MIGS: Multi-Identity Gaussian Splatting via Tensor Decomposition

Aggelina Chatziagapi[1], Grigorios G. Chrysos[2], and Dimitris Samaras[1]

[1] Stony Brook University {aggelina,samaras}@cs.stonybrook.edu
[2] University of Wisconsin-Madison chrysos@wisc.edu

Abstract. We introduce MIGS (Multi-Identity Gaussian Splatting), a novel method that learns a single neural representation for multiple identities, using only monocular videos. Recent 3D Gaussian Splatting (3DGS) approaches for human avatars require per-identity optimization. However, learning a multi-identity representation presents advantages in robustly animating humans under arbitrary poses. We propose to construct a high-order tensor that combines all the learnable 3DGS parameters for all the training identities. By assuming a low-rank structure and factorizing the tensor, we model the complex rigid and non-rigid deformations of multiple subjects in a unified network, significantly reducing the total number of parameters. Our proposed approach leverages information from all the training identities and enables robust animation under challenging unseen poses, outperforming existing approaches. It can also be extended to learn unseen identities. Project page: https://aggelinacha.github.io/MIGS/.

Keywords: Human Body Animation · Multiple Identities · 3D Gaussian Splatting

1 Introduction

Synthesizing and animating digital humans has broad applications, ranging from immersive telepresence in AR/VR to video games and the movie industry. However, this task presents significant challenges, since humans are highly deformable entities in the real world. Capturing their large variety of spatio-temporal (4D) rigid and non-rigid deformations is a widely studied topic in computer vision and graphics [48, 51, 92]. A vast number of works [48, 49, 51, 61, 67, 75] have proposed to learn parametric models of the human body, like SMPL [48], using large datasets of human scans. These models can be used as a prior, conditioning neural representations of human avatars [56, 72, 92, 95].

Advances in implicit representations and neural rendering, including Neural Radiance Fields (NeRFs) [54], have enabled photo-realistic modeling of the appearance and geometry of humans [15, 56, 64, 92, 95, 99]. The non-rigid and rigid deformations of the human body are represented by one or more multi-layer perceptrons (MLPs). The learned dynamic representation can be used for novel view synthesis and animation under novel poses [56, 92, 95]. More recently, 3D Gaussian Splatting (3DGS) [72] achieves real-time rendering of radiance fields

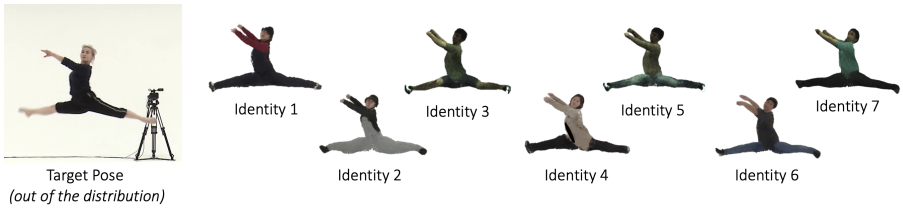


Fig. 1: We introduce MIGS (multi-identity Gaussian splatting) that learns a *single* neural representation for *multiple identities* based on tensor decomposition. MIGS enables robust animation of human avatars under novel poses, out of the training distribution.

with a high quality. By representing a scene with a set of 3D Gaussians and proposing a fast rasterization method, 3DGS significantly outperforms NeRFs that require expensive optimization and long inference times. Recent works extend 3DGS to human avatars [24, 25, 72, 98], by conditioning on a human articulation prior [48]. However, given a monocular setting, these approaches can only learn single-identity representations. Only a few works have proposed more generic neural representations, but require multiple views for training [40,73,97].

In this work, we address the less studied problem of learning a single neural representation of multiple identities from monocular videos. In multi-view settings, related methods usually learn blending mechanisms, extract features from nearby views, and infer a novel view in feed-forward manner [10,40,84,97]. Constrained by the input views, such multi-view approaches cannot animate humans under novel poses. In contrast, in a monocular setting, we cannot leverage nearby views as input. Learning a single representation of multiple subjects is an interesting research topic, since we need to preserve individual characteristics in a generic model, and presents the following advantages. (a) By leveraging information from multiple identities, it learns a better representation of the human body articulation, enabling robust animation under unseen, even out-of-distribution, poses. (b) Even though recent neural rendering methods have become faster, they still require per-identity optimization [25,72,92], which can lead to a prohibitive number of learnable parameters for a large number of identities.

We introduce MIGS (Multi-Identity Gaussian Splatting), a novel method that learns a *multi-identity* 3DGS representation of human avatars. To the best of our knowledge, this is the first method that extends 3DGS to multiple identities, learned in a monocular setting. Inspired by ideas that go back to TensorFaces [82], we construct a high-order tensor of all the learnable parameters of our 3DGS representation. The core premise of our method is to learn a tensor decomposition [36, 63] of the high-order tensor. The intuition is that different identities share common characteristics. Thus, we assume a *low-rank structure* of the high-order tensor and decompose it into a sum of rank-one tensors [23,36]. These rank-one tensors correspond to the only learnable parameters of our optimization procedure. In this way, we learn a *single* unified representation of multiple subjects, and use only a *fraction* of the total parameters that would be required by training multiple per-identity 3DGS representations. By leveraging information from multiple identities, MIGS enables robust animation under chal-

lenging out-of-distribution poses. We also demonstrate how it can be extended to learn novel identities.

In brief, our contributions are as follows:

- We propose to construct a high-order tensor that combines all the learnable parameters of a 3DGS representation for multiple human avatars.
- By factorizing this tensor with a tensor decomposition, we learn a single unified representation for multiple identities, and significantly reduce the total number of learnable parameters compared to single-identity representations.
- By leveraging information from multiple identities, our proposed MIGS enables robust animation under novel poses, outperforming existing approaches.

2 Related Work

Human Body Modeling. Reconstructing and animating humans has been a long standing problem in computer vision and graphics. Earlier approaches develop stereo techniques, relying on multi-view studio setups [7, 29, 52]. They model the non-rigid and rigid deformations of humans with mesh [83] or volume [13] representations. Motion retargeting has also been explored [18,42]. With the advent of generative adversarial networks [20], many approaches propose image-to-image translation methods [26, 87, 88] for 2D motion transfer [9, 58, 70], based on pose detection systems [5, 6, 77, 90]. A vast number of works [48, 49, 51, 61, 67, 75] have proposed to learn parametric models of the human body, like SMPL [48], using large datasets of 3D human scans. These models can represent a human with a set of parameters for the body shape, joint position and rotation. They have been beneficial for reconstructing and tracking humans over time in videos [4, 30, 74]. They have also been used as a human articulation prior to condition implicit neural representations of human avatars [56, 72, 92, 95].

Neural Rendering of Human Avatars. Recent advances in implicit representations and neural rendering have enabled photo-realistic modeling of the appearance and geometry of humans. Neural Radiance Fields (NeRFs) [54] have been first proposed for static scenes, showing remarkable results for novel view synthesis [2, 3, 46, 54]. They have been extended to dynamic scenes [44, 45, 71], making them a popular choice for modeling human bodies [16, 27, 31, 47, 59, 68, 69, 93] and faces [15, 64, 65]. While many of those NeRF-based approaches use multi-view videos, recent works attempt to capture the 4D dynamics and appearance of humans from monocular videos [15, 21, 28, 56, 78, 86, 92, 95]. Most of these works align the observed poses of different video frames in a single canonical space, and learn a conditional 3D representation. The non-rigid and rigid deformations of the human body are represented by one or more multi-layer perceptrons (MLPs). The learned dynamic representation can be used for novel view synthesis and animation under novel poses [56, 92, 95]. However, all these approaches require expensive per-identity optimization. A few works have proposed more generic representations [10, 34, 40, 56, 79, 84, 94], but these require static settings or multiple views for training. In contrast, we propose to learn a single representation for multiple identities, from just monocular videos of challenging poses.

3D Gaussian Splatting. More recently, several approaches propose techniques for faster training and rendering of radiance fields, with 3D Gaussian Splatting (3DGS) becoming very popular [72]. 3DGS achieves real-time rendering of complex static scenes with a high visual quality. It represents a scene with a set of 3D Gaussians. Each Gaussian is associated with learnable parameters, namely a position (mean), an anisotropic covariance matrix, an opacity, and a color. The 3D Gaussians are projected to 2D splats in the image space through a fast rasterization procedure. Recent works extend 3DGS to dynamic human avatars [24, 25, 72, 98], by conditioning the 3DGS representation on a human articulation prior [48]. However, these approaches only learn single-identity representations. Learning a more generic neural representation would require multiple views for training [40, 73, 97]. In this work, we extend 3DGS to multiple identities, in a monocular setting, demonstrating the benefits of such a multi-identity representation in robustly animating human avatars under novel poses.

Tensor Structures. Tensors are multi-dimensional arrays that can naturally represent data of multiple dimensions [36, 63]. They have been used in related fields to represent visual data and capture higher-order similarities or interactions [12, 17]. TensorFaces [82] is an early approach that approximates different modes of variation using a multilinear tensor decomposition. Tensor-based operations have been proposed within deep neural networks to leverage the topological structure in the data [37, 38]. Tensor decomposition extends the concept of matrix decomposition [81] to higher-order tensors for dimensionality reduction and data compression. It has been applied to neural networks to significantly reduce the number of learnable parameters and mitigate the curse of dimensionality [1, 41, 60]. A widely used tensor decomposition is the CANDECOMP/PARAFAC (CP) decomposition [8, 22, 23, 36] that decomposes a tensor into a sum of rank-one tensors. Computing the CP decomposition of a tensor requires optimization algorithms, like Alternating Least Square (ALS) [55]. In this work, we propose to construct a high-order tensor of all the learnable parameters of a 3DGS representation, and learn a CP tensor decomposition to model multiple identities in a unified network and use a reduced number of parameters. To the best of our knowledge, this is the first approach that applies such a decomposition to 3DGS and learns a multi-identity 3DGS representation.

3 Preliminary

Notation. We follow the notation introduced by Kolda and Bader [36]. We denote scalars as $w \in \mathbb{R}$, vectors as $\boldsymbol{w} \in \mathbb{R}^{I_1}$, and matrices as $\boldsymbol{W} \in \mathbb{R}^{I_1 \times I_2}$. Tensors are symbolized by calligraphic letters, *e.g.* $\boldsymbol{\mathcal{W}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_d}$. Tensor fibers $\boldsymbol{w}_{i,j,:}$ are higher-order analogues of matrix rows and columns. Tensor slices $\boldsymbol{W}_{i,:,:}$ are 2-dimensional sections of a tensor.

3.1 3D Gaussian Splatting

3D Gaussian Splatting (3DGS) [32] has been first proposed to represent static scenes, achieving fast and high-quality novel view synthesis. A scene is repre-

sented by a set of anisotropic 3D Gaussians. Each Gaussian is defined by a position (mean) $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$:

$$G(\boldsymbol{x}) = e^{(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})} , \tag{1}$$

and associated with an opacity α and a color $\boldsymbol{c}$ that are used for α-blending. The 3D Gaussians are projected to 2D splats in the image space following EWA splatting [100], where Zicker *et al.* demonstrate how the corresponding 2D covariance matrix can be derived from $\boldsymbol{\Sigma}$. The whole representation is differentiable, optimized via gradient descent. In order to ensure that $\boldsymbol{\Sigma}$ has a physical meaning, it is constrained to be positive semi-definite, by defining a parametric ellipsoid with a diagonal scaling matrix $\boldsymbol{S}$ and a rotation matrix $\boldsymbol{R}$:

$$\boldsymbol{\Sigma} = \boldsymbol{R}\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{R}^T . \tag{2}$$

In practice, the parameters of each Gaussian are stored as a position vector $\boldsymbol{\mu} \in \mathbb{R}^3$, a 3D scaling vector $\boldsymbol{s} \in \mathbb{R}^3$ and a quaternion $\boldsymbol{q} \in \mathbb{R}^4$ to represent the rotation [32]. The locations of the Gaussians are initialized by a sparse point cloud, obtained by Structure-from-Motion [76]. Then, they are optimized based on successive iterations of rendering, guided by a photometric loss and using an adaptive densification scheme [32]. The color C of a pixel is computed by blending all Gaussians g that overlap the pixel, after sorting them by depth:

$$C = \sum_g \boldsymbol{c}_g \alpha_g \prod_{j=1}^{g-1} (1 - \alpha_j) , \tag{3}$$

where $\boldsymbol{c}_g$ is the color and α_g is the learned opacity weighted by the probability density of the g-th projected 2D Gaussian splat. The color c_g can be a view-dependent color computed from learned spherical harmonic (SH) coefficients [14, 32, 57] or can be a learned RGB color vector [50, 72].

3.2 Deformable 3D Gaussian Splatting for Humans

Many recent works extend 3DGS to represent human avatars [24, 25, 72, 98], combining the 3D Gaussians with a human articulation prior. They follow a widely adopted paradigm, mapping the observation to the canonical space and learning a representation of the human body in the canonical pose [28,85,91,92]. They use linear blend skinning (LBS) to deform the human body under arbitrary poses, based on the SMPL model [49].

Given a monocular video of a human, corresponding SMPL parameters and foreground masks per frame, we learn a set of 3D Gaussians $\{G^{(g)}\}_{g=1}^{N_g}$ that represent the human body in the canonical pose. We initialize them by randomly sampling N_g points on the SMPL mesh surface under canonical pose. Following HumanNeRF [92] and 3DGS-Avatar [72], we decompose the complex human deformation to rigid and non-rigid parts, which model the skeletal motion and non-rigid cloth deformation correspondingly (see overview in Fig. 2 right).

Non-rigid Deformation. First, we learn a non-rigid deformation module f_d that maps the set of 3D Gaussians under canonical pose $\{G_c\}$ to a non-rigidly deformed space $\{G_d\}$. More specifically, given the parameters of a Gaussian in the canonical space (position $\boldsymbol{\mu}_c$, scale $\boldsymbol{s}_c$, and rotation quartenion $\boldsymbol{q}_c$) and a latent code $\boldsymbol{z}_p$, f_d outputs offsets and a feature vector $\boldsymbol{z}$ as follows:

$$(\delta\boldsymbol{\mu}, \delta\boldsymbol{s}, \delta\boldsymbol{q}, \boldsymbol{z}) = f_d(\boldsymbol{\mu}_c; \boldsymbol{z}_p) \ , \tag{4}$$

where $\boldsymbol{\mu}_d = \boldsymbol{\mu}_c + \delta\boldsymbol{\mu}$, $\boldsymbol{s}_d = \boldsymbol{s}_c \cdot e^{\delta\boldsymbol{s}}$, $\boldsymbol{q}_d = \boldsymbol{q}_c \cdot [1; \delta\boldsymbol{q}]$, and $\boldsymbol{z}_p$ is a latent code that encodes the SMPL parameters for pose and shape as the output of an hierarchical encoder [53, 72].

Rigid Transformation. The non-rigidly deformed Gaussians $\{G_d\}$ are mapped to the observation space via a rigid transformation module. The rigid transformation is based on LBS [49], where B rigid bone transformations $\{\boldsymbol{B}_b\}_{b=1}^{B}$ are linearly blended via a set of skinning weights learned by a neural skinning field represented as an MLP f_r [11, 72, 92]:

$$\boldsymbol{T} = \sum_{b=1}^{B} f_r(\boldsymbol{\mu}_d)_b \boldsymbol{B}_b \ . \tag{5}$$

Then, the position and rotation of the Gaussians in the observation space are computed as: $\boldsymbol{\mu}_o = \boldsymbol{T}\boldsymbol{\mu}_d$ and $\boldsymbol{R}_o = \boldsymbol{T}_{:3,:3}\boldsymbol{R}_d$, where $\boldsymbol{R}_d$ is the rotation matrix derived from the quartenion $\boldsymbol{q}_d$.

Color. Following 3DGS-Avatar [72], we predict the color $\boldsymbol{c}$ of a Gaussian using an MLP f_c, conditioned on a per-Gaussian learnable feature vector $\boldsymbol{f} \in \mathbb{R}^{32}$, the output $\boldsymbol{z}$ of the non-rigid network f_d, and the SH basis $\boldsymbol{\gamma}(\hat{\boldsymbol{d}})$:

$$\boldsymbol{c} = f_c(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\gamma}(\hat{\boldsymbol{d}})) \ , \tag{6}$$

where $\boldsymbol{d}$ is the viewing direction wrt the camera center and $\hat{\boldsymbol{d}} = T_{:3,:3}^{-1}\boldsymbol{d}$ is the canonicalized viewing direction to ensure consistency under the applied transformations [72]. Note that in contrast to 3DGS-Avatar, we do not learn any per-frame latent code, in order to avoid any overfitting to the training frames. Our goal is to learn a more general representation of the human body, given multiple identities.

4 Method

We present MIGS, a novel method that learns a multi-identity 3DGS representation for human avatars, based on tensor decomposition. An overview of our approach is illustrated in Fig. 2. Given only monocular videos of different subjects, we learn a unified neural representation that models their appearance and articulation. Inspired by ideas that go back to TensorFaces [82], we construct a high-order tensor of the Gaussian parameters of all our training identities. Given that different identities share common characteristics, we assume a low-rank structure of the high-order tensor. Applying a tensor decomposition [36, 63],

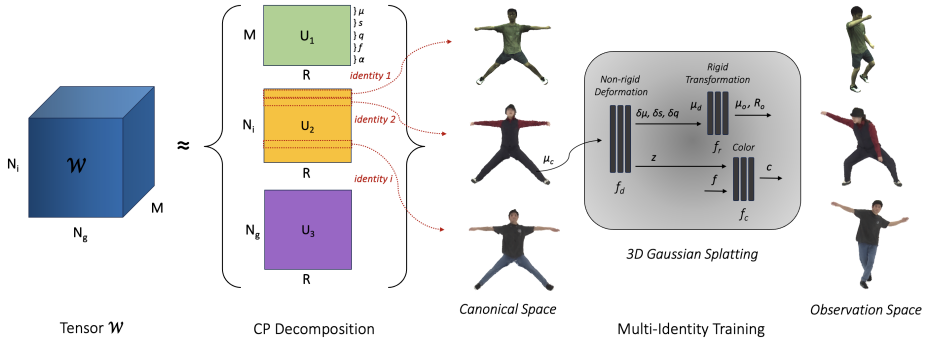


Fig. 2: Overview of MIGS. Given monocular videos of multiple identities, we learn a unified 3DGS representation for human avatars based on CP tensor decomposition. We construct a tensor $\boldsymbol{\mathcal{W}} \in \mathbb{R}^{N_i \times N_g \times M}$, where N_i is the number of identities, N_g the number of 3D Gaussians and M the number of parameters per Gaussian. In practice, we assume a low-rank structure of the tensor $\boldsymbol{\mathcal{W}}$ and thus, we only learn the matrices $\boldsymbol{U}_1 \in \mathbb{R}^{M \times R}$, $\boldsymbol{U}_2 \in \mathbb{R}^{N_i \times R}$, $\boldsymbol{U}_3 \in \mathbb{R}^{N_g \times R}$ that approximate $\boldsymbol{\mathcal{W}}$ through the CP decomposition with $R << N_g$. By leveraging information from the diverse deformations of different subjects, MIGS enables robust animation under novel challenging poses.

we decompose it into a sum of rank-one tensors. In this way, we share weights between different identities, using a minimum number of total learnable parameters for our Gaussian splatting. By learning a unified representation, MIGS enables robust animation of the training subjects under completely unseen poses. We also demonstrate how MIGS can be extended to novel identities.

4.1 Multi-Identity Representation

Previous works [24, 25, 72, 98] propose single-identity 3DGS representations for human avatars, where the parameters of the 3D Gaussians are learned from a monocular video of a single subject. Section 3.2 defines the 3DGS representation for a single identity. In this section, we demonstrate how we construct our *multi-identity* representation based on high-order tensors. Tensors are a multi-dimensional generalization of matrices [36,63] that can represent data of multiple dimensions and capture high-order similarities or interactions.

For each identity i, we assume a set of 3D Gaussians $\{G^{(i,g)}\}_{g=1}^{N_g}$ that model the human avatar under canonical pose. Each Gaussian $G^{(i,g)}$ is defined by the following learnable parameters: a position (mean) $\boldsymbol{\mu}^{(i,g)} \in \mathbb{R}^3$, a scaling vector $\boldsymbol{s}^{(i,g)} \in \mathbb{R}^3$, a quartenion $\boldsymbol{q}^{(i,g)} \in \mathbb{R}^4$, a feature vector $\boldsymbol{f}^{(i,g)} \in \mathbb{R}^{32}$ and an opacity $\alpha^{(i,g)} \in \mathbb{R}$. The color $\boldsymbol{c}^{(i,g)}$ of the Gaussian is given by the output of a color MLP as in Eq. (6). Given videos of N_i identities, we construct a tensor:

$$\boldsymbol{\mathcal{W}} \in \mathbb{R}^{N_i \times N_g \times M}, \text{where } \boldsymbol{w}_{i,g,:} = \left[\boldsymbol{\mu}^{(i,g)}; \boldsymbol{s}^{(i,g)}; \boldsymbol{q}^{(i,g)}; \boldsymbol{f}^{(i,g)}; \alpha^{(i,g)}\right] , \quad (7)$$

where M derives by adding the dimensions of the parameters for a Gaussian g and an identity i: $M = 3 + 3 + 4 + 32 + 1 = 43$.

4.2 Tensor Decomposition

Increasing the number of identities leads to a prohibitive number of learnable parameters in the tensor $\boldsymbol{\mathcal{W}}$. We assume a low-rank structure of $\boldsymbol{\mathcal{W}}$ and apply CANDECOMP/PARAFAC (CP) tensor decomposition [23, 36] to decompose it into a sum of a finite number of rank-one tensors. More specifically, we first matricize the tensor $\boldsymbol{\mathcal{W}}$ using the unfolding in the second dimension, such that $\boldsymbol{W}_{(2)} \in \mathbb{R}^{N_g \times (N_i M)}$. Then, we set:

$$\boldsymbol{W}_{(2)} \approx \boldsymbol{U}_3 \left(\boldsymbol{U}_2 \odot \boldsymbol{U}_1\right)^T \ , \tag{8}$$

where $\odot$ denotes the Khatri-Rao product [33], $\boldsymbol{U}_1 \in \mathbb{R}^{M \times R}$, $\boldsymbol{U}_2 \in \mathbb{R}^{N_i \times R}$, $\boldsymbol{U}_3 \in \mathbb{R}^{N_g \times R}$ are the factor matrices, and R is a positive integer that corresponds to the selected rank of the decomposition. In an exact CP decomposition, where there is equality in Eq. (8), the smallest number of rank-one tensors that can generate $\boldsymbol{\mathcal{W}}$ as their sum is equal with the rank of the tensor $\boldsymbol{\mathcal{W}}$ [36].

In practice, we only learn the matrices $\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3$. Thus, we learn $(M + N_i + N_g)R$ parameters, instead of MN_iN_g parameters, which leads to a significant decrease for a large number of identities N_i, when $R << N_g$. In our experiments, $R = 100$ and $N_g = 5 \times 10^4$. For $N_i = 30$ identities, MIGS learns only $(M + N_i + N_g)R \approx 5 \times 10^6$ parameters, compared to $MN_iN_g \approx 6.5 \times 10^7$ that would be required by single-identity 3DGS representations, leading to a decrease *by at least one order of magnitude* in the total number of learnable parameters. During each iteration, we use Eq. (8) to derive $\boldsymbol{W}_{(2)}$ and then rewrite it as a tensor $\boldsymbol{\mathcal{W}}$ (see Eq. (9)). Each fiber $\boldsymbol{w}_{i,g,:}$ corresponds to the parameters of a 3D Gaussian $G^{(i,g)}$ for an identity $i = 1, \ldots, N_i$ and $g = 1, \ldots, N_g$.

4.3 Optimization

Initialization. We initialize $\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3$ by using the parameters of only one identity. More specifically, for the identity $i = 1$, we randomly sample N_g points on the SMPL mesh surface under canonical pose to initialize the locations $\boldsymbol{\mu}^{(1,g)}$ for $g = 1, \ldots, N_g$. We follow [32] to initialize the rest of the parameters $\boldsymbol{s}^{(1,g)}$, $\boldsymbol{q}^{(1,g)}$, $\boldsymbol{f}^{(1,g)}$, $\alpha^{(1,g)}$ and we construct the tensor slice $\boldsymbol{W}_{1,:,:}$. We then compute the CP decomposition of $\boldsymbol{W}_{1,:,:}$ by applying Robust Tensor Power Iteration (CP-Power) using TensorLy [39] [3]. This gives us the initial matrices $\boldsymbol{U}_1$, $\boldsymbol{U}_3$, and the first row of $\boldsymbol{U}_2$ that we populate to all the rows to get our initial $\boldsymbol{U}_2$. We absorb the output weights of CPPower in $\boldsymbol{U}_3$.

Training. We follow the 3DGS implementation for human avatars as described in Sec. 3.2. We jointly optimize $\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3$, as well as the parameters of f_d, f_r, and f_c, by iteratively rendering frames from each training identity. Given the estimated $\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3$ at each iteration, we derive the parameters of the 3D Gaussians for a specific identity, as described in Sec. 4.2. We use the same loss function as 3DGS-Avatar [72] that consists of a photometric RGB loss, a mask

[3] `https://tensorly.org/stable/modules/generated/tensorly.decomposition.CPPower.html`

loss, a skinning weight regularization, and an as-isometric-as-possible regularization loss (see supplementary document for more implementation details).
Per-identity Updates. Our proposed approach provides the option to update the parameters for only a specific identity i at a given iteration, by using only the row i of $\boldsymbol{U}_2$ in Eq. (8). In our experiments, we observe similar results in both cases, *i.e.* using either the full $\boldsymbol{U}_2$ at each iteration or the corresponding row of $\boldsymbol{U}_2$ of the current identity i. However, we note that individual per-identity updates can be useful in other cases.

4.4 Personalization

Trained on multiple identities simultaneously, MIGS learns a large variety of rigid and non-rigid deformations. To enhance the output visual quality for a particular subject, we can further fine-tune our color MLP f_c (see Eq. (6)), keeping the rest of the parameters frozen, using a short video of that subject. We observe that this "personalization" procedure is useful when we train our network for a large number of identities. Increasing the number of identities leads to an increase in robustness under novel poses, but a smoother result, where our generic network might not capture high-frequency details. This is expected, since for a given number of parameters, we learn a single unified representation of a large number of identities with diverse characteristics. We address this with the personalization that, with only a few iterations, successfully captures identity-specific details.
Novel Identity. Learning a novel identity (that is not part of the initial training set) is also possible, by just adding an additional row to $\boldsymbol{U}_2$. Given a short video of the new identity, we optimize only the corresponding row of $\boldsymbol{U}_2$ and the color MLP f_c, keeping the rest of our representation frozen. In this way, we learn the characteristics of the new identity, without forgetting the large variety of the human body deformations captured by our multi-identity representation.

5 Experiments

5.1 Datasets

To evaluate our method, we use 2 datasets, namely ZJU-MoCap [69] and AIST++ [43, 80]. ZJU-MoCap is a widely used dataset that contains multi-view recordings of human subjects, performing motions such as kicking and arm swings. In our experiments, we assume a monocular setting and only use the camera 1 for training. Similarly with other works [56, 72, 92, 95], we use 6 subjects (377, 386, 387, 392, 393, 394) and the corresponding camera parameters, SMPL [48] body poses, and segmentation masks provided by the dataset.

AIST++ [43, 80] is a dance motion dataset that includes a large variety of human body movements. In contrast to ZJU-MoCap that consists of mostly repetitive sequences with limited variation, AIST++ contains more diverse and challenging human body articulations. It contains 30 subjects and 10 dance genres. Each subject performs a specific number of *basic* and *advanced* dance

Fig. 3: Animation of human avatars under novel out-of-distribution poses. Qualitative comparison with state-of-the-art approaches, namely HumanNeRF [92], MonoHuman [95], GauHuman [25], and 3DGS-Avatar [72]. The training subjects are from the ZJU-MoCap dataset [69] and the target poses (column 1) are from the AIST++ dataset [43, 80]. Since the target poses are out of the training distribution (first 3 come from the advanced dance videos of AIST++), animation under these poses is challenging. Our method demonstrates significant robustness.

choreographies. We randomly select 1 basic (∼12 sec.) and 1 advanced (∼30 sec.) video per subject. We assume a monocular setting by only using the camera 1 for training. To get more accurate and smooth tracking over time, we use 4DHumans [19] to get the corresponding SMPL annotations, camera parameters, and segmentation masks for this dataset. We construct 2 test sets: (a) a "Basic Test" set that consists of 10 basic dance videos, and (b) an "Advanced Test" set of 10 advanced dance videos, specifically chosen to evaluate the robustness of our method to novel and more challenging (out-of-distribution) poses.

5.2 Evaluation

Baselines. We compare our method against recent approaches for animatable human avatars, learned from monocular videos. We include both NeRF-based

Table 1: Quantitative evaluation on ZJU-MoCap. We compare our method with state-of-the-art approaches (HumanNeRF [92], MonoHuman [95], GauHuman [25], 3DGS-Avatar [72]) on novel view synthesis, using the standard test set of ZJU-MoCap. We report PSNR, SSIM and LPIPS* = LPIPS $\times 10^3$ on 4 subjects (377, 386, 392, 394).

	377			**386**			**392**			**394**		
Method	PSNR↑	SSIM↑	LPIPS*↓	PSNR↑	SSIM↑	LPIPS*↓	PSNR↑	SSIM↑	LPIPS*↓	PSNR↑	SSIM↑	LPIPS*↓
HumanNeRF	30.41	0.9743	24.06	33.20	0.9752	28.99	31.04	0.9705	32.12	30.31	0.9642	32.89
MonoHuman	29.12	0.9727	26.58	32.94	0.9695	36.04	29.50	0.9635	39.45	29.15	0.9595	38.08
GauHuman	29.48	0.9667	22.40	29.51	0.9648	28.80	27.86	0.9578	32.45	28.37	0.9533	34.16
3DGS-Avatar	30.64	0.9774	20.88	33.63	0.9773	25.77	31.66	0.9730	30.14	30.54	0.9661	31.21
Ours	**32.85**	**0.9847**	**19.67**	**34.98**	**0.9775**	**24.56**	**33.88**	**0.9805**	**25.73**	**32.28**	**0.9761**	**25.48**

(HumanNeRF [92] and MonoHuman [95]) and 3DGS-based (GauHuman [25] and 3DGS-Avatar [72]) methods. HumanNeRF [92] is a well-performing NeRF-based approach that achieves high-quality reconstruction of a human subject, given a monocular video. MonoHuman [95] introduces a bi-directional deformation field to improve rendering of human subjects under novel poses. It requires keyframe selection of frontal and back views from a monocular video. GauHuman [25] and 3DGS-Avatar [72] are very recent approaches that apply 3DGS for humans, achieving fast optimization and real-time rendering. In our preliminary experiments, we observe that 3DGS-Avatar achieves higher-quality results than GauHuman, and thus we build our proposed multi-identity approach upon the single-identity 3DGS-Avatar, but omitting any per-frame latent codes to avoid any overfitting to the training frames (see Sec. 3.2). We evaluate the performance of all the methods qualitatively and quantitatively when rendering a subject under novel poses and novel views.

Evaluation Metrics. We measure the visual quality of the generated videos, using standard reconstruction metrics, namely peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) [89], and learned perceptual image patch similarity (LPIPS) [96]. Following related works, we report LPIPS* = LPIPS $\times 10^3$.

Results on the ZJU-MoCap dataset. In our first evaluation setting, we train a model using the 6 subjects from ZJU-MoCap dataset [69]. Tab. 1 shows the quantitative comparisons on novel view synthesis on the standard test set. We train all methods on the same training set (camera view 1) and test on the rest 22 camera views for each subject. For fair comparison, we do the same for GauHuman and report the corresponding values (in their original code they used the refined dataset of ZJU-MoCap). See suppl. for more comparisons.

We are mostly interested in the benefit of our multi-identity representation in animating the training subjects under novel poses. Figure 10 demonstrates the effectiveness of MIGS in such a setting. More specifically, we test our model on target poses from the AIST++ dataset [43, 80] (column 1 in Fig. 10). These target poses are completely unseen and out of the training distribution, since in this case our model is trained on the 6 subjects from ZJU-MoCap. As mentioned in Sec. 5.1, ZJU-MoCap includes more repetitive and common human motion compared to the dance sequences of AIST++. In addition, some of the target

Fig. 4: Animation of human avatars under novel poses. Our method robustly animates all the identities under challenging unseen poses, from unseen camera views and advanced dance videos, outperforming the other methods, HumanNeRF [92], MonoHuman [95], and 3DGS-Avatar [72]. The subjects are from AIST++ [43, 80].

poses in Fig. 10 are from the advanced dance videos (first 3 rows). Thus, animating the subjects under those novel poses is more challenging. Artifacts under the arms and legs are visible in all other state-of-the-art methods, HumanNeRF [92], MonoHuman [95], GauHuman [25], and 3DGS-Avatar [72]. Our proposed method demonstrates significant robustness, outperforming all the methods.

Results on the AIST++ dataset. In our second evaluation setting, we train a model using all the 30 subjects of the AIST++ dataset [43,80]. Our training set includes both basic and advanced dance motion videos. We evaluate our model on both our "Basic Test" and "Advanced Test" sets. Our quantitative results are shown in Tab. 2. In both settings, MIGS outperforms the existing approaches.

Table 2: Quantitative evaluation on AIST++. We construct 2 test sets, a "Basic Test" and an "Advanced Test" set that consists of more advanced dance videos, specifically chosen to evaluate the robustness of our method to novel poses. We compare with state-of-the-art approaches (HumanNeRF [92], MonoHuman [95], 3DGS-Avatar [72]) and report PSNR, SSIM, and LPIPS* = LPIPS $\times 10^3$. Our method significantly outperforms the other methods.

	Basic Test			Advanced Test		
Method	PSNR↑	SSIM↑	LPIPS*↓	PSNR↑	SSIM↑	LPIPS*↓
HumanNeRF [92]	24.58	0.9732	29.20	22.01	0.9654	39.01
MonoHuman [95]	21.02	0.8302	42.63	18.89	0.7540	85.50
3DGS-Avatar [72]	28.89	0.9846	18.20	25.51	0.9704	28.86
Ours	**29.82**	**0.9865**	**17.73**	**26.54**	**0.9741**	**26.02**

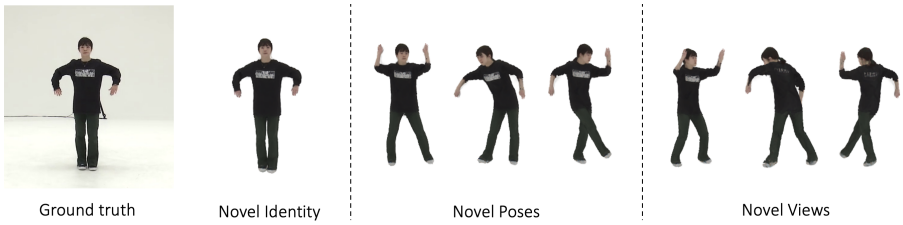


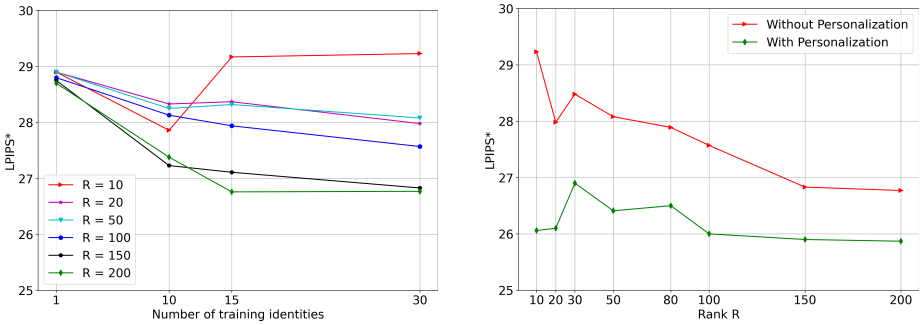
Fig. 5: Learning a novel identity. Our method can be extended to learn a novel identity, and then animate it under novel poses and novel views.

Figure 4 demonstrates the corresponding qualitative results, comparing with ground truth. The target poses (column 1) are unseen during training and come from unseen camera views. Poses in rows 2-4 are from challenging, more advanced dance combinations, and thus, are out of the training distribution. During training, the model has only seen more basic movements for those subjects. HumanNeRF [92] and MonoHuman [95] present significant artifacts in this case. Especially MonoHuman requires selection of frontal and back views, which might not be available in the monocular setting in the basic dance videos, leading to even wrong color (see first 2 rows in Fig. 4). 3DGS-Avatar [72] is more robust but also produces artifacts in case of unseen limb articulations. Our method demonstrates robustness in animating all identities under arbitrary poses.

Figure 5 shows the results of our method when learning a novel identity (see Sec. 4.4), animating it under novel poses and rendering it under novel views. In this case, we learn an additional row in $\boldsymbol{U}_2$, given a short video of the new identity (10 sec.) with only basic movements. We optimize only the corresponding row of $\boldsymbol{U}_2$ and the color MLP f_c, keeping the rest of our representation frozen.

5.3 Ablation Study

We conduct an ablation study on the rank R of our tensor decomposition. Figure 6a shows the average visual quality on the "Advanced Test" set for different number of training identities and for different values of R (without any person-



(a) LPIPS* vs number of training identities **(b)** LPIPS* vs rank R

Fig. 6: Ablation study. Visual quality (LPIPS* = LPIPS $\times 10^3$) on the "Advanced Test" set for different values R of our tensor decomposition. **(a)** LPIPS* (lower the better) vs the number of training identities (without any personalization). **(b)** LPIPS* vs rank R for the model trained on 30 identities, with or without personalization. We observe that $R = 10$ is not enough to capture a larger number of identities with diverse articulations, while $R = 100$ is sufficient.

alization). We observe that $R = 10$ is not enough to capture a larger number of identities. For larger values of R, we notice that increasing the number of identities leads to (a) an increase in robustness under novel poses (lower LPIPS), since our model leverages information from more subjects with diverse human body articulation, and (b) a smoother result, where our generic network might not capture high-frequency details. We address the described smoothing with our personalization procedure that, with only a few iterations, successfully captures such identity-specific details (see Fig. 6b). With personalization, we observe similar results for $R = 100$ and $R = 200$. We conclude that with just $R = 100$ for our learned tensor decomposition, we achieve robust animation under challenging novel poses. Further ablation studies and discussion about limitations and ethical considerations are included in the suppl. document.

6 Conclusion

In conclusion, we introduce MIGS, a novel method that learns a unified 3DGS representation for multiple identities. We propose to construct a high-order tensor that combines all the learnable parameters for multiple human avatars. We approximate this tensor with a tensor decomposition, significantly reducing the total number of learnable parameters. We quantitatively and qualitatively evaluate our method using a standard benchmark dataset and a more challenging dance motion dataset. By leveraging information from multiple identities, MIGS enables robust animation of any training subject under novel challenging poses, out of the training distribution. In the future, we envision extending our approach to learn a generic model from thousands of identities.

Acknowledgements. This work was partially supported by Amazon Prime Video. GC acknowledges travel support from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951847.

References

1. Astrid, M., Lee, S.I.: Cp-decomposition with tensor power method for convolutional neural networks compression. In: 2017 IEEE International Conference on Big Data and Smart Computing (BigComp). pp. 115–118. IEEE (2017)
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields (2021)
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. CVPR (2022)
4. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Springer International Publishing (Oct 2016)
5. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
6. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)
7. Carranza, J., Theobalt, C., Magnor, M.A., Seidel, H.P.: Free-viewpoint video of human actors. ACM transactions on graphics (TOG) **22**(3), 569–577 (2003)
8. Carroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. Psychometrika **35**(3), 283–319 (1970)
9. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5933–5942 (2019)
10. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. arXiv preprint arXiv:2103.15595 (2021)
11. Chen, X., Zheng, Y., Black, M.J., Hilliges, O., Geiger, A.: Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11594–11604 (2021)
12. Chrysos, G., Georgopoulos, M., Panagakis, Y.: Conditional generation using polynomial expansions. Advances in Neural Information Processing Systems **34**, 28390–28404 (2021)
13. De Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video. In: ACM SIGGRAPH 2008 papers, pp. 1–10 (2008)
14. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5501–5510 (2022)
15. Gafni, G., Thies, J., Zollhöfer, M., Nießner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction (2020)
16. Gao, X., Yang, J., Kim, J., Peng, S., Liu, Z., Tong, X.: MPS-NeRF: Generalizable 3D Human Rendering From Multiview Images. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
17. Georgopoulos, M., Chrysos, G., Pantic, M., Panagakis, Y.: Multilinear latent conditioning for generating unseen attribute combinations. In: International Conference on Machine Learning. pp. 3442–3451. PMLR (2020)

18. Gleicher, M.: Retargetting motion to new characters. In: Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques. p. 33–42. SIGGRAPH '98, Association for Computing Machinery, New York, NY, USA (1998). https://doi.org/10.1145/280814.280820, https://doi.org/10.1145/280814.280820
19. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa*, A., Malik*, J.: Humans in 4D: Reconstructing and tracking humans with transformers. In: International Conference on Computer Vision (ICCV) (2023)
20. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
21. Guo, Y., Chen, K., Liang, S., Liu, Y.J., Bao, H., Zhang, J.: AD-NeRF: Audio driven neural radiance fields for talking head synthesis. In: IEEE International Conference on Computer Vision (ICCV). pp. 5784–5794 (2021)
22. Harshman, R.A., et al.: Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis (1970)
23. Hitchcock, F.L.: The expression of a tensor or a polyadic as a sum of products. J. of Math. and Phys. **6**(1-4), 164–189 (1927). https://doi.org/10.1002/sapm192761164
24. Hu, L., Zhang, H., Zhang, Y., Zhou, B., Liu, B., Zhang, S., Nie, L.: Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. arXiv preprint arXiv:2312.02134 (2023)
25. Hu, S., Liu, Z.: Gauhuman: Articulated gaussian splatting from monocular human videos. arXiv preprint arXiv:2312.02973 (2023)
26. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CVPR (2017)
27. Jayasundara, V., Agrawal, A., Heron, N., Shrivastava, A., Davis, L.S.: Flexnerf: Photorealistic free-viewpoint rendering of moving humans from sparse views. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
28. Jiang, W., Yi, K.M., Samei, G., Tuzel, O., Ranjan, A.: Neuman: Neural human radiance field from a single video. In: European Conference on Computer Vision. pp. 402–418. Springer (2022)
29. Kanade, T., Rander, P., Narayanan, P.: Constructing virtual worlds from real scenes. In: ACM Multimedia. vol. 1 (1997)
30. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Recognition (CVPR) (2018)
31. Kania, K., Yi, K.M., Kowalski, M., Trzciński, T., Tagliasacchi, A.: CoNeRF: Controllable Neural Radiance Fields. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
32. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
33. Khatri, C., Rao, C.R.: Solutions to some functional equations and their applications to characterization of probability distributions. Sankhyā: The Indian Journal of Statistics, Series A pp. 167–180 (1968)
34. Kim, J., Wee, D., Xu, D.: You only train once: Multi-identity free-viewpoint neural human rendering from monocular videos (2023), https://arxiv.org/abs/2303.05835
35. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

36. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM review **51**(3), 455–500 (2009)
37. Kossaifi, J., Khanna, A., Lipton, Z., Furlanello, T., Anandkumar, A.: Tensor contraction layers for parsimonious deep nets. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 26–32 (2017)
38. Kossaifi, J., Lipton, Z.C., Kolbeinsson, A., Khanna, A., Furlanello, T., Anandkumar, A.: Tensor regression networks. Journal of Machine Learning Research **21**(123), 1–21 (2020)
39. Kossaifi, J., Panagakis, Y., Anandkumar, A., Pantic, M.: Tensorly: Tensor learning in python. arXiv preprint arXiv:1610.09555 (2016)
40. Kwon, Y., Kim, D., Ceylan, D., Fuchs, H.: Neural human performer: Learning generalizable radiance fields for human performance rendering. Advances in Neural Information Processing Systems **34**, 24741–24752 (2021)
41. Lebedev, V., Ganin, Y., Rakhuba, M., Oseledets, I., Lempitsky, V.: Speeding-up convolutional neural networks using fine-tuned cp-decomposition. arXiv preprint arXiv:1412.6553 (2014)
42. Lee, J., Shin, S.Y.: A hierarchical approach to interactive motion editing for human-like figures. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. p. 39–48. SIGGRAPH '99, ACM Press/Addison-Wesley Publishing Co., USA (1999). `https://doi.org/10.1145/311535.311539`, `https://doi.org/10.1145/311535.311539`
43. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Learn to dance with aist++: Music conditioned 3d dance generation (2021)
44. Li, T., Slavcheva, M., Zollhoefer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Newcombe, R., et al.: Neural 3d video synthesis from multi-view video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5521–5531 (2022)
45. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6498–6508 (2021)
46. Lindell, D.B., Van Veen, D., Park, J.J., Wetzstein, G.: Bacon: Band-limited coordinate networks for multiscale scene representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16252–16262 (2022)
47. Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: Neural free-view synthesis of human actors with pose control. ACM Trans. Graph.(ACM SIGGRAPH Asia) (2021)
48. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
49. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023)
50. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In: 3DV (2024)
51. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision. pp. 5442–5451 (Oct 2019)
52. Matusik, W., Buehler, C., Raskar, R., Gortler, S.J., McMillan, L.: Image-based visual hulls. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. pp. 369–374 (2000)

53. Mihajlovic, M., Zhang, Y., Black, M.J., Tang, S.: LEAP: Learning articulated occupancy of people. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (Jun 2021)
54. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
55. Mohlenkamp, M.J.: Musings on multilinear fitting. Linear Algebra and its Applications **438**(2), 834–852 (2013)
56. Mu, J., Sang, S., Vasconcelos, N., Wang, X.: ActorsNeRF: animatable few-shot human rendering with generalizable nerfs pp. 18391–18401 (2023)
57. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)
58. Neverova, N., Guler, R.A., Kokkinos, I.: Dense pose transfer. In: Proceedings of the European conference on computer vision (ECCV). pp. 123–138 (2018)
59. Noguchi, A., Sun, X., Lin, S., Harada, T.: Neural Articulated Radiance Field. In: International Conference on Computer Vision (ICCV) (2021)
60. Novikov, A., Podoprikhin, D., Osokin, A., Vetrov, D.P.: Tensorizing neural networks. Advances in neural information processing systems **28** (2015)
61. Osman, A.A.A., Bolkart, T., Black, M.J.: STAR: A sparse trained articulated human body regressor. In: European Conference on Computer Vision (ECCV). pp. 598–613 (2020), `https://star.is.tue.mpg.de`
62. Pan, X., Yang, Z., Ma, J., Zhou, C., Yang, Y.: Transhuman: A transformer-based human representation for generalizable neural human rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3544–3555 (October 2023)
63. Panagakis, Y., Kossaifi, J., Chrysos, G.G., Oldfield, J., Nicolaou, M.A., Anandkumar, A., Zafeiriou, S.: Tensor methods in computer vision and deep learning. Proceedings of the IEEE **109**(5), 863–890 (2021)
64. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. ICCV (2021)
65. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. arXiv preprint arXiv:2106.13228 (2021)
66. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
67. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019)
68. Peng, S., Geng, C., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Zhou, X., Bao, H.: Implicit Neural Representations with Structured Latent Codes for Human Body Modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
69. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: CVPR (2021)

70. Prokudin, S., Black, M.J., Romero, J.: Smplpix: Neural avatars from 3d human models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1810–1819 (January 2021)
71. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10318–10327 (2021)
72. Qian, Z., Wang, S., Mihajlovic, M., Geiger, A., Tang, S.: 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. arXiv preprint arXiv:2312.09228 (2023)
73. Raj, A., Zollhofer, M., Simon, T., Saragih, J., Saito, S., Hays, J., Lombardi, S.: Pixel-aligned volumetric avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11733–11742 (2021)
74. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Malik, J.: Tracking people by predicting 3d appearance, location and pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2740–2749 (2022)
75. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6) (Nov 2017)
76. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
77. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: CVPR (2017)
78. Su, S.Y., Yu, F., Zollhöfer, M., Rhodin, H.: A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. Advances in Neural Information Processing Systems **34**, 12278–12291 (2021)
79. Trevithick, A., Yang, B.: Grf: Learning a general radiance field for 3d representation and rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15182–15192 (2021)
80. Tsuchida, S., Fukayama, S., Hamasaki, M., Goto, M.: Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In: Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019. Delft, Netherlands (Nov 2019)
81. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of cognitive neuroscience **3**(1), 71–86 (1991)
82. Vasilescu, M.A.O., Terzopoulos, D.: Multilinear analysis of image ensembles: Tensorfaces. ECCV (1) **2350**, 447–460 (2002)
83. Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. In: Acm Siggraph 2008 papers, pp. 1–9 (2008)
84. Wang, Q., Wang, Z., Genova, K., Srinivasan, P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: CVPR (2021)
85. Wang, S., Schwarz, K., Geiger, A., Tang, S.: Arah: Animatable volume rendering of articulated human sdfs. In: European conference on computer vision. pp. 1–19. Springer (2022)
86. Wang, T., Sarafianos, N., Yang, M.H., Tung, T.: Neural rendering of humans in novel view and pose from monocular video. arXiv preprint arXiv:2204.01218 (2022)
87. Wang, T.C., Liu, M.Y., Zhu, J.Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. arXiv preprint arXiv:1808.06601 (2018)

88. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
89. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
90. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: CVPR (2016)
91. Weng, C.Y., Curless, B., Kemelmacher-Shlizerman, I.: Vid2actor: Free-viewpoint animatable person synthesis from video in the wild. arXiv preprint arXiv:2012.12884 (2020)
92. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: Humannerf: Free-viewpoint rendering of moving people from monocular video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition. pp. 16210–16220 (2022)
93. Xu, H., Alldieck, T., Sminchisescu, C.: H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. Advances in Neural Information Processing Systems **34**, 14955–14966 (2021)
94. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelNeRF: Neural radiance fields from one or few images. In: CVPR (2021)
95. Yu, Z., Cheng, W., Liu, X., Wu, W., Lin, K.Y.: Monohuman: Animatable human neural field from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16943–16953 (2023)
96. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
97. Zheng, S., Zhou, B., Shao, R., Liu, B., Zhang, S., Nie, L., Liu, Y.: Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. arXiv (2023)
98. Zielonka, W., Bagautdinov, T., Saito, S., Zollhöfer, M., Thies, J., Romero, J.: Drivable 3d gaussian avatars (2023)
99. Zielonka, W., Bolkart, T., Thies, J.: Instant volumetric head avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4574–4584 (2023)
100. Zwicker, M., Pfister, H., Van Baar, J., Gross, M.: Ewa volume splatting. In: Proceedings Visualization, 2001. VIS'01. pp. 29–538. IEEE (2001)

Appendix

The appendix is organized as follows:

- Additional Ablation Study in Sec. A.
- Additional Results in Sec. B.
- Implementation Details in Sec. C.
- Limitations in Sec. D.
- Ethical Considerations in Sec. E.

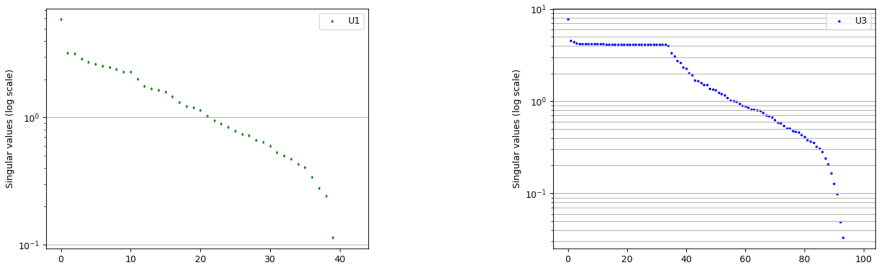
We strongly encourage the readers to watch our supplementary video on our project page: `https://aggelinacha.github.io/MIGS/`.

A Additional Ablation Study

Learned Matrices. In Fig. 7, we plot the singular values of our learned matrices $\boldsymbol{U}_1 \in \mathbb{R}^{M \times R}$ and $\boldsymbol{U}_3 \in \mathbb{R}^{N_g \times R}$, where $M = 43$ is the number of parameters per Gaussian, $N_g = 5 \times 10^4$ is the number of the 3D Gaussians, and $R = 100$ is the rank of our CP tensor decomposition (see Sec. 4.2 of the main paper). We use SVD from PyTorch [66] [4] and plot them in logarithmic scale. These matrices correspond to the model trained on the 30 subjects from the AIST++ dataset [43, 80].

Personalization. As mentioned in Sec. 4.4 of the main paper, we can personalize our generic model for a particular subject, in order to capture individual details (*e.g.* face or cloth details). We show an example output of our generic model trained on 30 identities and the corresponding personalized result in Fig. 8a. We notice that this personalization procedure is needed only for models trained with identities more than 20. We only fine-tune the color MLP for 5×10^3 iterations, keeping the rest of the parameters frozen, using a short

[4] https://pytorch.org/docs/stable/generated/torch.svd.html



(a) Singular values of $\boldsymbol{U}_1$ **(b)** Singular values of $\boldsymbol{U}_3$

Fig. 7: Singular values. The singular values of our learned $\boldsymbol{U}_1 \in \mathbb{R}^{M \times R}$ and $\boldsymbol{U}_3 \in \mathbb{R}^{N_g \times R}$, where $M = 43$ is the number of parameters per Gaussian, $N_g = 5 \times 10^4$ is the number of 3D Gaussians, and $R = 100$ is the rank of our CP tensor decomposition.

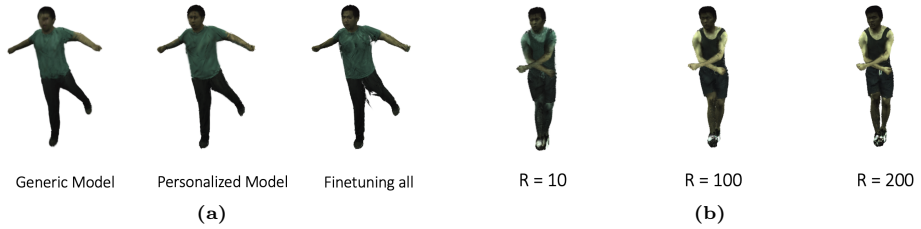
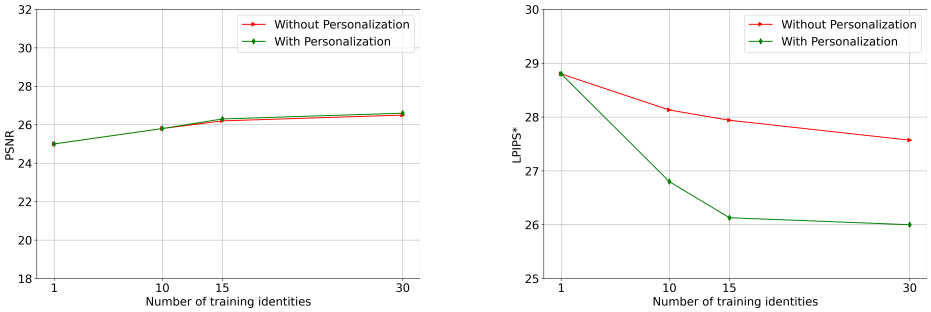


Fig. 8: Ablation study. (a) Ablation study on our proposed personalization procedure. In order to capture individual details, we fine-tune our color MLP of our generic model (left), leading to the personalized model for a particular subject (middle). We do not fine-tune all parameters (see artifacts on the right). **(b)** Ablation study on the rank R of our tensor decomposition. $R = 10$ leads to a mixture of identities (notice different shirt colors), while $R = 100$ is sufficient to capture the training identities.



(a) PSNR vs number of training identities **(b)** LPIPS* vs number of training identities

Fig. 9: Ablation study. Visual quality (PSNR and LPIPS*) on the "Advanced Test" set, with or without personalization, for different number of training identities and $R = 100$.

video of the target subject. We do not fine-tune all parameters, since in this case our network would forget the large variety of human body deformations learned from multiple subjects, leading to artifacts in novel poses (see Fig. 8a right).

Ablation Study on the Rank R. In Fig. 8b, we show a qualitative comparison for different values R of our tensor decomposition (see also Sec. 5.3). We observe that $R = 10$ is not enough to capture a larger number of identities. For example, it can lead to a mixture of colors (notice the shirt colors for $R = 10$). On the other hand, $R = 100$ seems sufficient to capture all the training identities.

Ablation Study on the Number of Identities. Figure 9 shows the visual quality for different number of training identities, with and without personalization. As also mentioned in the main paper (see also Fig. 6), increasing the number of identities leads to an increase in robustness under novel poses. Further personalization captures identity-specific details, enhancing the output visual quality.

Fig. 10: Animation of human avatars under novel poses. Qualitative comparison with state-of-the-art approaches, namely HumanNeRF [92], MonoHuman [95], GauHuman [25], and 3DGS-Avatar [72]. The training subjects and the target poses are from the ZJU-MoCap dataset [69]. Our method demonstrates significant robustness.

B Additional Results

Figure 10 demonstrates additional qualitative results for our model trained on subjects from the ZJU-MoCap dataset [69]. Animating them under novel poses leads to artifacts under the arms and legs in all other methods, namely HumanNeRF [92], MonoHuman [95], GauHuman [25], and 3DGS-Avatar [72]. Our proposed method demonstrates significant robustness.

Figure 14 demonstrates additional qualitative results when the target poses are from the AIST++ dataset [43, 80]. In this case, the poses are more challenging, completely unseen during training (out of the training distribution). Again, MIGS outperforms the other methods, robustly animating the identities under novel poses.

Figure 11 shows qualitative comparisons for novel view synthesis on the test set of the ZJU-MoCap dataset [69]. Corresponding quantitative results are shown in Table 1 of the main paper and Tab. 3. Our method demonstrates comparable performance with 3DGS-Avatar on novel view synthesis, while trained on multiple identities simultaneously.

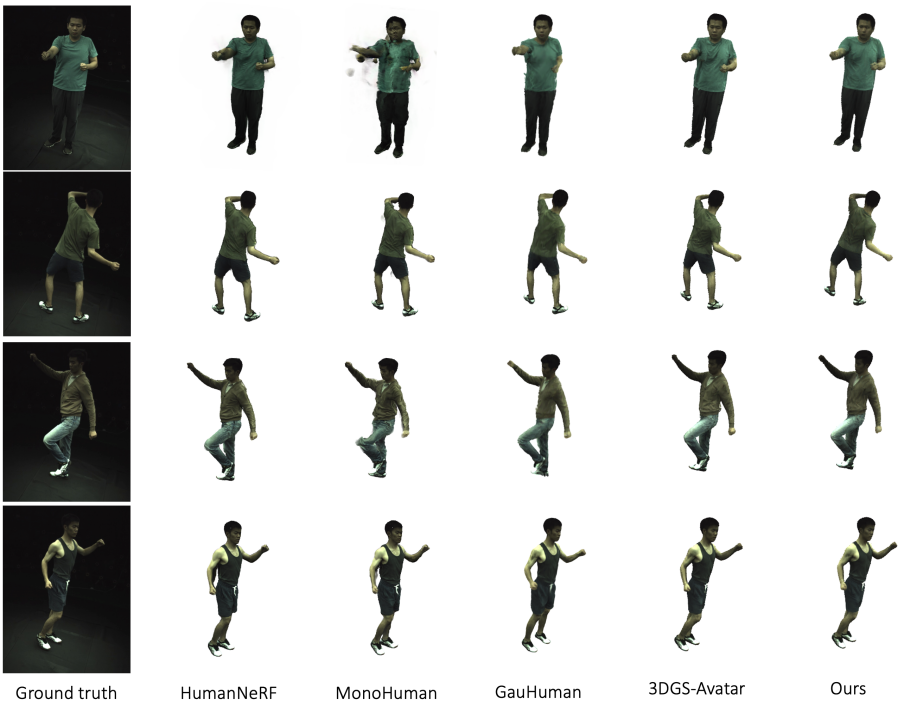


Fig. 11: Novel view synthesis on ZJU-MoCap. Qualitative comparison with state-of-the-art approaches, namely HumanNeRF [92], MonoHuman [95], GauHuman [25], and 3DGS-Avatar [72] on novel view synthesis on the test set of the ZJU-MoCap dataset [69].

Figure 12 shows qualitative comparisons with approaches that use multiple views as input. These works address the problem in a different way. They extract features from nearby views and infer a novel view in a feed-forward manner. Most of them cannot animate humans under novel poses, *e.g.* see results of NHP [45] and TransHuman [62] (we use black background to be consistent with their pretrained models - compare with row 2 in Fig. 10). ActorsNeRF [56] does not have code available. GPS-Gaussian [97] does not use body pose (SMPL parameters) as input at all, and thus cannot render novel poses.

C Implementation Details

In this section, we include implementation details of our proposed method (see also Tab. 4). Our implementation is based on PyTorch [66]. We built our architecture upon 3DGS-Avatar [72], but with some important modifications in order to learn multiple identities (see also Sec. 3.2 of the main paper). We use $N_g = 5{\times}10^4$ 3D Gaussians, that we initialize by randomly sampling N_g points on

Table 3: Quantitative evaluation on ZJU-MoCap. We compare our method with state-of-the-art approaches (HumanNeRF [92], MonoHuman [95], GauHuman [25], 3DGS-Avatar [72]) on novel view synthesis, using the standard test set of ZJU-MoCap. We report PSNR, SSIM, and LPIPS* = LPIPS $\times 10^3$ on 2 subjects (387, 393). See Table 1 in the main paper for the other 4 subjects (377, 386, 392, 394).

	387			**393**		
Method	PSNR↑	SSIM↑	LPIPS*↓	PSNR↑	SSIM↑	LPIPS*↓
HumanNeRF	28.18	0.9632	35.58	28.31	0.9603	36.72
MonoHuman	27.93	0.9601	41.76	27.64	0.9566	43.17
GauHuman	27.95	0.9608	40.70	27.88	0.9578	43.01
3DGS-Avatar	28.33	0.9642	**34.24**	28.88	0.9635	35.26
Ours	**30.70**	**0.9643**	35.33	**31.57**	**0.9640**	**30.44**

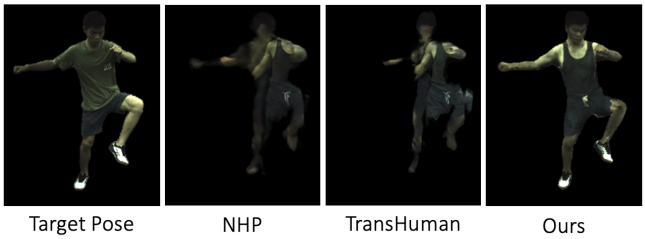


Fig. 12: Comparison with multi-view approaches: Neural Human Performer [40] and TransHuman [62].

the canonical SMPL mesh surface of the first identity. We initialize the matrices $\boldsymbol{U}_1$, $\boldsymbol{U}_2$, $\boldsymbol{U}_3$ as described in Sec. 4.3 of the main paper.

The positions $\boldsymbol{\mu}_c$ of the Gaussians in the canonical space are first encoded into a multi-level hash grid and passed through the non-rigid deformation MLP f_d. The hash grid has 16 levels of 2-dimensional features each, hash table size 2^{16}, coarse and fine resolution 16 and 2048 correspondingly [72]. The non-rigid MLP f_d is also conditioned on a latent code $\boldsymbol{z}_p$ that is the output of an hierarchical encoder [53,72]. The rigid MLP f_r inputs the non-rigidly deformed positions $\boldsymbol{\mu}_d$ and outputs the skinning weights that sum up to 1 through a softmax layer [72]. Similarly with 3DGS-Avatar [72], we normalize the coordinates in the canonical space by proportionally padding the bounding box enclosing the canonical SMPL mesh of the identity. The color MLP is conditioned on the output $\boldsymbol{z}$ of the non-rigid network, a per-Gaussian feature vector $\boldsymbol{f}$ and the canonicalized viewing direction (see Sec. 3.2). We do not learn any per-frame latent codes, in order to avoid any overfitting to the training frames. We use a larger color MLP compared to 3DGS-Avatar, of 3 layers and 256 hidden units each, in order to learn the diverse colors of multiple identities.

We observed that our network is very sensitive to the learning rate of the different components and the initialization, similarly with other 3DGS methods [32]. Our initialization is described in Sec. 4.3. We experimentally chose the learning rates depicted in Tab. 4. As shown, we use different learning rates for different rows of $\boldsymbol{U}_1$ that roughly correspond to scaling, rotation, features for

color, and opacity (see our tensor construction in Sec. 4.1). These learning rates are similar with the learning rates used for scaling, rotation, color, and opacity by other 3DGS implementations [32, 72]. However, in our case, we get the corresponding values after multiplying the matrices $\boldsymbol{U}_1$, $\boldsymbol{U}_2$, $\boldsymbol{U}_3$ using Eq. (8). We use Adam optimizer [35]. The rest of the Adam hyper-parameters are set at their default values ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$).

We train our multi-identity network for 5×10^4 iterations for 15 subjects, and 10^5 iterations for 30 subjects, that need about 2 and 4 hours correspondingly on a single NVIDIA Quadro RTX 6000 GPU. Following prior work [72, 92], we freeze everything in the first 10^3 iterations and train only the rigid MLP. In this way, we better initialize the rigid mapping from the canonical to the observation space based on estimated SMPL parameters, and avoid any noisy gradients in the beginning. After the first 10^3 iterations, we enable optimization to all the components, except for the non-rigid network that is frozen until the first 3×10^3 iterations. In constrast to 3DGS-Avatar [72], we do not add any learnable module for pose correction for the estimated SMPL parameters, in order to avoid overfitting to the training body poses.

We use the same loss function as 3DGS-Avatar [72]:

$$\begin{aligned} \mathcal{L} = \lambda_{l1}\mathcal{L}_{l1} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{mask}\mathcal{L}_{mask} + \lambda_{skin}\mathcal{L}_{skin} + \\ \lambda_{isopos}\mathcal{L}_{isopos} + \lambda_{isocov}\mathcal{L}_{isocov} \ , \end{aligned} \tag{9}$$

where $\mathcal{L}_{l1}$ is the L1 photometric loss and $\mathcal{L}_{perc}$ is the perceptual (LPIPS) loss with VGG as backbone. The mask loss $\mathcal{L}_{mask}$ corresponds to the L1 loss between the ground truth foreground mask and the predicted mask by accumulating the predicted opacities [72]. The skinning loss provides a regularization on the non-rigid MLP [72]. The as-isometric-as possible losses $\mathcal{L}_{isopos}$ and $\mathcal{L}_{isocov}$ restrict the 3D Gaussian positions to preserve a similar distance after deformation to the observed space, with also similar covariance matrices [72]. We set $\lambda_{l1} = 1$, $\lambda_{perc} = 0.01$, $\lambda_{mask} = 0.1$, $\lambda_{isopos} = 1$, $\lambda_{isocov} = 100$, and $\lambda_{skin} = 10$ for the first 10^3 iterations that is then decreased to $\lambda_{skin} = 0.1$ [72].

In order to avoid any non-differentiable gradient updates, we do not apply any densification or pruning of the 3D Gaussians during training, in contrast to the original 3DGS implementation [32]. Instead, we keep N_g Gaussians throughout our optimization, which are moved and deformed according to our network. In this way, the learnable matrices $\boldsymbol{U}_1$, $\boldsymbol{U}_2$, $\boldsymbol{U}_3$ are directly optimized with gradient descent, without any non-differentiable updates, and include all parameters for all the training identities. In our preliminary experiments, we observed that we achieve similar results for a single identity with or without densification and pruning, using $N_g = 5 \times 10^4$ (see Fig. 13). However, including an adaptive density control of the Gaussians can be explored as future work.

For $N_i = 30$ identities, MIGS learns only $(M + N_i + N_g)R = (43 + 30 + 5 \times 10^4) \times 100 \approx 5 \times 10^6$ parameters, compared to $MN_iN_g \approx 6.5 \times 10^7$ that would be required by single-identity 3DGS representations, leading to a decrease *by at least one order of magnitude* in the total number of learnable parameters.

Number of 3D Gaussians N_g	5×10^4
Rank of CP tensor decomposition R	100
Dimension of feature vector $\boldsymbol{f}$	32
Dimension of non-rigid output vector $\boldsymbol{z}$	32
Non-rigid deformation MLP f_d: Linear layers	3
Non-rigid deformation MLP f_d: Hidden units	128
Rigid transformation MLP f_r: Linear layers	3
Rigid transformation MLP f_r: Hidden units	128
Color MLP f_c: Linear layers	3
Color MLP f_c: Hidden units	256
Activation	ReLU
Optimizer	Adam
Learning rate for $\boldsymbol{U}_{1_{3:6,:}}$ (scaling)	5×10^{-3}
Learning rate for $\boldsymbol{U}_{1_{6:10,:}}$ (rotation)	10^{-3}
Learning rate for $\boldsymbol{U}_{1_{10:42,:}}$ (feature)	2.5×10^{-3}
Learning rate for $\boldsymbol{U}_{1_{42:43,:}}$ (opacity)	5×10^{-2}
Initial learning rate for $\boldsymbol{U}_{1_{:3,:}}$, $\boldsymbol{U}_2$, $\boldsymbol{U}_3$	1.6×10^{-4}
Final learning rate or $\boldsymbol{U}_{1_{:3,:}}$, $\boldsymbol{U}_2$, $\boldsymbol{U}_3$	1.6×10^{-6}
Initial learning rate for rigid MLP	10^{-4}
Initial learning rate for non-rigid and color MLP	10^{-3}
Final learning rate for MLPs	10^{-6}
Learning rate schedule	exponential decay
Max iterations	10^5

Table 4: Hyper-parameters of our architecture.

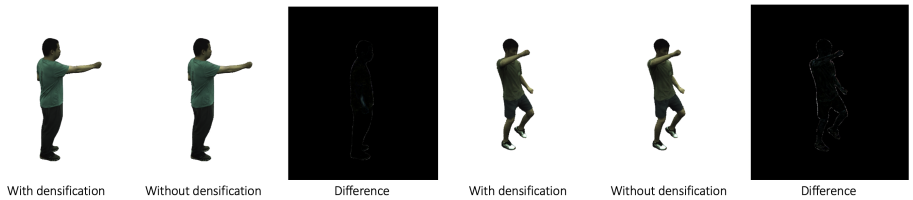


Fig. 13: With or without adaptive densification scheme for the 3D Gaussians [32].

D Limitations

We observe that in some cases, our multi-identity network may fail to capture fine-grained details, such as high-frequency texture in clothes or facial details. This tends to happen more when the number of identities increases beyond 20, since the network leverages information from multiple identities for learning, thus smoothing the result. In our work, we tackle this smoothing with the personalization procedure (see Sec. 4.4 and Fig. 8a). In the future, we plan to further enhance high-frequency details by using higher-resolution data and exploring other tensor structures. In addition, in the technical component, we have no theoretical proof that the CP decomposition is the optimal way to factorize the parameters, but we notice empirically that this suffices in our particular case.

E Ethical Considerations

We would like to note the potential misuse of video synthesis methods. With the advances in neural rendering, recent methods can generate photo-realistic human avatars. Our research focuses on human body animation and we presented as main application the animation of human avatars under challenging dance sequences. In contrast to deep fakes, we believe this application cannot be used to spread misinformation or for other harmful purposes. However, we would like to emphasize that there is still a risk of using such methods to generate misleading content. Thus, research on fake content detection and forensics is crucial. We intend to share our source code to help improving such research.

Fig. 14: Animation of human avatars under novel poses. Qualitative comparison with state-of-the-art approaches, namely HumanNeRF [92], MonoHuman [95], and 3DGS-Avatar [72]. The subjects are from ZJU-MoCap [69] and AIST++ [43, 80] datasets. The target poses (column 1) are unseen during training, from unseen camera views and advanced dance videos. Our method robustly animates all the identities under challenging novel poses, outperforming the other methods.