J. Chiang, W. Ma, C. Wang, and P. Debevec. 2010. The Digital Emily Project: Achieving a Photorealistic Digital Actor. *IEEE Computer Graphics and Applications* 30, 4 (2010), 20–31. <https://doi.org/10.1109/MCG.2010.65>
- Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. *Computer Graphics Forum* 39, 2 (2020), 487–496.
- Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. 2019. Mitigating Posterior Collapse in Strongly Conditioned Variational Autoencoders. (2019).
- T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. 2018. Detailed Human Avatars from Monocular Video. In *Proceedings of International Conference on 3D Vision (3DV)*, 98–109.
- Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018. Video Based Reconstruction of 3D People Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: Shape Completion and Animation of People. *ACM Trans. Graph.* 24, 3 (July 2005), 408–416. <https://doi.org/10.1145/1073204.1073207>
- T. Bagautdinov, C. Wu, J. Saragih, P. Fua, and Y. 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>
- Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. 2018. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062* (2018).
- Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.
- 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., USA, 187–194. <https://doi.org/10.1145/311535.311556>
- Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision*. Springer, 561–578.
- M. Botsch and O. Sorkine. 2008. On Linear Variational Surface Deformation Methods. *IEEE Transactions on Visualization and Computer Graphics* 14, 1 (2008), 213–230.
- Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. 2017. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine* 34, 4 (2017), 18–42.
- Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. 2018. Understanding disentangling in *beta*-VAE. *arXiv preprint arXiv:1804.03599* (2018).
- Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. 2019. Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5933–5942.
- Patrick Esser, Johannes Haux, Timo Milbich, et al. 2018. Towards learning a realistic rendering of human behavior. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 0–0.
- Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. 2009. Motion capture using joint skeleton tracking and surface estimation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 1746–1753.
- S. Galliani, K. Lasinger, and K. Schindler. 2015. Massively Parallel Multiview Stereopsis by Surface Normal Diffusion. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 873–881. <https://doi.org/10.1109/ICCV.2015.106>
- S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik. 2019. Learning Individual Styles of Conversational Gesture. In *Computer Vision and Pattern Recognition (CVPR)*. Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. 2018. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 216–224.
- P. Guan, L. Reiss, D. Hirshberg, A. Weiss, and M. J. Black. 2012. DRape: DRessing Any PErsone. *ACM Trans. on Graphics (Proc. SIGGRAPH)* 31, 4 (July 2012), 35:1–35:10.
- Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework. (2016).
- Stephen Hill, Stephen McAuley, Laurent Belcour, Will Earl, Niklas Harrysson, Sébastien Hillaire, Naty Hoffman, Lee Kerley, Jasmin Patry, Rob Piekk, Igor Skliar, Jonathan Stone, Pascal Barla, Mégane Bati, and Iliyan Georgiev. 2020. Physically Based Shading in Theory and Practice. In *ACM SIGGRAPH 2020 Courses*.
- Alec Jacobson and Olga Sorkine. 2011. Stretchable and Twistable Bones for Skeletal Shape Deformation. *ACM Transactions on Graphics (proceedings of ACM SIGGRAPH ASIA)* 30, 6 (2011), 165:1–165:8.
- B. Jiang, J. Zhang, J. Cai, and J. Zheng. 2020. Disentangled Human Body Embedding Based on Deep Hierarchical Neural Network. *IEEE Transactions on Visualization and Computer Graphics* 26, 8 (2020), 2560–2575.
- H. Joo, T. Simon, and Y. Sheikh. 2018. Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8320–8329. <https://doi.org/10.1109/CVPR.2018.00868>
- Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. 2008. Geometric Skinning with Approximate Dual Quaternion Blending. *ACM Trans. Graph.* 27, 4, Article 105 (Nov. 2008), 23 pages. <https://doi.org/10.1145/1409625.1409627>
- Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
- Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. 2020. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9799–9808.
- Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic DE NOYER, and Marc'Aurelio Ranzato. 2017. Fader Networks: Manipulating Images by Sliding Attributes. In *Advances in Neural Information Processing Systems*, Vol. 30. 5967–5976.
- Manfred Lau, Jinxiang Chai, Ying-Qing Xu, and Heung-Yeung Shum. 2009. Face Poser: Interactive Modeling of 3D Facial Expressions Using Facial Priors. *ACM Trans. Graph.* 29, 1, Article 3 (Dec. 2009), 17 pages. <https://doi.org/10.1145/1640443.1640446>
- J. P. Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Fred Pighin, and Zhigang Deng. 2014. Practice and Theory of Blendshape Facial Models. In *Eurographics 2014 - State of the Art Reports*, Sylvain Lefebvre and Michela Spagnuolo (Eds.). The Eurographics Association. <https://doi.org/10.2312/egst.20141042>
- J. P. Lewis, Matt Cordiner, and Nickson Fong. 2000. Pose Space Deformation: A Unified Approach to Shape Interpolation and Skeleton-Driven Deformation. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*. ACM Press/Addison-Wesley Publishing Co., USA, 165–172. <https://doi.org/10.1145/344779.344862>
- Lingjie Liu, Weipeng Xu, Michael Zollhöfer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. 2019c. Neural Rendering and Reenactment of Human Actor Videos. *ACM Trans. Graph.* 38, 5, Article 139 (Oct. 2019), 14 pages. <https://doi.org/10.1145/3333002>
- Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. 2019a. Soft Rasterizer: A Differentiable Renderer for Image-based 3D Reasoning. *The IEEE International Conference on Computer Vision (ICCV)* (Oct 2019).
- Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. 2019b. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5904–5913.
- Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep appearance models for face rendering. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–13.
- Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* 34, 6 (2015), 1–16.
- Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. 2020. Learning to Dress 3D People in Generative Clothing. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 6468–6477.
- N. Magnenat-Thalmann, R. Lapierre, and D. Thalmann. 1989. Joint-Dependent Local Deformations for Hand Animation and Object Grasping. In *Proceedings on Graphics Interface '88* (Edmonton, Alberta, Canada). Canadian Information Processing Society, CAN, 26–33.

- Pierre-Alexandre Mattei and Jes Frellsen. 2019. MIWAE: Deep generative modelling and imputation of incomplete data sets. In *International Conference on Machine Learning*. PMLR, 4413–4423.
- Gavin Miller. 1994. Efficient Algorithms for Local and Global Accessibility Shading. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*. 319–326.
- Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. 2020. DeepHandMesh: A Weakly-supervised Deep Encoder-Decoder Framework for High-fidelity Hand Mesh Modeling. In *Proceedings of European Conference on Computer Vision (ECCV)*.
- Evronne Ng, Hanbyul Joo, Shiry Ginosar, and Trevor Darrell. 2020. Body2Hands: Learning to Infer 3D Hands from Conversational Gesture Body Dynamics. *arXiv preprint arXiv:2007.12287* (2020).
- Ahmed A A Osman, Timo Bolkart, and Michael J. Black. 2020. STAR: A Sparse Trained Articulated Human Body Regressor. In *European Conference on Computer Vision (ECCV)*. <https://staris.tue.mpg.de>
- Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. 2021. NPMs: Neural Parametric Models for 3D Deformable Shapes. *arXiv preprint arXiv:2104.00702* (2021).
- Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 165–174.
- Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 10975–10985. <http://smpl-x.is.tue.mpg.de>
- Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *CVPR*.
- Sergey Prokudin, Michael J. Black, and Javier Romero. 2021. SMPLpix: Neural Avatars from 3D Human Models. In *Proceedings of Winter Conference on Applications of Computer Vision (WACV)*. 1810–1819.
- Albert Pumarola, Antonio Agudo, Alberto Sanfelix, and Francesc Moreno-Noguer. 2018. Unsupervised person image synthesis in arbitrary poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8620–8628.
- Neng Qian, Jiayi Wang, Franziska Mueller, Florian Bernard, Vladislav Golyanik, and Christian Theobalt. 2020. HTML: A Parametric Hand Texture Model for 3D Hand Reconstruction and Personalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer.
- Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. 2018. Generating 3D Faces using Convolutional Mesh Autoencoders. In *European Conference on Computer Vision (ECCV)*, Vol. Lecture Notes in Computer Science, vol 11207. Springer, Cham, 725–741.
- Edoardo Remelli, Artem Lukoianov, Stephan Richter, Benoît Guillard, Timur Bagautdinov, Pierre Baque, and Pascal Fua. 2020. MeshSDF: Differentiable Iso-Surface Extraction. *Neural Information Processing Systems (NeurIPS)* (2020).
- Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (Nov. 2017).
- 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)*. Springer, 234–241.
- Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. 2021. SCANimate: Weakly Supervised Learning of Skinned Clothed Avatar Networks. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. 2020. Neural Re-Rendering of Humans from a Single Image. In *European Conference on Computer Vision (ECCV)*.
- 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 Transactions on Graphics (TOG)* 39, 4 (2020), 91–1.
- Aliaksandra Shysheya, Egor Zakharov, Kara-Ali Aliev, Renat Bashirov, Egor Burkov, Karim Iskakov, Aleksei Ivakhnenko, Yury Malkov, Igor Pasechnik, Dmitry Ulyanov, Alexander Vakhitov, and Victor Lempitsky. 2019. Textured Neural Avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Chenyang Si, Wei Wang, Liang Wang, and Tieniu Tan. 2018. Multistage adversarial losses for pose-based human image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 118–126.
- Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning Structured Output Representation using Deep Conditional Generative Models. In *Advances in Neural Information Processing Systems*. C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc., 3483–3491. <https://proceedings.neurips.cc/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf>
- O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H.-P. Seidel. 2004. Laplacian Surface Editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing* (Nice, France) (*SGP ’04*). Association for Computing Machinery, New York, NY, USA, 175–184. <https://doi.org/10.1145/1057432.1057456>
- Carsten Stoll, Juergen Gall, Edilson de Aguiar, Sebastian Thrun, and Christian Theobalt. 2010. Video-Based Reconstruction of Animatable Human Characters. In *ACM SIGGRAPH Asia 2010 Papers* (Seoul, South Korea) (*SIGGRAPH ASIA ’10*). Association for Computing Machinery, New York, NY, USA, Article 139, 10 pages. <https://doi.org/10.1145/1866158.1866161>
- Mingxing Tan, Ruoming Pang, and Quoc V Le. 2020. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10781–10790.
- J. Rafael Tena, Fernando De la Torre, and Iain Matthews. 2011. Interactive Region-Based Linear 3D Face Models. *ACM Trans. Graph.* 30, 4, Article 76 (July 2011), 10 pages. <https://doi.org/10.1145/2010324.1964971>
- Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–12.
- Arash Vahdat and Jan Kautz. 2020. NVAE: A Deep Hierarchical Variational Autoencoder. In *Neural Information Processing Systems (NeurIPS)*.
- Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
- Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. 2005. Face Transfer with Multilinear Models. *ACM Trans. Graph.* 24, 3 (July 2005), 426–433. <https://doi.org/10.1145/1073204.1073209>
- Ting-Chun Wang, Ming-Yi Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. Video-to-video synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 1152–1164.
- 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 (2019).
- Chenglei Wu, Derek Bradley, Markus Gross, and Thabo Beeler. 2016. An Anatomically-Constrained Local Deformation Model for Monocular Face Capture. *ACM Trans. Graph.* 35, 4, Article 115 (July 2016), 12 pages. <https://doi.org/10.1145/2897824.2925882>
- Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geeyuk Lee. 2020. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity. *ACM Transactions on Graphics (TOG)* 39, 6 (2020).
- Keyang Zhou, Bharat Lal Bhatnagar, and Gerard Pons-Moll. 2020a. Unsupervised Shape and Pose Disentanglement for 3D Meshes. In *The European Conference on Computer Vision (ECCV)*.
- Yi Zhou, Chenglei Wu, Zimo Li, Chen Cao, Yuting Ye, Jason Saragih, Hao Li, and Yaser Sheikh. 2020b. Fully Convolutional Mesh Autoencoder using Efficient Spatially Varying Kernels. In *Advances in Neural Information Processing Systems*.