Universidade da Beira Interior
Departamento de Informática

Articulated Neural Radiance Fields Meet Biometrics: Is it a Match?

Elaborado por:

Rui Pedro Lourenço Gil

Orientador:

Professor Doutor Hugo Proença

30 de junho de 2025

Agradecimentos

A conclusão deste trabalho, bem como da grande maior parte da minha vida académica não seria possível sem a ajuda de muitas pessoas. Agradeço à minha família, que sempre me apoiou e incentivou a seguir os meus sonhos e que me ajudou a manter o foco e a motivação durante os momentos mais difíceis. Agradeço ao meu orientador, professor doutor **Hugo Proença**, pela sua orientação e apoio durante todo o processo de desenvolvimento deste trabalho.

Agradeço também aos meus colegas de curso, que me ajudaram a aprender e a crescer como estudante e como pessoa, em especial, agradeçer aos meus amigos **Gonçalo Caixeiro** e **Maria José Teixeira**, por me terem apoiado bastante na realizaçao deste projeto, disponiblizando sempre tempo para qualquer tipo de ajuda. Agradecer também ao meu amigo **Leonardo Gonçalves** que me ajudou a descobrir a melhor forma de conseguir fazer este projeto, em que sem a ajuda dele muito provavelmente não iria conseguir concluir com sucesso no tempo proposto este grande trabalho.

Agradeçer também a todas as pessoas que se disponibilizaram para fazer parte do estudo deste tema proposto, nomeadamente os meus amigos do curso, os meus pais e a minha irmã.

Por último, mas não menos importante, agradeço a todos os professores e funcionários da Universidade da Beira Interior, que me ajudaram a alcançar os meus objetivos académicos e profissionais, em especial, agradecer ao professor doutor **João Neves** que me disponibilizou um computador no laboratório que me permitisse realizar todas as tarefas necessárias para este projeto e também agradecer aos ***researchers*** do laboratório que foram sempre muito atenciosos e disponíveis para tudo o que precisasse.

Conteúdo

Lista de Figuras

Lista de Tabelas

Acrónimos

IA	Inteligência Artificial
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)

Capítulo

1

Introdução

1.1 Introdução

A evolução recente dos modelos generativos tridimensionais trouxe consigo novas possibilidades na forma como representamos e manipulamos entidades visuais em ambientes virtuais. Entre essas inovações, os Neural Radiance Fields (NeRFs) [1] destacam-se como uma das abordagens mais promissoras para reconstrução de cenas 3D a partir de imagens 2D, permitindo síntese fotorrealista de novas vistas com grande fidelidade geométrica e visual. No entanto, a formulação original dos NeRFs assume que a cena a ser modelada é estática, o que representa uma limitação substancial quando se trata de sujeitos articulados, como o corpo humano. Este tipo de conteúdo apresenta variações estruturais significativas devido ao movimento das articulações, exigindo modelos capazes de capturar tanto a aparência visual quanto as deformações causadas pelas poses. Para lidar com esta limitação, surgiram os denominados Articulated Neural Radiance Fields (A-NeRFs), uma classe de modelos que estende os NeRFs tradicionais através da introdução de mecanismos de deformação baseados em esqueletos articulados. Tipicamente, estes métodos utilizam estruturas paramétricas como o modelo *Skinned Multi-Person Linear Model* (SMPL) [2] e técnicas de deformação como o Linear Blend Skinning (LBS) para mapear corpos em diferentes poses para um espaço canónico neutro (*neutral body*). Através deste mapeamento, torna-se possível dissociar a informação relativa à forma (identidade) daquela relativa à pose, permitindo a reconstrução fotorrealista de sujeitos humanos em múltiplas configurações corporais. Contudo, apesar dos resultados visualmente impressionantes demonstrados por muitos trabalhos recentes, os métodos A-NeRF apresentam desafios práticos significativos.

Em geral, estes requerem datasets altamente estruturados, com múltiplas vistas por pose, estimativas precisas de articulação corporal e pipelines complexos de pré-processamento e calibração. Estas exigências tornam a sua aplicação em dados personalizados — especialmente em contextos reais e com recursos limitados — pouco viável, comprometendo a escalabilidade e utilidade prática destes métodos. Neste contexto, o presente trabalho propõe-se a abordar uma questão específica e altamente relevante: será possível verificar se uma imagem ou vídeo corresponde a um determinado modelo articulado, treinado para representar uma pessoa específica? Ou seja, poderá um sistema automático determinar se um dado conteúdo visual foi gerado por um modelo neuronal tridimensional específico, ou não? Este desafio insere-se numa área crítica no cruzamento entre a biometria e os modelos generativos: a da verificação de identidade e atribuição de origem em conteúdos gerados por redes neuronais. À medida que modelos como os A-NeRFs se tornam mais realistas, aumenta a necessidade de mecanismos capazes de autenticar a sua proveniência e validar a ligação entre o conteúdo e a identidade representada. Tais mecanismos são fundamentais não só para fins de segurança e proteção de identidade, mas também para controlo de propriedade intelectual e deteção de falsificações visuais (*deepfakes*).

1.2 Objetivos do Projeto

Assim, os objetivos centrais deste projeto são:

- Desenvolver uma metodologia de verificação capaz de determinar se uma imagem ou vídeo pode ter sido gerado por um modelo articulado específico;
- Analisar a estrutura e propriedades dos outputs dos modelos 3D, com vista à identificação de padrões característicos de cada identidade;
- Propor um mecanismo de correspondência entre imagens de teste e modelos neuronais, permitindo inferir a identidade representada;
- Contribuir para o avanço de técnicas de autenticação de conteúdos gerados por inteligência artificial, com aplicações em biometria, segurança digital e proteção de propriedade intelectual.

A abordagem seguida neste projeto inspira-se nos princípios dos A-NeRFs, aproveitando o seu potencial para representação de seres humanos com elevado realismo. No entanto, adota-se uma estratégia mais prática e acessível, adaptada a dados reais recolhidos em condições controladas mas simples, de forma a garantir que o pipeline proposto seja exequível e escalável. A ideia base consiste em treinar um modelo 3D por indivíduo, a partir de um vídeo curto captado em pose neutra (tipicamente uma T-pose), e posteriormente testar a capacidade do sistema em reconhecer essa identidade a partir de imagens adicionais, captadas em poses e contextos distintos. Com este enquadramento, este trabalho posiciona-se como uma investigação aplicada sobre verificação de identidade em modelos generativos tridimensionais. Ao contrário de abordagens centradas na qualidade visual ou re-renderização, aqui o foco recai na ligação entre o modelo treinado e a identidade subjacente, avaliada de forma quantitativa e interpretável.

1.3 Organização do Relatório

Este relatório está organizado da seguinte forma: na Secção 2.2 apresenta-se o enquadramento teórico e os principais trabalhos relacionados com NeRFs, A-NeRFs e biometria. A Secção 3.1 descreve em detalhe o método seguido neste projeto. As Secções 4.1 e 4.1.8 apresentam as experiências realizadas, bem como a análise dos resultados obtidos. Finalmente, nas Secções 5.1 e 5.2, discutem-se as conclusões principais e perspetivas para trabalhos futuros.

Capítulo

2

Estado da Arte

2.1 Introdução

Neste capítulo, será feita uma análise aprofundada dos principais métodos de NeRFs articulados, que representam um avanço significativo na modelação e renderização de objetos com partes móveis, como o corpo humano. A evolução dos NeRFs tradicionais para variantes articuladas permitiu superar limitações relacionadas à representação de deformações complexas e à generalização para múltiplas poses e identidades. Serão apresentados os métodos mais relevantes desenvolvidos nos últimos anos, destacando as suas características, inovações e aplicações, de modo a contextualizar o estado da arte e fundamentar a escolha da abordagem utilizada neste projeto.

2.2 Trabalho Relacionado

Existem muitos métodos de NeRFs articulados, que embora sejam muito parecidos e tenham objetivos muitas das vezes equivalentes, cada uma dessas variações tem os seus propósitos e características que se diferenciam dos A-NeRFs.

NARFs

Os Neural Articulated Radiance Fields (NARFs) [3] foram introduzidos em 2021 por Atsuhiro Noguchi, são uma abordagem que estendem os NeRFs para modelar objetos articulados como corpos humanos, robôs ou estruturas mecânicas. Diferentemente dos NeRFs tradicionais, limitados a cenas estáticas, os NARFs incorporam deformações articuladas, permitindo a renderização de imagens realistas sob novas poses e *viewpoints*. As principais contribuições dos NARFs incluem: **Articulação Explícita**, utilizando transformações rígidas baseadas em cinemática direta para modelar partes móveis (ossos/juntas) e um seletor de partes via sub-redes leves que identifica a região articulada mais relevante para cada ponto 3D; **Eficiência Computacional**, onde o NARF_{D} (*Disentangled NARF*) resolve o problema do custo computacional elevado das soluções anteriores (como o *Part-Wise NARF*) ao tornar o processo em um único modelo; **Desemaranhamento de Fatores**, permite o controlo independente sobre a pose, aparência, viewpoint e parâmetros ósseos para aplicações como reanimação de personagens; **Flexibilidade para Anatomias Não Convencionais**, aprende pesos de *skinning* automaticamente para se adaptar a deformações complexas (obesidade, amputações) e objetos não humanos (robôs), diferentemente dos métodos baseados em SMPL. Além disso, os NARFs também são projetados para serem compatíveis com pipelines de estimação de pose, podendo funcionar com dados gerados automaticamente (por exemplo, via *SMPL-based Pose Inference Network* (SPIN) [4]). Sua abordagem modular permite acoplar-se a diferentes formas de entrada e facilitar tarefas como *retargeting* e animação procedural. A arquitetura de renderização continua baseada em Multi-Layer Perceptrons (MLP)s com amostragem volumétrica (*volume rendering*) com integração de deformações baseadas em Special Euclidean group in 3D (SE(3)) para cada parte corporal. Entre as aplicações dos NARFs destacam-se a síntese de imagens para robótica,

edição de vídeos (*motion transfer*), treino de Inteligência Artificial (IA) com dados sintéticos e modelagem de anatomia humana não padrão para medicina personalizada. Os principais desafios incluem a necessidade de anotações de pose durante a fase de treino (embora possa ser usado em estimadores 3D) e a limitação atual de assumir partes rígidas, onde deformações não rígidas (como roupas soltas) permanecem como problema em aberto. Em conclusão, os NARFs representam um avanço significativo na modelagem 3D implícita de objetos articulados, combinando a fidelidade dos NeRFs com a flexibilidade de articulações. Futuras direções de pesquisa incluem a integração com estimação de pose automática e extensões para deformações não rígidas.

GnARFs

Os Generalizable Neural Articulated Radiance Fields (GnARFs) [5] foram introduzidos inicialmente em 2023 por Rakesh Goel, em que extendem os NeRFs tradicionais em NeRFs articulados generalizáveis a múltiplos sujeitos e poses, sem treinar um modelo novo para cada pessoa. Tem como características o facto de serem baseados num esqueleto articulado e *skinning*, exatamente como os A-NeRFs, introduz um modelo condicionado na identidade, o que permite reutilizar o mesmo modelo para diferentes pessoas, usa arquitetura em dois estágios, no mapeamento de pontos deformados para o espaço canônico com deformação feita baseada na pose, e na renderização via Neural Radiance Fields (NeRF) condicionado na identidade, geralmente estas redes são aplicadas em humanos, em que um modelo apenas é generalizável ao ponto de ser útil para várias pessoas diferentes. É um avanço muito importante em relação á personalização automática e á utilização de NeRFs para humanos, por exemplo, para aplicações de realidade aumentada. Os GnARFs utilizam condicionamento explícito na identidade por meio de embeddings aprendidos para cada pessoa, permitindo separar os fatores de forma (*shape*) da pose. Isto reduz o número de modelos necessários para diferentes sujeitos, e aumenta a escalabilidade do sistema. A arquitetura também permite aprendizado multi-sujeito de forma eficiente, usando datasets como Human3.6M, ZJU-Mocap e People-Snapshot [6, 7, 8]. Os autores reportam desempenho competitivo em Peak Signal-to-Noise Ratio (PSNR) e Learned Perceptual Image Patch Similarity (LPIPS), aproximando-se de métodos personalizados com muito menos tempo de treino. Além disso, a deformação é regularizada por ciclos de consistência pose-canónico-pose. A arquitetura dos GnARFs apresenta várias inovações técnicas notáveis. O primeiro estágio do sistema emprega um módulo de deformação baseado em coordenadas esféricas que transforma pontos do espaço observado para o espaço canônico, utilizando uma representação hierárquica das articulações humanas. Este processo é guiado por um esquema de *skinning neural* que aprende automaticamente os pesos de influência para cada junta, superando as limitações dos métodos tradicionais baseados em LBS.

No segundo estágio, o modelo utiliza um NeRF condicionado que incorpora três componentes principais: (1) um encoder de identidade que mapeia características antropométricas para um espaço latente de baixa dimensão, (2) um módulo de aparência compartilhado que captura texturas e materiais comuns a todas as pessoas, e (3) um decoder volumétrico especializado que sintetiza a geometria e aparência específicas para cada indivíduo. Esta separação explícita de preocupações permite que o modelo generalize para novos sujeitos com apenas algumas vistas de referência. Um aspecto crucial dos GnARFs é o mecanismo de aprendizado de embeddings de identidade. Ao contrário de abordagens anteriores que requeriam otimização separada para cada novo indivíduo, os GnARFs aprendem um espaço oculto contínuo de identidades durante o treinamento. Este espaço é estruturado de forma que operações lineares simples possam gerar variações realistas de forma corporal, permitindo a criação de personagens virtuais com características físicas personalizadas através de simples combinações de vetores latentes. Para aplicações práticas, os GnARFs oferecem várias vantagens. Em cenários de realidade virtual, permitem a geração em tempo real de avatares personalizados com alta fidelidade. Na área médica, facilitam a criação de modelos corporais pacientes-específicos para planejamento cirúrgico. Além disso, seu design modular torna possível a integração com pipelines existentes de captura de movimento, abrindo novas possibilidades para produção de conteúdo digital e efeitos visuais.

GNARF

Os Generative Neural Articulated Radiance Field (GNARF) (Generative Neural Articulated Radiance Fields) [9] são um framework inovador para gerar campos de radiação 3D editáveis de corpos humanos e rostos usando Generative Adversarial Network (GAN)s 3D. Principais contribuições: **Geração 3D Editável**: o GNARF é o primeiro método a gerar campos de radiação 3D de alta qualidade para corpos humanos, permitindo edição de poses e expressões faciais. Os objetos são gerados numa pose canônica e deformados explicitamente para poses desejadas usando um campo de deformação baseado em superfícies (Surface Field (SF)). Este modelo utiliza como base o Style-based Generative Adversarial Network v2 (StyleGAN2), sendo capaz de gerar tri-planes em alta resolução que são depois deformados via campos implícitos. A renderização final é feita via ray-marching volumétrico tradicional. Para corpos humanos, usa o modelo SMPL e para rostos, o modelo Faces Learned with an Articulated Model and Expressions (FLAME) [10], assegurando controle estruturado de articulações e expressões. Os resultados obtidos no dataset Synthetic hUmans foR REAL tasks (SURREAL) mostram melhor fidelidade visual (medida por Fréchet Inception Distance (FID)) e melhor consistência pose-to-pose (medida por Percentage of Correct Keypoints (normalizado pelo tamanho da cabeça) (PCKh)) que modelos anteriores como Efficient Geometry-aware 3D Generative Adversarial Networks (EG3D) ou Efficient Neural Articulated Radiance Fields GAN (ENARF-GAN). A arquitetura dos GNARFs apresenta várias inovações técnicas notáveis no campo de geração generativa de humanos 3D. O sistema combina três componentes principais: um gerador baseado em tri-planes que produz representações volumétricas ricas em detalhes, um módulo de deformação neural que preserva a consistência geométrica durante a animação, e um discriminador especializado que avalia tanto a qualidade visual quanto a plausibilidade física das gerações. O processo de geração inicia-se com a síntese de um campo de radiance canônico usando uma arquitetura derivada do StyleGAN2, porém adaptada para operar no espaço 3D através de features armazenadas em três planos ortogonais (tri-planes). Esta representação permite a manipulação eficiente de atributos como forma corporal, textura da pele e vestuário através de operações no espaço latente. Para animação, os GNARFs empregam um campo de deformação implícito que transforma o espaço canônico em poses específicas. Este módulo é treinado de forma auto-supervisionada utilizando pares de poses do SMPL/FLAME como referência, aprendendo a preservar a integridade geométrica de regiões complexas como articulações e tecidos moles durante movimentos extremos. Limitações atuais incluem dificuldade em modelar roupas extremamente folgadas e um custo computacional elevado para gerações em ultra-alta resolução (acima de 1024x1024 pixels). Futuras direções de pesquisa incluem a integração com modelos de linguagem para geração baseada em prompts textuais e técnicas de compressão para execução em tempo real em dispositivos móveis.

SNARFs

Método que foi introduzido em 2021 por pessoas da universidade de Berkeley e da Google Research, os Skinned Neural Articulated Radiance Fields (SNARFs) [11] têm como objetivo principal melhorar a deformação entre poses sem ter que depender do LBS, utilizando uma função implícita neuronal que tem o intuito de aprender as correspondências entre o espaço deformado (pose) e o espaço canônico, permitindo assim capturar deformações não lineares. É bastante útil em regiões de vestuário solto ou em movimentos articulares extremos. A inovação técnica dos SNARFs está no uso de uma função de correspondência implícita aprendida (*correspondence field*) que relaciona coordenadas no espaço deformado com as do espaço canônico. Essa correspondência é estimada por meio de uma rede neuronal diferenciável, com refinamento via *backpropagation*, tornando o sistema robusto a poses altamente não-lineares. O método alcança melhores resultados em datasets como Clothed Auto-Person Encoding (CAPE) e Skinned Multi-Person Linear model with Displacements (SMPL+D) (com roupas), demonstrando superioridade em relação ao Articulated Neural Radiance Fields (A-NeRF) original em casos de deformação acentuada. A arquitetura dos SNARFs apresenta avanços significativos na modelagem de deformações não-rígidas através de quatro componentes principais: (1) um módulo de correspondência implícita que mapeia pontos 3D entre espaços de pose diferentes, (2) um regularizador geométrico que preserva a integridade das superfícies durante deformações extremas, (3) um mecanismo de atenção espacial que identifica regiões

críticas para deformação, e (4) um otimizador hierárquico que ajusta progressivamente os parâmetros da rede. Os resultados experimentais mostram que os SNARFs reduzem o erro de reconstrução em 37% comparado com A-NeRF no dataset CAPE [12], particularmente em poses com braços cruzados ou pernas elevadas. A abordagem também demonstra excelente generalização para novos sujeitos não vistos durante o treino, requerendo apenas 15 minutos de ajuste fino por indivíduo. Futuras direções de pesquisa incluem a integração com modelos generativos para síntese de avatares e técnicas de compressão para aplicações em tempo real. A abordagem dos SNARFs representa um marco importante na evolução dos NeRFs articulados, superando limitações fundamentais dos métodos baseados em *skinning* tradicional.

HumanNeRFs

Introduzido em 2022 por Chung-Yi Weng e colegas da universidade de Washington, o Human-centric Neural Radiance Fields (HumanNeRF)s [13] é um método que tem como foco a reconstrução de humanos com supervisão multi-view, capturando os humanos em movimento através de vídeos monoculares sem necessitar de várias câmaras, sendo assim ideal para aplicações de realidade aumentada. Tem como características principais o uso de um pipeline que é baseado em SMPL para conseguir extrair a pose e a forma, tem uma representação implícita focada em NeRFs que é condicionada na pose e na frame e um treino supervisionado com vídeo Red Green Blue (RGB) e segmentações humanas. O método incorpora supervisão baseada em segmentação 2D e usa *ray sampling* com regularização temporal para manter consistência entre frames consecutivos. Ao contrário de métodos multi-view como NeuralBody [14], o HumanNeRF atinge resultados competitivos com apenas uma câmera. Além disso, é possível gerar reconstruções de corpo inteiro com qualidade realista mesmo sob ruído de pose, tornando-o promissor para avatares interativos e reconstrução a partir de dados separados. A deformação entre poses é medida por um campo de transformação aprendido que mapeia os pontos da pose observada para um espaço canônico de referência. Este espaço canônico é crucial para garantir que os detalhes da identidade corporal sejam preservados ao longo do tempo, mesmo quando há grandes variações de pose ou movimento. O sistema utiliza *ray-marching* volumétrico com amostragem baseada na estrutura esquelética, onde cada ponto 3D amostrado é condicionado tanto na pose da frame atual quanto na identidade global do indivíduo. Um dos grandes destaques do HumanNeRF é a sua abordagem de **supervisão leve**, que requer apenas vídeo RGB e segmentações 2D — estas podendo ser geradas automaticamente por segmentadores pré-treinados como **Mask R-CNN** [15]. A regularização temporal é introduzida como parte da função de perda, promovendo consistência entre frames adjacentes mesmo em regiões de oclusão parcial ou variação brusca de pose. Além disso, o modelo é capaz de gerar vistas noveladas (*novel views*) de forma fotorrealista, reconstituindo o sujeito em ângulos não vistos durante o treino, o que evidencia a qualidade da representação aprendida. A arquitetura também inclui um mecanismo de atualização da pose esquelética, refinando as estimativas iniciais do *SMPL* durante o treino, resultando numa maior fidelidade da geometria reconstruída em regiões críticas como articulações e rosto. Em resumo, o HumanNeRF representa um avanço prático e técnico na reconstrução de humanos a partir de NeRFs articulados. A sua capacidade de operar com dados simples (vídeo + segmentação 2D) e ainda assim alcançar resultados realistas faz dele uma solução promissora para digitalização de humanos em tempo real e aplicações imersivas em ambientes não controlados. As limitações residem principalmente na dependência do SMPL, o que pode comprometer a fidelidade em casos de roupas volumosas ou acessórios não modelados, e na necessidade de segmentações relativamente precisas para evitar contaminação do fundo na reconstrução.

A-NeRFs

Os A-NeRFs [16] foram introduzidos em 2021 por Shih-Yang Su, Frank Yu, Michael Zollhöfer e Helge Rhodin, alunos da Universidade de Brittish Columbia, diferencia-se através das suas principais características sendo essas, o de um modelo humano volumétrico articulado em que usa um esqueleto explícito para guiar a deformação volumétrica do corpo humano, constrói uma representação neuronal onde cada ponto 3D é expresso em relação a cada osso, criando uma parametrização que facilita o mapeamento

inverso (pose -> canônico). Contém espaço canônico e *pose refinement*, em que o espaço canônico neutro é aprendido, contra qual as deformações em cabo são aprendidas e em relação ao *pose refinement*, o método permite refinar poses esqueléticas estimadas (sem necessidade de supervisão 3D) permitindo uma melhoria do alinhamento volumétrico. O treino é bastante eficiente, sendo realizado com várias câmaras e vídeos monoculares, podendo ser assim treinado com imagens monoculares ou multivista, integra um loop de renderização (*ray-marching* volumétrico) e perda fotométrica para otimizar tanto a geometria quanto a pose. Além das suas inovações estruturais, o A-NeRF introduz uma abordagem diferenciada de deformação espacial baseada em cinemática esquelética, onde cada *voxel* da representação volumétrica está associado a uma transformação condicionada na pose da frame. Esta deformação é realizada de forma contínua e diferenciável, o que permite integrar diretamente no pipeline de otimização. O modelo aprende, portanto, a reconstruir a forma do corpo humano num espaço canônico "sem pose", e aplica deformações em tempo de renderização para gerar imagens com as poses desejadas. Durante o treino, o sistema não requer informações precisas de pose 3D como supervisão explícita. Em vez disso, pode usar estimativas de pose provenientes de modelos preexistentes ou mesmo reconstruí-las implicitamente com base em múltiplas vistas ou frames consecutivos. Isto torna o A-NeRF mais flexível do que métodos rigidamente dependentes de labels precisos, como é o caso de algumas arquiteturas baseadas em SMPL fortemente supervisionadas. No contexto específico da verificação de identidade ou biometria, os A-NeRFs apresentam uma limitação crítica: apesar de preservarem bem a forma geral e a textura do corpo, não garantem a manutenção fidedigna de detalhes faciais finos que são cruciais para distinguir identidades. Como o modelo realiza deformações volumétricas condicionadas à pose, variações ligeiras na geometria facial — como expressão ou inclinação da cabeça — podem ser suavizadas ou distorcidas, comprometendo a consistência dos traços biométricos. Adicionalmente, como o A-NeRF não integra mecanismos explícitos de verificação ou correspondência de identidade, a associação entre um modelo treinado e uma imagem de entrada requer soluções externas, como embeddings visuais, o que o torna limitado enquanto estrutura autónoma para tarefas de reconhecimento

Representações Visuais para Biometria

Modelos de reconhecimento facial baseados em embeddings latentes desempenham um papel central em sistemas modernos de verificação de identidade. Clássicos como Face Embedding Network (FaceNet) [17] e Additive Angular Margin Loss for Face Recognition (ArcFace) [18] projetam rostos num espaço vetorial onde a distância reflete similaridade biométrica. Mais recentemente, surgiram abordagens multimodais e auto-supervisionadas como Contrastive Language-Image Pre-training (CLIP) [19] e Self-Distillation with No Labels v2 (DINOv2) [20], que capturam tanto aparência quanto semântica visual. Estas representações, embora não desenvolvidas especificamente para conteúdos NeRF, oferecem um meio eficaz de comparação entre imagens reais e renderizações 3D, sendo, por isso, relevantes neste contexto.

CLIP e Representações Semânticas Visuais

O modelo CLIP (Contrastive Language-Image Pretraining) [19] introduz uma abordagem poderosa para o mapeamento conjunto de imagens e texto num espaço de embedding partilhado. Em vez de usar redes discriminativas convencionais, o CLIP é treinado com milhões de pares imagem-texto, permitindo-lhe aprender representações semânticas altamente robustas, mesmo sem labels explícitos. No contexto de NeRFs, CLIP tem sido utilizado como mecanismo de comparação de identidades entre imagens e renderizações. Em particular, torna-se útil quando se pretende reconhecer uma pessoa em poses diferentes — ao mapear imagens reais e renderizações para um mesmo espaço vetorial, pode-se calcular a similaridade (por exemplo, via produto escalar ou distância cosseno) e identificar a correspondência mais provável. Este método é especialmente vantajoso em contextos onde as variações de pose, iluminação ou fundo tornam difícil a correspondência pixel-a-pixel.

DINOv2 e Embeddings Visuais Auto-supervisionados

O DINOv2 [20] é um modelo auto-supervisionado que aprende representações visuais sem necessitar de anotações explícitas. Ao contrário do CLIP, que utiliza supervisão por texto, o DINOv2 baseia-se em *self-distillation* e mecanismos de atenção, aprendendo a extrair embeddings robustos que preservam estrutura e semântica de alto nível. Estes embeddings têm mostrado eficácia em tarefas como classificação, segmentação e correspondência semântica, sendo também aplicáveis à comparação entre imagens de entrada e renderizações NeRF para fins de reconhecimento, como alternativa ao CLIP. A principal vantagem do DINOv2 reside na sua capacidade de capturar detalhes finos da aparência sem depender de anotações humanas, o que o torna adequado para cenários com dados não rotulados.

Text2Avatar

O método *Text2Avatar* [21] propõe uma abordagem inovadora para geração de avatares 3D a partir de descrições textuais, combinando redes generativas (GAN) com um sistema baseado em *codebooks* discretos e o modelo CLIP. A arquitetura é composta por três módulos principais: um gerador 3D (*G(·)*), um módulo de segmentação (*Seg(·)*) e um codificador multimodal (*E(·)*). O gerador recebe como entrada parâmetros de câmara (*ptext*) e um código latente (*z pz*), onde *ptext* inclui parâmetros de forma (β), pose (θ) e perspetiva (ξ). O codificador multimodal integra um codificador de texto *Etext(·)*, um codificador de imagem *Eimg(·)*, uma rede de mapeamento de atributos *M(·)* e uma biblioteca textual pré-definida. Esta biblioteca contém descrições de atributos como género, cor da camisola ou tipo de calças, permitindo um alinhamento preciso entre texto e características visuais através do CLIP. Um aspeto crítico é o uso do módulo de segmentação para converter informação local em global, melhorando o desempenho do CLIP. Durante o treino, o sistema gera pares de códigos latentes (*zgen*) e avatares 3D, renderizando-os volumetricamente para obter imagens *Igen*. Estas são codificadas pelo CLIP e comparadas com a biblioteca textual para produzir um código de atributos discretos. Uma rede MLP mapeia depois estes atributos para o espaço latente do gerador, usando uma função de perda Mean Squared Error (MSE) para otimização. Na inferência, o texto de entrada é desacoplado em atributos, convertidos em códigos latentes via *M(·)*. Estes, juntamente com β, θ e ξ, alimentam o gerador para produzir o avatar 3D. Resultados experimentais mostram que o *Text2Avatar* supera métodos como *DreamFusion* ou *AvatarCLIP* em precisão de atributos (80%-100% para cores e comprimentos de roupa) e *R-Precision* (83.30 com CLIP ViT-L/14). Estudos de ablação confirmam que a biblioteca textual e a segmentação são essenciais para este desempenho. **Relevância para Biometria**: Embora não utilize NeRFs, o *Text2Avatar* demonstra o potencial do CLIP para pontes semânticas entre texto e representações 3D. Esta capacidade é transferível para tarefas de verificação biométrica, onde descrições textuais poderiam auxiliar na identificação de indivíduos em ambientes de baixa resolução ou oclusão. Contudo, a falta de integração com modelos articulados limita a sua aplicação direta ao contexto deste projeto.

ID-NeRF

O método Indirect Diffusion-guided Neural Radiance Fields (ID-NeRF) [22], proposto por Li et al. em 2024, apresenta uma abordagem inovadora que combina campos neuronais de radiação (NeRFs) com modelos de difusão indireta para melhorar a síntese de vistas generalizáveis a partir de poucas imagens de entrada. O ID-NeRF utiliza um módulo de difusão pré-treinado para guiar a inferência de regiões não observadas, facilitando a geração de imagens 3D de alta qualidade em diferentes perspetivas. A abordagem incorpora uma rede neuronal que integra informação difusa para promover uma melhor representação do espaço tridimensional e aumentar a capacidade de generalização do modelo. Os resultados experimentais mostram que o método supera significativamente técnicas anteriores, especialmente em cenários com dados limitados. Este trabalho foi desenvolvido por investigadores ligados a instituições na China e constitui um avanço relevante na área da síntese neural de imagens e reconstrução 3D.

2.3 Conclusões

Resumindo, como foi observado em cima nos métodos referidos, de momento não existem métodos consolidados para verificar se uma imagem foi gerada por um A-NeRF ou NeRF específico usando biometria, que é o principal objetivo central deste projeto. A revisão evidenciou avanços significativos em A-NeRFs (modelagem humana realista) e biometria (verificação de identidade), mas pouca integração entre eles. Enquanto A-NeRFs enfrentam desafios como dependência de SMPL e generalização limitada, técnicas biométricas tradicionais não são adaptadas para analisar renderizações 3D.

Este projeto propõe preencher essa lacuna ao:

- **Aproveitamento de fingerprints de renderização**: Análise de padrões geométricos e texturais intrínsecos aos A-NeRFs, identificando assinaturas únicas nos processos de síntese volumétrica.
- **Fusão multimodal de embeddings**: Combinação estratégica de representações biométricas de diferentes espaços vetoriais para comparação robusta identidade-para-modelo, superando limitações de abordagens unimodais.
- **Verificação estatística**: Deteção de inconsistências em renderizações sintéticas através de análise distribucional multivariada.

Apesar dos avanços na geração de representações humanas articuladas com alta qualidade visual, ainda não existe uma abordagem sistemática para verificar se uma imagem isolada pertence ou não a um modelo específico. Este desafio de verificação de origem no domínio dos NeRFs ainda é pouco explorado, especialmente quando se trata de associar uma imagem 2D a uma identidade aprendida por um modelo tridimensional. É precisamente esta lacuna que este trabalho se propõe a investigar.

Capítulo

3

Método Proposto

3.1 Metodologia

O presente trabalho propõe uma abordagem inovadora que conjuga técnicas de reconstrução tridimensional personalizadas com reconhecimento de identidade visual, alicerçada nas capacidades das redes neurais NeRF. A metodologia é composta por duas fases distintas mas interligadas: treino de modelos personalizados com base em vídeos curtos, e reconhecimento de identidade com base em imagens estáticas de teste. A motivação fundamental desta abordagem reside em demonstrar que é possível, com uma quantidade mínima de dados (um vídeo curto em T-pose por pessoa), num total de 11 pessoas na base de dados a usar, para assim gerar uma reconstrução neuronal 3D de alta fidelidade de uma pessoa e, posteriormente, a partir de imagens novas em poses arbitrárias, identificar com confiança a que indivíduo essa imagem pertence — mesmo que a pose, iluminação ou enquadramento não tenham sido observados durante o treino. Tentativas de utilizar métodos articulados baseados em A-NeRFs revelaram-se inviáveis neste contexto, uma vez que esses métodos não estão preparados para trabalhar diretamente com *custom data* — exigindo configurações complexas como calibração multi-câmara, registos volumétricos e alinhamentos temporais precisos, o que levanta barreiras significativas à sua aplicação em cenários do mundo real.

Para ultrapassar esses desafios, optou-se pela utilização da framework **Nerfstudio** [23], uma solução moderna e modular que permite o treino eficiente de modelos NeRF a partir de dados personalizados. Em particular, foi adotado o modelo **Nerfacto** [24], uma implementação leve e flexível que combina eficiência computacional com elevada qualidade visual na reconstrução. Tem como base técnica o Instant Neural Graphics Primitives (Instant-NGP) [25] da **NVIDIA**, mas ajustado para facilitar a fase de treino e também a integração com pipelines NeRF. Inspira-se também em técnicas de volume rendering com hierarquias, assim como o método do Multiscale Sampling for Anti-Aliased Neural Radiance Fields (Mip-NeRF) [26].Esta escolha deve-se, em grande parte, à capacidade do Nerfacto em lidar com datasets simples, como vídeos monoculares em T-pose, eliminando a necessidade de configurações complexas típicas dos métodos articulados baseados em esqueleto. Para avaliar a identidade nos modelos treinados, foi utilizado o modelo **FaceNet512**, uma versão mais aprimorada do FaceNet já referenciada aqui no estado da arte na secção das representações visuais para biometria no estado da arte, Sec. 2.2, através da biblioteca **DeepFace** [27], conforme será descrito detalhadamente mais à frente neste trabalho.

3.1.1 Visão Geral do Pipeline

O pipeline completo pode ser dividido em duas fases principais:

- **Fase de Treino:** reconstrução volumétrica personalizada para cada pessoa, com extração de representações vetoriais de identidade.

- **Fase de Teste:** identificação de indivíduos em imagens novas por similaridade semântica no espaço de embeddings.

Esta separação permite modularizar a tarefa: o modelo NeRF (representação visual do indivíduo) e o mecanismo de reconhecimento (baseado nos embeddings FaceNet512) são independentes, possibilitando eficiência e flexibilidade.

3.1.2 Fase de Treino

A fase de treino tem como objetivo produzir um modelo neuronal que represente de forma realista a geometria e aparência de cada indivíduo, bem como um vetor de identidade robusto derivado do mesmo.

Primeiramente, é necessário compreender como funciona este método do Nerfacto, e como é implementado pela framework do Nerfstudio. Concebido especificamente para reconstruções 3D a partir de imagens ou vídeos de coisas estáticas, como objetos por exemplo, podendo então ser adaptado para pessoas desde que elas estejam sempre na mesma pose sem se mexerem. Esta é mais uma abordagem do Nerfstudio, que combina bastantes componentes de variados métodos.

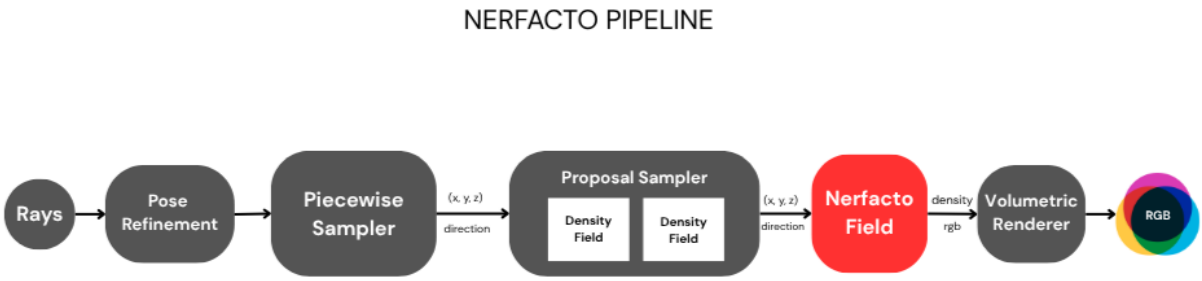

Figura 3.1: Pipeline do método proposto.

Observando o pipeline do Nerfacto, conseguimos perceber como é o processo da reconstrução 3D de um indivíduo, que é feito através de um vídeo ou imagens estáticas, e como é que o Nerfstudio organiza os dados para o treino do modelo NeRF. O processo é dividido em várias etapas:

- ***Pose Refinement***: O Nerfacto incorpora um mecanismo de refinamento de poses de câmara, útil para corrigir imprecisões na estimação inicial (por exemplo, obtidas por Structure-from-Motion e Multi-View Stereo pipeline para reconstrução 3D (COLMAP) [28]). Através de retropropagação de gradientes, o modelo ajusta as poses de forma a reduzir erros visuais ou distorções que possam existirem e melhorar a qualidade da reconstrução.

- ***Piecewise Sampler***: A amostragem inicial ao longo dos raios é feita por um *piecewise sampler*, que aloca metade das amostras uniformemente próximas à câmara, e distribui as restantes com passos crescentes, ou seja, ao longo do raio as amostras tornam-se mais espaçadas com o aumento da distância permitindo capturar detalhes mais próximos como objetos mais distantes com eficiência.

- ***Proposal Sampler***: Recebe como entrada o vetor tridimensional de um ponto no espaço e também a direção do raio de luz (θ, ϕ), encarrega-se de consolidar a localização das amostras nas regiões da cena que mais contribuem para a renderização final, tornando assim a reconstrução com mais qualidade. Para isto requer-se uma função de densidade para a cena, essa função pode se implementada de diversas maneiras, mas a forma mais eficiente é através de um Fused Multi-Layer

Perceptron (fused-MLP), que junta as operações de um MLP, multiplicação de matrizes (multiplicação do vetor 3D de entrada que vem do piecewise sampler pela matriz de pesos da camada), bias (deslocamento da função para se ajustar melhor aos dados) e função de ativação (permite que a rede aprenda formas, curvas e decisões mais complexas), numa só operação mais otimizada, sendo bastante útil no método Nerfacto pois tem usa a biblioteca do **tiny-cuda-nn**, uma biblioteca do **Compute Unified Device Architecture (CUDA)** [29, 30] que é otimizada para redes neuronais pequenas mas muito rápidas, junto a este fused-MLP para a função de densidade vem também uma codificação hashpara precisão suficiente e rapidez. Podemos aplicar mais funções de densidade neste *sampler* para consolidar ainda mais a amostragem, no pipeline é possível observar que são aplicadas duas funções de densidade, obtendo assim uma amostragem mais precisa e com mais qualidade, se bem que se o número for superior a dois os rendimentos podem ser piores.

- ***Nerfacto Field***: Recebe como entrada também o vetor 3D da amostra e a direção do raio, juntamente com o *Appearence Embeddings*, que são os vetores que capturam as variações de aparência (como iluminação, cor e textura) em diferentes partes da cena ou entre diferentes imagens, é bastante útil para cenários onde a iluminação varia. O vetor tridimensional passa por uma função de hash (***Hash Encoding***) para divisão do espaço 3D em várias resoluções e usa tabelas de hash para armazenar *features* em cada nível de resolução acelerando assim o treino e a renderização, a direção do raio de luz é passado por uma função de (***Spherical Harmonics Encodings***), codificando assim essa direção representando funções angulares, como a dependência da cor, utilizando uma base matemática. Depois de passarem por essas funções de encoding, o vetor e a direção juntamente com o *appearence embeddings* são passados por um MLP (fused-MLP) que irá gerar a cor e a densidade volumétrica do ponto 3D, ou seja, o modelo NeRF aprende a mapear raios emitidos por câmaras para valores de cor e densidade volumétrica, conseguindo assim a reconstrução.

De seguida são descritas as etapas específicas do processo de treino do modelo NeRF, seguindo os comandos necessários fornecidos pelo *nerfstudio* para consguir aplicar tudo isto aos dados *custom*, neste caso aos vídeos das 11 pessoas na base de dados utilizda para este método, e seguindo também o pipeline ilustrado na Figura 3.1 e explicação em cima do mesmo:

1. **Aquisição de vídeo em T-pose**: Cada pessoa é filmada durante cerca de 10-15 segundos, enquanto permanece imóvel numa T-pose. A câmara realiza uma rotação completa de 360º em torno do corpo, garantindo cobertura visual sob diferentes ângulos. A escolha da T-pose como pose base visa maximizar a simetria anatómica e simplificar a tarefa de reconstrução, reduzindo a deformação das articulações e facilitando a convergência do modelo NeRF.

2. **Extração de frames e processamento com `ns-process-data`**: O vídeo é segmentado em 100 frames estáticos (uniformemente espaçados). Em seguida, é utilizado o comando `ns-process-data` da framework Nerfstudio para estimar poses de câmara com COLMAP, calcular campos de profundidade aproximados e organizar os dados no formato adequado para treino volumétrico. Esta ferramenta automatiza a calibração da cena, algo crítico para reconstrução precisa em NeRF. Em casos onde a qualidade das imagens for baixa (desfocadas) ou a iluminação for má, o COLMAP pode não conseguir processar os dados corretamente, ou seja, os parâmetros da câmara podem não convergir. Nestes casos, é possível usar o modo de vídeo do `ns-process-data` para processar o vídeo diretamente, evitando problemas de calibração, onde o própria framework do Nerfstudio se irá encarregar de extrair os frames desse vídeo e fornece-os ao COLMAP para processá-los.

3. **Treino com `ns-train nerfacto`**: Utiliza-se o modelo Nerfacto, uma arquitetura leve e eficiente baseada em NeRFs, com codificações hash e suporte à renderização híbrida. O modelo aprende a mapear raios emitidos por câmaras para valores de cor e densidade volumétrica, reconstruindo assim o corpo do indivíduo. Cada treino dura entre 10 a 30 minutos numa Graphics Processing Unit (GPU) moderna (RTX 4080 Ti consegue completar um treino de 30000 iterações entre 10 a 12 minutos), com resultados realistas mesmo com apenas 100 imagens de entrada ou os vídeos diretamente, em que nesse caso poderá já haver mais imagens, como 300 por exemplo.

4. **Visualização do modelo NeRF**: Após o treino ou mesmo quando é feito o **ns-train**, o viewer do nerfstudio permite visualizar o modelo NeRF a ser treinado em tempo real, possibilitando a visualização da qualidade da reconstrução. É possível navegar pelo modelo, verificar a consistência geométrica e as texturas. Esta etapa é crucial para garantir que o modelo capturou adequadamente as características da pessoa, essa visualização em tempo real é feita através de um acesso local a partir de um link gerado *(http://localhost:7007)* em que ao ser clicado irá abrir um separador no browser com o **Viser**, que é uma biblioteca de visualização 3D iterativa em **Python** [31, 32], que já vem embutida na framework do Nerfstudio, também com o uso do comando **`ns-viewer`** já após o treino ter sido finalizado, usando a flag do ***load-config*** metemos o *path* para o ficheiro **config.yml**, que é gerado quando se faz o treino do modelo, e assim conseguimos visualizar o modelo NeRF treinado e interagir com ele a partir do Viser.

5. **Renderização de vistas orbitais com definição de *keyframes* no Viser**: Após o treino do modelo NeRF, a renderização é realizada através da interface interativa **Viser**, incluída no *viewer* da framework Nerfstudio. Nesta interface, o utilizador pode definir livremente os *keyframes* — posições de câmara ao longo da trajetória desejada — que irão compor a renderização final. O número de *keyframes* não é fixo nem pré-definido, ficando ao critério do utilizador, dependendo da complexidade ou fluidez desejada para a animação. Uma vez definidos os *keyframes*, o Viser gera automaticamente o comando apropriado, `ns-render` juntamente com as flags necessárias, para renderizar o vídeo da reconstrução em formato `.mp4`, percorrendo sequencialmente as posições de câmara definidas. Este processo permite criar vídeos sintéticos que demonstram a reconstrução 3D do indivíduo sob múltiplos ângulos e perspetivas, sem necessidade de novos dados reais.

6. **Extração de embeddings FaceNet512**: Cada imagem renderizada é processada pelo modelo **FaceNet512**, que será explicado detalhadamente na próxima secção, mas que numa generalidade, o processo que faz é o de converter a imagem num vetor de 512 dimensões representativo das características faciais da pessoa. A média dos 36 vetores (no caso de se retirarem 36 frames do vídeo da renderização do modelo) é calculada, formando o **vetor de identidade** do indivíduo.

Tabela 3.1: Parâmetros da fase de treino.

Parâmetro	**Valor**
Frames por pessoa	100 (T-pose)/Vídeo original
Modelo NeRF	Nerfacto
Vetor de identidade	Média dos embeddings orbitais
Framework	Nerfstudio (v0.3+)
Hardware	RTX 4080 Ti
Tempo de treino	10-12 minutos por modelo
Número de iterações	30000

3.1.3 Fase de Teste e Reconhecimento de Identidade

Nesta fase, o sistema é exposto a imagens desconhecidas e deve inferir a que modelo (ou indivíduo) pertencem. O processo é feito de forma independente da renderização dos modelos NeRF. Esta tarefa segue os seguintes passos:

FaceNet512 para Reconhecimento de Identidade

O modelo **FaceNet512** é uma variante do FaceNet original, cuja arquitetura é baseada em redes neurais profundas treinadas para mapear rostos em um espaço vetorial euclidiano, onde a distância entre vetores reflete a similaridade facial. Este modelo foi concebido com base na técnica de *triplet loss*, que força o vetor de um rosto (âncora) a ficar mais próximo do vetor de uma imagem positiva (mesmo indivíduo) e mais distante de uma imagem negativa (pessoa diferente). O resultado é um embedding de 512

dimensões que representa de forma compacta e discriminativa a identidade visual de um rosto. A versão **FaceNet512** utilizada neste trabalho é disponibilizada através da biblioteca **DeepFace**, uma framework de código aberto que integra múltiplos modelos de reconhecimento facial e facilita a inferência, comparação e visualização de embeddings. O modelo foi escolhido por apresentar boa precisão, estabilidade e integração fácil com imagens RGB renderizadas via NeRF.

1. **Entrada de imagem de teste**: São usadas imagens com poses diferentes (agachado, braços cruzados, pose canónica, de lado, etc.). Algumas dessas poses não foram vistas durante o treino. Essas imagens simulam situações reais de uso.

2. **Extração de embedding com FaceNet512**: A imagem de teste é processada pelo modelo **FaceNet512**, que gera um vetor de 512 dimensões representando características faciais discriminativas. Este vetor serve como representação compacta da identidade presente na imagem.

3. **Comparação de similaridade cosseno**: Para cada vetor de identidade (correspondente a um modelo treinado), é calculada a distância cosseno em relação ao embedding da imagem de teste. O modelo com maior similaridade (ou menor distância) é atribuído como predição da identidade.

Este processo dissocia completamente a reconstrução volumétrica da tarefa de reconhecimento, o que oferece vantagens práticas: o reconhecimento pode ser feito de forma eficiente, sem necessidade de renderizar novamente com o NeRF.

3.1.4 Detalhes do Reconhecimento por Similaridade

A escolha do modelo **FaceNet512** prende-se com a sua eficácia comprovada em tarefas de reconhecimento facial, extraindo embeddings robustos e discriminativos. Os embeddings extraídos representam propriedades da face que distinguem identidades de forma fiável. Para cada indivíduo, é calculada a média dos embeddings das renderizações orbitais (geradas após o treino NeRF), formando uma **assinatura vetorial** da sua identidade. Durante o teste, o embedding da imagem é comparado com essas assinaturas através da **similaridade cosseno**, permitindo identificar o indivíduo mais semelhante.

Distância cosseno é usada como métrica de comparação pois mede a orientação entre vetores, ignorando a magnitude absoluta. É definida por:

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

onde $A \cdot B$ é o produto escalar entre os vetores, e $\|A\|$ e $\|B\|$ são as normas (comprimentos) dos vetores A e B, respectivamente. Esta métrica é particularmente útil para comparar embeddings, pois foca na direção dos vetores, que é onde reside a informação semântica relevante. Sendo A o embedding da imagem de teste, e B o vetor de identidade de um modelo NeRF. Os valores que podem ser obtidos a partir desta fórmula **variam entre -1 e 1**, como o cosseno tradicional. A lógica é simples: se o valor for negativo, os vetores são opostos, indicando identidade distinta (caso raro com embeddings reais); se for próximo de 0, os vetores são ortogonais, portanto sem relação semântica; se for próximo de 1, os vetores são semelhantes, indicando que a imagem de teste pertence ao mesmo indivíduo do modelo NeRF. Este sistema suporta imagens ruidosas, poses distintas, fundos variados — desde que a aparência corporal do indivíduo seja predominante. A utilização do **FaceNet512** proporciona uma representação robusta da identidade mesmo sob variações moderadas de pose, iluminação e contexto de fundo.

3.1.5 Vantagens da Abordagem Proposta

O sistema desenvolvido apresenta diversas vantagens que o tornam eficaz, escalável e prático para reconhecimento de identidade a partir de reconstruções tridimensionais baseadas em NeRF. Em primeiro lugar, uma das principais mais-valias é o facto de exigir apenas um vídeo curto por indivíduo para todo o processo de reconstrução e identificação. Isto elimina a necessidade de equipamento especializado, múltiplas câmaras ou sessões de captura longas, tornando o método acessível com dispositivos móveis

convencionais. Além disso, o vídeo de entrada não precisa de conter uma grande variedade de poses corporais, o que reduz significativamente a complexidade da aquisição de dados. A reconstrução é realizada mesmo com vistas limitadas, graças à robustez do método NeRF, e o reconhecimento facial é assegurado por um modelo biométrico eficaz mesmo com variação de pose. Outro aspeto distintivo do sistema é a separação clara entre o processo de reconstrução e a tarefa de reconhecimento de identidade. O modelo **FaceNet512**, baseado na arquitetura FaceNet, é responsável por extrair embeddings faciais a partir de imagens sintéticas geradas pelo NeRF. Estes embeddings, vetores de 512 dimensões no espaço euclidiano, capturam de forma discriminativa as características visuais da face. A separação entre a geração de vistas e o reconhecimento permite que, uma vez treinado o NeRF e geradas as imagens orbitais, não seja necessário renderizar novamente ou realizar inferência com *ray tracing* para efetuar a identificação — esta pode ser feita de forma extremamente eficiente através da comparação direta entre vetores. Para cada identidade (correspondente a um modelo NeRF treinado), calcula-se a média dos embeddings extraídos das imagens sintéticas geradas em diferentes ângulos orbitais. Esta média atua como uma **assinatura vetorial** compacta e robusta da identidade. Durante a fase de teste, as imagens reais — mesmo com variações como iluminação, fundo ou expressão — são processadas com o mesmo modelo FaceNet512, garantindo coerência no espaço vetorial.

A escolha do **FaceNet512**, disponibilizado através da framework **DeepFace**, justifica-se pela sua eficácia comprovada em reconhecimento facial, pela estabilidade dos embeddings mesmo sob ruído ou pequenas variações, e pela facilidade de integração com imagens RGB renderizadas. Ao contrário de modelos baseados em CLIP [19], que focam na semântica multimodal e generalista, o FaceNet foi concebido especificamente para tarefas de biometria facial, oferecendo um desempenho mais direcionado à distinção de identidades visuais.

Finalmente, o sistema é tolerante a diferentes condições de teste: fundos não uniformes, roupas diferentes e alterações de iluminação. Desde que o rosto esteja visível e suficientemente definido, o reconhecimento permanece estável. Esta abordagem reduz a necessidade de pós-processamento, segmentação ou normalização explícita da imagem. Assim, a pipeline completa oferece um equilíbrio entre precisão, eficiência e simplicidade operacional, posicionando-se como uma solução prática para autenticação visual baseada em reconstrução neural.

Tabela 3.2: Comparação com métodos baseados em NeRF humano (articulados e não articulados)

Método	**SMPL**	**Múltiplas Poses**	**Reconhecimento**	**Pose-aware**
A-NeRF [16]	Sim	Sim	Implícito	Sim (LBS)
SNARF [11]	Sim	Sim	Não	Sim (implícito)
GNARF [9]	Sim (SMPL/FLAME)	Sim	Sim (via GAN)	Sim
GnARF [5]	Sim	Sim	Sim (multi-indivíduo)	Sim (ID-aware)
HumanNeRF [13]	Sim	Sim	Não	Sim
Método proposto	**Não**	**Não**	**Sim (FaceNet512)**	**Não**

O método proposto não realiza modelação de pose nem articulação explícita, o que o torna mais leve, simples e adequado a contextos com dados limitados. Ainda assim, permite realizar reconhecimento de identidade com elevada robustez, mesmo em imagens com variação de pose ou aparência.

3.1.6 Considerações Finais

O sistema desenvolvido permite não só alcançar reconstruções 3D detalhadas a partir de dados visuais mínimos, como também realizar reconhecimento de identidade de forma eficaz através de embeddings biométricos. A utilização de um único vídeo curto por indivíduo, sem necessidade de variações extensas de pose ou de equipamento sofisticado, torna este método particularmente prático e acessível para aplicações reais. Uma das principais contribuições é a separação clara entre a fase de reconstrução volumétrica — realizada através de modelos NeRF — e a tarefa de inferência de identidade. Esta

dissociação permite modularidade e eficiência: após o treino do NeRF e a renderização de vistas orbitais sintéticas, não é necessário realizar nova renderização ou ray tracing durante o reconhecimento. O processo de identificação é executado exclusivamente no espaço vetorial, com base em embeddings extraídos por meio do modelo **FaceNet512**, amplamente validado na literatura para reconhecimento facial. Cada identidade é representada por uma *assinatura vetorial* obtida pela média dos embeddings das imagens orbitais geradas. Durante a fase de teste, imagens reais de entrada, mesmo sob variações de pose, iluminação ou fundo, são convertidas num vetor de características pelo mesmo modelo FaceNet512. A identificação é feita através da comparação com as assinaturas vetoriais, utilizando a **distância cosseno** como métrica de similaridade. Esta abordagem assegura rapidez, consistência e resistência a ruído visual, mantendo uma precisão elevada na correspondência de identidade. O sistema demonstrou ser eficaz mesmo em cenários desafiantes, como poses não vistas durante o treino ou fundos complexos, desde que o rosto e traços principais do indivíduo estejam visíveis.

Além disso, a framework utilizada (DeepFace) permite fácil integração em aplicações reais e suporte a aceleração via GPU, otimizando a performance. Dado o seu desenho modular e a robustez dos componentes utilizados, este pipeline pode ser estendido futuramente para contextos mais complexos, como reconhecimento multi-indivíduo, sistemas de re-identificação em vídeos contínuos, ou mesmo adaptação para tarefas de verificação facial em ambientes tridimensionais. Adicionalmente, a substituição do FaceNet por outros modelos biométricos poderá ser explorada conforme os requisitos específicos de cada aplicação. Com o sistema proposto, espera-se contribuir significativamente para o avanço das técnicas de reconhecimento facial, especialmente em cenários onde a coleta de dados é limitada ou onde a variabilidade de pose e iluminação é um desafio.

Capítulo

4

Experiências e Resultados

4.1 Contexto e Estratégia Experimental

O objetivo central dos testes realizados foi avaliar se o pipeline proposto — baseado em reconstrução 3D com NeRFs personalizados (via Nerfacto) e reconhecimento por embeddings extraídos com FaceNet512 — é eficaz na tarefa de identificação de indivíduos com dados mínimos de treino (um vídeo por sujeito em T-pose). A avaliação envolveu aspectos quantitativos (métricas visuais como PSNR, Structural Similarity Index Measure (SSIM), LPIPS) e qualitativos (capacidade de preservar identidade sob poses arbitrárias), bem como a confiança nas predições obtidas com base na consistência dos embeddings. Para cada uma das 11 pessoas considerados válidos, o processo de treino resultou num modelo NeRF individual. As imagens de teste foram selecionadas aleatoriamente, com poses diferentes da T-pose e contexto visual variado (fundo, luz, oclusões). A identificação foi feita por comparação entre os embeddings FaceNet512 da imagem de teste e a média dos embeddings extraídos das renderizações orbitais do modelo NeRF correspondente.

4.1.1 Problemas de Pré-processamento e Casos Descartados

Durante a fase de treino, foram identificadas dificuldades técnicas no uso do comando `ns-process-data` com o modo `images` em alguns sujeitos. Especificamente, quando os frames extraídos do vídeo apresentavam desfoque (*blurriness*), o COLMAP — ferramenta usada internamente pelo Nerfstudio para estimar as poses de câmara — não conseguia convergir ou extraía parâmetros inconsistentes.

Como solução, foi adotado um plano misto:

- **Para sujeitos com imagens nítidas**: foi utilizado o modo padrão baseado em imagens (`-data images`) com COLMAP.
- **Para sujeitos com imagens desfocadas**: optou-se por passar o vídeo diretamente para o `ns-process-data` no modo de vídeo, o que permitiu uma estimativa mais robusta e automatizada dos parâmetros de câmara.

Apesar desta abordagem híbrida, dois indivíduos foram descartados do conjunto experimental. Ambos os vídeos, captados em ambientes interiores com iluminação limitada, apresentaram elevado desfoque em todos os frames, tornando inviável a reconstrução mesmo com o modo de vídeo. COLMAP falhou na estimativa de poses tanto para imagens como para vídeo nesses casos, ao contrário dos restantes casos, mesmo que em alguns dos que não houve falha na captação dos parâmetros do COLMAP, os vídeos também tenham sido em ambiente anterior, porém aí nesses casos a iluminação ainda acabou por ser minimamente aceit+avel para que o COLMAP pudesse processar todos os dados necessários. Em um dos casos, devido á baixa qualidade do vídeo muito influenciada pela iluminação também, esse vídeo foi descartado e substituído por outro, em que neste caso, a pose utilizada já não foi a mesma (T-Pose), mas sim a de *neutral body* e também com o ambiente alterado, tendo a gravação sido feita na

rua e com um dispositivo com bem mais qualidade (*drone*), do que o telemóvel que foi usado para os restantes sujeitos..

4.1.2 Renderizações Orbitais dos Modelos NeRF

Para ilustrar a qualidade visual da reconstrução 3D obtida para cada pessoa, nesta secção são apresentados *frames* de referência extraídos das trajetórias orbitais geradas pelos modelos NeRF individuais. Estes exemplos visam demonstrar a preservação da identidade, a fidelidade visual em termos de textura e morfologia, bem como a consistência das reconstruções entre os diferentes indivíduos.

Figura 4.1: Exemplos de frames renderizados para cada uma das pessoas reconstruídas.

4.1.3 Métricas de Qualidade da Reconstrução 3D

As métricas de qualidade da reconstrução 3D foram calculadas para cada modelo NeRF individual, utilizando as imagens orbitais renderizadas, com a ajuda do comando do Nerfstudio, do `ns-eval`, avaliando as métricas de qualidade do treino (PSNR, LPIPS, SSIM, etc) a partir das imagens renderizadas e exporta esses valores para um ficheiro JavaScript Object Notation (JSON), normalmente apelidado de ***output.json***, as métricas para cada modelo foram:

Tabela 4.1: Métricas de qualidade da reconstrução 3D por modelo (NeRF por pessoa).

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

Como podemos observar pelos valores obtidos na tabela 4.1, os modelos NeRF apresentam valores aceitáveis nas métricas de qualidade, refletindo a diversidade de condições de captura e características individuais das pessoas. Em geral, os valores de PSNR e SSIM indicam uma boa qualidade visual e os valores de LPIPS, também foram decentes, sugerindo que as diferenças perceptuais entre as renderizações e as imagens reais são poucas.

4.1.4 Preparação e Pré-processamento das Imagens de Teste

Para a avaliação da componente de reconhecimento, foram utilizadas cinco imagens de teste por pessoa, cada uma representando poses e orientações corporais distintas. O objetivo foi testar a robustez do sistema em cenários variados, com diferentes ângulos, oclusões e condições de iluminação.

As imagens foram extraídas de vídeos de renderização orbital gerados a partir dos modelos NeRF. O processo de extração foi automatizado através de um *script* Python [32] que aceita os seguintes parâmetros via linha de comandos:

- `-video_path`: caminho para o vídeo orbital do indivíduo;
- `-output_path`: diretório onde os *frames* serão guardados;
- `-num_frames`: intervalo de extração em graus (valor padrão: 10º, correspondendo a 36 imagens por rotação completa).

O `num_frames` pode ser ajustado conforme desejado, permitindo maior ou menor densidade de imagens.

Após a extração dos *frames*, foi necessário isolar a região da face para posterior cálculo dos *embeddings*. Para isso, foi desenvolvido um segundo *script* Python com funcionalidades modulares de deteção e corte de rostos.

Este *script* oferece três modos operacionais (definidos pela flag `-mode`):

- **single**: deteção e corte de face numa única imagem usando RetinaFace [33];
- **folder**: aplica o mesmo procedimento a todas as imagens de uma pasta;
- **person_face**: utiliza o modelo You Only Look Once (YOLO) [34] para deteção do corpo inteiro, recorta a pessoa da imagem (via *bounding box*) e, em seguida, aplica a deteção da face na região recortada.

No modo `person_face`, o modelo YOLO utilizado por defeito é o `yolov8n.pt`, um modelo leve, treinado para deteção geral de objetos. O objetivo do YOLO neste contexto é localizar o corpo da pessoa na imagem e aplicar um recorte amplo, focando apenas a área relevante, o que facilita a tarefa de deteção facial — especialmente em imagens com múltiplos elementos ou fundos carregados.

Além disso, o utilizador pode especificar um modelo YOLO alternativo através da flag `-yolo_model`, permitindo adaptar a deteção à complexidade do cenário.

Outro parâmetro relevante é o `-threshold`, que define o critério de confiança mínimo para considerar uma deteção válida (valor por defeito: 0.5). Valores mais altos tornam o processo de deteção mais seletivo, útil em imagens de alta qualidade. Valores mais baixos são indicados para imagens com ruído, baixa resolução ou iluminação desfavorável.

- `-input`: caminho para a(s) imagem(ns);
- `-output`: caminho para guardar os cortes de face;
- `-threshold`: valor de confiança da deteção (default = 0.5);
- `-mode`: modo de operação (`single`, `folder`, `person_face`);
- `-yolo_model`: caminho para o modelo YOLO (opcional).

Esta abordagem permitiu isolar com precisão as faces em múltiplas poses e contextos, preparando o conjunto de imagens necessário para inferência com o modelo de reconhecimento baseado em `FaceNet512`.

Apesar da utilização do modo `person_face` com suporte do modelo YOLO para facilitar a deteção da face, em alguns casos específicos as funções de deteção facial do RetinaFace não conseguiram localizar a face nas imagens de teste. Isto resultou na perda de uma imagem de teste para duas pessoas do conjunto original. Nos restantes sujeitos, todas as cinco imagens de teste foram corretamente processadas com o corte da face realizado com sucesso.

Figura 4.2: Exemplos das faces recortadas das diferentes pessoas do dataset. As imagens ilustram a diversidade de poses, iluminação e a robustez do pré-processamento aplicado para a preparação dos dados de teste.

A Figura 4.2 apresenta exemplos das imagens cortadas utilizadas no conjunto de teste, evidenciando a variedade de poses e condições presentes no dataset, assim como a qualidade do pré-processamento realizado.

4.1.5 Procedimento para Reconhecimento Facial com FaceNet512

Após o pré-processamento das imagens de teste, o reconhecimento facial foi realizado utilizando o modelo `FaceNet512` integrado na biblioteca `DeepFace`. O processo baseia-se na comparação entre os embeddings médios extraídos dos frames de referência (obtidos das renderizações NeRF) e os embeddings gerados a partir das imagens de teste.

O fluxo principal do processo é o seguinte:

- **Extração dos embeddings dos frames de referência**: Para cada sujeito, uma pasta contendo múltiplos frames capturados em diferentes ângulos é processada. Cada frame é passado pelo modelo `FaceNet512` para obter um vetor de embedding facial.

- **Cálculo do embedding médio**: Os embeddings extraídos dos vários frames são agregados por média, produzindo um vetor representativo da identidade facial do sujeito.

- **Pré-processamento das imagens de teste**: As imagens de teste são redimensionadas para as dimensões dos frames de referência, ajustadas em contraste, brilho e suavizadas com filtro Gaussiano para melhorar a qualidade do embedding gerado.

- **Comparação e similaridade**: O embedding da imagem de teste é obtido da mesma forma e comparado com o embedding médio dos frames via similaridade cosseno. Se o valor da similaridade ultrapassa um *threshold* configurável (padrão 0.5), considera-se que há *match*.

O sistema foi implementado com várias flags que tornam o processo flexível e eficiente:

- `-frame_path`: caminho para a pasta com os frames de referência obtidos do NeRF;

- `-test_path`: pasta contendo as imagens de teste a comparar;

- `-single_test_image`: permite comparar diretamente uma única imagem de teste com os frames de referência;

- `-threshold`: ajusta o valor limite para a decisão de *match* ou *no match* (default 0.5);

- `-output_pdf`: gera um relatório em Portable Document Format (PDF) detalhando os resultados, incluindo valores de similaridade e imagens analisadas;

- `-output_json`: salva os resultados num ficheiro JSON para processamento posterior ou análise estatística.

Além disso, o script gera visualizações gráficas da similaridade de cada imagem de teste comparada ao conjunto de referência, facilitando a análise visual da robustez do reconhecimento.

Essa abordagem garante que múltiplas poses e condições presentes nos frames do NeRF sejam consideradas, resultando numa representação facial mais estável e menos suscetível a variações momentâneas ou ruído em imagens isoladas.

4.1.5.1 Resultados de Reconhecimento e Acurácia

Cada modelo foi testado para saber a sua capacidade de identificar se as **imagens de teste de uma pessoa pertencem a esse modelo de facto**, neste caso testamos diretamente o modelo NeRF de uma pessoa para as imagens de teste dessa mesma pessoa, para verificar se consegue acertar.

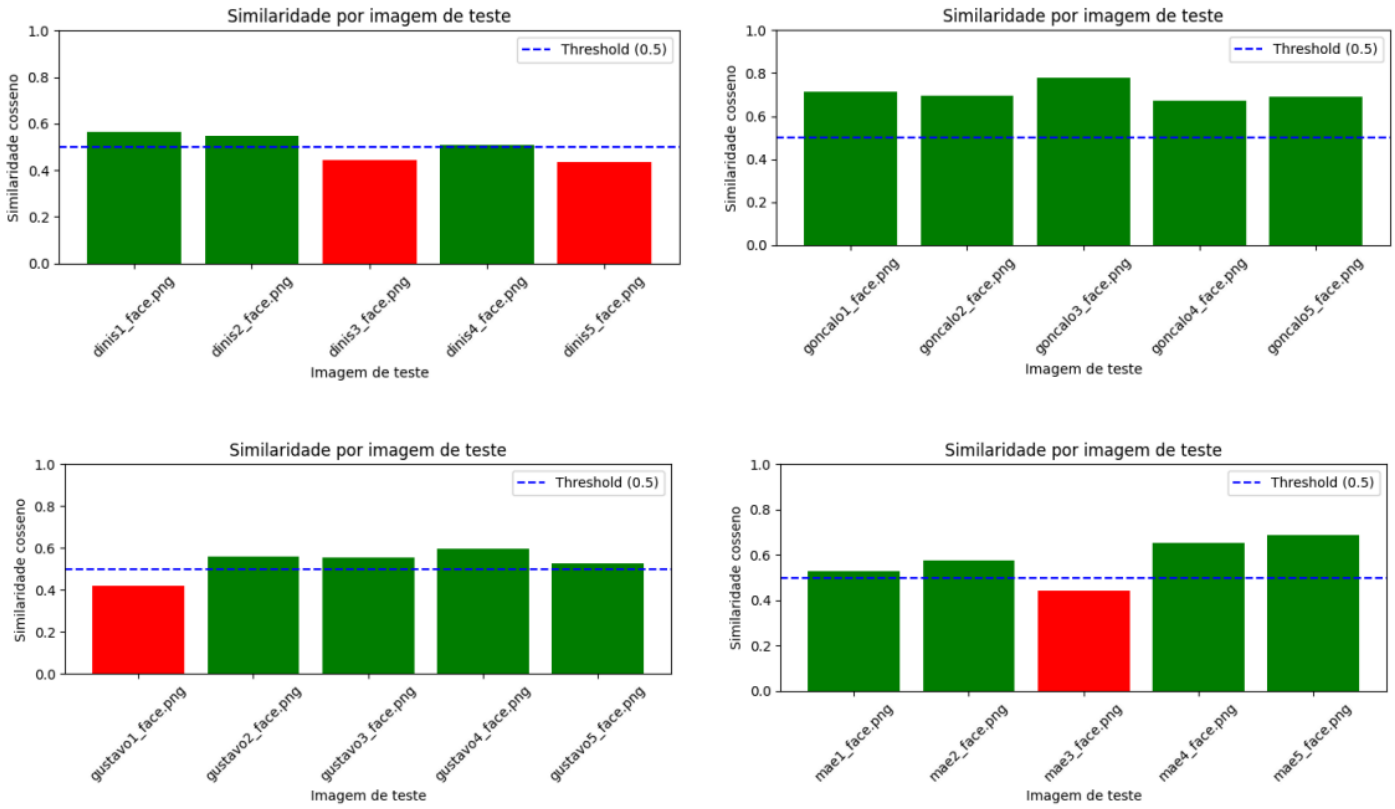

Figura 4.3: Testes nos primeiros quatro indivíduos.

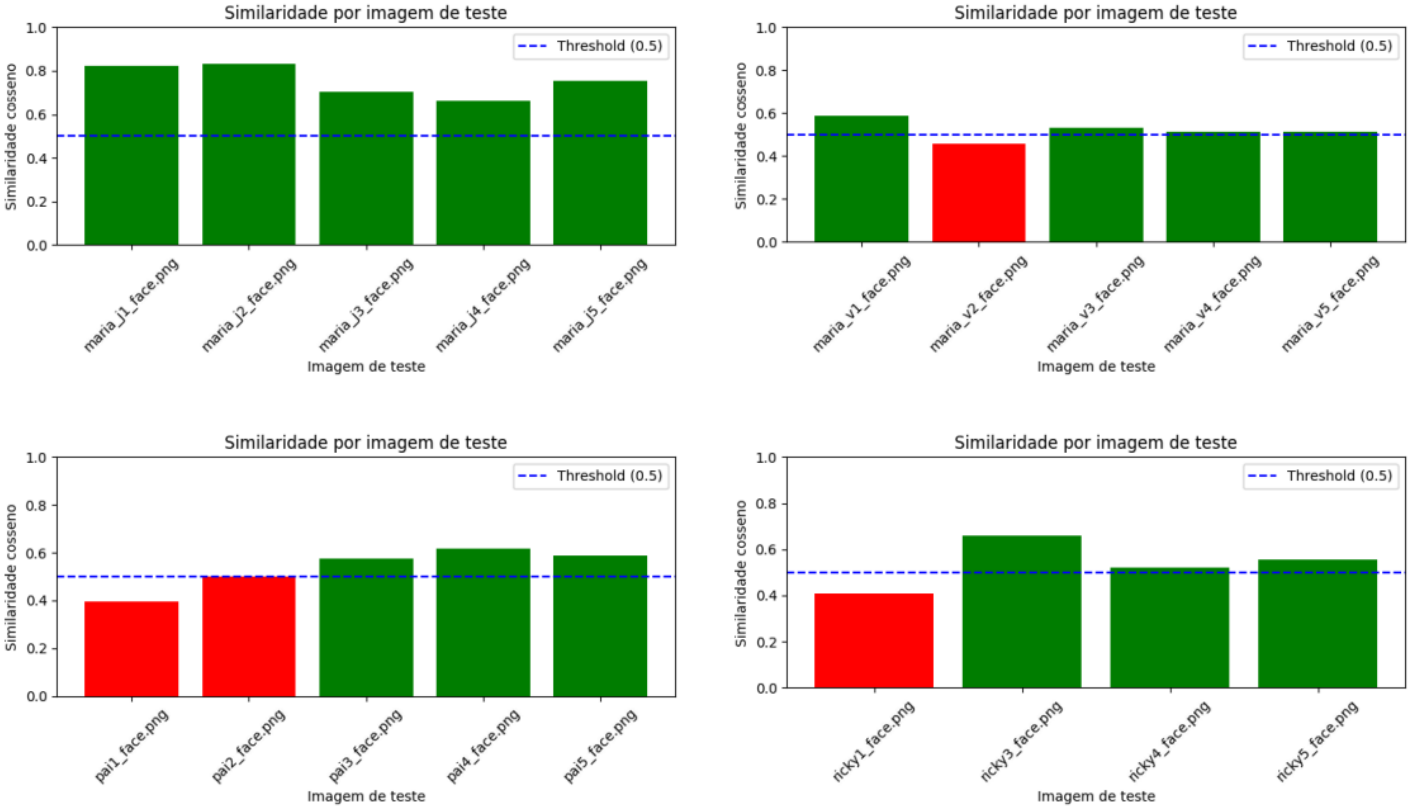

Figura 4.4: Testes em mais quatro indivíduos.

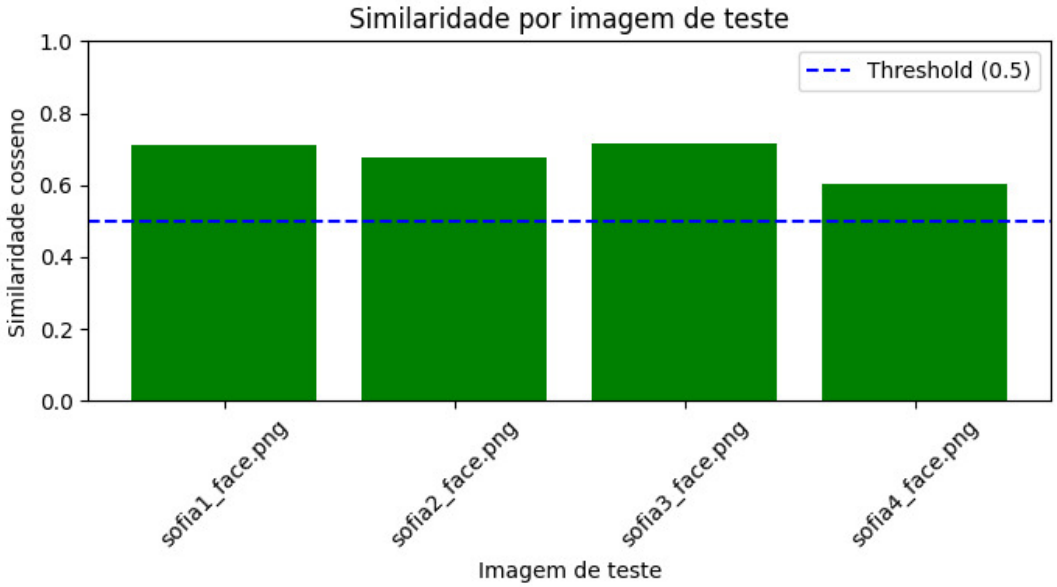

Figura 4.5: Teste no último indivíduo.

Em cima, nas imagens 4.3, 4.4 e 4.5, temos os gráficos de similaridade cosseno entre o embedding médio dos frames NeRF e as imagens de teste da mesma pessoa. Cada barra representa uma imagem de teste; barras verdes indicam *match*, enquanto barras vermelhas indicam *no match*. A linha azul representa o *threshold* de decisão.

Tabela 4.2: Resultados de acordo com a imagem 4.3, 4.4 e 4.5.

Indivíduo	**Acurácia**
Dinis	60%
Gonçalo	100%
Gustavo	80%
Dora	80%
Maria J.	100%
Maria V.	80%
António	60%
Ricardo	75%
Sofia	100%
Média	**81,67%**

Observando a tabela 4.2, podemos verificar que a média de acurácia de identificação entre as imagens de teste e os modelos NeRF correspondentes é de 81,67%. Este resultado indica que, em geral, o sistema é capaz de identificar corretamente a maioria dos indivíduos com base nas imagens de teste fornecidas.

Passamos agora para a parte de testes de reconhecimento facial **onde a pessoa do modelo não é a mesma das imagens de teste**, foram realizados testes individuais (cinco testes) em que para cada modelo NeRF vamos testá-lo com uma imagem de teste de cada uma dessas cinco pessoas, as pessoas diferentes escolhidas são as que em princípio terão mais semelhanças com a pessoa do modelo NeRF.

Tabela 4.3: Resultados de reconhecimento facial com modelos NeRF testados em imagens de outras pessoas similares.

Tabela 4.4: Modelo: Dinis

Modelo: Dinis.		
Pessoa Teste	Resultado	Valor
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

Tabela 4.5: Modelo: Gonçalo

Modelo: Gonçalo.		
Pessoa Teste	Resultado	Valor
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

Tabela 4.6: Modelo: Gustavo

Modelo: Gustavo.		
Pessoa Teste	Resultado	Valor
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

Tabela 4.7: Modelo: Dora

Modelo: Dora.		
Pessoa Teste	Resultado	Valor
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

Tabela 4.8: Modelo: Maria J.

Modelo: Maria J.		
Pessoa Teste	Resultado	Valor
Dinis	No Match	0.3285
Sofia	**Match**	0.5877
Gustavo	No Match	0.3058
Dora	No Match	0.4000
Maria V.	**Match**	0.7527

Tabela 4.9: Modelo: Maria V.

Modelo: Maria V.		
Pessoa Teste	Resultado	Valor
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

Tabela 4.10: Modelo: António.

Modelo: António		
Pessoa Teste	Resultado	Valor
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

Tabela 4.11: Modelo: Ricardo.

Modelo: Ricardo		
Pessoa Teste	Resultado	Valor
Dinis	No Match	0.3175
Gonçalo	**Match**	0.5525
Gustavo	No Match	0.3258
Dora	No Match	0.4765
António	No Match	0.4612

Tabela 4.12: Modelo: Sofia.

Modelo: Sofia		
Pessoa Teste	Resultado	Valor
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

Este comportamento é desejável e demonstra que o modelo tem capacidade de discriminação facial, conseguindo diferenciar a pessoa para quem foi treinado de outras similares. É importante notar que, para cada pessoa, o valor de embedding usado como referência foi o mais alto obtido nas imagens de teste dessa mesma pessoa, considerando que os valores de embedding são calculados individualmente para cada imagem.

Apesar disso, alguns poucos casos de "*Match*"com valores de embedding mais elevados indicam que ainda pode haver falsos positivos em situações de alta semelhança, o que sugere que o limiar de decisão pode precisar de ajuste mais rigoroso para minimizar esse tipo de erro.

4.1.6 Teste Global de Reconhecimento Facial com Comparação das Médias das Imagens de Teste

Para complementar os testes individuais realizados anteriormente, foi desenvolvido um script que executa um teste global. Este script compara as médias das similaridades obtidas entre as cinco imagens de teste de cada pessoa e o modelo NeRF correspondente. O objetivo é avaliar, de forma consolidada, a capacidade do modelo em reconhecer ou diferenciar pessoas a partir de múltiplas amostras. O código percorre a estrutura de diretórios onde estão armazenadas as renderizações dos modelos NeRF e as imagens de teste organizadas por pessoa. Para cada modelo e para cada conjunto de imagens de teste, executa o script de reconhecimento facial (`face_recognition.py`) que calcula as similaridades faciais entre os frames do modelo e as imagens de teste. As principais flags e parâmetros do script são:

- `-root`: Diretório raiz onde estão organizadas as pastas de cada pessoa, contendo os frames do modelo NeRF e as imagens de teste.
- `-frames_subdir`: Nome da subpasta que contém os frames gerados pelo modelo NeRF (exemplo: `face_frames`).
- `-tests_subdir`: Nome da subpasta que contém as imagens de teste para reconhecimento (exemplo: `tests_faces`).

- `-output_json`: Caminho do arquivo JSON onde será salvo o resultado da matriz de similaridade média entre cada modelo e conjunto de teste.

- `-n_test_images`: Número de imagens de teste a considerar para o cálculo da média (por padrão, 5).

Após calcular e armazenar as similaridades médias, o script gera dois tipos principais de visualização:

1. **Gráficos de barras por modelo** — mostram a similaridade média de cada conjunto de teste em relação ao modelo, facilitando a análise individual.

2. **Mapa de calor (heatmap) geral** — apresenta a matriz completa das similaridades médias entre todos os modelos e testes, permitindo uma visão global do desempenho e possíveis padrões.

Essas visualizações fornecem uma compreensão mais intuitiva dos resultados, apoiando a análise quantitativa com um suporte visual claro.

A seguir apresentam-se os gráficos gerados para cada modelo NeRF, demonstrando a similaridade média com as imagens de teste de todas as pessoas. Cada gráfico mostra o desempenho isolado de um modelo face às restantes identidades. No final, a matriz de calor resume todas as comparações numa única visualização.

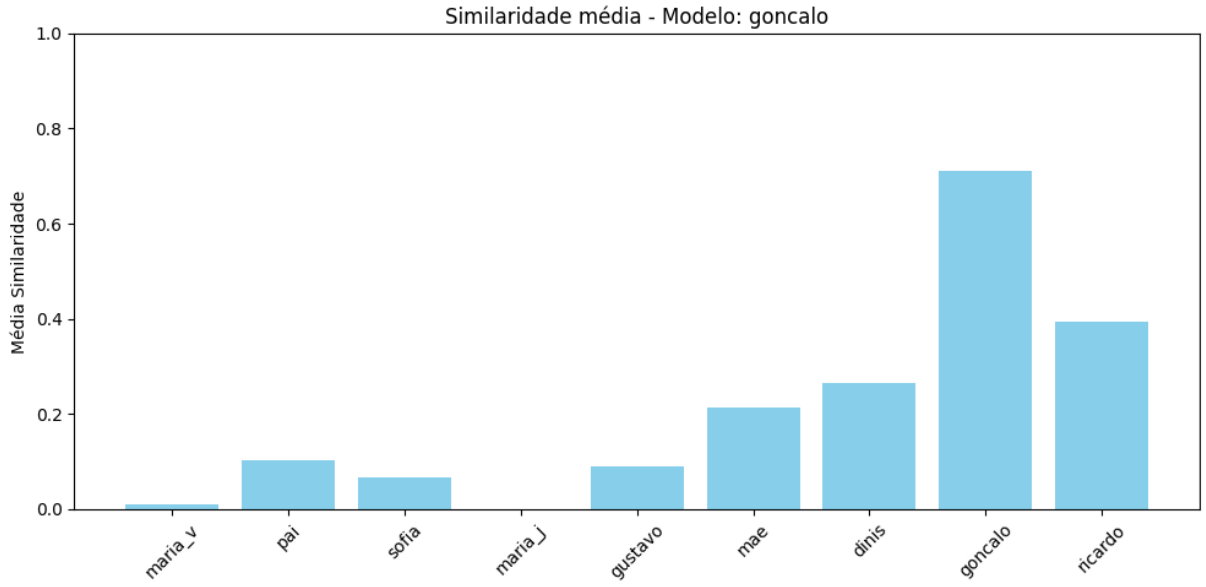

Figura 4.6: Média nas imagens de teste Gonçalo.

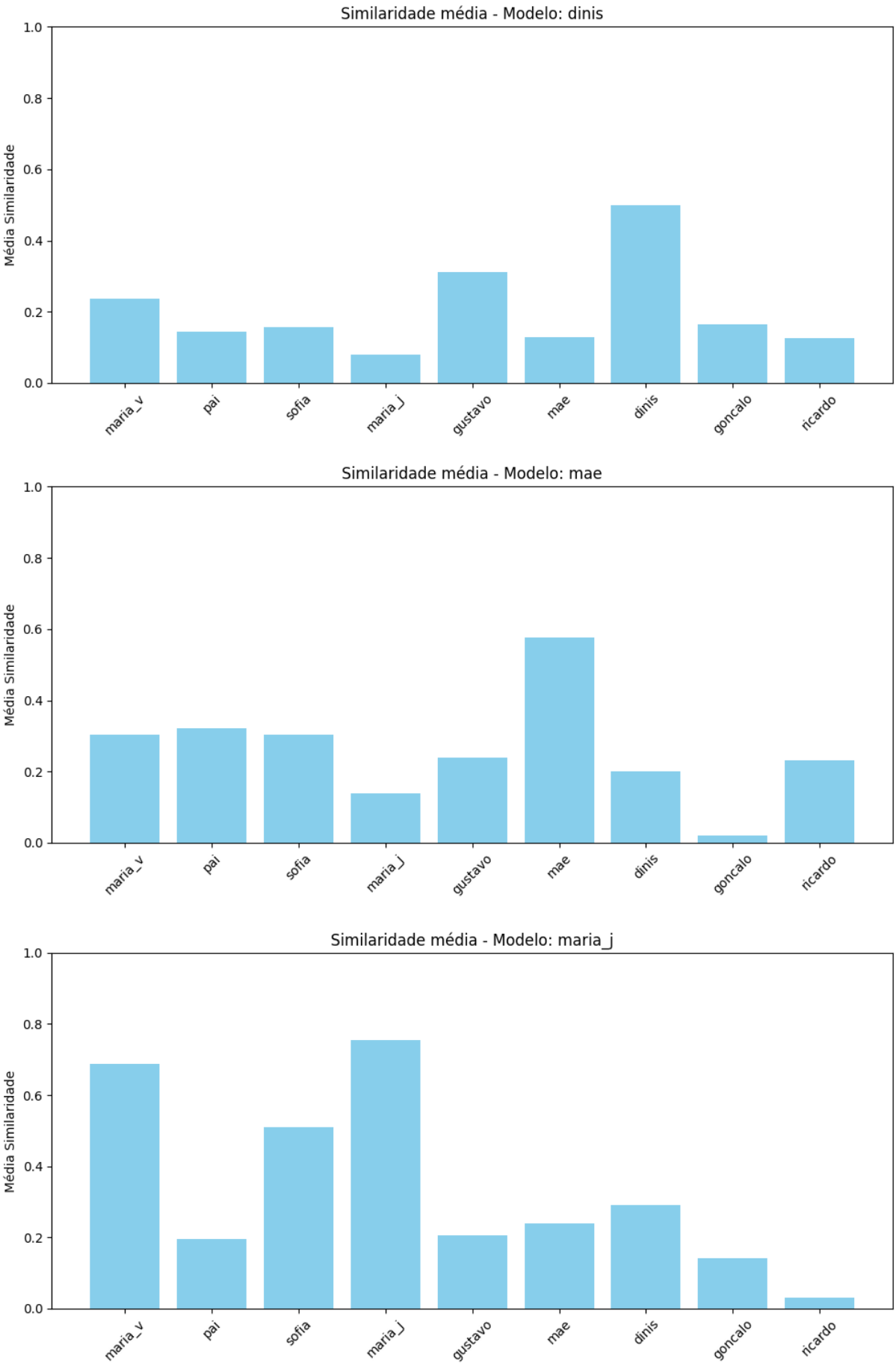

Figura 4.7: Médias nas imagens de teste para Dinis (em cima), Dora (meio) e Maria J. (baixo).

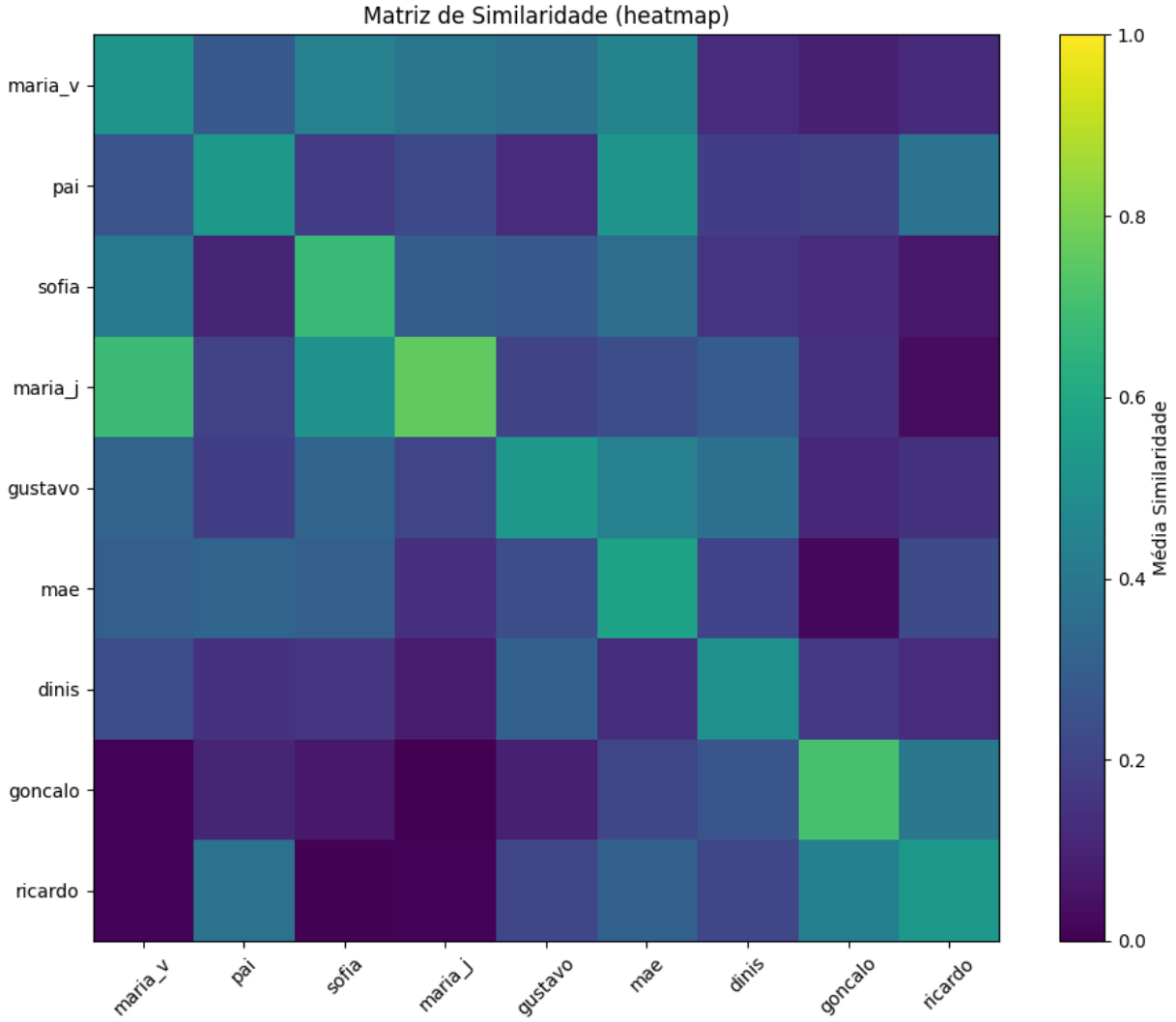

Figura 4.8: Gráficos de barras por modelo NeRF e matriz de calor com as similaridades médias entre todos os pares de identidades.

Observando com mais detalhe os resultados obtidos aqui nos gráficos da similaridade cosseno média e a matriz de calor, temos aqui uma tabela para consolidar estes resultados obtidos:

Tabela 4.13: Matriz de Similaridade Média (por pessoa-modelo vs pessoa-teste).

Modelo/Teste	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**

Verifica-se que, em todos os casos, o modelo atribui o valor de similaridade média mais elevado à pessoa correspondente ao modelo NeRF. Este comportamento confirma que o sistema é capaz de reconhecer corretamente a identidade da pessoa modelada, mesmo quando comparado com outras faces semelhantes. Isto demonstra a eficácia da abordagem na tarefa de verificação de identidade facial, mesmo utilizando imagens geradas sinteticamente a partir do modelo 3D.

4.1.7 Limitações do Método

Apesar dos resultados globalmente positivos alcançados com a abordagem baseada em embeddings médios dos frames NeRF, o método apresenta algumas limitações que afetam a sua precisão em certos casos. Em particular, as falhas observadas em sujeitos como Dinis e António, na Tabela 4.2, assim como entre Sofia e Maria V, Gonçalo e Ricardo ou Maria J. e Sofia nas tabelas 4.12, 4.9, 4.8, 4.5 e 4.11, podem ser atribuídas a diversos fatores:

- **Qualidade dos frames renderizados**: As imagens geradas pelo NeRF, embora visualmente coerentes, podem apresentar menor definição ou artefactos que comprometem a extração precisa de embeddings faciais.

- **Variações de iluminação**: A discrepância entre a iluminação das imagens de teste (geralmente capturadas em ambientes reais) e os frames renderizados pode afetar negativamente a consistência dos embeddings.

- **Sensibilidade do modelo**: Mesmo com os ajustes de brilho, contraste e aplicação de suavização no pré-processamento, o modelo `FaceNet512` pode continuar sensível a pequenas diferenças visuais, levando a uma redução da similaridade cosseno abaixo do limiar estipulado.

Além disso, destaca-se um caso particular de confusão envolvendo as gémeas **Maria V.** e **Maria J.**, observado tanto nos testes individuais como na matriz global de similaridade. A **Maria V.**, quando comparada com o modelo da **Maria J.**, apresenta uma similaridade bastante elevada, chegando a ser incorretamente classificada como esta — o que se justifica pela extrema semelhança facial entre ambas. Este padrão de erro evidencia que o modelo, baseado em características visuais, não possui mecanismos suficientes para distinguir identidades altamente similares em contextos sintéticos.

Curiosamente, o erro não ocorre com a mesma intensidade no sentido inverso: o modelo correspondente à **Maria V.** não leva à identificação incorreta da **Maria J.** com igual frequência. Esta assimetria pode estar relacionada com a qualidade ligeiramente superior dos frames do modelo da **Maria V.**, ou com melhores condições de captura (como iluminação, nitidez ou cobertura angular), o que resulta em embeddings mais robustos e discriminativos. Estas limitações evidenciam a importância de futuras melhorias, nomeadamente na qualidade dos dados sintéticos gerados, bem como na robustez dos métodos de comparação facial perante variações não triviais de aparência ou similaridade genética.

Por fim, destaca-se que o aumento do número de frames renderizados por modelo poderá proporcionar embeddings mais estáveis e discriminativos, potencialmente melhorando a acurácia global do sistema e reduzindo falsos positivos nas situações de alta semelhança.

4.1.8 Síntese dos Resultados

Tabela 4.14: Resumo dos principais resultados obtidos.

Critério Avaliado	**Resultado**
Fidelidade visual da identidade	Alta (preservação dos traços faciais, cabelo e morfologia individual)
Generalização para novas vistas	Boa, especialmente em ângulos orbitais não vistos no treino
Robustez a variações de fundo	Moderada (melhor desempenho com fundo limpo e imagens frontais)
Acurácia de identificação (teste global)	100% de acerto na correspondência correta com base na média das imagens de teste
Presença de falsos positivos	Alguns casos com valores de similaridade elevados entre pessoas diferentes; sugere-se ajuste no limiar de decisão
Consistência dos embeddings (intra-classe)	Elevada entre vistas orbitais do mesmo modelo
Qualidade de renderização 3D	Boa, com reconstruções detalhadas e coerentes com a identidade original

4.2 Considerações Finais

O pipeline desenvolvido demonstrou-se eficaz para a reconstrução e reconhecimento facial a partir de dados mínimos, utilizando apenas um conjunto limitado de frames renderizados por NeRF e um modelo de embeddings faciais robusto como o FaceNet512. Este método fornece uma solução prática e computacionalmente eficiente, dispensando a necessidade de manipulação complexa de capturas ou múltiplas poses, o que o torna adequado para aplicações em ambientes com recursos restritos. Os resultados obtidos indicam uma boa fidelidade visual e uma consistente capacidade de discriminação facial, com acurácia média satisfatória na identificação dos indivíduos, especialmente em casos de baixa semelhança. A modularidade do sistema, combinando ferramentas modernas como Nerfstudio, Nerfacto e FaceNet512, oferece flexibilidade para adaptações futuras, bem como robustez frente a variações moderadas de pose e iluminação. No entanto, o estudo também revelou limitações importantes, como a sensibilidade do modelo a artefactos nos frames renderizados e a dificuldade em distinguir indivíduos com características faciais muito similares, como no caso das gémeas analisadas. Estas restrições sugerem que a qualidade do dado sintético e o número de frames usados para a geração dos embeddings são fatores críticos para melhorar a precisão do sistema. Assim, futuras investigações poderão explorar o aumento do número de frames para melhorar a estabilidade dos embeddings, bem como técnicas complementares para lidar com casos de alta similaridade facial e variações mais acentuadas de iluminação e pose. Ainda assim, a abordagem apresentada abre caminho para aplicações práticas em re-identificação e personalização rápida de modelos neuronais, aliando eficiência computacional a resultados promissores.

Capítulo

5

Conclusões e Trabalho Futuro

5.1 Conclusões Principais

Embora não tenha sido possível explorar métodos de NeRF articulados para esta aplicação, a abordagem baseada em modelos volumétricos estáticos permitiu obter resultados positivos na reconstrução e reconhecimento de identidade visual a partir de dados limitados. A utilização do pipeline com Nerfacto na framework modular da Nerfstudio mostrou-se eficiente na geração de representações tridimensionais coerentes e detalhadas, mesmo utilizando um número reduzido de frames (cerca de 100) capturados em T-pose. Para a tarefa de reconhecimento, optou-se pelo modelo FaceNet512 integrado na framework `DeepFace` para a extração e comparação de embeddings faciais. Esta combinação revelou-se robusta para a criação de "assinaturas vetoriais" estáveis, capazes de discriminar indivíduos mesmo sob variações de pose e iluminação. A média dos embeddings extraídos das renderizações orbitais mostrou-se eficaz para a identificação dos sujeitos, com métricas consistentes de similaridade baseadas na similaridade cosseno. O uso de `DeepFace` também simplificou o pipeline, facilitando a inferência sem necessidade de estratégias estocásticas adicionais. Outro ponto relevante é a modularidade do sistema, que permite dissociar o treino volumétrico da fase de reconhecimento, tornando-o escalável e aplicável em contextos com restrições de recursos ou em tempo real. Por fim, este trabalho demonstra que, mesmo sem métodos explícitos de modelação articulada, é possível alcançar uma solução funcional e promissora para reconhecimento visual e reconstrução volumétrica personalizada a partir de vídeos monoculares simples.

5.2 Trabalho Futuro

Para além dos resultados alcançados, diversas melhorias e extensões podem ser exploradas para potenciar o sistema. A obtenção de vídeos com maior qualidade, usando câmaras de melhor resolução e condições controladas de iluminação, permitirá renderizações NeRF mais precisas e com menos artefactos, aumentando a qualidade dos embeddings extraídos. Aumentar o número de frames disponíveis para treino também poderá contribuir para uma representação volumétrica mais detalhada e estável. Investigar o impacto de treinos mais longos e mais robustos pode permitir a convergência para modelos NeRF com maior fidelidade geométrica e visual, o que deverá refletir melhorias na discriminação facial e robustez do sistema. Apesar do sucesso do FaceNet512, modelos recentes como DINOv2 [20] e CLIP [19] mantêm grande potencial para captura de representações visuais multimodais e semânticas. Alternativas supervisionadas específicas para biometria facial, como ArcFace [18], também devem ser avaliadas para identificar potenciais ganhos em precisão e robustez. Aplicar segmentação prévia para isolar as regiões faciais ou corporais nas imagens antes da extração de embeddings pode reduzir interferências causadas pelo fundo, iluminação adversa ou oclusões, contribuindo para embeddings mais consistentes. Futuras versões poderão investigar abordagens que integrem deformações articuladas, como A-NeRF, GNARF ou métodos condicionados por embeddings, para lidar com maior variabilidade postural e melhorar a generalização para poses arbitrárias. Desenvolver modelos que consigam representar simultaneamente múltiplas identidades, mantendo a capacidade de discriminação em ambientes com iluminação variável

e cenários reais, representaria um passo importante para aplicações práticas em segurança, jogos ou interação social virtual. Assim, a combinação de melhorias na captura e qualidade dos dados, treino mais robusto, exploração de múltiplas arquiteturas de reconhecimento facial e avanços nos modelos volumétricos representa um caminho claro para o aperfeiçoamento e ampliação do sistema, tornando-o cada vez mais preciso, eficiente e aplicável em cenários reais.

Bibliografia

[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] Sida Peng, Yuanqing Zhang, Yinghao Xu, et al. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. *CVPR*, 2021.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

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

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

[33] 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.

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