Multi-view 3D People Reconstruction combining Parametric and Non-parametric models

Òscar Lorente Corominas

Abstract

3D reconstruction of human bodies from multiple images has been a long-standing problem in computer vision. It is typically addressed using statistical models of the human body, which describe the geometry by a small number of parameters encoding 3D pose and shape. Non-parametric representations are alternatives that gain expressiveness for cloth capture, but have difficulties in recovering reasonable 3D human shapes when camera views are too sparse. In this dissertation, we aim to leverage the advantages of parametric and non-parametric models by extending the parametric Skinned Multi-Person Linear Model (SMPL) with Implicit Differentiable Renderer (IDR), an architecture that implicitly represents the geometry as a zero level-set of a neural network. The neural surface of IDR is typically initialized as a sphere, which allows rendering objects of all types. However, our work focuses on the reconstruction of human bodies, so we explore the contribution of parametric 3D human models such as SMPL as priors. The evaluation has been performed on a subset of the Renderpeople dataset, using as metrics for 3D reconstruction the Chamfer-L1 and point-to-surface distances, as well as PSNR for the corresponding renderings. The obtained results confirm that in scenarios where the camera views are too sparse, using an SMPL model as a prior improves 3D reconstruction and accelerates convergence. Finally, we propose a strategy based on an attention mechanism for IDR to improve the results on the head of the person, where the original IDR pipeline struggles to achieve a detailed reconstruction.

Index Terms

3D Surface Reconstruction, Human Model, Implicit Neural Representation, Differential Renderer, Attention.

I. Introduction

In recent years, interest and research in 3D vision and artificial intelligence have grown exponentially, and many of the applications in this field have been gradually integrated into our daily lives. This is the case of augmented/virtual reality, animation, or virtual dressing, to name a few. In many of them, we find a common factor: the need for a 3D model of a human body. Building these models from scratch is very time-consuming and labor-intensive for a designer, so researching alternatives to do it automatically is a must.

Many of the existing systems for building 3D models rely on specialized and expensive hardware, so reconstructing 3D shapes from data obtained with cheaper devices, such as RGB cameras, has become a common alternative. More specifically, 3D reconstruction of human bodies from RGB images is a fundamental problem in computer vision, and several works have been developed to approach it from different points of view. One of the most common is the use of parametric models of the entire human body [1], which describe the underlying geometry using a small number of parameters that encode 3D pose and shape. While fitting these low-dimensional parametric models allows for efficient approaches robust to in-the-wild images, the estimated shape corresponds to an undressed body and lacks the detail for capturing clothes. An alternative to gain expressiveness and detail in cloth capture are non-parametric representations. However, in these cases, it is more difficult to recover correct 3D human shapes when the camera views are too sparse, as shown in [2].

In this work, we take advantage of the robustness of parametric models and the flexibility of non-parametric representations by extending the Skinned Multi-Person Linear Model (SMPL) [3], which is smooth and corresponds to an undressed body that lacks the detail for capturing clothes, with Implicit Differentiable Renderer (IDR) [4], an architecture that implicitly represents geometry as a zero-level set of a Signed Distance Function (SDF) modeled by a neural network. The implicit surface from which IDR starts is usually initialized as a sphere, as this allows to adapt it to objects of all types. However, if we know a priori that the shape to be reconstructed is a person, we can initialize this surface with a parametric 3D human model such as SMPL. See Figure 1 for a big picture of the inputs and outputs of our system.

Hence, the main objective of our work is to investigate whether by using an initial shape more in line with the object to be reconstructed, we can improve the quality of the results in situations with very sparse camera views. We expect the 3D reconstruction to require fewer iterations to accurately represent the complex details of a human body, clothing, and hair, while also improving the rendered images. The latter is important because, although this network stands out for its 3D reconstructions, it is not able to achieve the same rendering quality as other state-of-the-art works such as NeRF and follow-ups [5, 6]. In this

Author: Òscar Lorente Corominas, oscar.lorente.co@gmail.com
Advisor 1: Xavier Giró-i-Nieto, Image Processing Group, Universitat Politècnica de Catalunya
Advisor 2: Francesc Moreno-Noguer, Perception and Manipulation Group, Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
Advisor 3: Enric Corona, PhD Student at Perception and Manipulation Group, Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
Thesis dissertation submitted: September 2021

(a) RGB images (b) SMPL model (c) Geometry (d) Appearance (e) Cameras

Fig. 1: Our system takes as inputs (a) RGB images and (b) SMPL models, which are undressed and lack the detail of clothes, and outputs (c, d) a high-detail clothed reconstruction and (e) refined camera parameters.

work, we aim to reconstruct 3D human bodies with a higher level of detail than IDR by using SMPL models as priors, and improve the quality of the renderings so that they are not so far from NeRF standards.

In parallel, we analyze the part of the human body where IDR has more difficulties in producing a detailed reconstruction: the head, and we propose a strategy to solve it based on an attention mechanism. This is described in detail in Section VII, but the main idea is to give emphasis to the parts where IDR is less confident about the results it provides. In this paper, we present some simple experiments to corroborate that these difficult parts have less detail because they need more attention and not because IDR does not have the capacity to reach higher quality.

This document presents the research conducted to corroborate the hypothesis formulated in the thesis. First, Section II reviews the literature of multi-view 3D surface reconstruction of human bodies, as well as recent related work. Section III describes the dataset used and discusses the proposed method in detail. Sections IV and V present the experiments performed and the corresponding results, respectively. The latter are analyzed in Section VI, where we determine whether the hypothesis has been corroborated and the implications of our work. Finally, in Section VII we discuss future work on this topic and propose a strategy on which we are already working to improve IDR results based on an attention mechanism.

II. State of the art

In this section, a review of the classical multi-view surface reconstruction techniques is presented, followed by the modern methods and the advantages that IDR presents with respect to them. Then, we explain what it means to represent a surface implicitly and how neural networks perform this task. Finally, some of the most recent works for reconstructing clothed humans are described to understand the motivation and advantages of our strategy.

A. Multi-view Surface Reconstruction

Reconstructing 3D objects from multiple images has been investigated for many years in the field of computer vision [7]. Long ago, before the era of deep learning, classic multi-view stereo (MVS) methods [8] attempted to extract the depth of a scene by point feature matching across neighboring views [9, 10, 11]. In these cases, a lossy post-processing step (e.g. volumetric fusion [12]) together with Poisson Surface Reconstruction algorithm [13] are needed to obtain a 3D watertight surface reconstruction from depth, which makes this process highly tedious and inefficient. Another alternative is to represent 3D shapes with a voxel grid [14, 15], but the resolution is limited due to the high memory requirements of three-dimensional voxel grids.

Early deep learning-based approaches proposed training neural models to perform subtasks of the MVS pipeline, such as feature matching [16, 17, 18] or depth fusion [19, 20], or even inferring depth maps directly in an end-to-end approach [21, 22]. However, a prior calibration of the cameras is usually required, which limits their potential to very controlled scenarios. Normally the camera parameters are unknown, so Structure-from-motion (SfM) methods [23] are used to estimate them and produce a sparse 3D reconstruction.

In contrast to these works, IDR is a neural network that provides an accurate watertight 3D surface reconstruction only using weak 2D supervision during optimization. The architecture also optimizes the camera parameters given a noisy linear initialization (e.g. obtained with COLMAP [24, 25]), so no prior camera calibration is strictly needed. Thus, IDR refines the camera projection matrices in order to represent the surface of an object accurately in an end-to-end approach, without low-resolution problems due to memory requirements and only needing one post-processing step: the Marching Cubes algorithm [26], which retrieves the reconstructed surface from the implicit representation.

B. Implicit Neural Representation

Explicit 3D representations such as point clouds [27, 28], meshes [29, 30], volumetric grids (voxels) [31, 32], or octrees [33], are the most intuitive ways to represent three-dimensional space, but have limited geometric quality due to the discrete nature of their underlying representations. On the other hand, the implicit representation of a surface defined as a zero-level set of a function allows flexibility and expressiveness without suffering the effects of discretization.

For this reason, neural implicit functions have recently emerged as an effective representation of 3D geometry [34, 35, 36, 37, 38] and appearance [39, 40, 41, 42, 5]. The main idea is to represent the geometry as a zero-level set of a function modeled by a neural network, in order to describe the information of a 3D point as the output of that network, which allows representing surfaces with arbitrary shapes and topologies. Most of these methods require 3D supervision, but several recent works have used differentiable rendering to train directly with 2D images [5, 42, 4]. This is the case of IDR, which is composed of a geometry network that models a Signed Distance Function (SDF) to its zero-level set to represent the surface implicitly, together with a differentiable neural renderer. This allows the renderer to backpropagate the loss to the implicit network parameters, and optimize the geometry using only 2D supervision and obtaining high-detailed 3D reconstructions.

C. Clothed Human Reconstruction

There are numerous efforts devoted to reconstructing 3D human bodies from multi-view cameras. One way to approach this problem is to use human statistical models of the full-body, like SCAPE [43], Total Capture [44] or SMPL and follow-ups [3, 45, 46], that encode 3D pose and shape with a small number of parameters. Accordingly, several deep learning approaches have been proposed to predict these parameters from 2D images [47, 48, 49, 50]. These low-dimensional parametric models allow for efficient approaches robust to in-the-wild images, but the geometry they describe is limited to represent an undressed body and lacks the detail for capturing clothes. To address this limitation, a number of works extend SMPL with displacement maps to represent clothes [51, 52, 53]. However, these approaches are too restricted by the SMPL body topology, and the cloth displacements they can predict are typically for relatively tight clothes.

In contrast, non-parametric representations such as silhouette convex hulls [54], geometry images [55], front/back depth map composition [56] or implicit function representations [37, 57], are an alternative to gain expressiveness for cloth capture. Those approaches based on implicit functions, PIFu [37] and PIFuHD [57], are arguably the current state-of-the-art. However, despite providing highly detailed clothing reconstructions, they are still constrained to relatively controlled scenarios with humans standing in simple poses. Recent works have combined differentiable renderers with deep neural networks [5, 42, 4], which allows obtaining impressive reconstruction results without needing so many input camera views. Nonetheless, when camera views are too sparse, these methods struggle to reconstruct 3D humans accurately, as shown in [2].

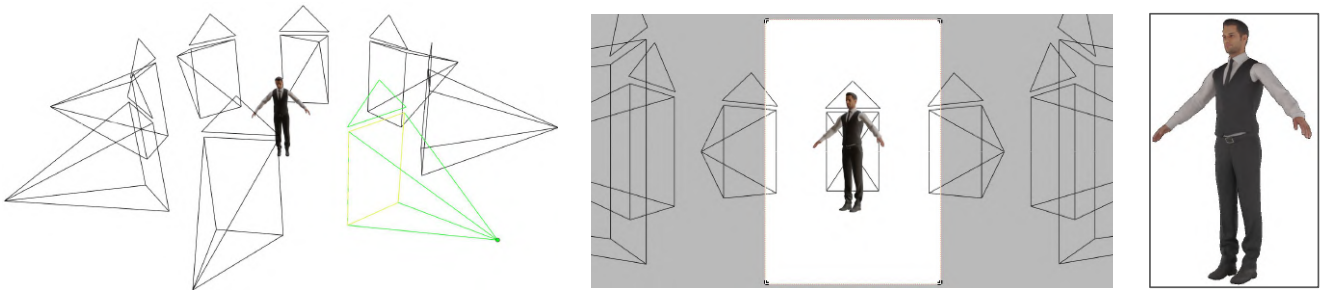
An interesting line of research is to combine the robustness of parametric models and the flexibility of non-parametric representations. Along this line, we find approaches that extend SMPL with voxels [58, 59], and more recently, with implicit functions [60, 61, 46]. The idea of our work is completely related to the latter, since we intend to leverage the advantages of parametric and non-parametric models by extending SMPL with an implicit function: IDR. Although these previous works retrieve rich geometric detail, their architectures are not as precise as IDR, and none of them generates novel views. In our case, we take advantage of IDR's ability to produce high-detailed 3D reconstructions and enhance it in the specific application of human body reconstruction, using SMPL as a prior.

III. Method

This section first describes the dataset used in the project, as well as the preprocessing steps followed to prepare the data for use. Then, the proposed method is exposed, explaining in detail how IDR works, and how its behavior changes when adding

(a) Dennis (b) Mei (c) Eric (d) Claudia

Fig. 2: 3D models of the Renderpeople subset used in this work.



(a) Blender setup (b) Camera view (of green camera) (c) Render

Fig. 3: Rendering process with Blender: given a setup (a), for each camera view (b) we render an RGB image (c).

an SMPL model as a prior. Finally, we describe the attention mechanism we propose (mostly as future work), but which has been superficially explored in this thesis.

A. Data

The data used in this work consists of two women and two men from the Renderpeople dataset [62]. Figure 2 shows the corresponding names, which will be used during the paper. Renderpeople provides 3D textured models, which are useful for further evaluation of the IDR surface reconstruction, but not for IDR training. As aforementioned, IDR needs RGB images from different viewpoints, masks indicating the object to be reconstructed, and a coarse initialization of the corresponding cameras. Therefore, the first step in our project is to prepare the data properly.

1) Image Rendering: We use the Blender software tool [63] to render multi-view RGB images from the 3D model provided by RenderPeople, which allows us to control the scene in terms of lighting, number of cameras, and position, among others. Specifically, we used four light sources to maintain an approximately constant illumination throughout the body, and rendered 32 images per 3D model with a resolution of 667x1002 pixels. This process is shown in Figure 3 with only eight cameras for visualization purposes.

2) 2D Masks: Once the images have been rendered, the next step is to create a 2D mask of the object we want to reconstruct. To do this, as we control the rendering of the images with Blender, we simply assign an *eccentric* color to the background of the image and apply a color-based thresholding to segment each person.

3) Coarse Camera Initialization: An approximate initialization of the camera parameters used to render the images is also needed. For this purpose, we use COLMAP [24, 25], a general-purpose structure-from-motion and multi-view stereo pipeline that allows us to obtain an estimation of the camera projection matrices.

4) Camera Normalization: The last step in order to use vanilla IDR is to normalize the cameras so that the visual hull of each person is contained in the unit sphere. The reason for this is explained in detail in Section III-B. By using Blender to create synthetic data we could have directly placed the cameras appropriately, but we wanted to simulate a real scenario and explore the difficulties this would pose and how to overcome them.

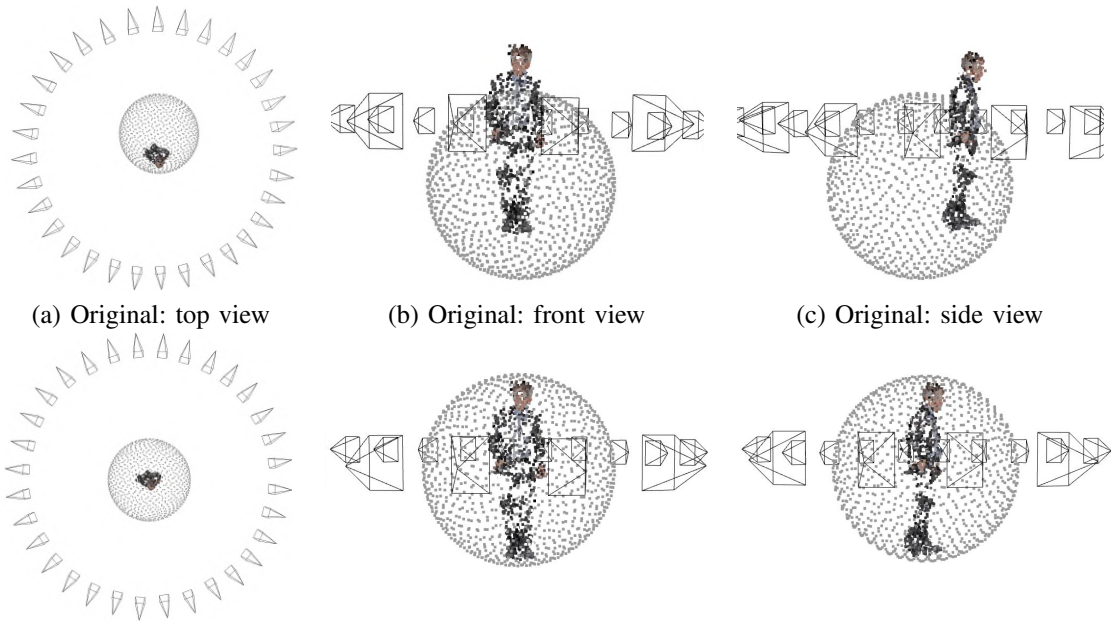
To accomplish that, we normalize the 3D body point cloud that COLMAP provides when estimating the cameras so that all the points of the cloud are contained inside the unit sphere. Specifically, we compute a similarity transformation T, consisting of a translation and a scaling, that takes the original points $p_i = [x_i, y_i, z_i]^T$ to a new set of points $\hat{p}_i = [\hat{x}_i, \hat{y}_i, \hat{z}_i]^T$ so that their centroid is the coordinate origin $[0, 0, 0]^T$ and the Euclidean distance to the farthest point from the origin is 1. This transformation is represented as

$$T = \begin{bmatrix} s & 0 & 0 & t_x \\ 0 & s & 0 & t_y \\ 0 & 0 & s & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad s = \frac{1}{max(d(p_i, c))}, \quad t = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} = -s \begin{bmatrix} c_x \\ c_y \\ c_z \end{bmatrix} = -sc, \tag{1}$$

where $c = [c_x, c_y, c_z]^T$ is the centroid of the original points p_i, $d(p_i, c)$ is the Euclidean distance between each point and the centroid, s is the scaling factor as the inverse of the distance from c to the farthest point p_i, and t is the corresponding translation vector that takes the cloud center to the coordinate origin.

To better understand how this transformation works, we present the result of applying T to a point $\tilde{p}_i = [p_i, 1]^T$ in homogeneous coordinates:

$$T\tilde{p}_i = \begin{bmatrix} s & 0 & 0 & t_x \\ 0 & s & 0 & t_y \\ 0 & 0 & s & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ z_i \\ 1 \end{bmatrix} = \begin{bmatrix} s & 0 & 0 & -sc_x \\ 0 & s & 0 & -sc_y \\ 0 & 0 & s & -sc_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ z_i \\ 1 \end{bmatrix} = \begin{bmatrix} s(x_i - c_x) \\ s(y_i - c_y) \\ s(z_i - c_z) \\ 1 \end{bmatrix} = \begin{bmatrix} s(p_i - c) \\ 1 \end{bmatrix} = \tilde{\hat{p}}_i \tag{2}$$



(a) Original: top view (b) Original: front view (c) Original: side view

(d) Normalized: top view (e) Normalized: front view (f) Normalized: side view

Fig. 4: Normalization of a cloud and camera matrices (a, b, c) so that the result (d, e, f) is contained inside the unit sphere.

As observed, multiplying the points p_i by T is equivalent to first translating p_i to the origin and then scaling the resulting points by a scaling factor s, finally obtaining the desired $\hat{p}_i$.

Given a projection matrix P, $\hat{P} = PT^{-1}$ is a new camera matrix such that the visual hull of each person is contained inside the unit sphere. To prove it, we denote by x the two-dimensional point that corresponds to the projection of a 3D point X with a matrix P, such that $\tilde{x} = P\tilde{X}$ in homogeneous coordinates. Then

$$\tilde{x} = P\tilde{X} = PT^{-1}T\tilde{X} = \hat{P}\tilde{\hat{X}}, \tag{3}$$

where $\hat{P} = PT^{-1}$ and $\tilde{\hat{X}} = T\tilde{X}$, being $\hat{X}$ the 3D point normalized inside the unit sphere, as demonstrated in Equation 2. This normalization process is visually presented in Figure 4.

B. Implicit Differentiable Renderer

IDR is an end-to-end neural architecture system that learns unknown geometry from masked 2D images and a noisy camera initialization. To accomplish that, Yariv et al. [4] represent the surface implicitly with an SDF modeled by a neural network, and the RGB color of each pixel as a differentiable function that depends on three unknowns: geometry, with parameters $\theta \in \mathbb{R}^m$; appearance, with $\gamma \in \mathbb{R}^n$; and cameras, as $\tau \in \mathbb{R}^k$. The notations used in the explanation below are depicted in Figure 5.

1) Geometry: The geometry S_θ is implicitly represented as the zero level set of an MLP f:

$$S_\theta = \left\{x \in \mathbb{R}^3 | f(x;\theta) = 0)\right\} \tag{4}$$

We refer to this network f as ImplicitNet, and it models a Signed Distance Function (SDF) to its level zero set. This means that the main objective is to optimize the ImplicitNet parameters θ so that the network output is approximately 0 when the input point x is on the surface of interest.

2) Appearance: The appearance is described as the factors that define the surface light field excluding the geometry, *i.e*. the surface bidirectional reflectance distribution function (BRDF) describing the reflectance and color properties of the surface, and the scene's lighting conditions. See [4] for more details.

To understand how IDR represents the appearance of a scene, being p a pixel from an input image I, the ray through p is denoted by $R_p(\tau) = \{c_p + tv_p | t \geqslant 0\}$, where $c_p = c_p(\tau)$ is the unknown center of the respective camera and $v_p = v_p(\tau)$ the vector pointing from c_p towards pixel p. The first intersection of the ray R_p and the surface S_θ is represented as $\hat{x}_p = \hat{x}_p(\theta, \tau)$. The incoming radiance along R_p, that is, the amount of light reflected from S_θ at $\hat{x}$ in direction -v reaching c, determines the rendered color of the pixel L_p. Note that for this to work, the object of interest must be contained in the unit sphere, since being the surface S_θ initialized as such, this first intersection between the ray R_p and S_θ will be on the surface of the sphere, and then deformed to adapt to the desired shape.

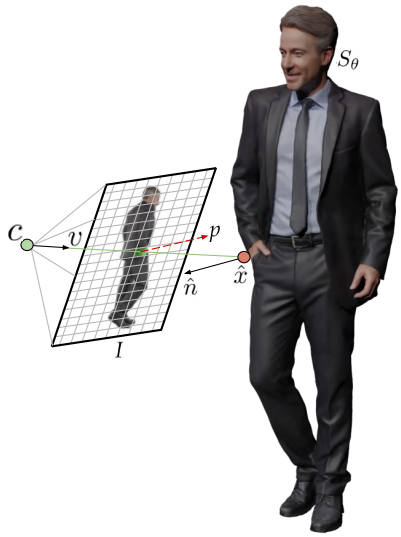


Fig. 5: Setup and notations for a given camera, RGB image and 3D surface.

L_p is a function of the surface properties and the incoming radiance at $\hat{x}_p$, which in turn are represented by the surface point $\hat{x}_p$, its corresponding surface normal $\hat{n}_p = \hat{n}_p(\theta)$, the viewing direction v_p, and a global geometry feature vector $\hat{z}_p = \hat{z}_p(\hat{x}_p; \theta)$. This feature vector allows the renderer to reason globally about the geometry S_θ, as it encodes the geometry relative to the surface sample x.

The surface light field that determines the rendered color of a pixel p is therefore modeled as

$$L_p(\theta, \gamma, \tau) = M(\hat{x}_p, \hat{n}_p, \hat{z}_p, v_p; \gamma), \tag{5}$$

where M is another MLP that we name RenderNet. L_p is used in a loss comparing it with the pixel input color I_p to simultaneously train the model's parameters θ, γ, τ. To backpropagate this loss from RenderNet to ImplicitNet, and thus modify the geometry only using mask 2D images, the intersection point $\hat{x}(\theta, \tau)$ is represented with parameters θ, τ by slightly modifying f. Then, the differentiable intersection of the ray $R(\tau)$ and the surface S_θ can be represented by the formula:

$$\hat{x}(\theta, \tau) = c + t_0 v - \frac{v}{\nabla_x f(x_0; \theta_0) v_0} f(c + t_0 v; \theta) \tag{6}$$

3) Loss: Let I_p, O_p be the RGB and mask values, respectively, corresponding to a pixel p in an image taken with camera $c_p(\tau)$ and direction $v_p(\tau)$, where $p \in P$ indexes all pixels in the input collection of images, and $\tau \in R^k$ represents the parameters of all the cameras in scene. The loss function has the form:

$$loss(\theta, \gamma, \tau) = loss_{RGB}(\theta, \gamma, \tau) + \rho loss_{MASK}(\theta, \tau) + \gamma loss_E(\theta) \tag{7}$$

This loss is trained on mini-batches of pixels in P, and for simplicity we denote by P the current mini-batch. For each $p \in P$ the sphere-tracing algorithm [12, 20] is used to obtain the first intersection point, $c_p + t_{p,0} v_p$, of the ray $R_p(\tau)$ and S_θ. Let $P^{in} \subset P$ be the subset of pixels p where intersection has been found and $O_p = 1$ (i.e. corresponds to the foreground mask). Being L_p defined as in Equation 5, the RGB loss is

$$loss_{RGB} = \frac{1}{|P|} \sum_{p \in P^{in}} |I_p - L_p(\theta, \gamma, \tau)|, \tag{8}$$

where $|\cdot|$ represents the L_1 norm. Let P^{out} denote the indices in the mini-batch for which no ray-geometry intersection or $O_p = 0$ (i.e. do not correspond to the foreground mask). The mask loss is

$$loss_{MASK}(\theta, \tau) = \frac{1}{\alpha |P|} \sum_{p \in P^{out}} CE(O_p, S_{p,\alpha}(\theta, \tau)), \tag{9}$$

where CE is the cross-entropy loss. Lastly, f is enforced to be approximately an SDF as in [64], incorporating the Eikonal regularization:

$$loss_E(\theta) = \mathbb{E}_x(\|\nabla_x f(x; \theta)\| - 1)^2, \tag{10}$$

where x is distributed uniformly in a bounding box of the scene.

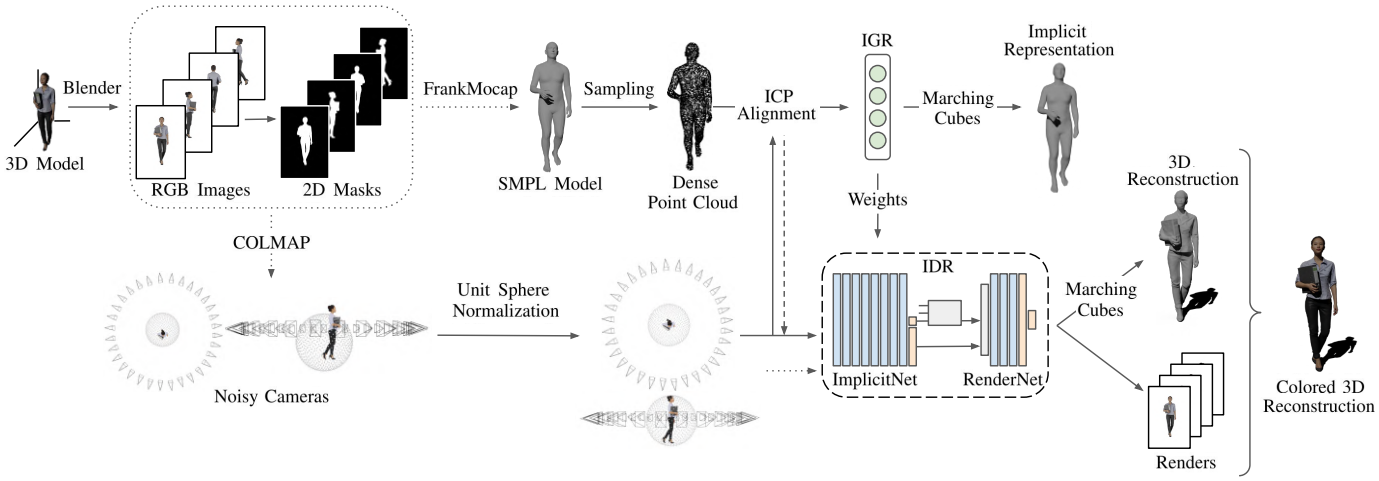


Fig. 6: Overview of our system.

4) Training: Each multi-view image collection is trained in an iterative process. For each iteration, 2048 pixels $p \in P$ are randomly sampled from each image, and the corresponding rays R_p are traced for each pixel. Then, from the intersections between these rays and the surface S_θ, the loss in Equation 7 is minimized to optimize the parameters of ImplicitNet and RenderNet. After training, the Marching Cubes (MC) algorithm [26] is used to retrieve the reconstructed surface from f. MC is a simple but intuitive algorithm that generates a triangular mesh by iterating (marching) over a uniform grid of cubes superimposed over a region of an implicit function. If the eight vertices of a cube are positive (negative), the cube is completely above (below) the surface and no triangular faces are generated. Otherwise, the surface crosses the cube, so some triangular faces and vertices are extracted. After performing the same process for all cubes, the resulting triangular faces are joined to generate the reconstructed mesh.

C. Proposed Method

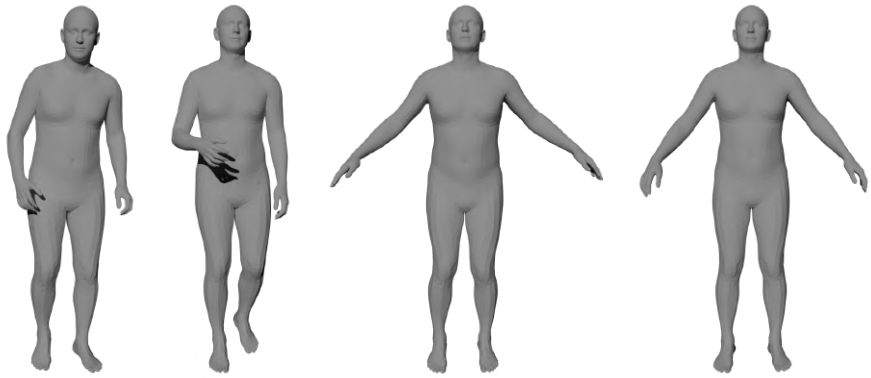
In this work, we propose a method to reconstruct 3D human bodies from 2D images based on IDR and SMPL-X [45], a parametric model that combines the realistic 3D vertex-based model SMPL [3] with the FLAME [65] head and MANO [66] hand models. Specifically, we want to explore the contribution of this parametric model to the ability of IDR to accurately represent 3D shapes in difficult situations with sparse views. The pipeline of our system is presented in Figure 6.

IDR has been designed to reconstruct objects of all types, so the geometric initialization of ImplicitNet causes the SDF to start representing approximately a unit sphere. In our case, however, the objects to be reconstructed are human bodies, so the network can be initialized to represent a coarse human model instead, and thus help IDR to converge to the desired shape faster and in more difficult situations. To do so, the first step is to obtain the 3D human model from our data.

1) Skinned Multi-Person Linear Model (SMPL): SMPL [3] is a skinned and realistic 3D vertex-based model that accurately represents body shapes in natural human poses. One of its follow-ups is SMPL-X [45], which extends SMPL with an expressive face and fully articulated hands. Specifically, SMPL-X combines SMPL with the FLAME head model [65] and the MANO

(a) Hand artifact (b) Body orientation error

Fig. 7: FrankMocap fails to estimate Dennis' hand (a) and Claudia's body orientation (b).



(a) Dennis (b) Mei (c) Eric (d) Claudia

Fig. 8: Final SMPL models estimated with FrankMocap.

hand model [66], and is the one used as a prior to initialize IDR, as it allows us to represent the 3D shape more accurately. For simplicity, we refer to SMPL-X as SMPL.

To estimate the parameters of the SMPL model we use FrankMocap [67, 68], a 3D motion capture system that provides 3D pose estimation from a single image. As shown in Figure 7, there are some artifacts in the estimations of FrankMocap. In the case of Dennis, the hand in the pocket is not well represented, so we estimate the parameters of the SMPL model for all 32 images and average their values, resulting in a better representation. Furthermore, FrankMocap completely fails to estimate Claudia's SMPL parameters in one image, so we cannot average the values of all images and we have to decide which estimation is more appropriate. Thus, obtaining the models is a semi-manual task that requires supervision to choose the most appropriate parameters for each person. Figure 8 shows the resulting SMPL model for each person, and the next step is to integrate them in IDR.

2) Cameras and SMPL Alignment: As detailed in Section III-A4, we need to normalize the camera parameters so that the visual hull of each person is contained in the unit sphere in order to use vanilla IDR. In this case, as the implicit surface is initialized with an SMPL model, if the visual hull is aligned with the model, IDR would not have to make extra effort during the early training stage to adjust the initial surface (SMPL model) to the desired person shape (which is determined by the visual hull). Although the initial shape is no longer a unit sphere, the visual hull still has to be contained in it, since IDR's algorithm for finding the intersecting point between a visual ray and the surface implicitly modeled by ImplicitNet assumes that this implicit surface is contained in the unit sphere.

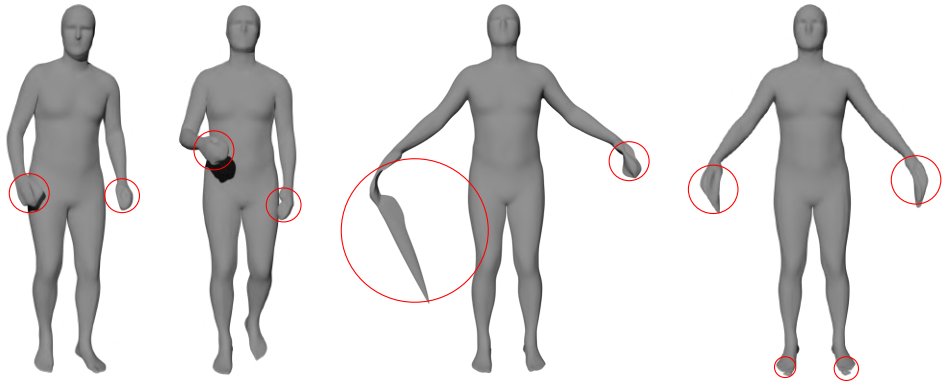
Therefore, the first step to normalize the cameras in our system is the same as when using vanilla IDR: make the visual hull centered and contained in the unit sphere, thus obtaining the projection matrices $\hat{P} = PT^{-1}$, as shown in Equation 3. In this step, recall that the point cloud provided by COLMAP when estimating the cameras is also normalized with T. Next, we sample the SMPL model to obtain a dense point cloud, and normalize it in the same way. Finally, we use the Iterative Closest Point (ICP) algorithm to align the two point clouds. This algorithm takes as an input two point clouds and an initial transformation that aligns them approximately, which in this case is an identity matrix, since both point clouds are centered and contained within the unit sphere. ICP then iteratively refines the transformation to adjust the two point clouds. Once this refined transformation matrix T_{align} is obtained, we normalize the camera projection matrices with $\hat{P}_{align} = \hat{P}T_{align}^{-1}$ so that their visual hull is aligned with the SMPL model.

3) Extending IDR with SMPL: ImplicitNet models the geometry as a zero level set of an SDF, i.e., the parameters of this network are optimized so that when the input is a point of the desired surface, the output (SDF) is approximately 0. With this in mind, one of our first ideas to integrate SMPL into the IDR architecture was to add the loss

$$loss_{SMPL} = \sum_{v \in V} |f(v;\theta)|\,, \tag{11}$$

to Equation 7, which would force the ImplicitNet output ($f(v;\theta)$) to be minimized at the vertices of the SMPL model $v \in V$. In this way, the surface would converge to these vertices. However, the SMPL model does not represent in detail the person we want to reconstruct in terms of clothes, hair, or face, so the final result would be too imprecise. A possible solution is to schedule the $loss_{SMPL}$ weight to be larger in the early stages of training, and thus force the network to converge to the SMPL model rapidly, and then lower this weight to let the IDR architecture take care of detailing the reconstruction.

This loss schedule involves another parameter to optimize and is not an efficient approach, so we propose another strategy: to directly initialize the ImplicitNet parameters so that, instead of representing a unit sphere, they describe the corresponding SMPL model. In this way, the network starts directly with an approximate shape of the person, and the objective of the



(a) Dennis (b) Mei (c) Eric (d) Claudia

Fig. 9: Implicit representation of the SMPL models with IGR, with small artifacts (a, b, d) and severe errors (c).

architecture is to add detail to it. This method achieves the same as proposed above with the extra $loss_{SMPL}$, but in an elegant and more intuitive way.

Yariv et al. [4] use [69] to initialize the ImplicitNet weights such that the SDF they approximate is that of a unit sphere. In our case, we use Implicit Geometric Regularization (IGR) [64], a deep learning architecture that learns implicit signed distance representations directly from raw point clouds. The IGR neural network is identical to ImplicitNet, except that the latter includes positional encoding. Therefore, we slightly modify the IGR network so that its parameters are the same as those of ImplicitNet, and we can load the weights from one to the other directly.

IGR tries to implicitly represent a surface that is contained inside the input point cloud, so instead of using only the (few) vertices of the SMPL model as input, we randomly sample 250,000 points to ensure an appropriate representation of the SMPL model. Figure 9 shows the result of using Marching Cubes on the implicit representation produced by IGR for each SMPL model. As can be seen, this representation is less detailed than Figure 8 (faces), and some artifacts appear. This affects the potential of our system, since if the prior used to initialize IDR presents severe artifacts, IDR will have to make an extra effort to correct these errors first before focusing on detailing the coarse model.

D. Attention Mechanism

Alongside the development and implementation of the proposed method, we also identify the part of the body where IDR struggles to produce a detailed reconstruction: the head, and we propose a strategy for the network to focus on it. As aforementioned, for each iteration IDR samples 2048 random pixels of each 2D image, trace the corresponding rays and minimize the loss at the intersection points between these rays and the surface. Thus, our idea is to sample more often the pixels corresponding to the head. This, although very simple, serves to identify whether by emphasizing the head we can improve the IDR reconstruction, or if on the contrary the detail of the head is limited by the power of the IDR architecture. As will be explained in Section V, we can indeed improve IDR reconstructions by focusing on parts such as the head. In Section VII we discuss a proposal for a more elaborate and robust attention mechanism for IDR to automatically detect the parts that need more attention, which is currently under development.

IV. Experiments

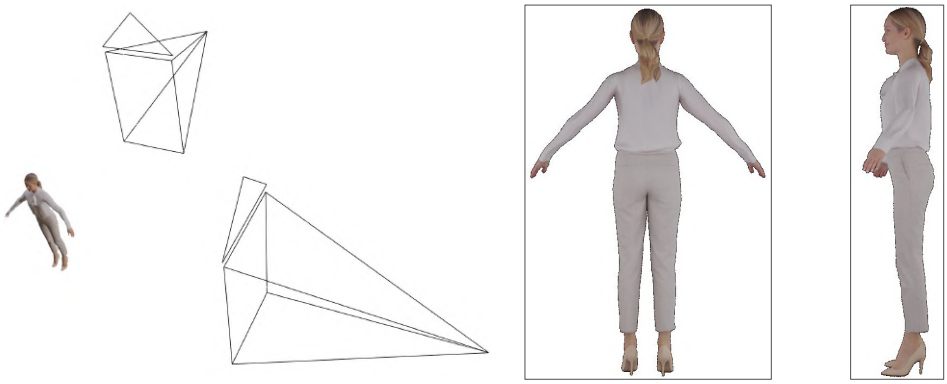
In this section, the experiments performed in the thesis, which aim to corroborate the formulated hypothesis, are described in detail. To do so, we consider not only 32, but also 16, 8, 4, 2, and 1 images per 3D model, and in all cases when we reduce the number of images by half, the symmetry of the cameras is maintained. We add another case: 2*, in which the two cameras are not symmetrical (and different for each person), as shown in Figure 10. The idea is to study the behavior of the architecture in a variety of scenarios with different levels of difficulty.

A. Setups

In order to maintain a structured results section, we number the experiments (e.g. Exp. 1) and add a title to them to understand their purpose. But before that, we expose the different configurations we used.

1) Vanilla IDR: The first configuration, named **IDR**, consists on using the vanilla IDR with unit sphere initialization.

2) SMPL Model Initialization: Our system configuration is named **IDR+SMPL**, as it uses the SMPL model to initialize IDR.



(a) Blender camera setup (b) Render (back) (c) Render (side)

Fig. 10: Claudia 2* scenario with two sparse cameras.

B. ***Exp. 1***: *IDR and IDR+SMPL using different number of views*

In this first experiment, we use both IDR and IDR+SMPL to reconstruct the four persons in the dataset for the 32, 16, 8, 4, 2, 2*, and 1 number of images. We want to evaluate the ability of vanilla IDR to produce detailed reconstructions of clothed 3D human bodies, and test whether initializing the IDR ImplicitNet with the corresponding SMPL models improves the performance of the network.

C. ***Exp. 2***: *IDR and IDR+SMPL in casual and A-pose subsets*

In the second experiment we compare the performance of IDR and IDR+SMPL when the 3D models are in a casual or in an A-pose, since in casual situations there are more occlusions and parts of the body that are not seen from some points of view, thus making the reconstruction more difficult. In our case, Dennis and Mei are in a casual pose, and Eric and Claudia are in an A-pose, so we test both systems in these two subsets.

D. ***Exp. 3***: *IDR and IDR+SMPL convergence speed*

The last experiment performed with these two configurations is intended to test whether using the SMPL model as prior (IDR+SMPL) the architecture converges faster. For this purpose, we evaluate the metrics at different time instants during the training stage.

E. ***Exp. 4***: *Attention mechanism*

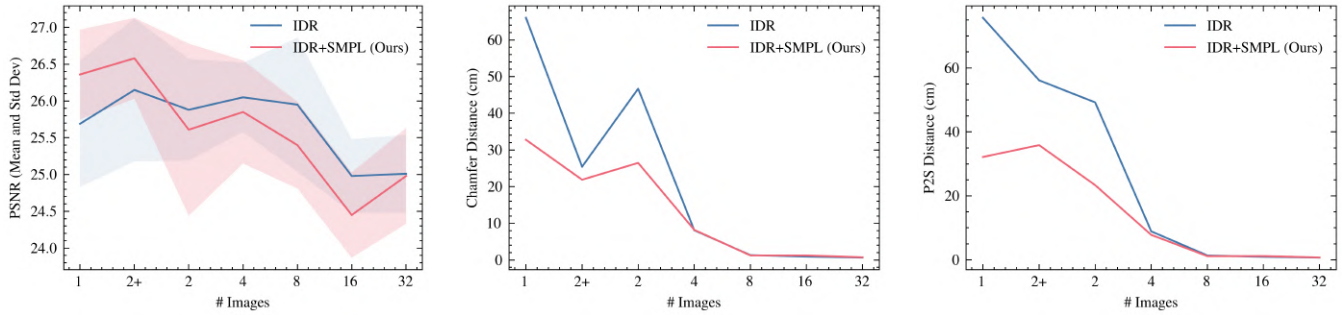
As will be shown in detail in Section V, the part of the body where IDR struggles to achieve a detailed reconstruction is the head. Therefore, we perform a series of experiments to test whether this lack of detail is due to the fact that IDR is not able to reconstruct complex areas such as the face, or if by emphasizing this area the quality improves. As mentioned, to corroborate this, a simple strategy is implemented to sample more points on the head.

(a) SMPL vertices (b) SMPL render (front) (c) SMPL render (side) (d) Head BB (front) (e) Head BB (side)

Fig. 11: FLAME vertices (a, in green) are projected onto the RGB images (b, c) to determine the head bounding boxes (d, e).

# Images	Experiment	PSNR (dB) ↑ Mean	PSNR (dB) ↑ Std Dev	Chamfer ↓ (cm)	P2S ↓ (cm)
32	IDR	**25.01**	0.53	**0.63**	**0.67**
	IDR+SMPL	24.98	0.65	0.72	0.72
16	IDR	**24.98**	0.50	**0.93**	**0.94**
	IDR+SMPL	24.45	0.58	1.25	1.17
8	IDR	**25.95**	0.91	1.33	1.36
	IDR+SMPL	25.40	0.59	**1.24**	**1.14**
4	IDR	**26.05**	0.47	**8.10**	8.89
	IDR+SMPL	25.85	0.70	8.21	**7.79**
2	IDR	**25.88**	0.69	46.70	49.23
	IDR+SMPL	25.61	1.17	**26.46**	**23.35**
2+	IDR	26.15	0.97	25.42	56.14
	IDR+SMPL	**26.58**	0.55	**21.88**	**35.87**
1	IDR	25.69	0.86	66.01	75.68
	IDR+SMPL	**26.36**	0.61	**32.76**	**32.16**

TABLE I: **Exp. 1 results**. Comparison between IDR and IDR+SMPL performance using different number of views. Higher PSNR, lower Chamfer and lower Point-to-Surface (P2S) are better.



(a) PSNR (dB) ↑ (b) Chamfer (cm) ↓ (c) P2S (cm) ↓

Fig. 12: Our system performs better than vanilla IDR in scenarios with very sparse views in terms of (b) Chamfer and (c) Point-to-Surface distances (in centimeters), while maintaining a similar (a) PSNR.

First, to identify the pixels that correspond to the head for each image automatically, a specific face detection network could be used, but instead, we use the SMPL model that we have already obtained for each person and image. SMPL uses the FLAME model for the head, so we can obtain the head by simply extracting the vertices of the model that correspond to FLAME. By projecting these vertices on each image, the corresponding bounding box can be easily determined, as shown in Figure 11.

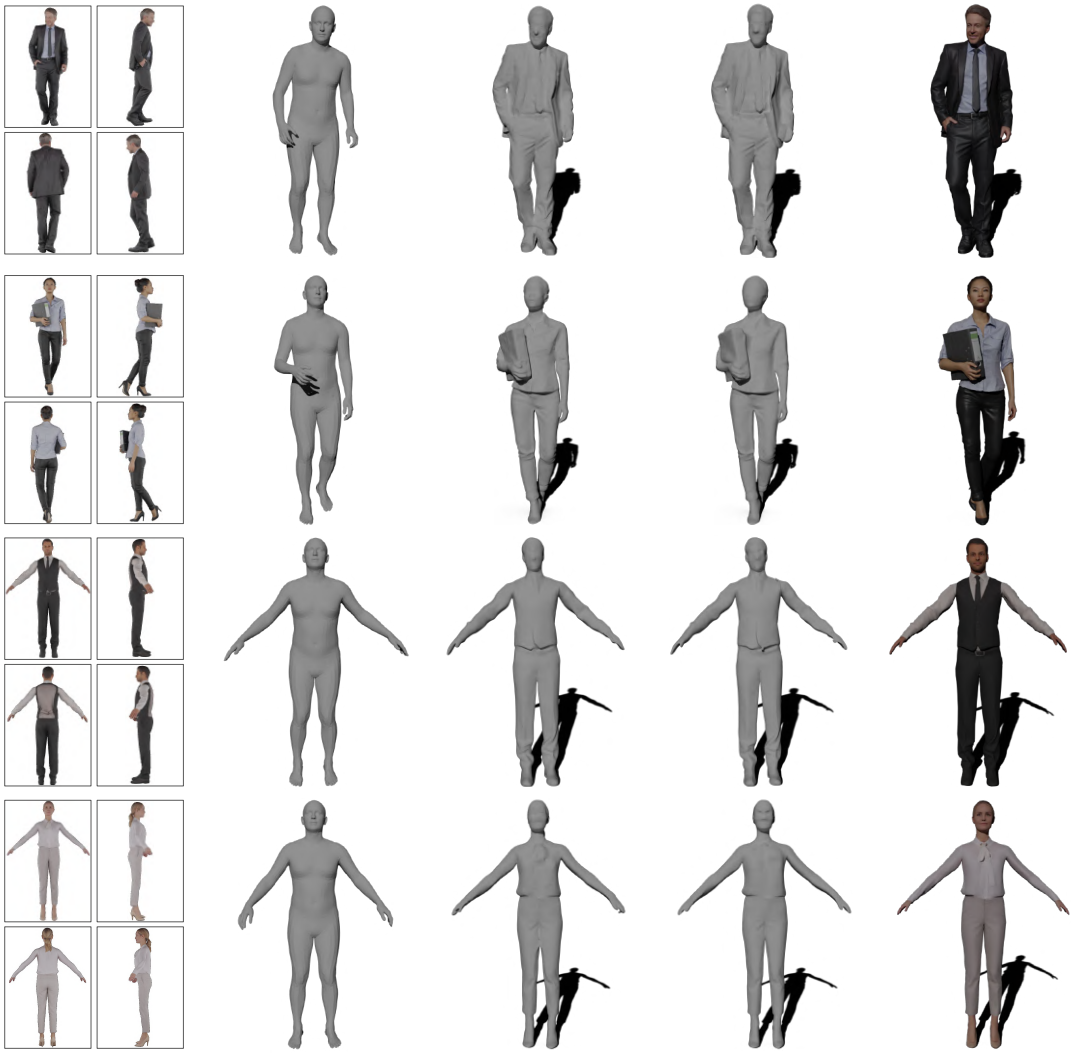
In Section III-B we explain that IDR samples 2048 random pixels from each image in each iteration and optimizes the network parameters by minimizing the loss for each intersection point. To sample more points from the head, we limit the sampling range to the head bounding box (rather than the entire image) every X iterations, to ensure that the network parameters are optimized specifically for the head periodically. Thus, the goal of this last experiment is to determine how much the 3D reconstruction of a person improves by emphasizing the head, so we sample this area every 2, 4, 8, 16, 32, and 64 iterations (defined as epoch step). The corresponding experiments are referred to as **HE** (Head Emphasis), being HE_64 the experiment with an epoch step of 64 iterations.

V. Results

The results obtained in the proposed experiments are presented in this section. First of all, we describe the metrics used in the evaluation, which help us to corroborate the hypothesis formulated. Then, the results are presented in various formats: tables, graphs, image collages, etc., as we consider it appropriate to better understand the implications of our system from different points of view. All experiments have been performed with a GTX 2080 Ti with 11GB of RAM, so the training times provided are obtained using this GPU.

A. Metrics

The evaluation is carried out both qualitatively and quantitatively, using as metrics the Peak-Signal-to-Noise-Ratio (PSNR) for the renders and the Chamfer-L1 and Point-to-Surface (P2S) distances for the 3D reconstructions.



(a) RGB images (b) SMPL model (c) IDR (d) IDR+SMPL (e) Ground truth

Fig. 13: Results obtained with (c) IDR and (d) IDR+SMPL using 32 (a) RGB images. With so many views, the results of vanilla IDR are already accurate, and IDR+SMPL helps to add a few detail in some cases (or smooth the result in others).

1) Peak-Signal-to-Noise-Ratio (dB): PSNR is used to evaluate the quality of the renders provided by IDR (f) with respect to the corresponding ground truth images (g) as

$$PSNR = 20log_{10}(\frac{MAX_f}{\sqrt{MSE}}), \quad MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\|f(i,j) - g(i,j)\|^2, \tag{12}$$

where $MAX_f = 1$ and $m = 1002, n = 667$ are the number of rows and columns of each image respectively. We provide the mean and standard deviation (Std Dev) values from the PSNR computed over each rendered image. The higher the PSNR the better.

2) Chamfer distance (centimeters, cm): One of the metrics used to evaluate the 3D reconstructions of our system is the Chamfer distance, computed between the vertices of the predicted reconstruction (F) and the vertices of the ground truth 3D model (G) with

$$D_{chamfer}(F;G) = D_{chamfer}(F,G) + D_{chamfer}(G,F), \tag{13}$$

where $D_{chamfer}(F;G)$ represents the total Chamfer distance between the two point clouds F and G, $D_{chamfer}(F,G)$ is the unidirectional (Chamfer) distance from point cloud F to G as

$$D_{chamfer}(F,G) = \frac{1}{|F|}\sum_{f\in F} d_G(f), \tag{14}$$

where $d_G(f)$ is the distance between a point $f \in F$ and its closest point in G; and $D_{chamfer}(G, F)$ is the same but in the other direction. Lower Chamfer distance is better.

3) Point-to-Surface distance (centimeters, cm): The other metric used to evaluate the quality of the 3D reconstructions is the P2S distance, which computes the distance between the vertices of the predicted reconstruction (V) and the mesh of the ground truth 3D model (M) as

$$D_{P2S}(V; M) = D_{P2F}(V, M) + D_{F2P}(M, V), \tag{15}$$

where $D_{P2F}(V, M)$ is the mean squared distance of each point in V to its closest triangular face in M, and $D_{F2P}(M, V)$ is the mean squared distance of each triangular face in M to its closest point in V. Lower P2S distance is better.

B. 3D Reconstruction Alignment

In order to evaluate the 3D reconstructions accurately, we use the Iterative Closest Point (ICP) algorithm to align the IDR output with the ground truth of the 3D model. As explained in Section III-C2, this technique starts from two roughly aligned point clouds, so the first step is to normalize both of them to be contained in the unit sphere, and then align them. Finally, we denormalize both point clouds to have the dimensions of the 3D ground truth model, which are in centimeters, so that we can calculate the Chamfer and P2S distances in centimeters.

Another thing to keep in mind is that the ground truth of each 3D model has a different number of vertices, so we have sampled 100,000 equidistant points from the ground truth surface to make the evaluation more consistent regardless of the reconstructed person.

*C. **Exp. 1 results**: IDR and IDR+SMPL using different number of views*

In Table I we present a comparison of the evaluation of vanilla IDR and our system (IDR+SMPL) in the different scenarios proposed, with different numbers of images in each. As observed, the PSNR remains more or less constant in all cases, and even improves (although with more variance) in situations with fewer images. This is due to the fact that, since the IDR training is based on supervision with RGB images, using very few views the network overfits more to them and generates better results. The opposite happens in the case of 3D reconstruction, since having fewer views of the object, it is more difficult to adapt the surface properly, so the Chamfer and P2S distances increase as the number of images decreases.

In these difficult situations, it seems that using an SMPL model as a prior improves the 3D reconstruction considerably, as the distances decrease by up to half in some cases. This is even more evident in the case of a single view, since IDR is not able to generate the surface correctly because it does not have information about the area that is not seen in the image. However, the SMPL model has a very similar shape to the person in question, so using it as a prior implies that in these areas where there is no visibility, the surface will be more in line with the person. It is the same for the 2 and 2* image scenarios, where using the SMPL model greatly improves the accuracy of the 3D reconstruction.

This reasoning is shown visually in Figure 12, where we can see how the Chamfer and P2S distances are larger in the cases where fewer views are available, but using the SMPL model as a prior, the 3D reconstruction improves. We also observe that the renderings do not depend so much on the number of images in question, as the PSNR remains similar in all cases, and even drops a little more when 16 or more images are used.

On the other hand, with 32 and 16 images the results are better when the SMPL model is not used, so it seems that in these cases IDR has a sufficient number of views to reconstruct a 3D model accurately without any extra help, as shown in

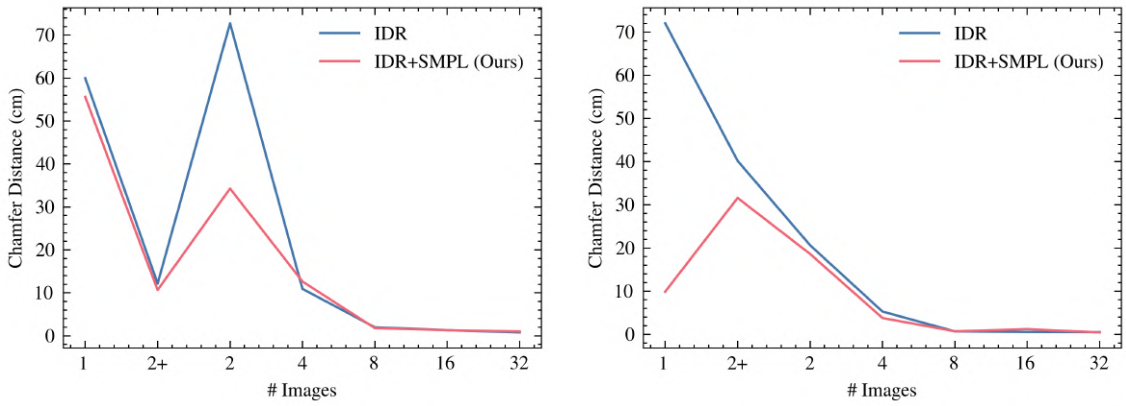
# Images	Experiment	PSNR ↑ (dB)	Chamfer ↓ (cm)	P2S ↓ (cm)
32	IDR	**25.45**	**0.76**	**0.54**
	IDR+SMPL	25.42	1.00	0.73
16	IDR	**25.05**	1.31	1.06
	IDR+SMPL	25.00	**1.29**	**1.00**
8	IDR	**26.77**	1.97	1.77
	IDR+SMPL	25.69	**1.78**	**1.36**
4	IDR	**26.02**	**10.91**	**10.57**
	IDR+SMPL	25.88	12.63	12.21
2	IDR	25.46	72.75	71.76
	IDR+SMPL	**26.46**	**34.28**	**29.61**
2+	IDR	26.36	12.15	11.83
	IDR+SMPL	**27.01**	**10.65**	**9.93**
1	IDR	25.75	59.97	58.58
	IDR+SMPL	**26.34**	**55.64**	**54.29**

(a) Casual pose

# Images	Experiment	PSNR ↑ (dB)	Chamfer ↓ (cm)	P2S ↓ (cm)
32	IDR	**24.57**	0.5	0.8
	IDR+SMPL	24.53	**0.43**	**0.71**
16	IDR	**24.92**	0.54	0.82
	IDR+SMPL	23.9	**1.21**	**1.34**
8	IDR	**25.14**	0.69	0.95
	IDR+SMPL	25.1	**0.7**	**0.91**
4	IDR	**26.07**	5.28	7.2
	IDR+SMPL	25.81	**3.79**	**3.36**
2	IDR	**26.3**	20.66	26.7
	IDR+SMPL	24.77	**18.64**	**17.09**
2+	IDR	25.94	40.19	100.46
	IDR+SMPL	**26.16**	**31.6**	**61.8**
1	IDR	25.62	72.06	92.79
	IDR+SMPL	**26.37**	**9.88**	**10.03**

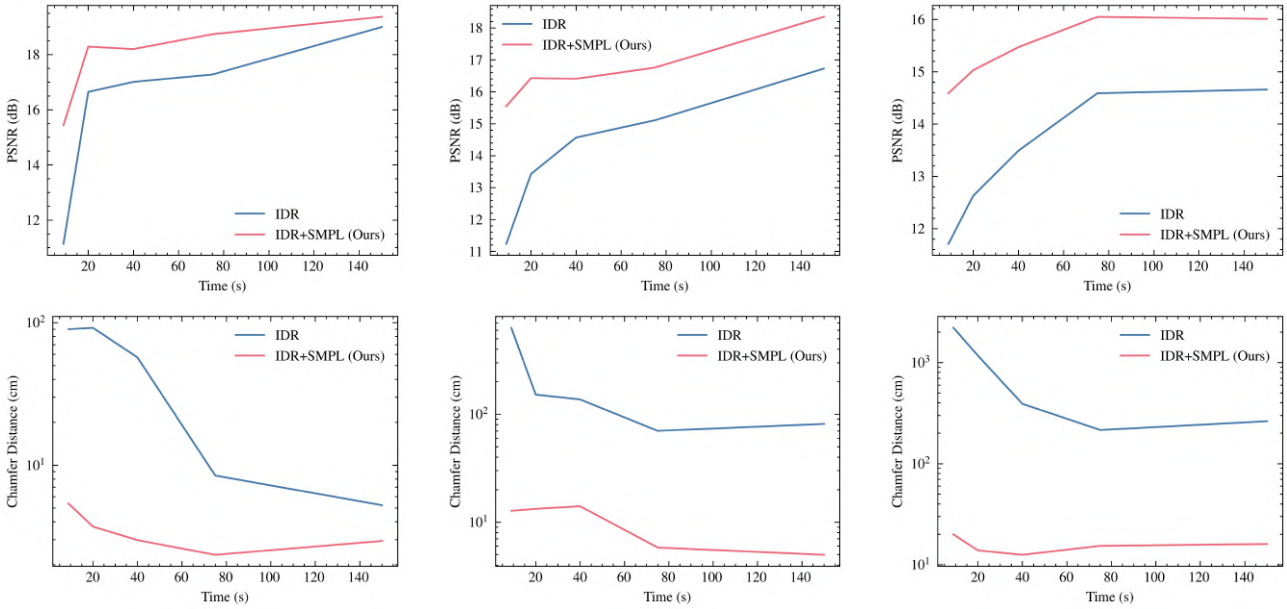
(b) A-pose

TABLE II: **Exp. 2 results**. Comparison of IDR and IDR+SMPL performance on casual and A-pose subsets.



(a) Casual pose (b) A-pose

Fig. 14: Chamfer distance of IDR and IDR+SMPL for the (a) casual and (b) A-pose subsets. Our system works better with A-pose people, substantially improving the results in cases of very sparse views.



(a) 32 RGB Images (b) 4 RGB Images (c) 1 RGB Image

Fig. 15: Top: PSNR (dB), higher is better. Bottom: Chamfer distance (cm), lower is better. Measuring the speed convergence of IDR and IDR+SMPL in three scenarios: (a) 32, (b) 4, and (c) 1 RGB images.

Figure 13. This is not a problem, since our interest is in improving the results in the cases where IDR struggles the most, i.e., with sparse camera views.

*D. **Exp. 2 results**: IDR and IDR+SMPL in casual and A-pose subsets*

We present the evaluation for the casual and A-pose subsets in Tables IIa and IIb, respectively.

As can be seen, with a casual pose the 3D reconstruction results are worse in general, and although improved by using the SMPL model as a prior, in the 1 or 2 image scenarios the reconstruction quality is not good at all. In the specific case of 1 image, if the person has self-occlusions, IDR loses a lot of visibility and is not able to adapt the surface properly. As the SMPL model does not represent the pose of the person accurately, it is also not able to compensate for the lack of views. On the other hand, it seems that the A-pose results are considerably better when using the SMPL model, probably because the SMPL model of a person in an A-pose is much simpler to estimate than that of a casual pose, so the prior more closely resembles the shape of the person.

Figure 14 visually presents the improvement of our system with respect to vanilla IDR in the specific case of the Chamfer distance in the casual pose and the A-pose. Specifically, we observe how this improvement is indeed more substantial in the case of the A-pose, since it is easier to accurately obtain the corresponding SMPL model.

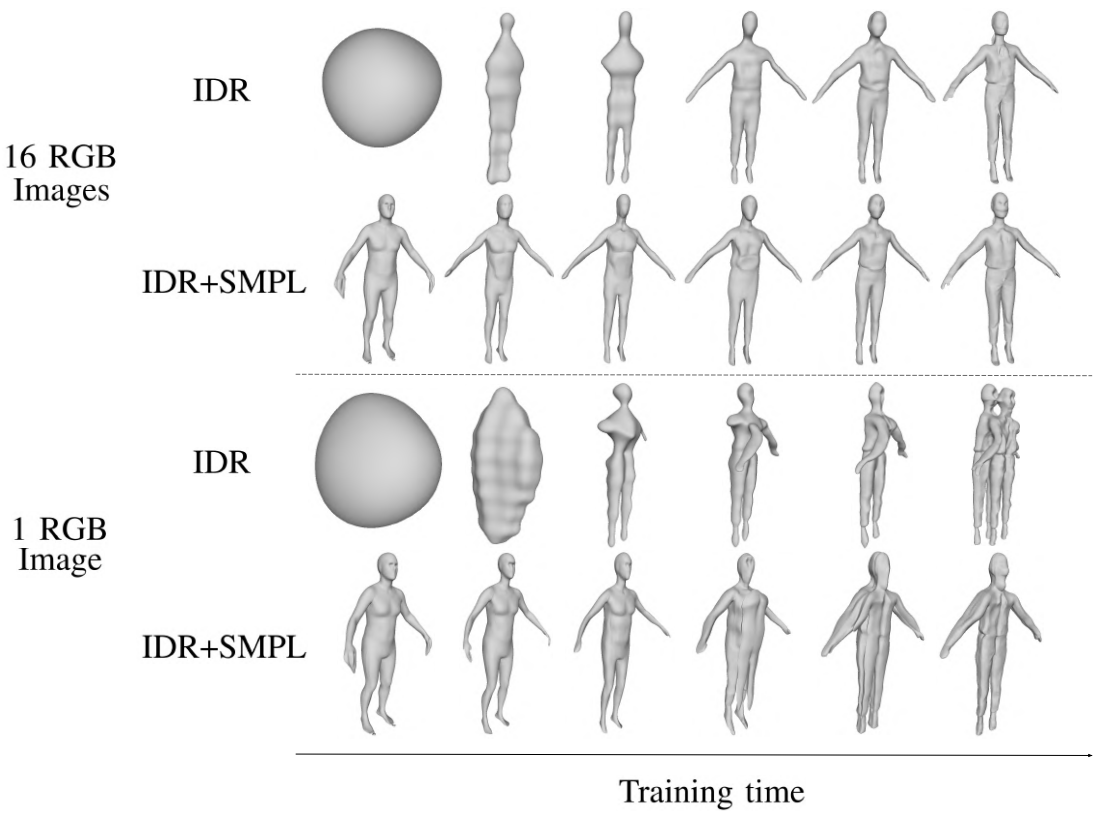


Fig. 16: Evolution of IDR and IDR+SMPL implicit surfaces for (top) 16 and (bottom) 1 RGB images. Using an SMPL model as prior helps with the reconstruction, specially in sparse-view scenarios.

*E. **Exp. 3 results**: IDR and IDR+SMPL convergence speed*

In Figure 15 we present the PSNR and Chamfer evaluation of IDR and IDR+SMPL at different time instants for 32, 4 and 1 views. The P2S distance is not exposed since its behavior is very similar to that of the Chamfer.

In the case of PSNR we observe how using the SMPL model as a prior effectively accelerates the convergence of the model regardless of the number of images, since as we have commented above, the renderer is not as affected by the available views. This improvement of up to 4 dB using the SMPL model is maintained over a long period of time, so the convergence is faster. Nonetheless, it must be taken into account that in the case of IDR+SMPL the SMPL model has been estimated a priori, and IGR has been used to represent it implicitly, which implies an extra time that is not being considered in this experiment.

In the case of the Chamfer distance, the difference between using or not the SMPL model as a prior is substantial, since after a few seconds we already find cases in which the distance is more than 100 times smaller in IDR+SMPL than in IDR. This is because initializing the surface with a model that roughly represents the desired surface greatly reduces the distance between the two. If instead the initialization is a sphere, the system takes much longer to reduce the error, especially in cases with sparse views. Using 32 views the Chamfer distance of IDR ends up converging to a value similar to that of IDR+SMPL (although requiring more time), but when the views are sparse IDR+SMPL presents better results.

An example of the evolution of IDR and IDR+SMPL in the case of 16 and 1 number of images is presented visually in

Experiment	Epoch Step	PSNR (dB) ↑		Chamfer ↓	P2S ↓
		Mean	Std Dev	(cm)	(cm)
IDR	-	25.01	0.53	0.63	0.67
IDR+SMPL	-	24.98	0.65	0.72	0.72
HE	64	**25.66**	1.22	0.71	0.79
HE	32	25.39	0.88	0.75	0.79
HE	16	25.07	1.31	0.69	0.73
HE	8	24.79	0.96	**0.62**	**0.65**
HE	4	24.29	0.46	0.64	0.69
HE	2	24.64	0.88	0.74	0.78

(a) Full-body evaluation

Experiment	Epoch Step	PSNR (dB) ↑		Chamfer ↓	P2S ↓
		Mean	Std Dev	(cm)	(cm)
IDR	-	21.16	0.28	0.45	0.77
IDR+SMPL	-	21.22	0.31	0.51	0.79
HE	64	**22.40**	0.70	0.53	0.89
HE	32	22.17	0.22	0.52	0.85
HE	16	22.05	0.52	0.49	0.81
HE	8	21.63	0.39	**0.43**	**0.69**
HE	4	21.39	1.55	0.53	0.81
HE	2	22.26	0.50	0.55	0.82

(b) Head evaluation

TABLE III: **Exp. 4 results.** Comparison between (a) full-body and (b) head evaluations for the different epoch steps.

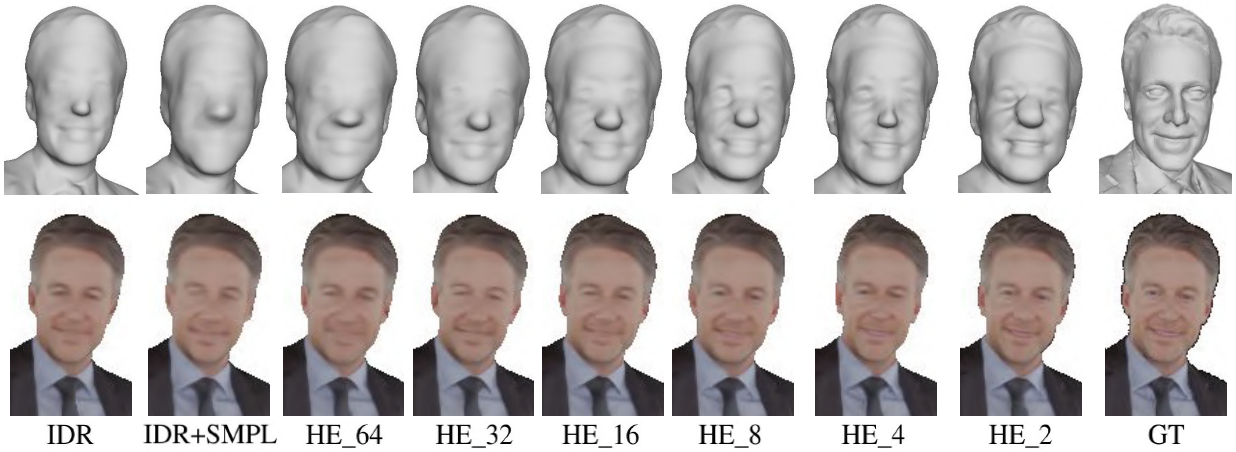


Fig. 17: Head emphasis effect in the (top) 3D reconstruction and (bottom) renderings of Dennis face using 32 views. HE_64 means the attention is performed every 64 iterations (epoch step).

Figure 16. As can be seen, using an SMPL model as prior allows for convergence to the desired shape in less time and with more accuracy. This is especially true in the case of 1 view, as the final IDR reconsturction does not resemble a person, and IDR+SMPL maintains this shape. Note how in the case of 1 view and IDR+SMPL, the shape of the person starting with the SMPL model deteriorates over time, causing the result not to be even better. This could be solved by performing a schedule of the weights that control ImplicitNet, or by freezing its parameters so that IDR would not try to modify them in the first iterations and would focus more on rendering.

F. ***Exp. 4****: Attention mechanism*

Even if we obtain satisfactory 3D reconstructions using many points of view, there are details that IDR does not represent properly, such as those of the face. As mentioned above, we performed some experiments to see if sampling the head more often it would gain definition in the reconstruction and renderings. The corresponding results are shown in Table IIIa. Although minimal improvements in the 3D reconstruction are observed when using an epoch step of 8, the other results do not seem to indicate that there is any considerable improvement in the renderings or 3D reconstruction by emphasizing the head.

However, it should be noted that we have presented the evaluation of the whole body, so it is normal that if we focus more on the face, the other parts of the body lose detail. To check whether this strategy specifically improves the area of the face, we present in Table IIIb the results obtained by evaluating only the head region. In this case, it is observed that the 3D reconstruction is indeed slightly improved in some cases, again with epoch step 8, and the PSNR is improved in all cases.

To better understand the effect of the epoch step in IDR, we show the reconstructed surface and a rendering of each case in Figure 17. Indeed, the more often we sample the face, the more details we obtain both at the appearance and geometry level, being the cases *HE_2* and *HE_4* the most similar to the ground truth (GT). Even so, there does not seem to be a clear improvement at the quantitative level in terms of 3D reconstruction, which we think may be due to the fact that the surface reconstructed by IDR is slightly misaligned with respect to ground truth, so that although visually the result is more defined, this is not reflected in the metrics.

VI. Conclusions

In this work, we present a 3D people reconstruction system that effectively combines parametric and non-parametric models. For this purpose, we make use of an emerging technology in the field of computer vision: implicit neural representation with differentiable rendering, which allows for reconstructing 3D geometry with only 2D weak supervision. We propose to combine Implicit Differentiable Renderer (IDR), one of the state-of-the-art architectures that takes advantage of this emerging technology, with the parametric Skinned Multi-person Linear Model (SMPL), and explore whether we can obtain highly detailed 3D reconstructions of people with our system.

To do so, we use Implicit Geometric Regularization (IGR) to initialize the parameters of the IDR implicit network so that they represent the SMPL model instead of a sphere. At this point, we prepared a dataset consisting of two women and two men from the Renderpeople dataset in order to experiment with them. Using different numbers of viewpoints, we managed to study the behavior of both architectures, IDR and our system (IDR+SMPL), in a variety of scenarios with different levels of difficulty, evaluating the PSNR (dB), Chamfer and Point-to-Surface (P2S) distances (cm).

The first results show us that using this parametric model as a prior we achieve a considerable improvement in 3D reconstructions and renderings, and even more in situations with sparse camera views. Not only that, but our system converges

faster to these results, although the previous steps to obtain the SMPL model and represent it implicitly with IGR must be taken into account. Starting from the SMPL model, our system is even able to roughly reconstruct parts that are not seen from any point of view, because although IDR cannot reconstruct what it does not see, initializing the surface so that it is aligned to the person allows to obtain an approximate reconstruction.

On the other hand, we have performed experiments with subsets in casual and A-poses, which show that our system performs better when people are in simple A-poses, as the estimated SMPL model is more accurate in these cases. Casual poses present self-occlusions that complicate the task, so vanilla IDR also performs better in the A-pose subset, since the surface is seen from most viewpoints.

Finally, we have explored whether giving emphasis to the parts of the body with less detail, i.e., the face, we can improve the IDR reconstruction. To do so, we decided to sample this area more often, and although quantitatively there does not seem to be significant improvements, it is visually observed that more details are obtained both in the renderings and in the 3D reconstruction. This difference in results may be due to a misalignment of the surface, which makes the metrics obtained not so good. In short, this last experiment has helped us to deduce that IDR can improve its results if it focuses automatically on the parts with less details, as we propose as a future work.

VII. Future work

Although we have concluded the thesis with the expected results and corroborating the hypotheses formulated, we would like to propose a number of improvements for future systems based on ours, to further improve 3D reconstructions of people with IDR.

To begin with, as mentioned above, IGR produces artifacts when representing a point cloud implicitly, which affects our system, because IDR must make an extra effort to correct these errors. IGR optimizes its parameters so that the implicit surface (modeled by a signed distance function) is contained in the point cloud, but does not take into account what is outside this cloud. Therefore, we propose to sample points on a uniform grid of the 3D space and add an extra loss that forces the IGR output for each 3D point to be equivalent to the distance from the point to the point cloud of interest. In this way, we understand that the surface will be forced to be zero outside the input point cloud.

On the other hand, it seems that in situations with very sparse camera views, although our system starts with a SMPL model very close to the person, IDR modifies this surface deteriorating it by not having enough points of view. A possible solution would be to freeze the parameters of the IDR geometry network during the first stage of the training, in order to adapt the parameters of the rendering network keeping the same surface, which we already know a priori that it is relatively accurate.

Finally, we propose a more advanced strategy of the attention mechanism explained in this work, so that IDR automatically detects areas of the body with less detail and emphasizes to improve them. This idea is related to [70], whose work is based on quantifying the uncertainty in 3D implicit representations. In our case, we propose to modify IDR so that, instead of producing an RGB color output from a point in 3D space (obtained by intersecting a ray and the implicit surface), two values are produced: the mean and the variance of the color. In this way, IDR could identify which points have more variance, i.e., are more uncertain, and devote more attention to improve them. We are currently working on this new project and hope to improve the reconstruction of 3D people even further.

Acknowledgment

I would like to thank Francesc and Enric, for the time and effort they have dedicated to the project and for their innumerable advice. I look forward to continuing to work with them to achieve my goals.

I would also like to thank my academic supervisor: Xavier, who has helped me in everything I have needed and has shown great interest in both the work done and my learning.

Last but not least, thanks to my family, partner, friends and classmates for the support they have given me, thanks to which I have been able to meet my goals and enjoy this project.

References

[1] Zhi-Quan Cheng, Yin Chen, Ralph Martin, Tong Wu, and Zhan Song. Parametric modeling of 3d human body shape—a survey. *Computers Graphics*, 71, 12 2017.

[2] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans, 12 2020.

[3] M. Loper, Naureen Mahmood, J. Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34:248:1–248:16, 2015.

[4] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33, 2020.

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

[6] Frank Dellaert and Lin Yen-Chen. Neural volume rendering: Nerf and beyond, 12 2020.

[7] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2004.

[8] S.M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 519–528, 2006.

[9] S. E. Chen and L. Williams. View interpolation for image synthesis. *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, 1993.

[10] R.T. Collins. A space-sweep approach to true multi-image matching. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 358–363, 1996.

[11] Steven Seitz and Charles Dyer. View morphing. *SIGGRAPH Conf Proc*, 09 1999.

[12] B. Curless and M. Levoy. A volumetric method for building complex models from range images. *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996.

[13] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*, SGP '06, page 61–70, Goslar, DEU, 2006. Eurographics Association.

[14] J. Bonet. Poxels: Probabilistic voxelized volume reconstruction. 1999.

[15] M. Agrawal and L.S. Davis. A probabilistic framework for surface reconstruction from multiple images. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II, 2001.

[16] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. 04 2015.

[17] Wenjie Luo, A. Schwing, and R. Urtasun. Efficient deep learning for stereo matching. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5695–5703, 2016.

[18] Wilfried Hartmann, Silvano Galliani, Michal Havlena, and Luc Van Gool. Learned multi-patch similarity. pages 1595–1603, 10 2017.

[19] G. Riegler, Ali O. Ulusoy, H. Bischof, and Andreas Geiger. Octnetfusion: Learning depth fusion from data. *2017 International Conference on 3D Vision (3DV)*, pages 57–66, 2017.

[20] Simon Donne and Andreas Geiger. Learning non-volumetric depth fusion using successive reprojections. pages 7626–7635, 06 2019.

[21] Po-Han Huang, K. Matzen, J. Kopf, N. Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018.

[22] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. *ArXiv*, abs/1804.02505, 2018.

[23] Onur Ozyesil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A survey on structure from motion. *Acta Numerica*, 26, 01 2017.

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

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

[26] William Lorensen and Harvey Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM SIGGRAPH Computer Graphics*, 21:163–, 08 1987.

[27] Haoqiang Fan, Hao Su, and Leonidas Guibas. A point set generation network for 3d object reconstruction from a single image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2463–2471, 2017.

[28] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1679–1688, 2020.

[29] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, W. Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 2018.

[30] Chao Wen, Yinda Zhang, Zhuwen Li, and Yanwei Fu. Pixel2mesh++: Multi-view 3d mesh generation via deformation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1042–1051, 2019.

[31] V. Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, G. Wetzstein, and M. Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2441, 2019.

[32] Shubham Tulsiani, Tinghui Zhou, Alyosha Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1, 02 2019.

[33] Maxim Tatarchenko, A. Dosovitskiy, and T. Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2107–2115, 2017.

[34] Matan Atzmon, Niv Haim, Lior Yariv, Ofer Israelov, Haggai Maron, and Y. Lipman. Controlling neural level sets. In *NeurIPS*, 2019.

[35] Mateusz Michalkiewicz, J. K. Pontes, Dominic Jack, Mahsa Baktash, and Anders P. Eriksson. Implicit surface representations as layers in neural networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4742–4751, 2019.

[36] Jeong Joon Park, Peter R. Florence, J. Straub, Richard A. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019.

[37] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2304–2314, 2019.

[38] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, R. Mech, and U. Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *NeurIPS*, 2019.

[39] Michael Oechsle, Lars M. Mescheder, Michael Niemeyer, Thilo Strauss, and Andreas Geiger. Texture fields: Learning texture representations in function space. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4530–4539, 2019.

[40] V. Sitzmann, Michael Zollhoefer, and G. Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *NeurIPS*, 2019.

[41] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[42] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[43] Dragomir Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and James E. Davis. Scape: shape completion and animation of people. *ACM Trans. Graph.*, 24:408–416, 2005.

[44] H. Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018.

[45] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[46] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *CVPR*, 2021.

[47] Christoph Lassner, J. Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and P. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4704–4713, 2017.

[48] Nikos Kolotouros, G. Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2252–2261, 2019.

[49] Riza Alp Güler and I. Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10876–10886, 2019.

[50] Yu Rong, Takaaki Shiratori, and H. Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *ArXiv*, abs/2008.08324, 2020.

[51] Thiemo Alldieck, M. Magnor, Bharat Lal Bhatnagar, C. Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1175–1186, 2019.

[52] Qianli Ma, Jinlong Yang, A. Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3d people in generative clothing. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6468–6477, 2020.

[53] Hayato Onizuka, Zehra Hayirci, Diego Thomas, A. Sugimoto, H. Uchiyama, and Rin ichiro Taniguchi. Tetratsdf: 3d human reconstruction from a single image with a tetrahedral outer shell. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6010–6019, 2020.

[54] R. Natsume, S. Saito, Zeng Huang, Weikai Chen, Chongyang Ma, H. Li, and S. Morishima. Siclope: Silhouette-based clothed people. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4475–4485, 2019.

[55] Albert Pumarola, Jordi Sanchez, G. Choi, A. Sanfeliu, and F. Moreno-Noguer. 3dpeople: Modeling the geometry of dressed humans. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2019.

[56] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, C. Schmid, and Grégory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2232–2241, 2019.

[57] S. Saito, Tomas Simon, Jason M. Saragih, and H. Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 81–90, 2020.

[58] Gül Varol, Duygu Ceylan, Bryan C. Russell, Jimei Yang, Ersin Yumer, I. Laptev, and C. Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *ECCV*, 2018.

[59] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7738–7748, 2019.

[60] Zeng Huang, Yuanlu Xu, Christoph Lassner, H. Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3090–3099, 2020.

[61] Bharat Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. *Combining Implicit Function Learning and Parametric Models for 3D Human Reconstruction*, pages 311–329. 11 2020.

[62] Lukas Lamprecht, Mark Geilhausen, Dennis Siebertz, David Tomicic, Niklas Köhler-Prediger, Tim Kluthe, Andreas Köhler, Sedef Avcioglu, and Johannes Weber. Renderpeople, https://renderpeople.com, 2018.

[63] Blender Online Community. *Blender - a 3D modelling and rendering package, http://www.blender.org*. Blender Foundation, Blender Institute, Amsterdam, 2021.

[64] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of Machine Learning and Systems 2020*, pages 3569–3579. 2020.

[65] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017.

[66] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017.

[67] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. *arXiv preprint arXiv:2004.03686*, 2020.

[68] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*, 2021.

[69] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. pages 2562–2571, 06 2020.

[70] Jianxiong Shen, Adria Ruiz, Antonio Agudo, and Francesc Moreno. Stochastic neural radiance fields:quantifying uncertainty in implicit 3d representations, 2021.