University of Beira Interior

Department of Computer Science

FACULDADE ENGENHARIA

Departamento de Informática

Neural Radiance Fields Meet Biometrics: Is it a Match?

Prepared by:

Rui Pedro Lourenço Gil

Advisor:

Professor Hugo Proença, PhD

July 10, 2025

Acknowledgements

The completion of this work, as well as most of my academic life, would not have been possible without the help of many people. I am grateful to my family, who always supported and encouraged me to follow my dreams and helped me stay focused and motivated during the most difficult times. I would like to thank my supervisor, Professor Hugo Proença, for his guidance and support throughout the development of this work. I would also like to thank my classmates, who helped me learn and grow as a student and as a person. In particular, I would like to thank my friends **Gonçalo Caixeiro** and **Maria José Teixeira** for their tremendous support in carrying out this project, always making time to help in any way they could. I would also like to thank my friend **Leonardo Gonçalves**, who helped me discover the best way to carry out this project. Without his help, I would most likely not have been able to successfully complete this major work within the proposed time frame.

I would also like to thank all the people who volunteered to take part in the study of this proposed topic, namely my friends from the course, my parents and my sister.

Last but not least, I would like to thank all the teachers and staff at the University of Beira Interior, who helped me achieve my academic and professional goals. In particular, I would like to thank Professor **João Neves**, who provided me with a computer in the laboratory that allowed me to carry out all the necessary tasks for this project, and also to the ***researchers*** in the laboratory, who were always very attentive and available for anything I needed.

Contents

List of Figures

List of Tables

Acrónimos

AI	Artificial Intelligence
SF	Surface Field
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
MLP	Multi-Layer Perceptrons
LBS	Linear Blend Skinning
PDF	Portable Document Format
FID	Fréchet Inception Distance
RGB	Red Green Blue
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity Index Measure
MSE	Mean Squared Error
YOLO	You Only Look Once
CLIP	Contrastive Language-Image Pre-training
CUDA	Compute Unified Device Architecture
DINOv2	Self-Distillation with No Labels v2
CAPE	Clothed Auto-Person Encoding
JSON	JavaScript Object Notation
GNARF	Generative Neural Articulated Radiance Field
SPIN	*SMPL-based Pose Inference Network*
SMPL	*Skinned Multi-Person Linear Model*
NeRF	Neural Radiance Fields
SE(3)	Special Euclidean group in 3D
EG3D	Efficient Geometry-aware 3D Generative Adversarial Networks
FaceNet	Face Embedding Network
NARFs	Neural Articulated Radiance Fields
NeRFs	Neural Radiance Fields
A-NeRF	Articulated Neural Radiance Fields
ID-NeRF	Indirect Diffusion-guided Neural Radiance Fields
SNARFs	Skinned Neural Articulated Radiance Fields
A-NeRFs	Articulated Neural Radiance Fields
GnARFs	Generalizable Neural Articulated Radiance Fields
LPIPS	Learned Perceptual Image Patch Similarity
SMPL+D	Skinned Multi-Person Linear model with Displacements
StyleGAN2	Style-based Generative Adversarial Network v2

Mip-NeRF	Multiscale Sampling for Anti-Aliased Neural Radiance Fields
SURREAL	Synthetic hUmans foR REAL tasks
FLAME	Faces Learned with an Articulated Model and Expressions
ArcFace	Additive Angular Margin Loss for Face Recognition
COLMAP	Structure-from-Motion e Multi-View Stereo pipeline para reconstrução 3D
HumanNeRF	Human-centric Neural Radiance Fields
ENARF-GAN	Efficient Neural Articulated Radiance Fields GAN
fused-MLP	Fused Multi-Layer Perceptron
Instant-NGP	Instant Neural Graphics Primitives
PCKh	Percentage of Correct Keypoints (normalizado pelo tamanho da cabeça)

Chapter

1

Introduction

1.1 Introduction

Recent developments in three-dimensional generative models have brought new possibilities in how we represent and manipulate visual entities in virtual environments. Among these innovations, Neural Radiance Fields (NeRFs) [1] stand out as one of the most promising approaches for reconstructing 3D scenes from 2D images, enabling photorealistic synthesis of new views with high geometric and visual fidelity. However, the original formulation of NeRFs assumes that the scene to be modelled is static, which represents a substantial limitation when it comes to articulated subjects, such as the human body. This type of content presents significant structural variations due to joint movement, requiring models capable of capturing both visual appearance and deformations caused by poses. To address this limitation, so-called Articulated Neural Radiance Fields (A-NeRFs) have emerged, a class of models that extends traditional NeRFs by introducing deformation mechanisms based on articulated skeletons. Typically, these methods use parametric structures such as the *Skinned Multi-Person Linear Model* (SMPL) model [2] and deformation techniques such as Linear Blend Skinning (LBS) to map bodies in different poses to a neutral canonical space (*neutral body*). Through this mapping, it becomes possible to dissociate information related to shape (identity) from that related to pose, allowing for the photorealistic reconstruction of human subjects in multiple body configurations. However, despite the visually impressive results demonstrated by many recent works, Articulated Neural Radiance Fields (A-NeRF) methods present significant practical challenges.

In general, they require highly structured datasets with multiple views per pose, accurate body joint estimates, and complex pre-processing and calibration pipelines. These requirements make their application to customised data — especially in real-world contexts with limited resources — impractical, compromising the scalability and practical utility of these methods. In this context, this paper proposes to address a specific and highly relevant question: is it possible to verify whether an image or video corresponds to a particular articulated model, trained to represent a specific person? In other words, can an automatic system determine whether a given visual content was generated by a specific three-dimensional neural model or not? This challenge falls within a critical area at the intersection of biometrics and generative models: that of identity verification and attribution of origin in content generated by neural networks. As models such as A-NeRFs become more realistic, there is a growing need for mechanisms capable of authenticating their provenance and validating the link between the content and the identity represented. Such mechanisms are fundamental not only for security and identity protection purposes, but also for intellectual property control and detection of visual forgeries (deepfakes).

1.2 Project Objectives

Thus, the central objectives of this project are:

- Develop a verification methodology capable of determining whether an image or video may have been generated by a specific articulated model;
- Analyse the structure and properties of 3D model outputs, with a view to identifying patterns characteristic of each identity;
- Propose a mechanism for matching test images and neural models, allowing the represented identity to be inferred;
- Contribute to the advancement of authentication techniques for content generated by artificial intelligence, with applications in biometrics, digital security, and intellectual property protection.

The approach followed in this project is inspired by the principles of A-NeRFs, taking advantage of their potential for highly realistic representation of human beings. However, a more practical and accessible strategy is adopted, adapted to real data collected under controlled but simple conditions, in order to ensure that the proposed pipeline is feasible and scalable. The basic idea is to train a 3D model per individual, based on a short video captured in a neutral pose (typically a T-pose), and then test the system's ability to recognise that identity from additional images, captured in different poses and contexts. Within this framework, this work positions itself as applied research on identity verification in three-dimensional generative models. Unlike approaches focused on visual quality or re-rendering, here the focus is on the connection between the trained model and the underlying identity, evaluated in a quantitative and interpretable way.

1.3 Organisation of the Report

This report is organised as follows: Section 2.2 presents the theoretical framework and the main works related to NeRFs, A-NeRFs and biometrics. Section 3.1 describes in detail the method followed in this project. Sections 4.1 and 4.1.8 present the experiments carried out, as well as the analysis of the results obtained. Finally, Sections 5.1 and 5.2 discuss the main conclusions and prospects for future work.

Chapter

2

State of the Art

2.1 Introduction

In this chapter, we will conduct an in-depth analysis of the main methods of articulated NeRFs, which represent a significant advance in the modelling and rendering of objects with moving parts, such as the human body. The evolution from traditional NeRFs to articulated variants has made it possible to overcome limitations related to the representation of complex deformations and generalisation to multiple poses and identities. The most relevant methods developed in recent years will be presented, highlighting their characteristics, innovations, and applications, in order to contextualise the state of the art and justify the choice of approach used in this project.

2.2 Related Work

There are many articulated Neural Radiance Fields (NeRF) methods, which, although very similar and often having equivalent objectives, each have their own purposes and characteristics that differ from A-NeRFs.

NARFs

Neural Articulated Radiance Fields (NARFs) [3] were introduced in 2021 by Atsuhiro Noguchi. They are an approach that extends NeRFs to model articulated objects such as human bodies, robots, or mechanical structures. Unlike traditional NeRFs, which are limited to static scenes, NARFs incorporate articulated deformations, allowing the rendering of realistic images under new poses and viewpoints. The main contributions of NARFs include: **Explicit Articulation**, using rigid transformations based on direct kinematics to model moving parts (bones/joints) and a part selector via lightweight sub-networks that identifies the most relevant articulated region for each 3D point; **Computational Efficiency**, where NARF$_\text{D}$ (*Disentangled NARF*) solves the problem of high computational cost of previous solutions (such as *Part-Wise NARF*) by making the process into a single model; **Factor Disentanglement**, allows independent control over pose, appearance, viewpoint, and bone parameters for applications such as character reanimation; **Flexibility for Unconventional Anatomies** automatically learns skinning weights to adapt to complex deformations (obesity, amputations) and non-human objects (robots), unlike methods based on SMPL. In addition, NARFs are also designed to be compatible with pose estimation pipelines, and can work with automatically generated data (e.g., via *SMPL-based Pose Inference Network* (SPIN) [4]). Its modular approach allows it to be coupled with different forms of input and facilitates tasks such as *retargeting* and procedural animation. The rendering architecture continues to be based on Multi-Layer Perceptrons (MLP)s with volumetric sampling (*volume rendering*) with integration of Special Euclidean group in 3D (SE(3))-based deformations for each body part. Among the applications of NARFs are image synthesis for robotics, video editing (*motion transfer*), Artificial Intelligence (AI) training with synthetic data, and modelling of non-standard human anatomy for personalised medicine. The main challenges include the need for pose annotations during the training phase (although it can be used in 3D estimators) and the current limitation of assuming rigid parts, where non-rigid deformations (such as loose clothing) remain an open problem. In conclusion, NARFs represent a significant advance in implicit 3D modelling of articulated objects, combining the fidelity of NeRFs with the flexibility of joints. Future research directions include integration with automatic pose estimation and extensions for non-rigid deformations.

GnARFs

Generalizable Neural Articulated Radiance Fields (GnARFs) [5] were first introduced in 2023 by Rakesh Goel, extending traditional NeRFs into articulated NeRFs generalisable to multiple subjects and poses, without training a new model for each person. They are characterised by the fact that they are based on an articulated skeleton and skinning, just like A-NeRFs, introduce an identity-conditioned model, which allows the same model to be reused for different people, use two-stage architecture, in the mapping of deformed points to canonical space with deformation based on pose, and in rendering via identity-conditioned NeRF. These networks are generally applied to humans, where a single model is generalisable to the point of being useful for several different people. This is a very important advance in relation to automatic personalisation and the use of NeRFs for humans, for example, for augmented reality applications. GnARFs use explicit identity conditioning through learned embeddings for each person, allowing the separation of shape factors from pose. This reduces the number of models needed for different subjects and increases the scalability of the system. The architecture also allows for efficient multi-subject learning using datasets such as Human3.6M, ZJU-Mocap, and People-Snapshot [6, 7, 8]. The authors report competitive performance in Peak Signal-to-Noise Ratio (PSNR) and Learned Perceptual Image Patch Similarity (LPIPS), approaching customised methods with much less training time. In addition, deformation is regularised by pose-canonical-pose consistency cycles. The architecture of GnARFs features several notable technical innovations. The first stage of the system employs a spherical coordinate-based deformation module that transforms points from the observed space to the canonical space, using a hierarchical representation of human joints. This process is guided by a neural skinning scheme that automatically learns the influence weights for each joint, overcoming the limitations of traditional LBS-based methods.

In the second stage, the model uses a conditioned NeRF that incorporates three main components: (1) an identity encoder that maps anthropometric features to a low-dimensional latent space, (2) a shared appearance module that captures textures and materials common to all people, and (3) a specialised volumetric decoder that synthesises the specific geometry and appearance for each individual. This explicit separation of concerns allows the model to generalise to new subjects with only a few reference views. A crucial aspect of GnARFs is the identity embedding learning mechanism. Unlike previous approaches that required separate optimisation for each new individual, GnARFs learn a continuous latent space of identities during training. This space is structured so that simple linear operations can generate realistic variations in body shape, allowing the creation of virtual characters with customised physical characteristics through simple combinations of latent vectors. For practical applications, GnARFs offer several advantages. In virtual reality scenarios, they enable the real-time generation of high-fidelity customised avatars. In the medical field, they facilitate the creation of patient-specific body models for surgical planning. In addition, their modular design makes it possible to integrate them with existing motion capture pipelines, opening up new possibilities for digital content production and visual effects.

GNARF

Generative Neural Articulated Radiance Field (GNARF) (Generative Neural Articulated Radiance Fields) [9] is an innovative framework for generating editable 3D radiance fields of human bodies and faces using Generative Adversarial Network (GAN)s 3D. Main contributions: **Editable 3D Generation**: GNARF is the first method to generate high-quality 3D radiance fields for human bodies, allowing for the editing of poses and facial expressions. Objects are generated in a canonical pose and explicitly deformed to desired poses using a surface-based (Surface Field (SF)) deformation field. This model is based on Style-based Generative Adversarial Network v2 (StyleGAN2) and is capable of generating high-resolution triplanes that are then deformed via implicit fields. The final rendering is done via traditional volumetric ray-marching. For human bodies, it uses the SMPL model and for faces, the Faces Learned with an Articulated Model and Expressions (FLAME) model [10], ensuring structured control of joints and expressions. The results obtained in the Synthetic hUmans foR REAL tasks (SURREAL) dataset show better visual fidelity (measured by Fréchet Inception Distance (FID)) and better pose-to-pose consistency (measured by Percentage of Correct Keypoints (normalizado pelo tamanho da cabeça) (PCKh)) than previous models such as Efficient Geometry-aware 3D Generative Adversarial Networks (EG3D) or Efficient Neural Articulated Radiance Fields GAN (ENARF-GAN). The GNARF architecture presents several notable technical innovations in the field of generative 3D human generation. The system combines three main components: a tri-plane-based generator that produces detailed volumetric representations, a neural deformation module that preserves geometric consistency during animation, and a specialised discriminator that evaluates both the visual quality and physical plausibility of the generations. The generation process begins with the synthesis of a canonical radiance field using an architecture derived from StyleGAN2, but adapted to operate in 3D space through features stored in three orthogonal planes (tri-planes). This representation allows for the efficient manipulation of attributes such as body shape, skin texture, and clothing through operations in latent space. For animation, GNARFs employ an implicit deformation field that transforms the canonical space into specific poses. This module is trained in a self-supervised manner using pairs of poses from SMPL/FLAME as reference, learning to preserve the geometric integrity of complex regions such as joints and soft tissues during extreme movements. Current limitations include difficulty in modelling extremely loose clothing and a high computational cost for ultra-high-resolution generations (above 1024x1024 pixels). Future research directions include integration with language models for text-prompt-based generation and compression techniques for real-time execution on mobile devices.

SNARFs

A method introduced in 2021 by researchers at the University of Berkeley and Google Research, Skinned Neural Articulated Radiance Fields (SNARFs) [11] , aims to improve deformation between poses without relying on LBS, using an implicit neural function that learns the correspondences between the deformed space (pose) and the canonical space, thus allowing non-linear deformations to be captured. It is particularly useful in areas of loose clothing or extreme joint movements. The technical innovation of SNARFs lies in the use of an implicit learned correspondence function (*correspondence field*) that relates coordinates in the deformed space to those in the canonical space. This correspondence is estimated using a differentiable neural network, with refinement via *backpropagation*, making the system robust to highly non-linear poses. The method achieves better results on datasets such as Clothed Auto-Person Encoding (CAPE) [12] and Skinned Multi-Person Linear model with Displacements (SMPL+D) (with clothing), demonstrating superiority over the original A-NeRF in cases of severe deformation. The architecture of SNARFs presents significant advances in modelling non-rigid deformations through four main components: (1) an implicit matching module that maps 3D points between different pose spaces, (2) a geometric regulariser that preserves surface integrity during extreme deformations, (3) a spatial attention mechanism that identifies critical regions for deformation, and (4) a hierarchical optimiser that progressively adjusts the network parameters. Experimental results show that SNARFs reduce reconstruction error by 37% compared to A-NeRF on the CAPE dataset, particularly in poses with crossed arms or raised legs. The approach also demonstrates excellent generalisation to new subjects not seen during training, requiring only 15 minutes of fine-tuning per individual. Future research directions include integration with generative models for avatar synthesis and compression techniques for real-time applications. The SNARFs approach represents an important milestone in the evolution of articulated NeRFs, overcoming fundamental limitations of traditional *skinning*-based methods.

HumanNeRFs

Introduced in 2022 by Chung-Yi Weng and colleagues at the University of Washington, Human-centric Neural Radiance Fields (HumanNeRF)s [13] is a method that focuses on the reconstruction of humans with multi-view supervision, capturing humans in motion through monocular videos without the need for multiple cameras, making it ideal for augmented reality applications. Its main features include the use of a pipeline based on SMPL to extract pose and shape, an implicit representation focused on NeRFs that is conditioned on pose and frame, and supervised training with Red Green Blue (RGB) video and human segmentations. The method incorporates supervision based on 2D segmentation and uses ray sampling with temporal regularisation to maintain consistency between consecutive frames. Unlike multi-view methods such as NeuralBody, HumanNeRF achieves competitive results with only one camera. Furthermore, it is possible to generate realistic-quality full-body reconstructions even under pose noise, making it promising for interactive avatars and reconstruction from separated data. Deformation between poses is measured by a learned transformation field that maps the observed pose points to a canonical reference space. This canonical space is crucial to ensure that body identity details are preserved over time, even when there are large variations in pose or movement. The system uses volumetric ray-marching with sampling based on the skeletal structure, where each sampled 3D point is conditioned on both the current frame's pose and the individual's global identity. One of the highlights of HumanNeRF is its **light supervision** approach, which requires only RGB video and 2D segmentations — these can be automatically generated by pre-trained segmenters such as **Mask R-CNN** [14]. Temporal regularisation is introduced as part of the loss function, promoting consistency between adjacent frames even in regions of partial occlusion or sudden pose variation. In addition, the model is capable of generating photorealistic novel views, reconstructing the subject at angles not seen during training, which highlights the quality of the learned representation. The architecture also includes a skeletal pose update mechanism, refining the initial estimates of *SMPL* during training, resulting in greater fidelity of the reconstructed geometry in critical regions such as joints and face. In summary, HumanNeRF represents a practical and technical advance in the reconstruction of humans from articulated NeRFs. Its ability to operate with simple data (video + 2D segmentation) and still achieve realistic results makes it a promising solution for real-time human digitisation and immersive applications in uncontrolled environments. The limitations lie mainly in the dependence on SMPL, which can compromise fidelity in cases of bulky clothing or unmodelled accessories, and in the need for relatively accurate segmentations to avoid background contamination in the reconstruction.

A-NeRFs

A-NeRFs [15] were introduced in 2021 by Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin, students at the University of British Columbia. They are distinguished by their main characteristics, which are an articulated volumetric human model that uses an explicit skeleton to guide the volumetric deformation of the human body, constructing a neural representation where each 3D point is expressed in relation to each bone, creating a parameterisation that facilitates inverse mapping (pose -> canonical). It contains canonical space and pose refinement, in which the neutral canonical space is learned, against which the cable deformations are learned, and in relation to pose refinement, the method allows for the refinement of estimated skeletal poses (without the need for 3D supervision), enabling an improvement in volumetric alignment. Training is very efficient, being performed with multiple cameras and monocular videos, and can thus be trained with monocular or multi-view images. It integrates a rendering loop (volumetric ray-marching) and photometric loss to optimise both geometry and pose. In addition to its structural innovations, A-NeRF introduces a differentiated approach to spatial deformation based on skeletal kinematics, where each *voxel* of the volumetric representation is associated with a transformation conditioned on the pose of the frame. This deformation is performed continuously and differentiably, which allows it to be directly integrated into the optimisation pipeline. The model therefore learns to reconstruct the shape of the human body in a canonical 'pose-free' space and applies deformations at render time to generate images with the desired poses. During training, the system does not require accurate 3D pose information as explicit supervision. Instead, it can use pose estimates from pre-existing models or even reconstruct them implicitly based on multiple views or consecutive frames. This makes A-NeRF more flexible than methods that are rigidly dependent on precise labels, as is the case with some strongly supervised SMPL-based architectures. In the specific context of identity verification or biometrics, A-NeRFs have a critical limitation: although they preserve the overall shape and texture of the body well, they do not guarantee the reliable maintenance of fine facial details that are crucial for distinguishing identities. As the model performs pose-conditional volumetric deformations, slight variations in facial geometry — such as expression or head tilt — can be smoothed or distorted, compromising the consistency of biometric features. Additionally, since A-NeRF does not integrate explicit identity verification or matching mechanisms, the association between a trained model and an input image requires external solutions, such as visual embeddings, which limits it as a standalone framework for recognition tasks.

Nerfacto

Nerfacto Neural Radiance Fields with Factored Components (Nerfacto) is an advanced method for static 3D reconstruction and rendering that combines the expressiveness of NeRFs with efficient light decomposition techniques, enabling realistic and fast renderings from 2D images. This framework proposes a hybrid architecture that merges implicit representations with factored components to separately model the geometry, appearance, and lighting of a scene. Main contributions: Explicit separation of lighting and appearance: Nerfacto introduces an explicit decomposition of the factors that make up the image, allowing specular reflections, diffuse shading, and surface texture to be modelled separately. This results in more interpretable reconstructions and facilitates applications such as relighting, material editing, and appearance transfer. Efficient rendering: Unlike traditional NeRFs that require extensive volumetric sampling and deep networks, Nerfacto uses compact representations and acceleration with hierarchical ray-marching, making the process faster and more feasible for practical use. The scene geometry is implicitly represented by a lightweight neural network that models opacity and density field, while the appearance is encoded in feature fields conditioned by view direction and lighting. Technical basis and architecture: Nerfacto's architecture is built on the principles of NeRFs, but introduces a new form of neural conditioning based on factorised components. The model incorporates a volumetric density field that determines the geometry of the scene, a surface-oriented reflectance field, and a separate lighting model that learns to estimate the lighting conditions in the scene. This approach allows both diffuse and reflective materials to be represented with high visual fidelity. Experimental results: Nerfacto was evaluated on benchmarks such as the RealEstate10K and DTU datasets [16], [17], showing better visual quality (measured by PSNR and LPIPS) and shorter inference time when compared to methods such as NeRF, Multiscale Sampling for Anti-Aliased Neural Radiance Fields (Mip-NeRF), and PlenOctrees [18], [19]. The separation between light and material also allows for realistic re-illumination of static scenes, something that traditional NeRFs cannot achieve without complete retraining. Limitations and future directions: limitations include the difficulty in dealing with highly specular or translucent scenes, and the need for multiple images with good angular coverage of the scene for high-fidelity reconstruction. Future work includes integration with depth sensors for initial geometry optimisation, zero-shot learning in unknown environments, and compression of factorised models for real-time execution in augmented reality and mobile gaming applications.

Visual Representations for Biometrics

Facial recognition models based on latent embeddings play a central role in modern identity verification systems. Classics such as Face Embedding Network (FaceNet) [20] and Additive Angular Margin Loss for Face Recognition (ArcFace) [21] project faces onto a vector space where distance reflects biometric similarity. More recently, multimodal and self-supervised approaches such as Contrastive Language-Image Pre-training (CLIP) [22] and Self-Distillation with No Labels v2 (DINOv2) [23] have emerged, which capture both appearance and visual semantics. These representations, although not developed specifically for NeRF content, offer an effective means of comparison between real images and 3D renderings, and are therefore relevant in this context.

CLIP and Visual Semantic Representations

The CLIP (Contrastive Language-Image Pretraining) model [22] introduces a powerful approach for jointly mapping images and text into a shared embedding space. Instead of using conventional discriminative networks, CLIP is trained with millions of image-text pairs, allowing it to learn highly robust semantic representations even without explicit labels. In the context of NeRFs, CLIP has been used as a mechanism for comparing identities between images and renderings. In particular, it becomes useful when one wants to recognise a person in different poses — by mapping real images and renderings to the same vector space, one can calculate similarity (e.g., via dot product or cosine distance) and identify the most likely match. This method is especially advantageous in contexts where variations in pose, lighting, or background make pixel-to-pixel matching difficult.

DINOv2 and Self-supervised Visual Embeddings

DINOv2 [23] is a self-supervised model that learns visual representations without requiring explicit annotations. Unlike CLIP, which uses text supervision, DINOv2 relies on self-distillation and attention mechanisms, learning to extract robust embeddings that preserve high-level structure and semantics. These embeddings have proven effective in tasks such as classification, segmentation, and semantic matching, and are also applicable to comparing input images and NeRF renderings for recognition purposes, as an alternative to CLIP. The main advantage of DINOv2 lies in its ability to capture fine details of appearance without relying on human annotations, making it suitable for scenarios with unlabelled data.

Text2Avatar

The *Text2Avatar* method [24] proposes an innovative approach for generating 3D avatars from textual descriptions, combining generative networks (GAN) with a system based on discrete *codebooks* and the CLIP model. The architecture consists of three main modules: a 3D generator (*G(·)*), a segmentation module (*Seg(·)*), and a multimodal encoder (*E(·)*). The generator receives camera parameters (*ptext*) and a latent code (*z pz*) as input, where *ptext* includes shape (β), pose (θ) and perspective (ξ) parameters. The multimodal encoder integrates a text encoder *Etext(·)*, an image encoder *Eimg(·)*, an attribute mapping network *M(·)*, and a predefined textual library. This library contains descriptions of attributes such as gender, shirt colour or trouser type, enabling precise alignment between text and visual features through CLIP. A critical aspect is the use of the segmentation module to convert local information into global information, improving the performance of CLIP. During training, the system generates pairs of latent codes (*zgen*) and 3D avatars, rendering them volumetrically to obtain *Igen* images. These are encoded by CLIP and compared with the textual library to produce a code of discrete attributes. An MLP network then maps these attributes to the latent space of the generator, using an Mean Squared Error (MSE) loss function for optimisation. In inference, the input text is decoupled into attributes, converted into latent codes via *M(·)*. These, together with β, θ and ξ, feed the generator to produce the 3D avatar. Experimental results show that *Text2Avatar* outperforms methods such as *DreamFusion* or *AvatarCLIP* in attribute accuracy (80%-100% for colours and clothing lengths) and *R-Precision* (83.30 with CLIP ViT-L/14). Ablation studies confirm that the textual library and segmentation are essential for this performance. Relevance to Biometrics: Although it does not use NeRF, Text2Avatar demonstrates the potential of CLIP for semantic bridges between text and 3D representations. This capability is transferable to biometric verification tasks, where textual descriptions could aid in identifying individuals in low-resolution or occluded environments. However, the lack of integration with articulated models limits its direct application to the context of this project.

ID-NeRF

The Indirect Diffusion-guided Neural Radiance Fields (ID-NeRF) [25] method, proposed by Li et al. in 2024, presents an innovative approach that combines neural radiation fields (NeRFs) with indirect diffusion models to improve the synthesis of generalisable views from few input images. ID-NeRF uses a pre-trained diffusion module to guide the inference of unobserved regions, facilitating the generation of high-quality 3D images from different perspectives. The approach incorporates a neural network that integrates diffuse information to promote a better representation of three-dimensional space and increase the model's generalisation capacity. Experimental results show that the method significantly outperforms previous techniques, especially in scenarios with limited data. This work was developed by researchers affiliated with institutions in China and constitutes a significant advance in the field of neural image synthesis and 3D reconstruction.

2.3 Conclusions

In summary, as noted above in the methods section, there are currently no established methods for verifying whether an image was generated by a specific A-NeRF or NeRF using biometrics, which is the main focus of this project. The review highlighted significant advances in A-NeRFs (realistic human modelling) and biometrics (identity verification), but little integration between them. While A-NeRFs face challenges such as dependence on SMPL and limited generalisation, traditional biometric techniques are not adapted to analyse 3D renderings.

This project proposes to fill this gap by:

- **Leveraging rendering fingerprints**: Analysis of geometric and textural patterns intrinsic to A-NeRFs, identifying unique signatures in volumetric synthesis processes.
- **Multimodal embedding fusion**: Strategic combination of biometric representations from different vector spaces for robust identity-to-model comparison, overcoming the limitations of unimodal approaches.
- **Statistical verification**: Detection of inconsistencies in synthetic renderings through multivariate distributional analysis.

Despite advances in the generation of articulated human representations with high visual quality, there is still no systematic approach to verify whether an isolated image belongs to a specific model. This challenge of origin verification in the domain of NeRFs is still largely unexplored, especially when it comes to associating a 2D image with an identity learned by a three-dimensional model. It is precisely this gap that this work aims to investigate.

Chapter

3

Proposed Method

3.1 Methodology

This paper proposes an innovative approach that combines customised three-dimensional reconstruction techniques with visual identity recognition, based on the capabilities of neural networks NeRF. The methodology consists of two distinct but interconnected phases: training customised models based on short videos, and identity recognition based on static test images. The fundamental motivation for this approach is to demonstrate that it is possible, with a minimum amount of data (one short T-pose video per person), with a total of 11 people in the database, to generate a high-fidelity 3D neural reconstruction of a person and, subsequently, from new images in arbitrary poses, to confidently identify which individual that image belongs to — even if the pose, lighting, or framing were not observed during training. Attempts to use articulated methods based on A-NeRFs proved unfeasible in this context, as these methods are not designed to work directly with custom data — requiring complex setups such as multi-camera calibration, volumetric registrations, and precise temporal alignments, which pose significant barriers to their application in real-world scenarios.

To overcome these challenges, we opted to use the **Nerfstudio** [26] framework, a modern and modular solution that allows for the efficient training of NeRF models from customised data. In particular, the **Nerfacto** [27] model was adopted, a lightweight and flexible implementation that combines computational efficiency with high visual quality in reconstruction. It is based on Instant Neural Graphics Primitives (Instant-NGP) [28] from **NVIDIA**, but adjusted to facilitate the training phase and also integration with NeRF pipelines. It is also inspired by volume rendering techniques with hierarchies, as well as the Mip-NeRF method. This choice is largely due to Nerfacto's ability to handle simple datasets, such as monocular T-pose videos, eliminating the need for complex configurations typical of articulated skeleton-based methods. To evaluate identity in the trained models, we used the **FaceNet512** model, a more refined version of FaceNet already referenced here in the state of the art in the section on visual representations for biometrics in the state of the art, Sec. 2.2, through the **DeepFace** library [29], as will be described in detail later in this paper.

3.1.1 Pipeline Overview

The complete pipeline can be divided into two main phases:

- **Training Phase:** personalised volumetric reconstruction for each person, with extraction of vector representations of identity.
- **Test Phase:** Identification of individuals in new images by semantic similarity in the embedding space.

This separation allows the task to be modularised: the NeRF model (visual representation of the individual) and the recognition mechanism (based on FaceNet512 embeddings) are independent, enabling efficiency and flexibility.

3.1.2 Training Phase

The training phase aims to produce a neural model that realistically represents the geometry and appearance of each individual, as well as a robust identity vector derived from it.

Firstly, it is necessary to understand how this Nerfacto method works and how it is implemented by the Nerfstudio framework. Designed specifically for 3D reconstructions from images or videos of static things, such as objects, it can then be adapted for people as long as they are always in the same pose without moving. This is yet another approach by Nerfstudio, which combines many components from various methods.

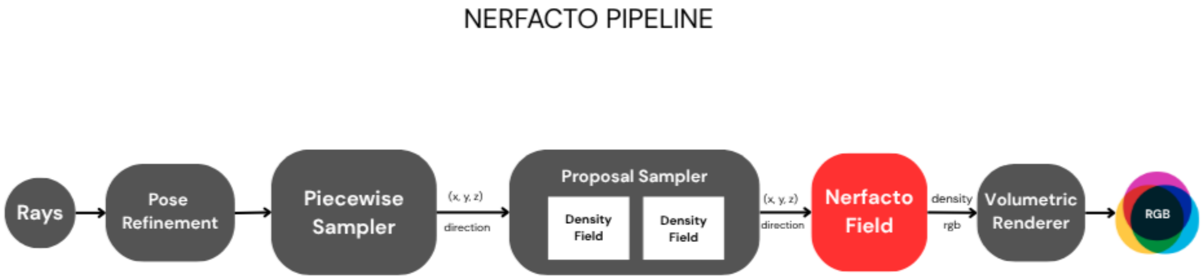

Figure 3.1: Proposed method pipeline.

By observing Nerfacto's pipeline, we can see how the 3D reconstruction of an individual is done using video or static images, and how Nerfstudio organises the data for training the NeRF model. The process is divided into several steps:

- **Pose Refinement**: Nerfacto incorporates a camera pose refinement mechanism, useful for correcting inaccuracies in the initial estimation (e.g., obtained by Structure-from-Motion e Multi-View Stereo pipeline para reconstrução 3D (COLMAP) [30]). Through gradient backpropagation, the model adjusts the poses in order to reduce visual errors or distortions that may exist and improve the quality of the reconstruction.

- **Piecewise Sampler**: The initial sampling along the rays is performed by a piecewise sampler, which allocates half of the samples uniformly close to the camera and distributes the rest in increasing steps, i.e., along the ray, the samples become more spaced out as the distance increases, allowing closer details and more distant objects to be captured efficiently.

- **Proposal Sampler**: It receives as input the three-dimensional vector of a point in space and also the direction of the light ray (θ, ϕ), and is responsible for consolidating the location of the samples in the regions of the scene that contribute most to the final rendering, thus improving the quality of the reconstruction. This requires a density function for the scene, which can be implemented in several ways, but the most efficient way is through a Fused Multi-Layer Perceptron (fused-MLP), which combines the operations of an MLP, matrix multiplication (multiplication of the 3D input vector from the piecewise sampler by the layer weight matrix), bias (shifting the function to better fit the data) and activation function (allows the network to learn more complex shapes, curves and decisions), into a single, more optimised operation. This is very useful in the Nerfacto method as it uses the **tiny-cuda-nn** library, a **Compute Unified Device Architecture (CUDA)** [31, 32] library that is optimised for small but very fast neural networks. Along with this fused-MLP for the density function, there is also hash encoding for sufficient accuracy and speed. We can apply more density functions in this sampler to further consolidate the sampling. In the pipeline, it is possible to observe that two density functions are applied, thus obtaining a more accurate and higher quality sampling, although if the number is greater than two, the results may be worse.

- **Nerfacto Field**: It also receives as input the 3D vector of the sample and the direction of the ray, along with Appearance Embeddings, which are vectors that capture variations in appearance (such as lighting, colour, and texture) in different parts of the scene or between different images. This is very useful for scenarios where lighting varies. The three-dimensional vector goes through a hash function (**Hash Encoding**) to divide the 3D space into several resolutions and uses hash tables to store *features* at each resolution level, thus accelerating training and rendering. The direction of the light ray is passed through a function (**Spherical Harmonics Encodings**), thus encoding this direction representing angular functions, such as colour dependence, using a mathematical basis. After passing through these encoding functions, the vector and direction, together with the appearance embeddings, are passed through an MLP (fused-MLP) that will generate the colour and volumetric density of the 3D point, i.e., the NeRF model learns to map rays emitted by cameras to colour and volumetric density values, thus achieving reconstruction.

The specific steps of the NeRF model training process are described below, following the necessary commands provided by *nerfstudio* to apply all of this to *custom* data, in this case to the videos of the 11 people in the database used for this method, and also following the pipeline illustrated in Figure 3.1 and explanation above it:

1. **T-pose video capture**: Each person is filmed for approximately 10-15 seconds while remaining motionless in a T-pose. The camera performs a complete 360º rotation around the body, ensuring visual coverage from different angles. The choice of the T-pose as the base pose aims to maximise anatomical symmetry and simplify the reconstruction task, reducing joint deformation and facilitating the convergence of the NeRF model.

2. **Frame extraction and processing with `ns-process-data`**: The video is segmented into 100 static frames (evenly spaced). Next, the `ns-process-data` command from the Nerfstudio framework is used to estimate camera poses with COLMAP, calculate approximate depth fields, and organise the data into the appropriate format for volumetric training. This tool automates scene calibration, which is critical for accurate reconstruction in NeRF. In cases where image quality is low (blurred) or lighting is poor, COLMAP may not be able to process the data correctly, i.e., the camera parameters may not converge. In these cases, it is possible to use the video mode of `ns-process-data` to process the video directly, avoiding calibration problems, where the Nerfstudio framework itself will extract the frames from that video and provide them to COLMAP for processing.

3. **Training with `ns-train nerfacto`**: The Nerfacto model is used, a lightweight and efficient architecture based on NeRFs, with hash encodings and hybrid rendering support. The model learns to map rays emitted by cameras to colour values and volumetric density, thus reconstructing the individual's body. Each training session lasts between 10 and 30 minutes on a modern Graphics Processing Unit (GPU) (RTX 4080 Ti can complete a training session of 30,000 iterations in 10 to 12 minutes), with realistic results even with only 100 input images or videos directly, in which case there may already be more images, such as 300, for example.

4. **NeRF model visualisation**: After training or even when **`ns-train`** is performed, the nerfstudio viewer allows you to view the NeRF model to be trained in real time, enabling you to visualise the quality of the reconstruction. It is possible to navigate through the model, check the geometric consistency and textures. This step is crucial to ensure that the model has adequately captured the person's characteristics. This real-time visualisation is done through local access from a generated

link *(http://localhost:7007)* which, when clicked, will open a tab in the browser with **Viser**, which is an iterative 3D visualisation library in **Python** [33, 34], which is already embedded in the Nerfstudio framework, also using the `ns-viewer` command after training has been completed, using the **load-config** flag, we put the *path* to the **config.yml** file, which is generated when the model is trained, and thus we can visualise the trained NeRF model and interact with it from Viser.

5. **Rendering orbital views with keyframe definition in Viser**: After training the NeRF model, rendering is performed through the interactive interface **Viser**, included in the Nerfstudio framework *viewer*. In this interface, the user can freely define the *keyframes* — camera positions along the desired trajectory — that will compose the final rendering. The number of keyframes is not fixed or predefined, but is left to the user's discretion, depending on the desired complexity or fluidity of the animation. Once the keyframes are defined, Viser automatically generates the appropriate command, ns-render, along with the necessary flags, to render the reconstruction video in `.mp4` format, sequentially traversing the defined camera positions. This process allows the creation of synthetic videos that demonstrate the 3D reconstruction of the individual from multiple angles and perspectives, without the need for new real data.

6. **Extraction of FaceNet512 embeddings**: Each rendered image is processed by the **FaceNet512** model, which will be explained in detail in the next section, but in general, the process it performs is to convert the image into a 512-dimensional vector representing the person's facial features. The average of the 36 vectors (in the case of removing 36 frames from the model's rendering video) is calculated, forming the individual's **identity vector**.

Table 3.1: Training phase parameters.

Parameter	**Value**
Frames per person	100 (T-pose)/Original video
Model NeRF	Nerfacto
Identity vector	Average orbital embeddings
Framework	Nerfstudio (v0.3+)
Hardware	RTX 4080 Ti
Training time	10-12 minutes per model
Number of iterations	30000

3.1.3 Testing and Identity Recognition Phase

At this stage, the system is exposed to unknown images and must infer which model (or individual) they belong to. The process is performed independently of the rendering of the NeRF models. This task follows the following steps:

FaceNet512 for Identity Recognition

The **FaceNet512** model is a variant of the original FaceNet, whose architecture is based on deep neural networks trained to map faces onto a Euclidean vector space, where the distance between vectors reflects facial similarity. This model was designed based on the triplet loss technique, which forces the vector of a face (anchor) to be closer to the vector of a positive image (same individual) and further away from a negative image (different person). The result is a 512-dimensional embedding that compactly and discriminatively represents the visual identity of a face. The **FaceNet512** version used in this work is available through the **DeepFace** library, an open-source framework that integrates multiple facial recognition models and facilitates the inference, comparison, and visualisation of embeddings. The model was chosen for its good accuracy, stability, and easy integration with RGB images rendered via NeRF.

1. **Test image input**: Images with different poses (crouching, arms crossed, canonical pose, sideways, etc.) are used. Some of these poses were not seen during training. These images simulate real-life situations.

2. **Embedding extraction with FaceNet512**: The test image is processed by the **FaceNet512** model, which generates a 512-dimensional vector representing discriminative facial features. This vector serves as a compact representation of the identity present in the image.

3. **Cosine similarity comparison**: For each identity vector (corresponding to a trained model), the cosine distance is calculated in relation to the embedding of the test image. The model with the highest similarity (or lowest distance) is assigned as the identity prediction.

This process completely separates volumetric reconstruction from the recognition task, which offers practical advantages: recognition can be performed efficiently, without the need for re-rendering with NeRF.

3.1.4 Details of Similarity Recognition

The choice of the **FaceNet512** model is due to its proven effectiveness in facial recognition tasks, extracting robust and discriminative embeddings. The extracted embeddings represent facial properties that reliably distinguish identities. For each individual, the average of the orbital renderings embeddings (generated after NeRF training) is calculated, forming a **vector signature** of their identity. During testing, the image embedding is compared to these signatures using **cosine similarity**, allowing the most similar individual to be identified.

Cosine similarity It is used as a comparison metric because it measures the orientation between vectors, ignoring the absolute magnitude. It is defined by:

$$\text{sim}(A, B) = \frac{A \cdot B}{\|A\| \|B\|}$$

where $A \cdot B$ is the scalar product between the vectors, and $\|A\|$ and $\|B\|$ are the norms (lengths) of vectors A and B, respectively. This metric is particularly useful for comparing embeddings, as it focuses on the direction of the vectors, which is where the relevant semantic information resides. Let A be the embedding of the test image, and B the identity vector of a NeRF model. The values that can be obtained from this formula **range from -1 to 1**, like the traditional cosine. The logic is simple: if the value is negative, the vectors are opposite, indicating distinct identity (a rare case with real embeddings); if it is close to 0, the vectors are orthogonal, therefore without semantic relation; if it is close to 1, the vectors are similar, indicating that the test image belongs to the same individual as the NeRF model. This system supports noisy images, distinct poses, and varied backgrounds — as long as the individual's bodily appearance is predominant. The use of **FaceNet512** provides a robust representation of identity even under moderate variations in pose, lighting, and background context.

3.1.5 Advantages of the Proposed Approach

The developed system has several advantages that make it effective, scalable, and practical for identity recognition from three-dimensional reconstructions based on NeRF. First, one of the main advantages is that it requires only a short video per individual for the entire reconstruction and identification process. This eliminates the need for specialised equipment, multiple cameras or long capture sessions, making the method accessible with conventional mobile devices. In addition, the input video does not need to contain a wide variety of body poses, which significantly reduces the complexity of data acquisition. Reconstruction is performed even with limited views, thanks to the robustness of the NeRF method, and facial recognition is ensured by an effective biometric model even with pose variation. Another distinctive aspect of the system is the clear separation between the reconstruction process and the identity recognition task. The **FaceNet512** model, based on the FaceNet architecture, is responsible for extracting facial embeddings from synthetic images generated by NeRF. These embeddings, 512-dimensional vectors in

Euclidean space, discriminatively capture the visual characteristics of the face. The separation between view generation and recognition means that, once NeRF has been trained and the orbital images generated, it is not necessary to render again or perform inference with ray tracing to perform identification — this can be done extremely efficiently by directly comparing vectors. For each identity (corresponding to a trained NeRF model), the average of the embeddings extracted from the synthetic images generated at different orbital angles is calculated. This average acts as a compact and robust **vector signature** of the identity. During the testing phase, real images — even with variations such as lighting, background, or expression — are processed with the same FaceNet512 model, ensuring consistency in the vector space.

The choice of **FaceNet512**, available through the **DeepFace** framework, is justified by its proven effectiveness in facial recognition, the stability of embeddings even under noise or small variations, and the ease of integration with rendered RGB images. Unlike CLIP based models [22], which focus on multimodal and generalist semantics, FaceNet was designed specifically for facial biometrics tasks, offering performance more geared towards distinguishing visual identities.

Finally, the system is tolerant to different testing conditions: non-uniform backgrounds, different clothing, and lighting changes. As long as the face is visible and sufficiently defined, recognition remains stable. This approach reduces the need for post-processing, segmentation, or explicit image normalisation. Thus, the complete pipeline offers a balance between accuracy, efficiency, and operational simplicity, positioning itself as a practical solution for visual authentication based on neural reconstruction.

Table 3.2: Comparison with human NeRF-based methods (articulated and non-articulated)

Method	SMPL	Multiple Poses	Acknowledgement	Pose-aware
A-NeRF [15]	Yes	Yes	Implicit	Yes (LBS)
SNARF [11]	Yes	Yes	No	Yes (Implicit)
GNARF [9]	Yes (SMPL/FLAME)	Yes	Yes (via GAN)	Yes
GnARF [5]	Yes	Yes	Yes (multi-individual)	Yes (ID-aware)
HumanNeRF [13]	Yes	Yes	No	Yes
Proposed Method	**No**	**No**	**Yes (FaceNet512)**	**No**

The proposed method does not perform pose modelling or explicit articulation, which makes it lighter, simpler, and suitable for contexts with limited data. Even so, it allows for highly robust identity recognition, even in images with variations in pose or appearance.

3.1.6 Final Thoughts

The developed system not only allows detailed 3D reconstructions to be achieved from minimal visual data, but also enables effective identity recognition through biometric embeddings. The use of a single short video per individual, without the need for extensive pose variations or sophisticated equipment, makes this method particularly practical and accessible for real-world applications. One of the main contributions is the clear separation between the volumetric reconstruction phase — performed using NeRF models — and the identity inference task. This dissociation allows for modularity and efficiency: after training the NeRF and rendering synthetic orbital views, no further rendering or ray tracing is required during recognition. The identification process is performed exclusively in vector space, based on embeddings extracted using the **FaceNet512** model, which has been widely validated in the literature for facial recognition. Each identity is represented by a *vector signature* obtained by averaging the embeddings of the generated orbital images. During the testing phase, real input images, even under variations in pose, lighting, or background, are converted into a feature vector by the same FaceNet512 model. Identification is performed by comparing the vector signatures, using the **cosine distance** as a similarity metric. This approach ensures speed, consistency, and resistance to visual noise, while maintaining high accuracy in identity matching. The system has proven to be effective even in challenging scenarios, such as poses not seen during training or complex backgrounds, as long as the individual's face and main features are visible.

In addition, the framework used (DeepFace) allows for easy integration into real applications and supports GPU acceleration, optimising performance. Given its modular design and the robustness of the components used, this pipeline can be extended in the future to more complex contexts, such as multi-individual recognition, re-identification systems in continuous videos, or even adaptation for facial verification tasks in three-dimensional environments. Additionally, the replacement of FaceNet with other biometric models can be explored according to the specific requirements of each application. The proposed system is expected to contribute significantly to the advancement of facial recognition techniques, especially in scenarios where data collection is limited or where pose and lighting variability is a challenge.

Chapter

4

Experiences and Results

4.1 Context and Experimental Strategy

The main objective of the tests was to evaluate whether the proposed pipeline — based on 3D reconstruction with customised NeRFs (via Nerfacto) and recognition by embeddings extracted with FaceNet512 — is effective in identifying individuals with minimal training data (one video per subject in T-pose). The evaluation involved quantitative aspects (visual metrics such as PSNR, Structural Similarity Index Measure (SSIM), LPIPS) and qualitative aspects (ability to preserve identity under arbitrary poses), as well as confidence in the predictions obtained based on the consistency of the embeddings. For each of the 11 people considered valid, the training process resulted in an individual NeRF model. The test images were selected randomly, with poses different from the T-pose and varied visual context (background, light, occlusions). Identification was performed by comparing the FaceNet512 embeddings of the test image with the average of the embeddings extracted from the orbital renderings of the corresponding NeRF model.

4.1.1 Pre-processing Problems and Discarded Cases

During the training phase, technical difficulties were identified in the use of the `ns-process-data` command with the `images` mode in some subjects. Specifically, when the frames extracted from the video were blurry, COLMAP — a tool used internally by Nerfstudio to estimate camera poses — was unable to converge or extracted inconsistent parameters.

As a solution, a mixed plan was adopted:

- **For subjects with clear images**: The standard image based mode (`–data images`) was used with COLMAP.
- **For subjects with blurred images**: It was decided to pass the video directly to `ns-process-data` in video mode, which allowed for a more robust and automated estimation of the camera parameters.

Despite this hybrid approach, two individuals were excluded from the experimental set. Both videos, captured indoors with limited lighting, showed high blurring in all frames, making reconstruction unfeasible even with the video mode. COLMAP failed to estimate poses for both images and video in these cases, unlike the other cases, even though in some of the cases where there was no failure in capturing the COLMAP parameters, the videos were also in a previous environment, however, in these cases, the lighting was still minimally acceptable for COLMAP to process all the necessary data. In one of the cases, due to the low quality of the video, which was also greatly influenced by the lighting, this video was discarded and replaced by another, in which case the pose used was no longer the same (T-Pose), but rather that of *neutral body* pose and also with a changed environment, the recording having been made on the street and with a device of much higher quality (*drone*) than the mobile phone that was used for the other subjects.

4.1.2 Orbital Renderings of NeRF Models

To illustrate the visual quality of the 3D reconstruction obtained for each person, this section presents reference frames extracted from the orbital trajectories generated by the individual NeRF models. These examples aim to demonstrate the preservation of identity, visual fidelity in terms of texture and morphology, as well as the consistency of reconstructions between different individuals.

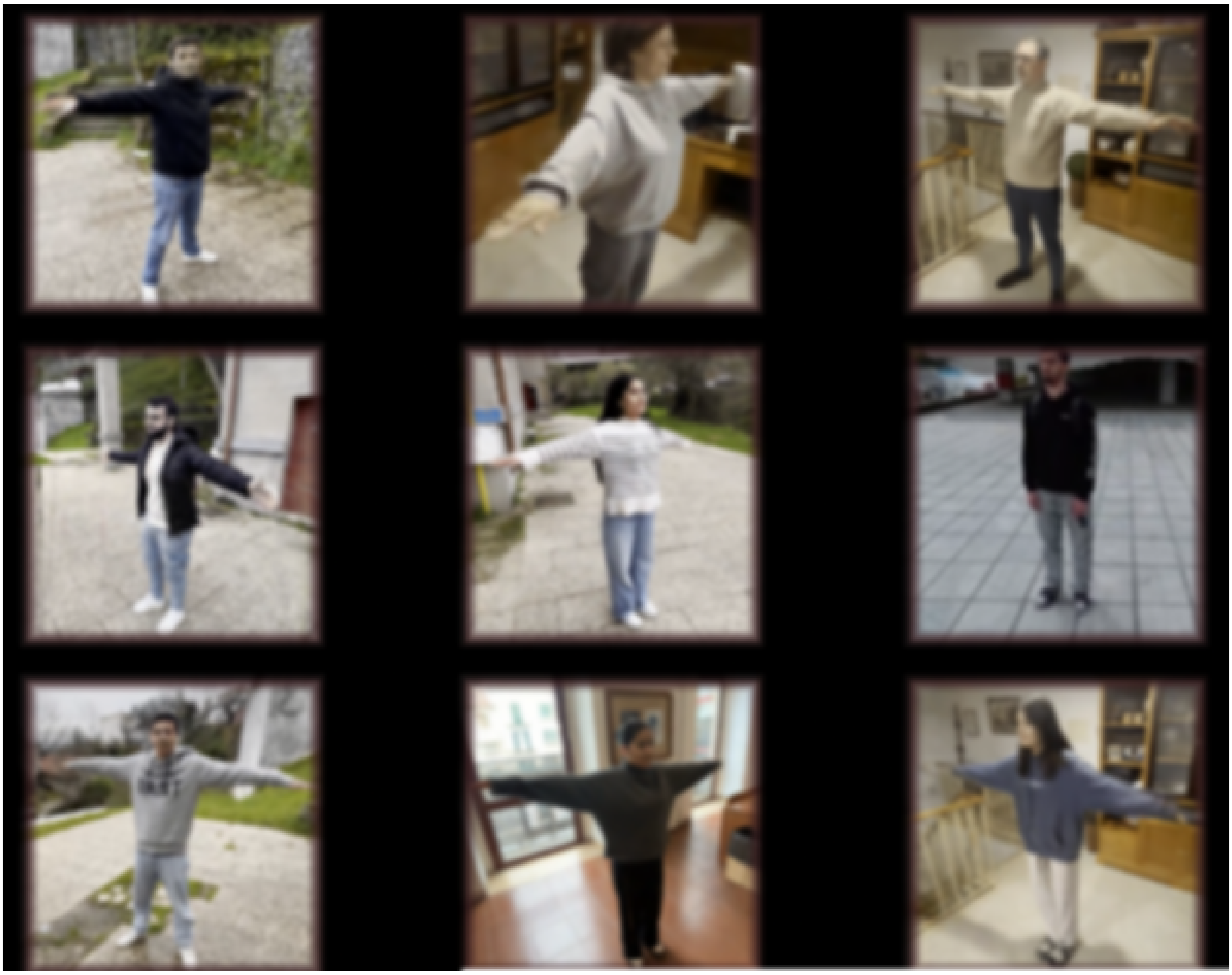

Figure 4.1: Examples of rendered frames for each of the reconstructed individuals.

4.1.3 3D Reconstruction Quality Metrics

The 3D reconstruction quality metrics were calculated for each individual NeRF model using the rendered orbital images, with the help of the Nerfstudio command from `ns-eval`, evaluating the training quality metrics (PSNR, LPIPS, SSIM, etc.) from the rendered images and exporting these values to a JavaScript Object Notation (JSON) file, usually named ***output.json***, the metrics for each model were:

Table 4.1: 3D reconstruction quality metrics per model (NeRF per person).

Sujeito	PSNR (↑)	SSIM (↑)	LPIPS (↓)
Dinis	21.447	0.639	0.167
Gonçalo	20.946	0.613	0.276
Gustavo	26.181	0.855	0.144
Dora	23.365	0.904	0.145
Maria J.	21.076	0.682	0.284
Maria V.	26.299	0.922	0.132
António	23.688	0.899	0.145
Ricardo	24.459	0.675	0.291
Sofia	22.499	0.888	0.157

As we can see from the values obtained in Table 4.1, NeRF models present acceptable values in quality metrics, reflecting the diversity of capture conditions and individual characteristics of people. In general, the PSNR and SSIM values indicate good visual quality, and the LPIPS values were also decent, suggesting that there are few perceptual differences between the renderings and the real images.

4.1.4 Preparation and Pre-processing of Test Images

For the recognition component evaluation, five test images per person were used, each representing different poses and body orientations. The objective was to test the robustness of the system in varied scenarios, with different angles, occlusions, and lighting conditions.

The images were extracted from orbital rendering videos generated from NeRF models. The extraction process was automated using a Python script [34] that accepts the following parameters via command line:

- `–video_path`: path to the individual's orbital video;
- `–output_path`: directory where the frames will be stored;
- `–num_frames`: Extraction interval in degrees (default value: 10º, corresponding to 36 images per complete rotation).

O `num_frames` can be adjusted as desired, allowing for greater or lesser image density.

After extracting the frames, it was necessary to isolate the face region for subsequent calculation of the embeddings. For this purpose, a second Python script was developed with modular face detection and cropping functionalities.

This script offers three operating modes (defined by the flag `–mode`):

- **single**: Face detection and cropping in a single image using RetinaFace [35];
- **folder**: applies the same procedure to all images in a folder;
- **person_face**: uses the You Only Look Once (YOLO) model [36] for whole-body detection, crops the person from the image (via *bounding box*), and then applies face detection to the cropped region.

In `person_face` mode, the default YOLO model used is `yolov8n.pt`, a lightweight model trained for general object detection. The goal of YOLO in this context is to locate the person's body in the image and apply a broad crop, focusing only on the relevant area, which facilitates the task of facial detection — especially in images with multiple elements or busy backgrounds.

In addition, the user can specify an alternative YOLO model using the `–yolo_model` flag, allowing detection to be adapted to the complexity of the scenario.

Another relevant parameter is `–threshold`, which defines the minimum confidence criterion for considering a detection valid (default value: 0.5). Higher values make the detection process more selective, which is useful in high-quality images. Lower values are recommended for images with noise, low resolution, or unfavourable lighting.

- `–input`: path to image(s);
- `–output`: path to store the face cuts;
- `–threshold`: confidence value of detection (default = 0.5);
- `–mode`: operating mode (`single`, `folder`, `person_face`);
- `–yolo_model`: path to the YOLO model (optional).

This approach allowed us to accurately isolate faces in multiple poses and contexts, preparing the set of images needed for inference with the recognition model based on `FaceNet512`.

Despite using the `person_face` mode with YOLO model support to facilitate face detection, in some specific cases RetinaFace's face detection functions were unable to locate the face in the test images. This resulted in the loss of one test image for two people from the original set. For the remaining subjects, all five test images were correctly processed with successful face cropping.

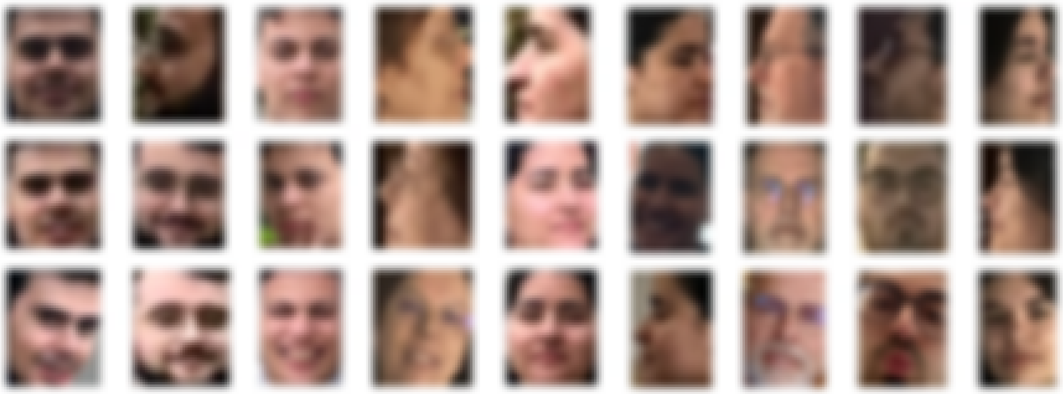

Figure 4.2: Examples of cropped faces of different people from the dataset.

Figure 4.2 shows examples of the cropped images used in the test set, highlighting the variety of poses and conditions present in the dataset, as well as the quality of the pre-processing performed.

4.1.5 Procedure for Facial Recognition with FaceNet512

After pre-processing the test images, facial recognition was performed using the `FaceNet512` model integrated into the `DeepFace` library. The process is based on comparing the average embeddings extracted from the reference frames (obtained from the NeRF renderings) with the embeddings generated from the test images.

The main flow of the process is as follows:

- **Extraction of embeddings from reference frames**: For each subject, a folder containing multiple frames captured at different angles is processed. Each frame is passed through the `FaceNet512` model to obtain a facial embedding vector.
- **Calculation of average embedding**: The embeddings extracted from the various frames are aggregated by averaging, producing a vector representative of the subject's facial identity.
- **Pre-processing of test images**: The test images are resized to the dimensions of the reference frames, adjusted in contrast and brightness, and smoothed with a Gaussian filter to improve the quality of the generated embedding.

- **Comparison and similarity**: The embedding of the test image is obtained in the same way and compared with the average embedding of the frames via cosine similarity. If the similarity value exceeds a configurable threshold (default 0.5), it is considered a match.

The system was implemented with several flags that make the process flexible and efficient:

- −frame_path: path to the folder containing the reference frames obtained from NeRF;
- −test_path: folder containing the test images to be compared;
- −single_test_image: allows direct comparison of a single test image with the reference frames;
- −threshold: sets the threshold value for the *match* or *no match* decision (default 0.5);
- −output_pdf: generates a report in Portable Document Format (PDF) detailing the results, including similarity values and analysed images;
- −output_json: saves the results in a JSON file for further processing or statistical analysis.

In addition, the script generates graphical visualisations of the similarity of each test image compared to the reference set, facilitating visual analysis of the robustness of recognition.

This approach ensures that multiple poses and conditions present in the NeRF frames are considered, resulting in a more stable facial representation that is less susceptible to momentary variations or noise in isolated images.

4.1.5.1 Recognition and Accuracy Results

Each model was tested to determine its ability to identify whether the test images of a person actually belong to that model. In this case, we directly tested the NeRF model of a person against test images of that same person to verify whether it could get it right.

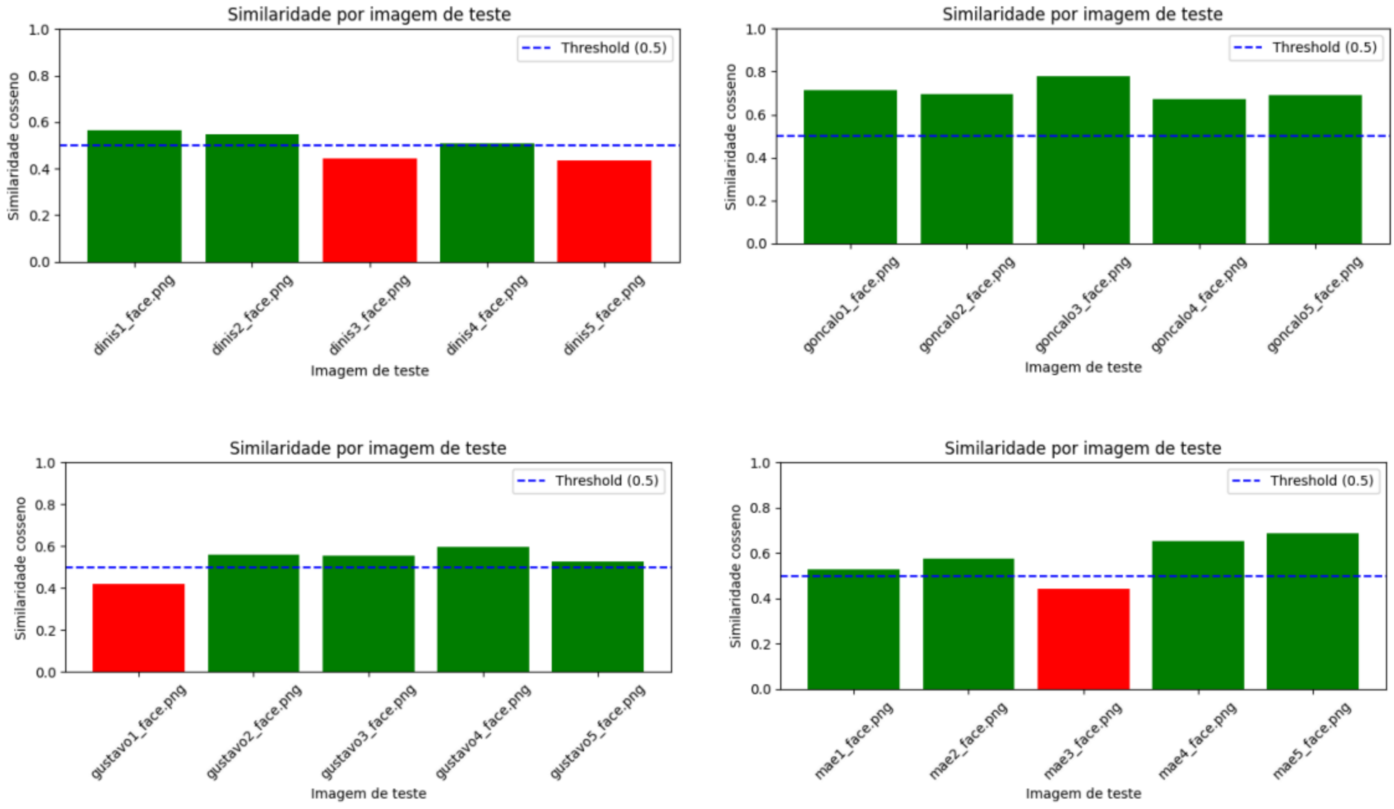

Figure 4.3: Tests on the first four individuals.

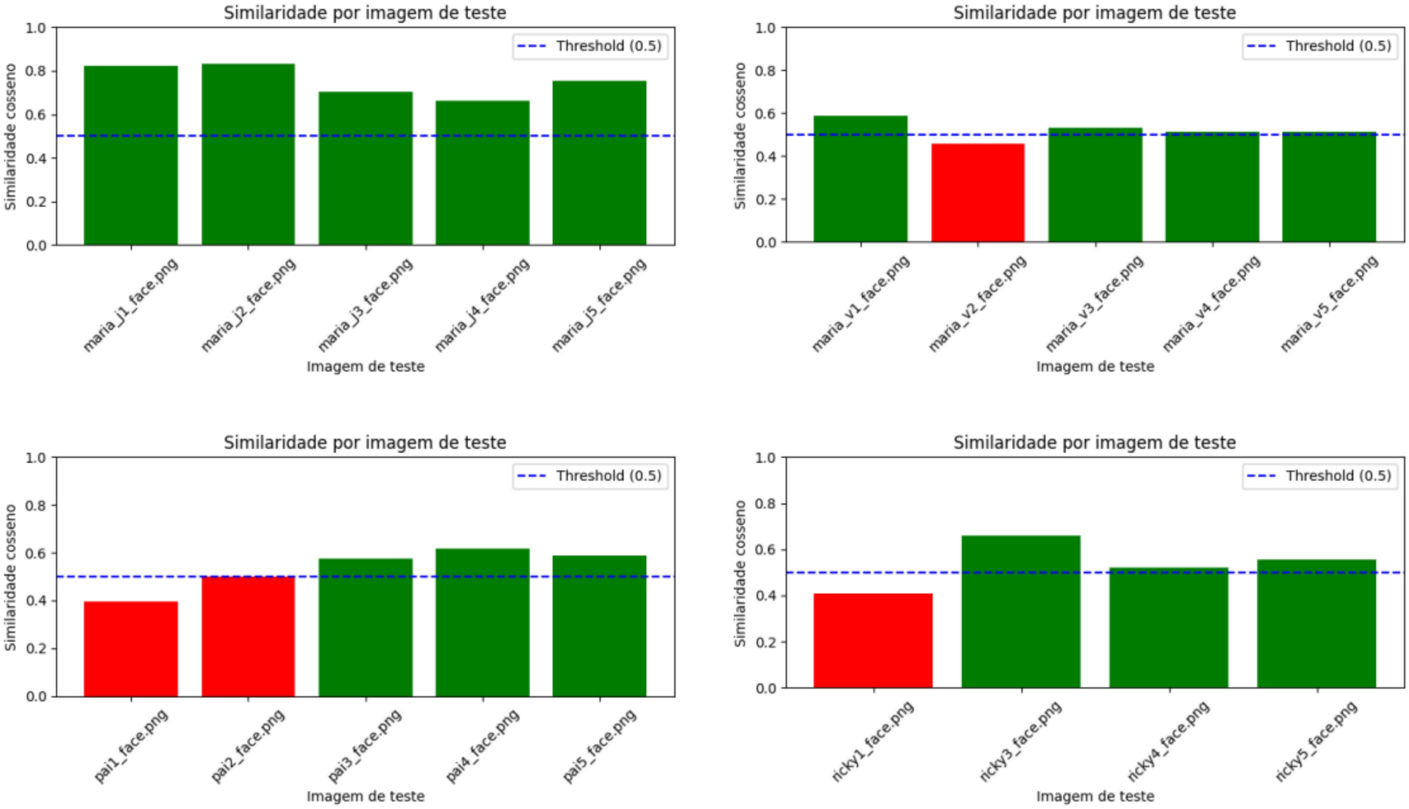

Figure 4.4: Tests on four more individuals.

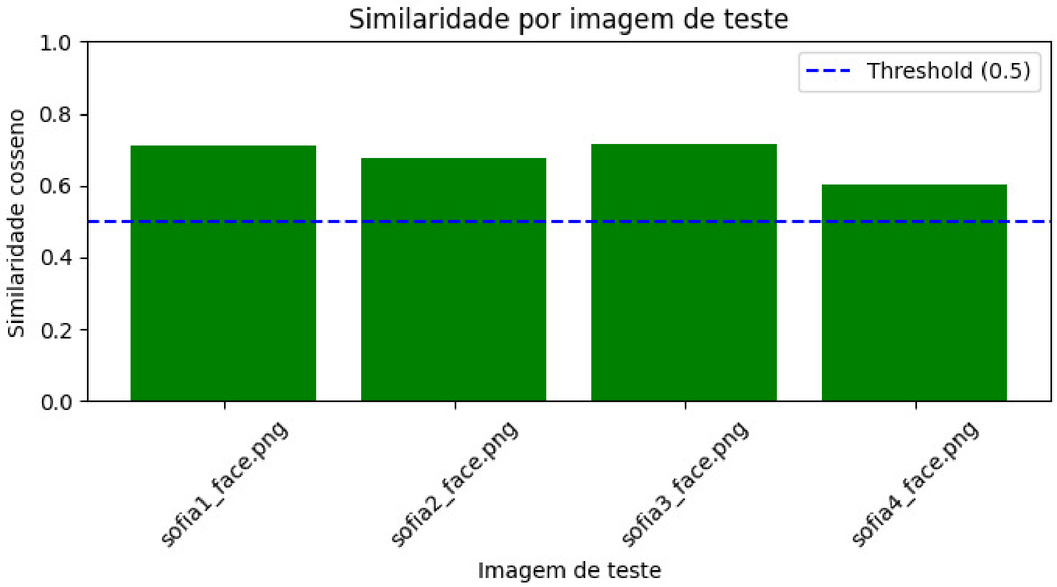

Figure 4.5: Test on the last individual.

Above, in images 4.3, 4.4 and 4.5, we have the cosine similarity graphs between the average embedding of the NeRF frames and the test images of the same person. Each bar represents a test image; green bars indicate a match, while red bars indicate no match. The blue line represents the decision threshold.

Table 4.2: Results according to images 4.3, 4.4 and 4.5.

Individual	**Accuracy**
Dinis	60%
Gonçalo	100%
Gustavo	80%
Dora	80%
Maria J.	100%
Maria V.	80%
António	60%
Ricardo	75%
Sofia	100%
Mean Value	**81,67%**

Looking at Table 4.2, we can see that the average identification accuracy between the test images and the corresponding NeRF models is 81.67%. This result indicates that, in general, the system is able to correctly identify most individuals based on the test images provided.

We now move on to the facial recognition tests, where the model person is not the same as the test images. Individual tests (five tests) were carried out in which each NeRF model was tested with a test image of each of these five people. The different people chosen are those who, in principle, will have the most similarities with the model person in the NeRF model.

Table 4.3: Facial recognition results with NeRF models tested on images of other similar people.

Table 4.4: Model: Dinis.

Model: Dinis		
Test Person	Result	Value
Gonçalo	No Match	0.3188
Gustavo	No Match	0.4800
Ricardo	No Match	0.2148
António.	No Match	0.2538
Sofia	No Match	0.2575

Table 4.5: Model: Gonçalo.

Model: Gonçalo		
Test Person	Result	Value
Dinis	No Match	0.3360
Gustavo	No Match	0.1482
António	No Match	0.1830
Maria J.	No Match	0.0508
Ricardo.	**Match**	0.5719

Table 4.6: Model: Gustavo.

Model: Gustavo		
Test Person	Result	Value
Dinis	No Match	0.4202
Gonçalo	No Match	0.2070
Ricardo	No Match	0.2172
António	No Match	0.3360
Maria V.	No Match	0.3542

Table 4.7: Model: Dora.

Model: Dora		
Test Person	Result	Value
Dinis	No Match	0.2274
Sofia	No Match	0.4001
Gustavo	No Match	0.3588
Maria J.	No Match	0.2061
Maria V.	No Match	0.3500

Table 4.8: Model: Maria J.

Model: Maria J.		
Test Person	Result	Value
Dinis	No Match	0.3285
Sofia	**Match**	0.5877
Gustavo	No Match	0.3058
Dora	No Match	0.4000
Maria V.	**Match**	0.7527

Table 4.9: Model: Maria V.

Model: Maria V.		
Test Person	Result	Value
Sofia	**Match**	0.5541
Dinis	No Match	0.1836
Gustavo	No Match	0.4019
Dora	No Match	0.4686
Maria J.	No Match	0.4329

Table 4.10: Model: António.

Model: António		
Test Person	Result	Value
Dinis	No Match	0.3008
Gonçalo	No Match	0.2479
Gustavo	No Match	0.2378
Ricardo	No Match	0.4831
Sofia	No Match	0.2887

Table 4.11: Model: Ricardo.

Model: Ricardo		
Test Person	Result	Value
Dinis	No Match	0.3175
Gonçalo	**Match**	0.5525
Gustavo	No Match	0.3258
Dora	No Match	0.4765
António	No Match	0.4612

Table 4.12: Model: Sofia.

Model: Sofia		
Test Person	Result	Value
Maria V.	**Match**	0.5876
António	No Match	0.2256
Gustavo	No Match	0.3605
Dora	No Match	0.4424
Maria J.	No Match	0.3804

This behaviour is desirable and demonstrates that the model has facial discrimination capabilities, being able to differentiate the person for whom it was trained from other similar people. It is important to note that, for each person, the embedding value used as a reference was the highest obtained in the test images of that same person, considering that the embedding values are calculated individually for each image.

Despite this, a few cases of '*Match*' with higher embedding values indicate that there may still be false positives in situations of high similarity, suggesting that the decision threshold may need to be adjusted more strictly to minimise this type of error.

4.1.6 Global Facial Recognition Test with Comparison of Test Image Averages

To complement the individual tests performed previously, a script was developed to perform a global test. This script compares the averages of the similarities obtained between the five test images of each person and the corresponding NeRF model. The objective is to evaluate, in a consolidated manner, the model's ability to recognise or differentiate people from multiple samples. The code traverses the directory structure where the NeRF model renderings and test images organised by person are stored. For each model and each set of test images, it runs the facial recognition script (`face_recognition.py`) that calculates the facial similarities between the model frames and the test images. The main flags and parameters of the script are:

- `−root`: Root directory where each person's folders are organised, containing the NeRF model frames and test images.
- `−frames_subdir`: Name of the subfolder containing the frames generated by the NeRF model (example: `face_frames`).
- `−tests_subdir`: Name of the subfolder containing the test images for recognition (example: `tests_faces`).
- `−output_json`: Path to the JSON file where the result of the average similarity matrix between each model and test set will be saved.
- `−n_test_images`: Number of test images to consider for calculating the average (by default, 5).

After calculating and storing the average similarities, the script generates two main types of visualisation:

1. **Bar charts by model** — show the average similarity of each test set in relation to the model, facilitating individual analysis.
2. **General heat map** — presents the complete matrix of average similarities between all models and tests, providing an overview of performance and possible patterns.

These visualisations provide a more intuitive understanding of the results, supporting quantitative analysis with clear visual support.

The following graphs were generated for each NeRF model, demonstrating the average similarity with the test images of all individuals. Each graph shows the isolated performance of a model compared to the other identities. Finally, the heatmap summarises all comparisons in a single visualisation.

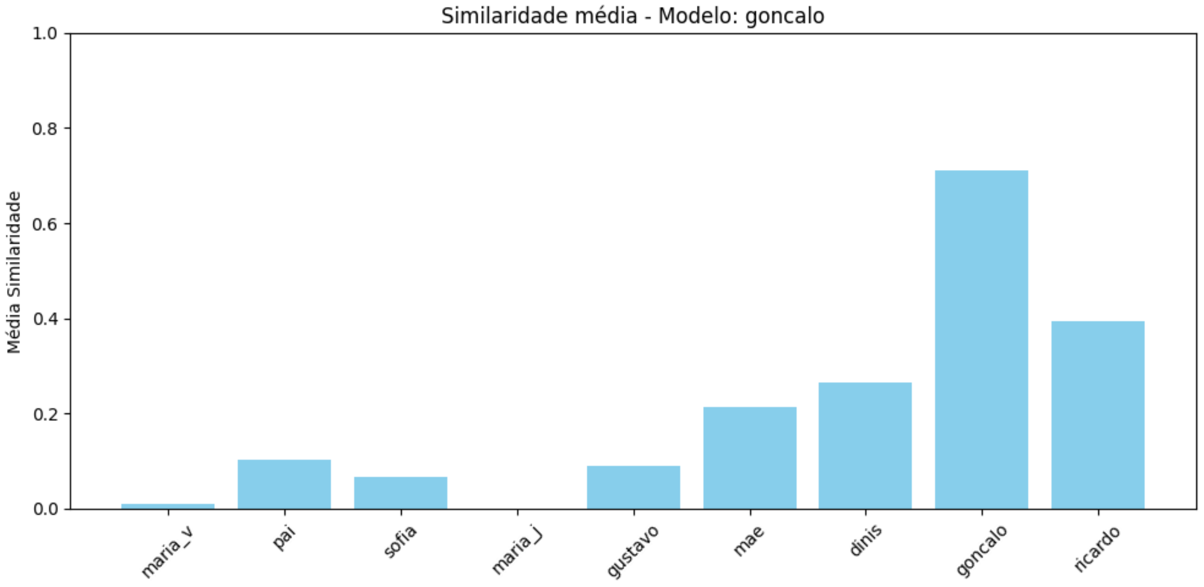

Figure 4.6: Average value in Gonçalo's test images.

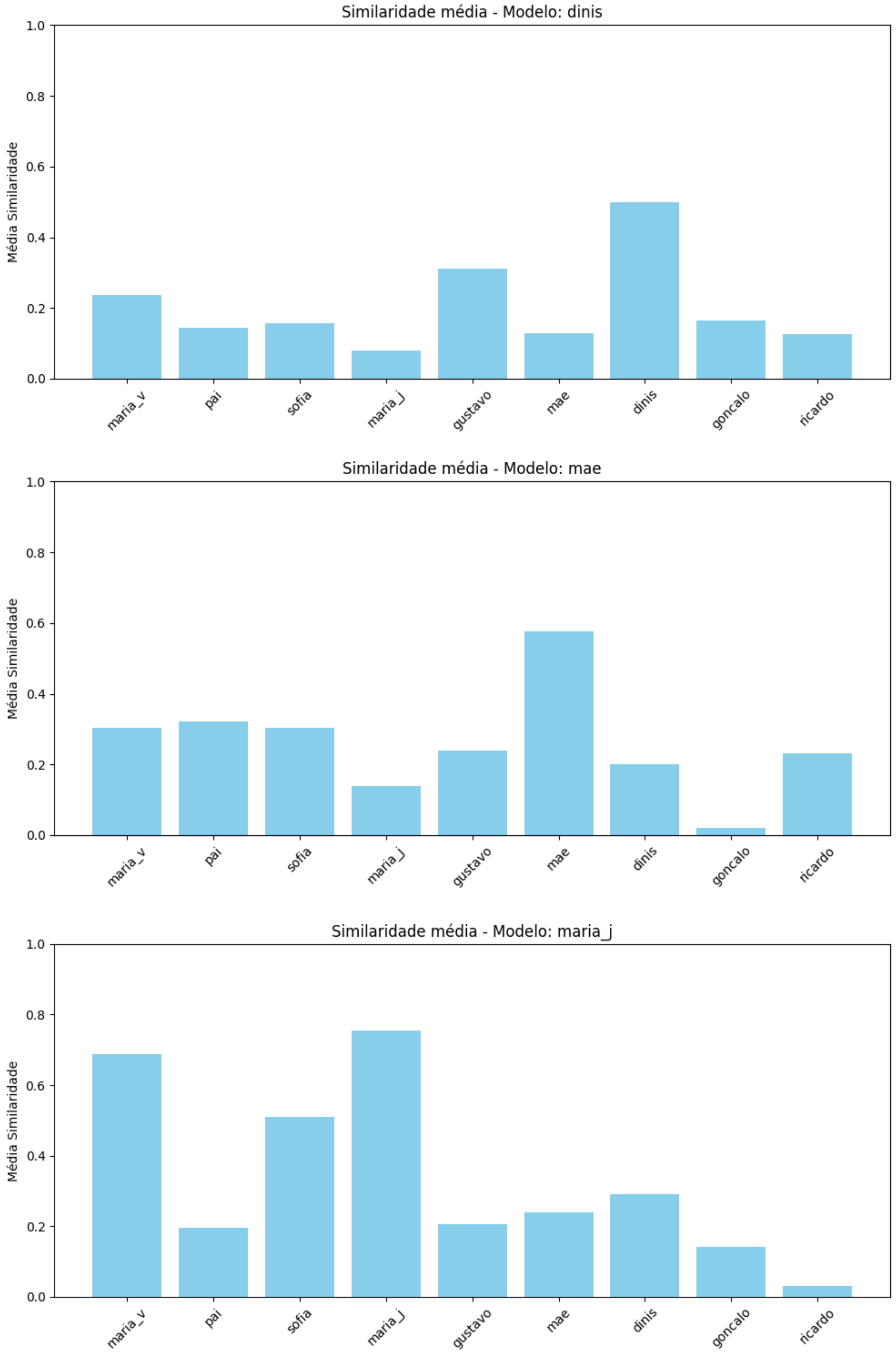

Figure 4.7: Averages value in test images for Dinis (top), Dora (middle) and Maria J. (bottom).

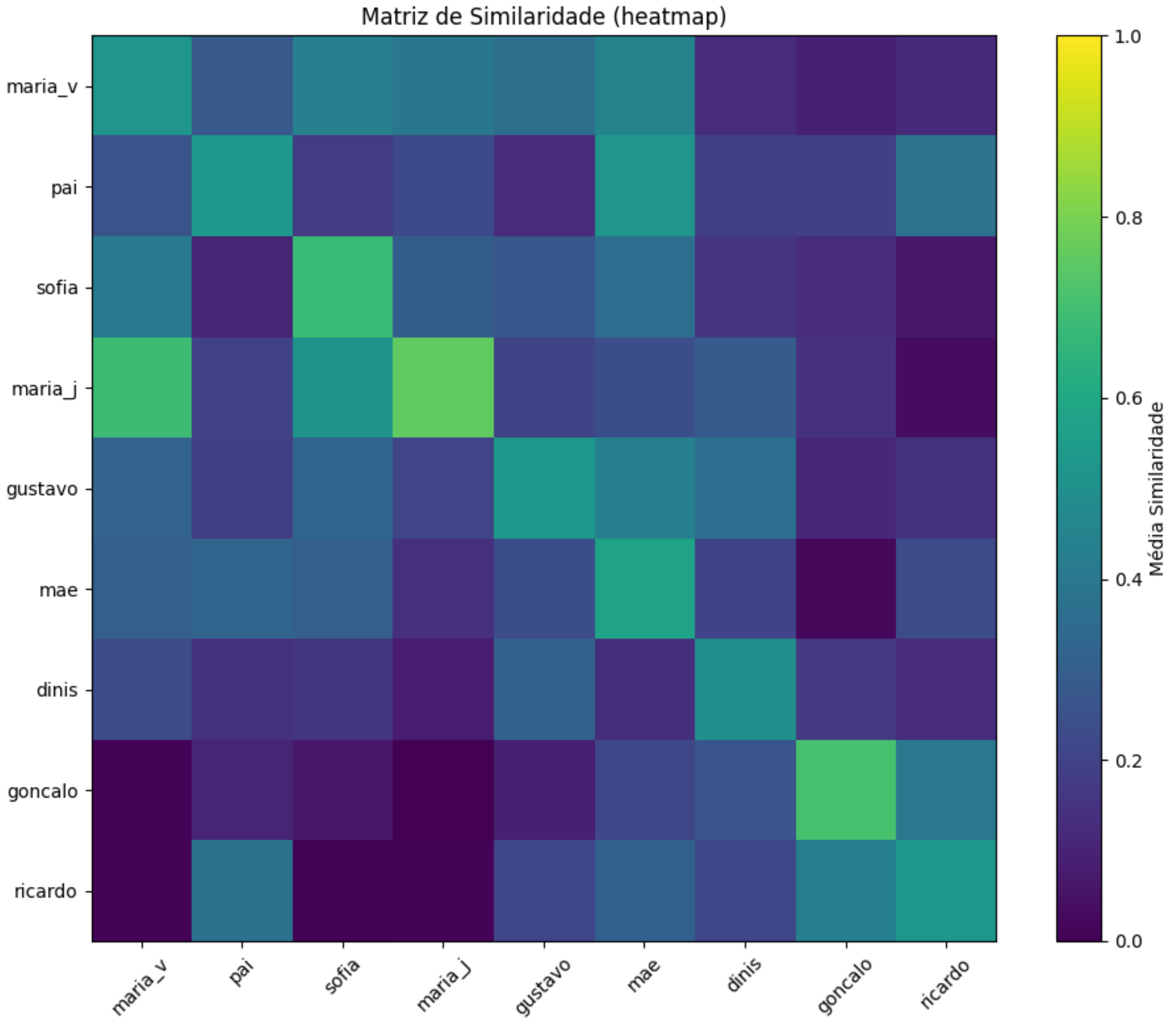

Figure 4.8: Bar charts by NeRF model and heat map showing the average similarities between all pairs of identities.

Looking more closely at the results obtained here in the average cosine similarity graphs and the heat matrix, we have a table here to consolidate these results:

Table 4.13: Average Similarity Matrix (per model person vs test person).

Model/Test	Maria V.	António	Sofia	Maria J.	Gustavo	Dora	Dinis	Gonçalo	Ricardo
Maria V.	**0.520**	0.278	0.435	0.393	0.368	0.447	0.126	0.088	0.123
António	0.254	**0.534**	0.175	0.223	0.126	0.522	0.187	0.190	0.381
Sofia	0.417	0.103	**0.677**	0.297	0.271	0.358	0.156	0.128	0.068
Maria J.	0.687	0.196	0.511	**0.754**	0.206	0.241	0.290	0.141	0.032
Gustavo	0.317	0.184	0.322	0.208	**0.532**	0.435	0.366	0.114	0.147
Dora	0.304	0.322	0.303	0.138	0.240	**0.577**	0.201	0.020	0.230
Dinis	0.236	0.145	0.158	0.079	0.312	0.129	**0.500**	0.165	0.126
Gonçalo	0.010	0.102	0.068	-0.006	0.089	0.213	0.265	**0.711**	0.395
Ricardo	0.010	0.375	0.008	0.010	0.214	0.312	0.214	0.431	**0.535**

It can be seen that, in all cases, the model assigns the highest average similarity value to the person corresponding to the NeRF model. This behaviour confirms that the system is able to correctly recognise the identity of the modelled person, even when compared to other similar faces. This demonstrates the effectiveness of the approach in the task of facial identity verification, even when using images generated synthetically from the 3D model.

4.1.7 Limitations of the Method

Despite the overall positive results achieved with the approach based on average embeddings of NeRF frames, the method has some limitations that affect its accuracy in certain cases. In particular, the failures observed in subjects such as Dinis and António, in Table 4.2, as well as between Sofia and Maria V, Gonçalo and Ricardo, or Maria J. and Sofia in Tables 4.12, 4.9, 4.8, 4.5, and 4.11, can be attributed to several factors:

- **Quality of rendered frames**: The images generated by NeRF, although visually coherent, may have lower definition or artefacts that compromise the accurate extraction of facial embeddings.

- **Lighting variations**: The discrepancy between the lighting in the test images (usually captured in real environments) and the rendered frames can negatively affect the consistency of the embeddings.

- **Model sensitivity**: Even with brightness and contrast adjustments and smoothing applied in preprocessing, the `FaceNet512` model may remain sensitive to small visual differences, leading to a reduction in cosine similarity below the stipulated threshold.

In addition, there is a particular case of confusion involving the twins **Maria V.** and **Maria J.**, observed both in individual tests and in the overall similarity matrix. When compared to the model of **Maria J.**, **Maria V.** shows a very high similarity, even being incorrectly classified as the latter — which is justified by the extreme facial similarity between the two. This error pattern shows that the model, based on visual characteristics, does not have sufficient mechanisms to distinguish highly similar identities in synthetic contexts.

Interestingly, the error does not occur with the same intensity in the opposite direction: the model corresponding to **Maria V.** does not lead to the incorrect identification of **Maria J.** with equal frequency. This asymmetry may be related to the slightly higher quality of the frames in the model for **Maria V.**, or to better capture conditions (such as lighting, sharpness, or angular coverage), which results in more robust and discriminative embeddings. These limitations highlight the importance of future improvements, particularly in the quality of the synthetic data generated, as well as in the robustness of facial comparison methods in the face of non-trivial variations in appearance or genetic similarity.

Finally, it should be noted that increasing the number of frames rendered per model may provide more stable and discriminative embeddings, potentially improving the overall accuracy of the system and reducing false positives in situations of high similarity.

4.1.8 Summary of Results

Table 4.14: Summary of the main results obtained.

Criterion Evaluated	**Result**
Visual fidelity of identity	High (preservation of facial features, hair and individual morphology)
Generalisation for new views	Good, especially in orbital angles not seen in training
Robustness to background variations	Moderate (best performance with clean background and front images)
Identification accuracy (global test)	100% correct match based on the average of the test images
Presence of false positives	Some cases with high similarity values between different people; adjustment of the decision threshold is suggested.
Consistency of embeddings (intra-class)	High consistency between orbital views of the same model
3D rendering quality	Good, with detailed reconstructions consistent with the original identity.

4.2 Final Thoughts

The developed pipeline proved to be effective for facial reconstruction and recognition from minimal data, using only a limited set of frames rendered by NeRF and a robust facial embedding model such as FaceNet512. This method provides a practical and computationally efficient solution, eliminating the need for complex manipulation of captures or multiple poses, making it suitable for applications in resource-constrained environments. The results obtained indicate good visual fidelity and consistent facial discrimination capability, with satisfactory average accuracy in identifying individuals, especially in cases of low similarity. The modularity of the system, combining modern tools such as Nerfstudio, Nerfacto, and FaceNet512, offers flexibility for future adaptations, as well as robustness against moderate variations in pose and lighting. However, the study also revealed important limitations, such as the model's sensitivity to artefacts in the rendered frames and the difficulty in distinguishing individuals with very similar facial features, as in the case of the twins analysed. These restrictions suggest that the quality of the synthetic data and the number of frames used to generate the embeddings are critical factors in improving the system's accuracy. Thus, future research may explore increasing the number of frames to improve the stability of embeddings, as well as complementary techniques to deal with cases of high facial similarity and more pronounced variations in lighting and pose. Nevertheless, the approach presented paves the way for practical applications in re-identification and rapid personalisation of neural models, combining computational efficiency with promising results.

Chapter

5

Conclusions and Future Work

5.1 Main Conclusions

Although it was not possible to explore articulated NeRF methods for this application, the approach based on static volumetric models yielded positive results in reconstruction and visual identity recognition from limited data. The use of the pipeline with Nerfacto in the Nerfstudio modular framework proved to be efficient in generating coherent and detailed three-dimensional representations, even using a small number of frames (about 100) captured in T-pose. For the recognition task, we chose the FaceNet512 model integrated into the `DeepFace` framework for the extraction and comparison of facial embeddings. This combination proved to be robust for the creation of stable 'vector signatures,' capable of discriminating individuals even under variations in pose and lighting. The average of the embeddings extracted from the orbital renderings proved effective for identifying subjects, with consistent similarity metrics based on cosine similarity. The use of `DeepFace` also simplified the pipeline, facilitating inference without the need for additional stochastic strategies. Another relevant point is the modularity of the system, which allows the volumetric training to be dissociated from the recognition phase, making it scalable and applicable in contexts with resource constraints or in real time. Finally, this work demonstrates that, even without explicit articulated modelling methods, it is possible to achieve a functional and promising solution for visual recognition and customised volumetric reconstruction from simple monocular videos.

5.2 Future Work

In addition to the results achieved, several improvements and extensions can be explored to enhance the system. Obtaining higher quality videos, using higher resolution cameras and controlled lighting conditions, will allow for more accurate NeRF renderings with fewer artefacts, increasing the quality of the extracted embeddings. Increasing the number of frames available for training may also contribute to a more detailed and stable volumetric representation. Investigating the impact of longer and more robust training may allow convergence to NeRF models with greater geometric and visual fidelity, which should reflect improvements in facial discrimination and system robustness. Despite the success of FaceNet512, recent models such as DINOv2 [23] and CLIP [22] hold great potential for capturing multimodal and semantic visual representations. Supervised alternatives specific to facial biometrics, such as ArcFace [21], should also be evaluated to identify potential gains in accuracy and robustness. Applying prior segmentation to isolate facial or body regions in images before embedding extraction can reduce interference caused by background, adverse lighting, or occlusions, contributing to more consistent embeddings. Future versions may investigate approaches that integrate articulated deformations, such as A-NeRF, GNARF, or embedding-conditioned methods, to handle greater postural variability and improve generalisation to arbitrary poses. Developing models that can simultaneously represent multiple identities while maintaining discrimination capability in variable lighting environments and real-world scenarios would represent an important step towards practical applications in security, gaming, or virtual social interaction. Thus, the combination of improvements in data capture and quality, more robust training, exploration of multiple facial recognition architectures, and advances in volumetric models represents a clear path for improving and expanding the system, making it increasingly accurate, efficient, and applicable in real-world scenarios.

Bibliography

[1] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421, 2020.

[2] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. In *ACM Transactions on Graphics (TOG)*, volume 34, pages 248:1–248:16. ACM, 2015.

[3] Atsuhiro Noguchi, Satoshi Shirakabe, Yusuke Sugano, and Yasuyuki Matsushita. Neural articulated radiance fields for novel view synthesis of articulated objects. In *British Machine Vision Conference (BMVC)*, 2021.

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

[5] Rakesh Goel, Karttikeya Jatavallabhula, Svetlana Lazebnik, and Varun Jampani. Gnarf: Generalizable neural articulated radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[6] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

[7] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Learning neural parametric models for human avatar from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5346–5355, 2021.

[8] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. People-snapshot: Dataset. `https://graphics.tu-bs.de/people-snapshot`, 2020. Accessed: 2025-06-14.

[9] Alexander W. Bergman, Petr Kellnhofer, Yifan Wang, Eric R. Chan, David B. Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[10] 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), 2017.

[11] Yufeng Zheng, Tao Yu, Yuliang Xiu, Qionghai Dai, and Sergey Tulyakov. Snarf: Differentiable forward skinning for animating neural fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[12] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Cape: Clothed auto-person encoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5451–5460. IEEE, 2021.

[13] Chung-Yi Weng, Brian Lin, Brian Curless Wang, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[15] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[16] Rasmus Jensen, Anders Boesen Lindbo Larsen Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 406–413, 2014.

[17] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9020–9028, 2018.

[18] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *arXiv preprint arXiv:2103.13415*, 2021.

[19] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, Angjoo Kanazawa, and Jonathan T. Barron. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5752–5761, 2021.

[20] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[21] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019.

[22] Alec Radford, Jong Wook Kim, Christopher Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pam Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Morgan Szafraniec, Yannis Kalantidis, Francisco Massa, Mathilde Caron, and Hervé Jegou. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[24] Jianfeng Zhang, Zihang Huang, Xiaoyu Shen, Yifan Ni, Tengfei Hu, Xinyi Liu, Yujun Liu, Jing Liu, Yu Liu, et al. Text2avatar: Text to 3d human avatar generation with codebook-driven body controllable attribute. *arXiv preprint arXiv:2305.02190*, 2023.

[25] Yaokun Li, Chao Gou, and Guang Tan. Id-nerf: Indirect diffusion-guided neural radiance fields for generalizable view synthesis. *arXiv preprint arXiv:2402.01217*, 2024.

[26] Matthew Tancik, Christian Reiser, Ethan Weber Miller, et al. Nerfstudio: A modular framework for neural radiance field development. *arXiv preprint arXiv:2302.04264*, 2023.

[27] Nerfstudio Contributors. Nerfacto — nerfstudio. `https://docs.nerf.studio/nerfology/methods/nerfacto.html`, 2023. Acessado em: 28 de junho de 2025.

[28] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. In *ACM Transactions on Graphics (TOG)*, volume 41, 2022.

[29] Sefik Ilkin Serengil and Alper Ozpinar. Deepface: A lightweight face recognition and facial attribute analysis framework for python. https://github.com/serengil/deepface, 2020. Acessado em: 28 de junho de 2025.

[30] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016.

[31] NVIDIA Corporation. CUDA Toolkit Documentation. Online documentation, 2023. https://developer.nvidia.com/cuda-toolkit.

[32] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. tiny-cuda-nn: A neurally-inspired cuda library for efficient neural networks. GitHub repository, 2021. https://github.com/NVlabs/tiny-cuda-nn.

[33] Nerfstudio Contributors. Viser - interactive 3d visualization tool. https://github.com/nerfstudio-project/nerfstudio/tree/main/nerfstudio/viewer/viser, 2023. Accessed: 2025-06-25.

[34] Python Software Foundation. Python programming language. https://www.python.org, 2023. Accessed: 2025-06-25.

[35] Jiankang Deng, Jiaji Guo, Yuxiang Zhou, Jun Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5203–5212, 2020.

[36] Ultralytics. YOLOv8 - Ultralytics. https://yolov8.com/, 2023. Acedido em junho de 2025.