UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ISABELA MARINHO RIBEIRO

Considerações éticas sobre algoritmos de anonimização facial

Recife

2025

ISABELA MARINHO RIBEIRO

Considerações éticas sobre algoritmos de anonimização facial

Trabalho apresentado ao Programa de Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Áreas de Concentração: Visão Computacional e IA Responsável

Orientador: Francisco Paulo Magalhães Simões

Coorientador: Willams de Lima Costa

Recife

2025

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Ribeiro, Isabela Marinho.
Considerações éticas sobre algoritmos de anonimização facial / Isabela Marinho Ribeiro. - Recife, 2025.
45 p. : il., tab.

Orientador(a): Francisco Paulo Magalhães Simões
Cooorientador(a): Willams de Lima Costa
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Informática, Ciências da Computação - Bacharelado, 2025.
Inclui referências.

1. Análise de justiça. 2. Inteligência artificial. 3. Visão computacional. 4. Privacidade. I. Simões, Francisco Paulo Magalhães. (Orientação). II. Costa, Willams de Lima. (Coorientação). IV. Título.

000 CDD (22.ed.)

ISABELA MARINHO RIBEIRO

Considerações éticas sobre algoritmos de anonimização facial

Trabalho apresentado ao Programa de Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em: 21 de Março de 2025

BANCA EXAMINADORA

Prof. Dr. Francisco Paulo Magalhães Simões (Orientador)

Universidade Federal de Pernambuco - UFPE

Prof. Dr. Cleber Zanchettin (Examinador Interno)

Universidade Federal de Pernambuco - UFPE

RESUMO

Este trabalho investiga o impacto da aplicação de técnicas de anonimização de faces em imagens de pessoas, com enfoque na preservação da estrutura facial de indivíduos pertencentes a diferentes grupos demográficos. O estudo analisa quatro técnicas de anonimização aplicadas a três conjuntos de dados anotados em categorias de raça e gênero, com o intuito de avaliar a existência de vieses e quantificar o nível de viés de cada um. A quantificação é feita através do uso de métricas de distorção visual e de *score* de qualidade perceptível da face. Os valores gerados são então avaliados por métodos de análise de dados. Os resultados mostram que as técnicas baseadas em redes adversárias generativas distorcem mais a face humana, comprometendo a utilidade da face gerada, enquanto a técnica baseada em modelo de difusão preserva mais a estrutura facial. Um dos algoritmos baseados em redes adversárias mostrou discrepâncias atípicas em valores de deformação de gêneros e raças diferentes, o que pode indicar a presença de vieses demográficos no modelo. Por fim, as análises apontaram para limitações em ambas as métricas adotadas, pela presença de vieses que podem influenciar a interpretação das análises das técnicas, demandando melhorias futuras.

Palavras-chaves: Anonimização. Imagem. Deformação. Reconhecimento facial. Privacidade.

ABSTRACT

This work investigates the impact of applying face anonymization techniques on images of people, with focus on the preservation of facial structure in individuals from different demographic groups. The study analyzes four anonymization techniques applied to three datasets annotated with race and gender categories to assess the existence of biases and quantify their level. The quantification is made using visual distortion metrics and perceptual face quality scores. The generated values are then evaluated through data analysis methods. The results show that the techniques based on generative adversarial networks distort the human face more, compromising the usability of the generated face, while the diffusion model-based technique preserves facial structure better. One of the algorithms based on adversarial networks showed atypical discrepancies in deformation values across different gender and race groups, which may indicate the presence of demographic biases in the model. Finally, the analyses highlighted limitations in both adopted metrics due to biases that can influence the interpretation of the techniques' assessments, requiring future improvements.

Keywords: Anonymization. Image. Deformation. Facial recognition. Privacy.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BRISQUE	*Blind/Referenceless Image Spatial Quality Evaluator*
DEX	*Deep EXpectation*
EPE	*Endpoint Error*
FADM	*Full-Body Anonymization using Diffusion Models*
FEI	Fundação Educacional Inaciana Padre Sabóia de Medeiros
FIQA	*Face Image Quality Assessment*
FR-IQA	*Full-Reference Image Quality Assessment*
GAN	*Generative Adversarial Networks*
GDPR	*General Data Protection Regulation*
IFQA	*Interpretable Face Quality Assessment*
IQA	*Image Quality Assessment*
JPEG	*Joint Photographic Experts Group*
LDFA	*Latent Diffusion Face Anonymization*
LGPD	Lei Geral de Proteção de Dados
NR-IQA	*No-Reference Image Quality Assessment*
PSNR	*Peak Signal-to-Noise Ratio*
RAD	*Realistic Anonymization using Diffusion*
SG-GAN	*Surface-Guided GAN*
SSIM	*Structural Similarity Index Measure*
V-SAM	*Variational Surface-Adaptive Modulation*

LISTA DE SÍMBOLOS

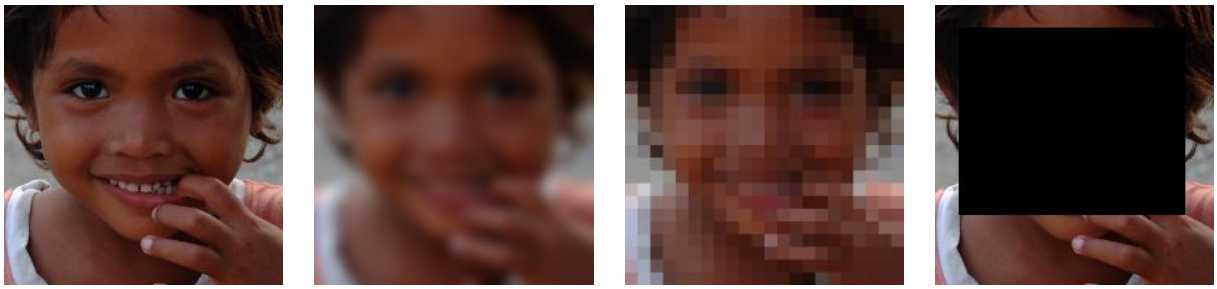
$\in$	Pertence
r	Coeficiente de correlação de Pearson
p	Valor-p
α	Nível de significância
σ	Desvio padrão

SUMÁRIO

1 INTRODUÇÃO

Com o rápido crescimento na quantidade de aplicações utilizando reconhecimento facial, essa tecnologia pode ser considerada ubíqua, ou seja, presente em vários contextos do dia a dia. Através de sistemas de câmeras de vigilância, autenticação em dispositivos móveis e até no controle de imigração em aeroportos, os cidadãos são cada vez mais expostos à coleta de dados biométricos sensíveis (SEO et al., 2015). Na maioria das vezes, o objetivo da coleta não transparece aos usuários de forma clara, de forma que eles não sabem com que propósito suas informações serão usadas (SMITH; MILLER, 2022).

Nesse contexto, algoritmos de anonimização surgiram como uma forma de preservar a privacidade do usuário. Essas técnicas envolvem desde abordagens elementares, como o borramento da face da pessoa, até métodos refinados, como as *Generative Adversarial Networks* (GAN). Os métodos elementares são amplamente usados em situações onde a preservação do anonimato é essencial. Além do borramento, outras técnicas também são aplicadas, como a ocultação por tarja e a pixelização, que corresponde à redução da resolução de um trecho da imagem, de forma a tornar os *pixels* visíveis. Apesar de simples de serem implementadas, essas técnicas provocam perda de informações visuais, de forma que certos detalhes se tornam indistinguíveis. Esses detalhes podem ser as bordas das faces, as expressões faciais e até informações demográficas, como a idade da pessoa anonimizada. Exemplos desses métodos são vistos na Figura 1.



(a) Imagem original (b) Imagem borrada (c) Imagem pixelizada (d) Imagem com tarja

Figura 1 – Métodos elementares de anonimização facial.

Os métodos mais recentes, por sua vez, utilizam arquiteturas baseadas em GANs (HELLMANN et al., 2024) e modelos de difusão (ZWICK et al., 2024) para gerar faces realistas, que substituem o rosto contido na imagem original por uma versão anonimizada. Dessa forma, os detalhes visuais da face e do seu entorno não são perdidos, como visto na Figura 2.

(a) Imagem original (b) Imagem gerada por GAN (c) Imagem gerada por difusão

Figura 2 – Métodos generativos de anonimização facial.

1.1 MOTIVAÇÃO

O estudo de justiça em sistemas automatizados aborda a forma como os algoritmos tomam decisões de maneira equitativa e seguindo princípios éticos, garantindo que não haja discriminação. Por exemplo, um sistema que nega empréstimos para mulheres com mais frequência, mesmo quando elas têm as mesmas condições que homens, está sendo injusto. A área de justiça (ou *Fairness*) teve origem no século 20 e ganhou força na última década, com a ascensão dos sistemas baseados em inteligência artificial. A presença de vieses em redes neurais pode ser provocada por limitações técnicas, como problemas no design dos algoritmos (NOIRET; LUMETZBERGER; KAMPEL, 2021), pela inexistência de boas práticas de coleta e processamento dos dados (BAROCAS; SELBST, 2016) e até por falhas no trabalho de pesquisa anterior ao desenvolvimento dos sistemas (SELBST et al., 2019). No contexto de modelos que preservam a privacidade de faces, eles também estão sujeitos a vieses, de modo que grupos demográficos diferentes podem ser anonimizados de forma desigual. A estrutura da face pode ser distorcida por um modelo a tal ponto que deixe de se assemelhar a um rosto humano, podendo, assim, preservar melhor a estrutura facial em certos grupos em detrimento de outros. A partir daí, sistemas que dependem dos dados anonimizados podem ter sua performance comprometida, o que pode gerar problemas éticos e legais, além de minimizar a utilidade dos dados anonimizados.

Nesse âmbito, foi encontrada uma lacuna na literatura com relação à análise de justiça em algoritmos de anonimização facial. Na raiz dessa lacuna, há um problema de escassez de métricas e *benchmarks* específicos para avaliação da qualidade e utilidade de imagens geradas artificialmente (BRANT, 2024). Por isso, ainda não compreendemos a extensão nem os mecanismos pelos quais essas técnicas podem estar reproduzindo vieses visuais, o que pode

impactar diretamente na aplicabilidade das imagens geradas. Dependendo do grau de deformação presente, as arquiteturas generativas podem não apresentar vantagens significativas com relação a algoritmos baseados em abordagens elementares, executando a função de ocultar a face, mas sem preservar detalhes visuais.

1.2 OBJETIVOS

O objetivo deste trabalho é avaliar a hipótese de que algoritmos de anonimização de faces podem apresentar vieses de gênero ou raça. Acerca dos vieses dos modelos, queremos avaliar em que nível eles estão presentes na tarefa de anonimização. Para isso, uma análise comparativa será realizada entre quatro diferentes algoritmos sobre a qualidade da reconstrução das faces e da deformação dos rostos. Essa análise será executada entre diferentes grupos demográficos, com o objetivo de avaliar a justiça dos modelos. Os resultados dos experimentos serão examinados para identificar padrões quantitativos e visuais que influenciam a utilidade das faces sintéticas e, durante a validação das hipóteses, descobrimos que as métricas avaliadoras dos algoritmos possuem seus próprios vieses, os quais também serão investigados e descritos nesta pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

Para compreender os desafios relacionados à anonimização de faces e seus vieses, é necessário revisar conceitos e trabalhos relacionados, o que faremos nesta seção.

2.1 ANONIMIZAÇÃO DE FACES

Anonimização é o processo de modificar ou remover características de um dado, de modo a impedir a identificação de indivíduos. Ela é utilizada como um método de pré-processamento de dados em aplicações com requisitos de privacidade, de forma que ela permite o uso dos dados sem comprometer a identidade das pessoas envolvidas, o que é importante em setores como saúde (CHEVRIER et al., 2019) e o campo jurídico (CSÁNYI et al., 2021), onde a análise de dados sensíveis pode impactar de forma significativa a vida das pessoas. A demanda por transparência nas instituições sobre práticas de anonimização é crescente, visto que a coleta de dados em escala cresce rapidamente e o risco de reidentificação dos dados anonimizados aumenta (PHILLIPS; DOVE; KNOPPERS, 2017). As práticas a favor da privacidade são defendidas por regulamentações de proteção de dados, como a Lei Geral de Proteção de Dados (LGPD) no Brasil e a *General Data Protection Regulation* (GDPR) na Europa.

No contexto de visão computacional, a anonimização está associada à modificação de dados que podem identificar uma pessoa em imagens e vídeos, através de características faciais ou corporais. As técnicas mais comumente utilizadas se baseiam no borramento ou no uso de outros efeitos visuais para ocultar a identidade da pessoa, o que provoca uma perda de informações para além dos dados sensíveis, impactando a utilidade dos dados em tarefas posteriores à anonimização. Apesar disso, esses métodos elementares ainda têm um papel importante na área de privacidade, especialmente em cenários em que soluções rápidas e diretas são necessárias. Entretanto, com o avanço das tecnologias baseadas em biometria, as preocupações com a privacidade também aumentaram: o ato de regulamentação do uso de inteligência artificial feito pela União Europeia proibiu o uso de ferramentas biométricas em contextos como trabalho e educação por riscos à privacidade (PARLIAMENT, 2024). Portanto, técnicas de anonimização podem ser úteis nesse contexto para minimizar os riscos de exposição, colaborando com a necessidade do desenvolvimento de algoritmos que preservam a privacidade em contextos mais complexos.

2.1.1 Redes Generativas

Nesse cenário, entram as redes generativas, que consistem em modelos de inteligência artificial projetados para criar novos dados com características semelhantes aos dados em que foram treinados. No contexto de anonimização de faces, esses modelos geram novas imagens contendo rostos sintéticos que substituem os originais e preservam informações da imagem além da face. Um exemplo de redes generativas são as GANs (MAXIMOV; ELEZI; LEAL-TAIXÉ, 2020; BARATTIN et al., 2023), que são compostas por dois modelos em competição: um gerador, que cria novos dados, e um discriminador, que avalia se os dados gerados são autênticos ou falsos. Esse treinamento melhora progressivamente a qualidade dos dados gerados. Outra abordagem relevante é o *Stable Diffusion* (KLEMP et al., 2023; MALM et al., 2023), um tipo de modelo baseado em difusão que transforma gradualmente um ruído em uma imagem realista, seguindo um processo iterativo de adição e remoção de ruído na imagem.

2.2 JUSTIÇA EM SISTEMAS AUTOMATIZADOS

Os sistemas autônomos ganharam muito espaço em processos que eram antes exclusivamente humanos, como a autenticação de transações bancárias por reconhecimento facial. No entanto, a definição de justiça nos sistemas automatizados nem sempre é clara, pois ela envolve considerações éticas que visam à transparência dos processos, o que permite a contestação das decisões por parte de seus usuários. Ou seja, a justiça depende, principalmente, da forma como a tecnologia é desenvolvida e supervisionada.

Na era da inteligência artificial, essa questão toma proporções maiores. Modelos treinados com grandes volumes de dados podem herdar vieses presentes nas informações, o que reforça comportamentos discriminatórios. Sendo assim, o uso desses sistemas por usuários que já estão inseridos em um contexto de vulnerabilidade social pode colocá-los em situações potencialmente piores. Para desenvolver sistemas que não só funcionem bem tecnicamente falando, mas que também sejam socialmente justos, é necessário compreender como as pessoas entendem o que é justiça (STARKE et al., 2022).

O estudo de equidade em inteligência artificial busca garantir que os modelos preditivos tomem decisões justas. Porém, estudos apontam que, mesmo em algoritmos que implementam técnicas de mitigação de vieses, o sistema ainda pode perpetuar discriminações sociais (CHENG et al., 2023). Dessa forma, lidar com justiça requer um entendimento de como funcionam as

estruturas sociais que contribuem para a desigualdade dentro de cada contexto, visto que os vieses são culturais. Uma das principais abordagens de estudo de justiça em inteligência artificial é o *Group Fairness*, ou justiça de grupo. Ela tem como propósito que grupos demográficos diferentes sejam tratados de maneira semelhante pelos modelos, independentemente de características sensíveis como raça, gênero ou idade.

O estudo de *Group Fairness* já conta com diversas métricas de avaliação de justiça estabelecidas na literatura, como a *Demographic Parity*, que representa a proporção igual de predições positivas em cada grupo, e a *Equalized Odds*, que representa a taxa de erro do modelo como sendo igual para todos os grupos. Porém, essas e outras métricas estabelecidas são voltadas apenas para modelos de classificação. Quando avaliamos as métricas de *Group Fairness* para modelos de regressão ou modelos generativos, encontramos poucas referências na literatura. Estudos anteriores (PERERA et al., 2022) mostram que o desenvolvimento de uma métrica de justiça para um modelo de regressão na área de saúde requer o uso de abordagens inovadoras que levam em consideração as especificidades dos dados e dos objetivos do modelo. Já na área de modelos generativos, o conceito de justiça é ainda mais ambíguo, porém igualmente crítico. Enquanto os vieses em tarefas de classificação são mensuráveis de forma mais fácil, os modelos generativos também podem contribuir com vieses representativos danosos, de forma que é essencial o desenvolvimento de métricas que avaliem quantitativamente a equidade deles (GOLDFARB-TARRANT et al., 2020).

2.3 MEDIÇÃO DE QUALIDADE DE IMAGENS

A área de estudo *Image Quality Assessment* (IQA) trata da avaliação da qualidade de imagens, sendo a qualidade interpretada de forma perceptiva ou técnica. A IQA com foco perceptivo mede a compreensão humana do que é a qualidade de uma imagem, o que constitui uma avaliação subjetiva relacionada à naturalidade das imagens. Já a IQA técnica possui um enfoque matemático e estatístico, de forma que ela mede distorções como ruído e compressão e é mais utilizada em processos de otimização de algoritmos de processamento de imagens.

Além da divisão entre perceptiva e técnica, a área de IQA também pode ser dividida entre *Full-Reference Image Quality Assessment* (FR-IQA) e *No-Reference Image Quality Assessment* (NR-IQA), categorias em que a imagem é avaliada em comparação a uma imagem de referência ou não, respectivamente.

As métricas mais utilizadas de FR-IQA são a *Peak Signal-to-Noise Ratio* (PSNR) (HUYNH-

THU; GHANBARI, 2008) e a *Structural Similarity Index Measure* (SSIM) (WANG et al., 2004). A primeira mede a relação entre o sinal e o ruído na imagem, indicando o quanto ela foi degradada em relação à referência (abordagem técnica), e a segunda mede a similaridade estrutural entre as duas imagens, considerando luminância e contraste (abordagem perceptiva). Embora sejam métricas muito utilizadas no contexto de restauração de imagens, ambas possuem limitações claras quando comparadas com a percepção humana das imagens (ZHANG et al., 2018), de forma que técnicas de aprendizado profundo já as ultrapassaram nesse aspecto.

Já com relação à categoria de NR-IQA, a métrica mais utilizada é a *Blind/Referenceless Image Spatial Quality Evaluator* (BRISQUE) (MITTAL; MOORTHY; BOVIK, 2011). Ela assume que uma imagem natural, capturada do mundo real, segue uma distribuição gaussiana específica e que distúrbios na imagem, como compressão e ruído, alteram essa distribuição. Coeficientes estatísticos são extraídos dos dados de luminância e contraste da imagem e um modelo de regressão é treinado para prever um *score* de qualidade da imagem. Porém, estudos apontam que, juntamente com outras métricas de NR-IQA, a BRISQUE não tem correlação suficiente com a percepção humana para ser utilizada em cenários reais (PINSON, 2022).

A partir da IQA, surgiu a área de *Face Image Quality Assessment* (FIQA), que concentra seus estudos na avaliação da qualidade de faces em imagens e vídeos. Geralmente, as técnicas de FIQA são usadas no contexto de reconhecimento facial, de forma que imagens de baixa qualidade podem impactar negativamente o desempenho desses sistemas. Por conta disso, os métodos de FIQA utilizam funções objetivas derivadas de métodos de reconhecimento facial, como a similaridade por cosseno das representações vetoriais da imagem (BABNIK; PEER; ŠTRUC, 2022). Porém, por estarem atreladas a esse contexto, as técnicas de FIQA são dificilmente generalizadas para outras tarefas de medição de qualidade facial além do reconhecimento.

3 METODOLOGIA

Este trabalho foi dividido em quatro partes: preparação dos dados e dos algoritmos para inferência, automatização e execução das anonimizações, cálculo das métricas e análise dos resultados obtidos. Nesta seção, serão descritas as etapas realizadas para a condução do trabalho.

3.1 SELEÇÃO DAS TÉCNICAS

Para garantir uma análise comparativa entre diferentes técnicas de anonimização facial, selecionamos quatro algoritmos de diferentes abordagens do estado da arte. Para cada algoritmo, foi criado um ambiente de experimentação utilizando o software Anaconda, com o intuito de testar suas inferências localmente e avaliar a viabilidade de seu uso na pesquisa. Os critérios de seleção foram a atualidade dos algoritmos, a viabilidade de modificação do código e a capacidade de rodar inferências em GPU, para otimizar o tempo de inferência com o grande volume de dados utilizados.

Além dos quatro selecionados, outros três foram testados e eliminados. O primeiro, *Surface-Guided GAN* (SG-GAN) (HUKKELÅS et al., 2023), utiliza uma rede generativa com uma abordagem guiada por superfície, que faz correspondências entre uma imagem e uma superfície 3D canônica através do método *Variational Surface-Adaptive Modulation* (V-SAM). Esta técnica foi executada e avaliada localmente, e seus resultados foram considerados obsoletos com relação a uma outra técnica publicada pelo mesmo autor, a DeepPrivacy2 (HUKKELÅS; LINDSETH, 2023). O segundo algoritmo foi o *Latent Diffusion Face Anonymization* (LDFA) (KLEMP et al., 2023), que possui arquitetura baseada em modelos de difusão e tem o intuito de anonimizar faces em contextos de veículos autônomos. Ele foi testado e se mostrou inviável para utilizar nos experimentos, pois toda a sua aplicação está encapsulada em um contêiner Docker, fazendo com que a automatização dos experimentos fosse consideravelmente mais complexa, ou até inviável, dado o tempo disponível no escopo da pesquisa. Por fim, o terceiro algoritmo testado e desconsiderado foi o *Realistic Anonymization using Diffusion* (RAD) (MALM et al., 2023), modelo de *Stable Diffusion* cujo código não estava em estado funcional. Alguns problemas foram tratados e resolvidos, mas não houve tempo suficiente para nos dedicarmos à resolução de todos.

3.2 SELEÇÃO DOS DADOS

Com o objetivo de analisar os algoritmos de anonimização pelo viés da justiça e da diversidade, foram selecionados conjuntos de dados de imagens contendo rostos de pessoas categorizadas em grupos demográficos de raça e gênero. Garantindo essa diversidade nos dados, podemos fazer análises de *Group Fairness* para avaliar como os métodos se comportam com dados pertencentes a cada uma das categorias.

Além disso, dois dos três conjuntos utilizados nesta pesquisa foram criados por pesquisadores estrangeiros, que empregam classificações de raça baseadas em perspectivas culturais e sociais diferentes das do Brasil. Por exemplo, categorias como *Latino* e *Indian* são consideradas raças nessas bases de dados, enquanto no Brasil, são vistas como identidades culturais ou nacionais. Por conta dessa subjetividade presente na anotação de dados raciais (BARRETT; CHEN; ZHANG, 2023) e com o intuito de manter a fidelidade às categorias originais, optou-se por utilizar as nomenclaturas originais em inglês, tanto para raça quanto para gênero, facilitando comparações diretas com outros estudos que empreguem os mesmos conjuntos de dados.

3.3 MÉTRICAS

Dado um conjunto de imagens I, executamos cada um dos algoritmos de anonimização em I para gerar um novo conjunto anonimizado $I' = f(I)$. Avaliamos esse novo conjunto por duas métricas diferentes, que são descritas a seguir.

3.3.1 Deformação

A deformação causada por um algoritmo de anonimização em uma face humana refere-se às alterações estruturais introduzidas na imagem para ocultar a identidade do indivíduo. Essas alterações podem incluir deslocamento de pixels, suavização de traços, distorção geométrica ou a geração de uma nova textura sintética. O objetivo é preservar a privacidade, mas a intensidade da deformação pode impactar a percepção do realismo da síntese, pois uma face muito distorcida pode não se assemelhar a um rosto humano real. Sendo assim, queremos mensurar o nível de deformação que os algoritmos aplicam em faces humanas.

Com o intuito de calcular a deformação, utilizamos uma ferramenta de detecção de marcadores faciais. Neste trabalho, aplicamos o MediaPipe (LUGARESI et al., 2019), que embora

não seja estado da arte, é uma ferramenta amplamente usada por desenvolvedores que trabalham com visão computacional, dada a sua eficiência e facilidade de uso. Os marcadores são aplicados em cada imagem $i \in I$ e em sua correspondente anonimizada $i' \in I'$, como visto na Figura 3. Sendo assim, para cada conjunto de dados, é calculada uma métrica de *Endpoint Error* (EPE), que quantifica a deformação das faces. EPE é definida como a distância euclidiana média entre todos os 478 pontos das malhas faciais de cada par de imagens i e i', para o conjunto inteiro. O cálculo é mostrado a seguir:

$$\text{EPE} = \frac{1}{N} \sum_{i=1}^{N} d_i$$

Onde N é a quantidade de imagens no conjunto e d_i é a distância euclidiana média calculada para cada imagem i:

$$d_i = \frac{1}{K} \sum_{j=1}^{K} \sqrt{(x_{2,j} - x_{1,j})^2 + (y_{2,j} - y_{1,j})^2}$$

Para d_i, $x_{1,j}$ e $x_{2,j}$ são as coordenadas horizontais dos *pixels* para cada um dos K pontos da imagem, enquanto $y_{1,j}$ e $y_{2,j}$ são as coordenadas verticais dos *pixels*. O valor de EPE resultante representa o nível de deformação estrutural aplicado no conjunto anonimizado pelo algoritmo.

Figura 3 – Marcadores faciais gerados pelo MediaPipe.

3.3.2 Qualidade Facial Perceptível

A segunda métrica utilizada é baseada em uma abordagem de avaliação de faces restauradas, que busca refletir melhor a percepção humana ao considerar características específicas das regiões primárias do rosto, como olhos e nariz. A *Interpretable Face Quality Assessment* (IFQA) (JO et al., 2023) utiliza uma arquitetura baseada em GAN, na qual o gerador restaura imagens de baixa qualidade (LQ) para alta qualidade (HQ). As imagens LQ são criadas a partir de imagens HQ corrompidas por desfoque, ruído gaussiano ou compressão *Joint Photographic Experts Group* (JPEG). Já o discriminador atribui pontuações por *pixel* para a imagem restaurada, agregando as pontuações em um *score* único. Esse *score* varia de 0 a 1 e, quanto mais próximo de 1, melhor é a qualidade perceptível da face. O treinamento supervisionado do discriminador considera apenas as regiões primárias do rosto como rótulos verdadeiros, o que ajuda a métrica a focar mais na região da face do que na imagem inteira. Essa abordagem mostrou maior correlação com a percepção humana do que outras métricas usadas na área, como a PSNR e a SSIM.

A IFQA é considerada uma métrica adaptável para vários tipos de tarefas relacionadas à análise de faces, sem depender de métodos de reconhecimento facial, como outras abordagens de FIQA. Neste trabalho, nós utilizamos o discriminador desacoplado do gerador para atribuir *scores* às faces anonimizadas pelos algoritmos em questão. O fluxo das avaliações pode ser visto na Figura 4.

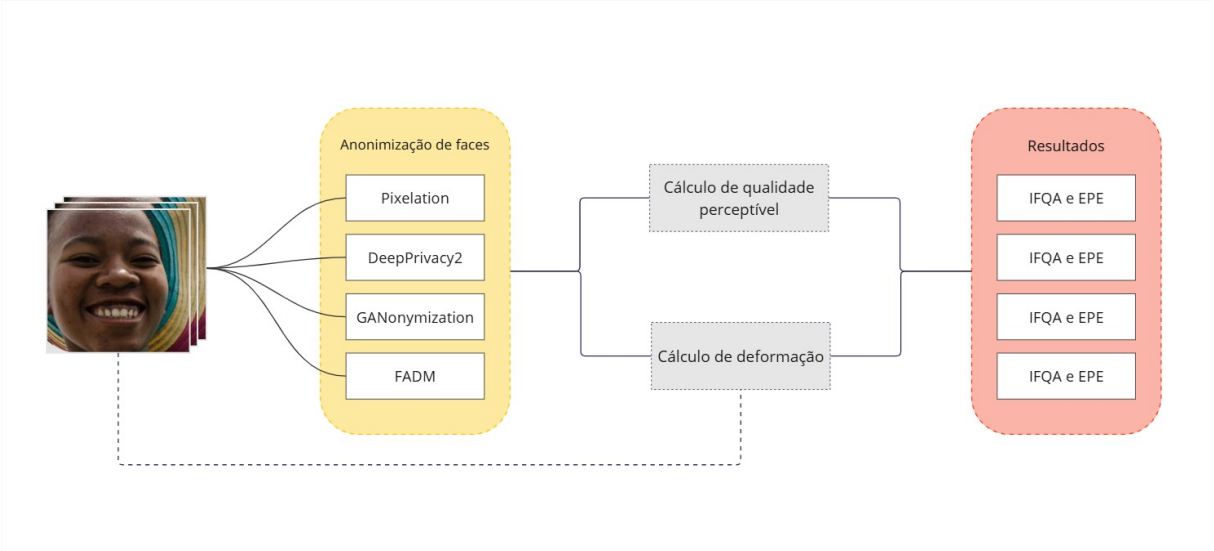


Figura 4 – Visão geral do nosso processo de avaliação. Na primeira etapa, o conjunto de dados é anonimizado por quatro algoritmos diferentes, resultando em quatro versões da base original. As cópias têm suas regiões primárias avaliadas para cálculo de qualidade perceptível, enquanto as malhas faciais de todos os conjuntos são extraídas para medição da deformação das faces, gerando valores de IFQA e EPE, respectivamente.

3.4 ANÁLISE ESTATÍSTICA

Durante a avaliação dos resultados de I', fizemos análises estatísticas de correlação de Pearson entre diferentes variáveis, com o intuito de entender como e quanto a mudança de uma variável impacta na outra. O cálculo de correlação pode ser visto a seguir:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Onde X_i e Y_i são as instâncias das variáveis X e Y e $\bar{X}$ e $\bar{Y}$ são as médias das variáveis. O valor de r está entre -1 e 1, de forma que $r = 1$ representa uma correlação positiva perfeita, $r =$ -1 representa uma correlação negativa perfeita e $r = 0$ representa nenhuma correlação.

Implementamos, também, a análise de valor-p para checar a significância estatística da correlação de Pearson. Ela faz isso testando a hipótese nula (H_0) de que não há correlação real entre as variáveis. Primeiramente, é calculado o teste t, que transforma o valor de r em uma distribuição t de Student:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Onde n é o número de observações. Depois, é calculado o valor-p:

$$p = 2 \times P(T > |t|)$$

Representando a probabilidade de observar um valor de t tão extremo quanto o encontrado, assumindo que a hipótese nula seja verdadeira. O valor-p é comparado com um nível de significância α pré-estabelecido, no valor de 0,05, de forma que, se $p < \alpha$, a correlação r é estatisticamente significativa, ou seja, não foi provocada pela força do acaso. Já se $p \geq \alpha$, significa que não há confiança suficiente para afirmar que as variáveis são relacionadas de forma significativa. Ou seja, não podemos descartar a hipótese nula. A implementação da análise foi feita usando a biblioteca SciPy da linguagem de programação Python.

3.5 TESTE COM FILTRO GAUSSIANO GRADUAL

Além da análise estatística aplicada em I', também avaliamos os dados sintéticos de forma qualitativa, com o objetivo de compreender os padrões visuais das imagens. Durante a avaliação, foi percebido um padrão nos dados, detalhado na seção de Resultados e Discussão, que levantou a necessidade de um experimento com a aplicação de filtro gaussiano a uma amostra de I'.

O filtro gaussiano é um operador de convolução, de forma que ele aplica uma máscara sobre os *pixels* de uma imagem para modificar seus valores com base em uma função matemática. Essa função é a Gaussiana, que aplica uma suavização na imagem através da redução de detalhes finos. Ela é definida por:

$$G(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right)$$

Onde x e y são as coordenadas do *pixel* em relação ao centro da máscara e σ é o desvio padrão da máscara. A implementação foi feita utilizando a biblioteca OpenCV na linguagem Python, com máscaras de tamanhos variados de 3×3, 9×9, 15×15 e 25×25 e desvio padrão igual a zero, o que implica que o OpenCV calcula automaticamente o valor de σ com base no tamanho da máscara.

4 TÉCNICAS E CONJUNTOS DE DADOS AVALIADOS

Quatro algoritmos e três bases de dados foram selecionados para esta pesquisa, de acordo com os critérios estabelecidos na Seção 3. Nesta seção, faremos uma explicação detalhada dos algoritmos avaliados e dos conjuntos de dados utilizados.

4.1 TÉCNICAS

4.1.1 Pixelation

O primeiro algoritmo selecionado foi o Pixelation, que consiste em uma técnica baseada na visão computacional clássica, na qual a resolução da imagem é reduzida e os *pixels* ficam visíveis. A técnica é amplamente usada por outros trabalhos de anonimização facial para estabelecer um ponto de comparação com abordagens mais avançadas e, apesar de não ser baseada em redes neurais e, consequentemente, não apresentar vieses provenientes de dados de treinamento e arquitetura do modelo, a pixelização não elimina problemas sistêmicos, como aqueles associados ao design da técnica ou ao contexto em que é aplicada. O tamanho do núcleo selecionado para a pixelização foi de 8×8, pois tamanhos maiores podem prejudicar a estrutura facial (FAN, 2018).

4.1.2 DeepPrivacy2

Outra técnica selecionada foi a DeepPrivacy2 (HUKKELÅS; LINDSETH, 2023). Sua arquitetura é composta por uma GAN *style-based* (KARRAS et al., 2020), que aplica mecanismos de regularização capazes de manipular as características visuais para gerar imagens com propriedades diferentes. Ela é uma melhoria de um trabalho anterior chamado DeepPrivacy (HUKKELÅS; MESTER; LINDSETH, 2019), proporcionando maior diversidade e qualidade nas imagens geradas, além de anonimizações de corpo inteiro.

4.1.3 GANonymization

A terceira técnica selecionada foi a GANonymization (HELLMANN et al., 2024), baseada em redes generativas *pix2pix*. Ao contrário das GANs *style-based*, que geram imagens realistas do

zero, esse tipo de arquitetura gera imagens a partir de outras imagens, utilizando re-síntese de rostos a partir de representações baseadas em marcadores faciais. Isso preserva a estrutura geométrica dos rostos, mas gera uma nova aparência que não corresponde ao rosto original, garantindo anonimato. A técnica propõe a remoção de diversos tipos de traços faciais, como acessórios e cores de cabelo, para garantia da ocultação da identidade do indivíduo.

4.1.4 Full-Body Anonymization using Diffusion Models

Por fim, o último algoritmo selecionado foi o *Full-Body Anonymization using Diffusion Models* (FADM) (ZWICK et al., 2024). Este método usa um modelo de *Stable Diffusion* para preencher com ruído as máscaras de segmentação detectadas nas imagens originais, que depois serão revertidas para gerar novas imagens. O nível de modificação é controlado por um parâmetro proposto na arquitetura da FADM, que balanceia o nível de realismo da face gerada e quais informações pessoais devem ser ocultadas. O uso de *prompts* positivos define as características desejadas para a imagem criada, enquanto os *prompts* negativos definem o que deve ser evitado. Os *prompts* usados em nossos experimentos utilizam a configuração padrão da técnica.

4.2 CONJUNTOS DE DADOS

4.2.1 UTKFace

O conjunto de dados UTKFace (ZHANG; SONG; QI, 2017) contém mais de 20.000 imagens categorizadas em raça, gênero e idade. Suas imagens foram coletadas da internet, sem especificar fontes, e capturadas em cenários não controlados, ou seja, sem condições padronizadas de iluminação, enquadramento ou ambiente. As anotações dos dados foram feitas pelo algoritmo *Deep EXpectation* (DEX) (ROTHE; TIMOFTE; GOOL, 2015) e verificadas por um anotador humano. Suas classes correspondentes à raça são cinco, e se dividem em *White*, *Black*, *Asian*, *Indian* e *Others*, sendo a última um agrupamento de diferentes categorias, como *Hispanic*, *Latino* e *Middle Eastern*. Já os gêneros são divididos em *Male* e *Female*. Amostras dos dados de UTKFace podem ser vistas na Figura 5.

Figura 5 – Exemplos de imagens do conjunto UTKFace.

4.2.2 FairFace

O segundo conjunto utilizado nos experimentos desta pesquisa foi o FairFace (KARKKAINEN; JOO, 2021). Ele foi criado com o objetivo de minimizar os vieses raciais existentes em bases de dados de faces públicas, que sub-representam faces não caucasianas. Sendo assim, o FairFace contém 108.501 imagens divididas em sete categorias raciais: *White*, *Black*, *Indian*, *East Asian*, *Southeast Asian*, *Middle Eastern* e *Latino*. Suas imagens foram coletadas do conjunto YFCC-100M Flickr (THOMEE et al., 2016) e de outras fontes como o Twitter e jornais *online*, e algumas delas podem ser vistas na Figura 6.

Devido ao UTKFace conter apenas 5 categorias raciais, nós processamos os dados do FairFace para igualar suas categorias às da base anterior, com o objetivo de facilitar a comparação entre os resultados dos experimentos. Sendo assim, os grupos *East Asian* e *Southeast Asian* foram agrupados em um único grupo *Asian*, enquanto *Middle Eastern* e *Latino* foram agrupados em uma única categoria *Others*. Além das categorias de raça, o FairFace também categoriza seus dados em gênero e idade, sendo os gêneros divididos em *Male* e *Female*.

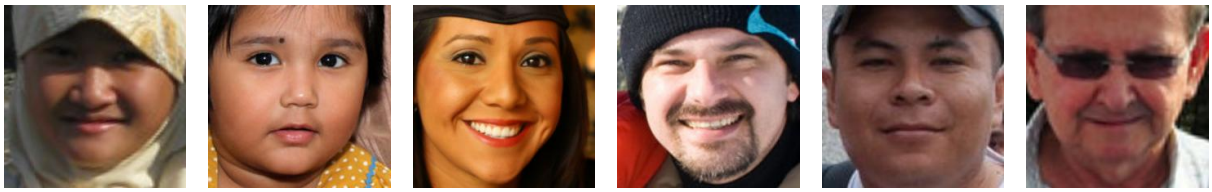


Figura 6 – Exemplos de imagens do conjunto FairFace.

4.2.3 FEI

A base de dados FEI (JUNIOR; THOMAZ, 2006) foi desenvolvida por pesquisadores brasileiros do Centro Universitário Fundação Educacional Inaciana Padre Sabóia de Medeiros (FEI) e contém imagens capturadas entre junho de 2005 e março de 2006. Participaram da coleta de dados 200 indivíduos, com 14 imagens capturadas cada um, totalizando 2.800 instâncias

de dados. Todas as fotos foram tiradas em cenário controlado, com fundo homogêneo, mas com variações na iluminação e nos ângulos das faces em relação à câmera. Os participantes foram estudantes e funcionários do Centro Universitário FEI, com idades entre 19 e 40 anos, e divididos igualmente entre gêneros, com 100 homens e 100 mulheres. Esse conjunto não possui classificação racial, de forma que o utilizamos somente para os experimentos de análise de gênero. Sendo assim, totalizamos duas bases de dados para análise racial e três para análise de gênero. Algumas de suas imagens podem ser vistas na Figura 7.

Figura 7 – Exemplos de imagens do conjunto FEI.

5 RESULTADOS E DISCUSSÃO

Nesta seção, apresentamos os resultados referentes à análise de *Group Fairness* dos algoritmos de anonimização de faces, com enfoque nos cálculos de deformação e qualidade perceptível obtidos pelas métricas EPE e IFQA. A partir dos resultados obtidos, novas análises são feitas sobre os dados, com o objetivo de entendê-los com maior profundidade, assim como avaliações qualitativas das faces anonimizadas para detectar padrões visuais.

5.1 RESULTADOS GERAIS

Na Tabela 1, observamos os resultados das distorções médias calculadas para cada experimento, em *pixels*. Os valores mostram que a técnica DeepPrivacy2 foi a que teve a maior deformação média para todos os conjuntos de dados, totalizando 15,83 *pixels*, seguida pela GANonymization, com 12,52 *pixels*. Já a FADM foi a que teve a menor distorção, abaixo até da Pixelation, com 2,18 e 3,88 *pixels*, respectivamente. Os algoritmos com menor deformação média podem ser considerados melhores em manter a estrutura original da face enquanto realizam a anonimização. A partir daí, foi formulada a hipótese de que quanto menor a deformação de uma face, maior a qualidade perceptível dela. Essa hipótese será testada e avaliada posteriormente nesta pesquisa e ela reflete diretamente nos valores das métricas calculadas, de forma que EPE e IFQA seriam inversamente proporcionais. Na Figura 8, vemos as imagens com maior deformação do FairFace com o DeepPrivacy2.

Tabela 1 – Valores de deformação média (EPE), em *pixels*. ↓

	FADM	**DeepPrivacy2**	**GANonymization**	**Pixelation**
FairFace	2,51	28,18	13,30	4,60
UTKFace	1,25	12,01	18,69	4,08
FEI	2,78	7,32	5,58	2,96
Média	2,18	15,83	12,52	3,88

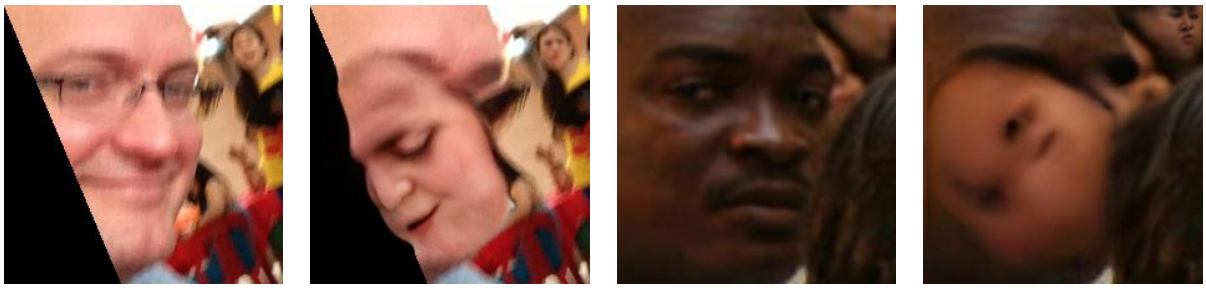
A Tabela 2 mostra os resultados dos *scores* referentes à qualidade perceptível das faces, calculados pela IFQA. A técnica que gerou imagens com melhores *scores* foi a Pixelation, com uma média de 0,27, seguida pela FADM, pela DeepPrivacy2 e, por fim, pela GANonymization, com *scores* médios de 0,19, 0,18 e 0,04, respectivamente. Esse resultado é surpreendente, considerando que a pixelização é uma técnica mais simples, geralmente associada à perda de qualidade visual. Em contrapartida, técnicas baseadas em GAN e difusão são projetadas para

preservar a naturalidade da imagem ao fazer a anonimização, o que geralmente as coloca em vantagem em métricas de qualidade visual. Sendo assim, pode-se concluir que a IFQA captura aspectos que a pixelização consegue preservar melhor, em detrimento das outras técnicas.

Tabela 2 – *Scores* médios da IFQA. ↑

	FADM	**DeepPrivacy2**	**GANonymization**	**Pixelation**
FairFace	0,16	0,15	0,05	0,26
UTKFace	0,10	0,10	0,04	0,27
FEI	0,30	0,30	0,04	0,28
Média	0,19	0,18	0,04	0,27

Outro fator que chama a atenção nos resultados é o fato de os *scores* médios serem muito baixos, menores ou iguais a 0,30. Tais dados indicam que todos os algoritmos podem estar degradando consideravelmente a qualidade das imagens, especialmente nas regiões primárias da face, que têm o maior peso no cálculo dos *scores* da IFQA. Essa possibilidade indica uma desvantagem em utilizar as técnicas mais sofisticadas para anonimização, de forma que elas podem estar priorizando a ocultação do indivíduo em detrimento da qualidade visual, da mesma forma que a pixelização faz. Outra possibilidade é a de que a própria métrica tenha problemas de calibração, de forma que ela esteja erroneamente penalizando os *scores* de forma severa. Na Figura 9, vemos anonimizações da técnica FADM com *scores* mais elevados e mais baixos.



(a) Original (b) Anonimizada (c) Original (d) Anonimizada

Figura 8 – Imagens do FairFace com deformações de 171 e 143 *pixels*, anonimizadas pelo DeepPrivacy2.

Com relação à hipótese de que os valores de EPE e IFQA sejam inversamente proporcionais, podemos observar que ambas as técnicas Pixelation e FADM tiveram os menores valores de deformação média e os maiores *scores* de qualidade perceptível, o que vai ao encontro da hipótese. Já a DeepPrivacy2 teve *scores* médios muito próximos a FADM, mas também teve deformações elevadas, o que contraria a hipótese, enquanto a GANonymization também teve deformações mais elevadas e *scores* mais baixos, o que condiz com a hipótese. Sendo assim,

fizemos uma análise de correlação de Pearson entre as duas métricas, com o objetivo de determinar se elas de fato estão relacionadas.

(a) Original (b) Anonimizada (c) Original (d) Anonimizada

Figura 9 – Imagens do conjunto FEI com *scores* de 0,60 e 0,02 respectivamente, anonimizadas pela FADM.

Juntando todas as imagens anonimizadas nos experimentos, sendo cada experimento a anonimização de um conjunto de dados por um algoritmo, totalizamos 355.557 imagens anonimizadas. Removendo todas as amostras em que a face não foi detectada pelo MediaPipe, sobraram 256.702 imagens com deformações calculadas. Com essa seleção, calculamos o coeficiente de correlação de Pearson e o valor-p para os valores de deformação e qualidade perceptível. O resultado de $r =$ -0,08 indica que existe uma associação fraca entre o aumento da deformação e a diminuição na qualidade perceptível, ao mesmo tempo em que o valor de $p =$ 0,0 ($p < \alpha$) indica que podemos rejeitar a hipótese de que não há correlação entre as variáveis, ou seja, de que a relação é fruto do acaso. Sendo assim, concluímos que a deformação impacta a qualidade perceptível calculada pela IFQA, mas não é seu principal determinante, e outros fatores devem ser considerados.

5.2 ANÁLISE DE JUSTIÇA POR GRUPO DEMOGRÁFICO

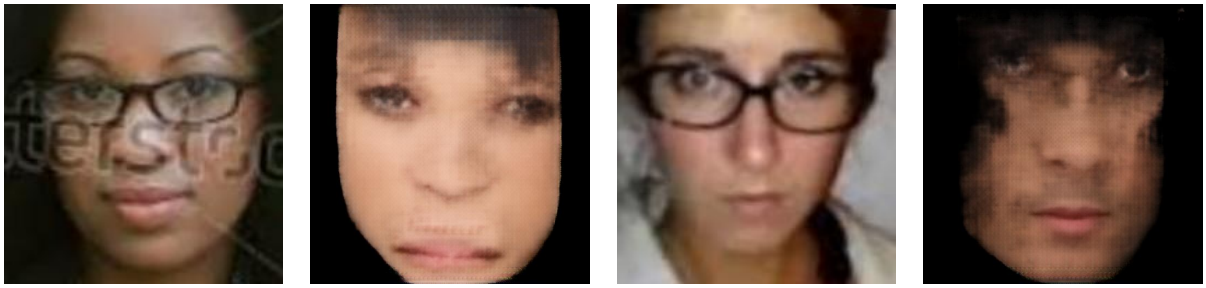
Após a análise geral dos resultados por algoritmo, examinamos os dados entre os grupos demográficos para avaliação de *Group Fairness*. Nas Tabelas 3 e 4, observamos os resultados por gênero referentes à deformação facial média e à qualidade facial perceptível, respectivamente. Já nas Tabelas 5 e 6, observamos os mesmos dados, mas distribuídos entre raças.

Na Tabela 3, estimamos as diferenças de deformação para os gêneros *Male* e *Female*, de forma que 11 dos 12 experimentos concentram suas diferenças entre 0 e 2 *pixels*. Essas variações são consideradas ínfimas para a determinação de vieses de gênero, de forma que elas podem ser resultantes de outros tipos de variações nos dados. Um único valor se mostrou *outlier*, com diferença de 5,26 *pixels* entre os gêneros, sendo ele resultante da anonimização do UTKFace pelo GANonymization. Esse resultado indica uma deformação maior para os dados

Tabela 3 – Deformação facial média (EPE) por gênero, em *pixels*. ↓

	FADM		**DeepPrivacy2**		**GANonymization**		**Pixelation**	
	Male	Female	Male	Female	Male	Female	Male	Female
FairFace	2,62	2,39	27,30	29,05	14,04	12,61	4,73	4,46
UTKFace	1,35	1,14	11,62	12,41	15,87	21,13	4,40	3,77
FEI	2,97	2,59	7,28	7,35	5,86	5,30	2,99	2,93

do gênero feminino, como podemos ver na Figura 10.



(a) Original (b) Anonimizada (c) Original (d) Anonimizada

Figura 10 – Imagens do UTKFace para o gênero feminino anonimizadas pelo GANonymization. O par da esquerda tem uma deformação média de 207 *pixels* e o da direita de 204 *pixels*.

Já na Tabela 4, os resultados são ainda mais homogêneos. Para todos os algoritmos baseados em inteligência artificial, as diferenças de qualidade perceptível das faces, entre gêneros, foram nulas ou com valor de 0,01, valores também desprezíveis para a determinação da presença de vieses. O Pixelation registrou diferenças maiores em dois experimentos, com valores de 0,02 e 0,04, sendo ambos de qualidade maior para o gênero masculino. Com relação ao resultado do GANonymization com o UTKFace, correspondente à maior disparidade de deformações entre gêneros discutida anteriormente, os *scores* de qualidade calculados foram muito próximos, com vantagem para a categoria *Female*, o que confirma que a diferença de distorções não impactou negativamente a qualidade calculada pela IFQA. Na Figura 11, vemos imagens do experimento com maior *score* médio da Tabela 4, correspondente ao DeepPrivacy2 com o gênero masculino do conjunto FEI.

Analogamente, vemos na Tabela 5 resultados majoritariamente homogêneos entre raças, com variações pequenas de distorções faciais médias. Em todos os experimentos, a classe *Black* teve distorções maiores, porém com variações pequenas quando comparada às outras classes. Os experimentos com maior disparidade nos resultados foram provenientes do GANonymization em ambos os conjuntos de dados, com uma diferença de 4,19 e 3,70 *pixels* entre a classe *Black* e a segunda colocada, para o FairFace e o UTKFace, respectivamente. Esses resultados são

Tabela 4 – *Scores* médios de IFQA, por gênero. ↑

	FADM		**DeepPrivacy2**		**GANonymization**		**Pixelation**	
	Male	Female	Male	Female	Male	Female	Male	Female
FairFace	0,16	0,16	0,15	0,16	0,04	0,05	0,27	0,25
UTKFace	0,10	0,10	0,10	0,10	0,03	0,04	0,29	0,25
FEI	0,30	0,30	0,31	0,30	0,04	0,04	0,28	0,28

(a) Original (b) Anonimizada (c) Original (d) Anonimizada

Figura 11 – Imagens com os maiores *scores* para o gênero masculino do conjunto FEI anonimizadas pelo DeepPrivacy2. O par da esquerda tem um *score* de 0,59 e o da direita de 0,56.

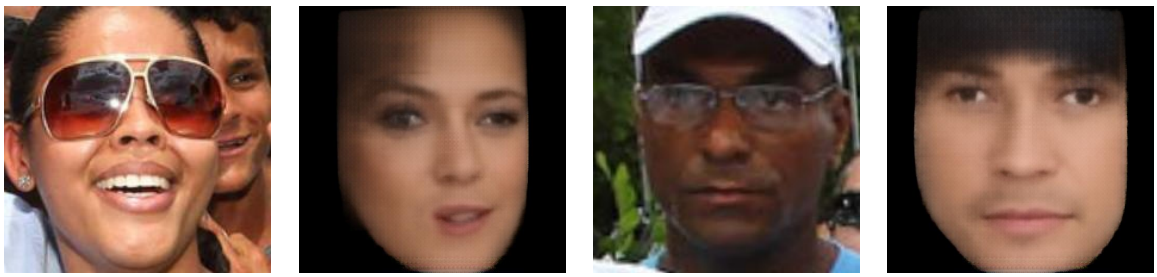
vistos na Figura 12, na qual podemos ver que a técnica altera os tons de pele dos indivíduos, clareando-os.

Já nos resultados relacionados à qualidade perceptível por raça, a Tabela 6 mostra que os *scores* da classe *Black* são variados. Em alguns casos, eles representam o maior *score* entre as classes, como nos experimentos do FADM com o FairFace e do DeepPrivacy2 com o FairFace, como também representam o menor valor entre as classes, como no experimento do GANonymization com o FairFace. Porém, as diferenças de *score* dessa classe para as outras são todas próximas, com valores médios de 0,01. Dessa forma, não temos indícios suficientes que apontem para a presença de vieses de qualidade perceptível nesses modelos, tanto para raça quanto para gênero. Na Figura 13, vemos as imagens com os melhores scores de IFQA para o conjunto FairFace anonimizadas pela técnica FADM. Podemos observar que a técnica consegue preservar expressões faciais parecidas com as originais e seus tons de pele, alterando apenas a idade perceptível em um dos casos.

Por outro lado, com relação à distorção das faces, alguns valores atípicos foram encontrados. Em ambas as análises de gênero e raça, percebemos que o GANonymization provocou pequenos desvios em classes específicas, para determinados conjuntos de dados. Com relação ao gênero, ao anonimizar o FairFace, o algoritmo distorceu mais as faces do gênero feminino, enquanto que, com relação às raças, ele distorceu mais as faces da categoria *Black*, em ambos os conjuntos. Essas classes estariam desfavorecidas, considerando que uma deformação mais

Tabela 5 – Deformação facial média (EPE) por raça, em *pixels*. ↓

		FairFace	UTKFace
FADM	Asian	2,53	1,20
	White	2,54	1,25
	Black	2,78	1,44
	Indian	2,36	1,09
	Other	2,41	1,19
DeepPrivacy2	Asian	28,11	12,14
	White	28,43	12,02
	Black	29,43	12,87
	Indian	27,37	11,21
	Other	27,92	11,32
GANonymization	Asian	13,45	18,84
	White	12,79	19,53
	Black	17,15	23,72
	Indian	12,22	15,97
	Other	12,74	13,43
Pixelation	Asian	4,35	3,74
	White	4,74	4,10
	Black	5,38	4,67
	Indian	4,26	3,93
	Other	4,38	3,77



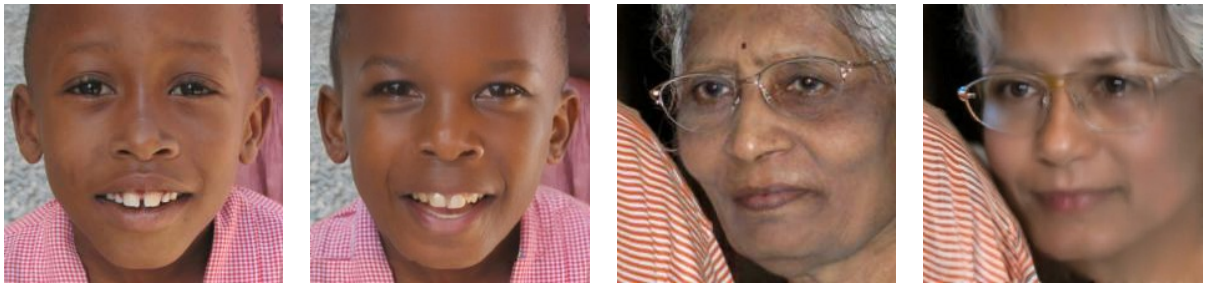
(a) Original (b) Anonimizada (c) Original (d) Anonimizada

Figura 12 – Imagens de pessoas de raça negra do conjunto FairFace anonimizadas pelo GANonymization, correspondentes às maiores deformações da classe. O par da esquerda tem uma deformação de 182 *pixels* e o da direita de 179.

alta indica uma falta de preservação da estrutura facial do indivíduo.

Tabela 6 – *Scores* médios de IFQA, por raça. ↑

		FairFace	UTKFace
FADM	Asian	0,16	0,09
	White	0,16	0,12
	Black	0,17	0,10
	Indian	0,16	0,07
	Other	0,16	0,09
DeepPrivacy2	Asian	0,15	0,08
	White	0,15	0,12
	Black	0,16	0,09
	Indian	0,15	0,07
	Other	0,15	0,09
GANonymization	Asian	0,05	0,03
	White	0,05	0,03
	Black	0,04	0,04
	Indian	0,05	0,03
	Other	0,05	0,04
Pixelation	Asian	0,27	0,28
	White	0,26	0,27
	Black	0,26	0,27
	Indian	0,25	0,28
	Other	0,26	0,26



(a) Original (b) Anonimizada (c) Original (d) Anonimizada

Figura 13 – Imagens com melhores *scores* gerados pela anonimização do FairFace com o FADM.

5.3 ANÁLISE DAS MÉTRICAS

Ao analisar os resultados dos cálculos de deformação, notamos que muitas imagens não tinham suas malhas faciais geradas pelo MediaPipe. Isso acontece quando o *framework* não consegue localizar a face na imagem, resultando na desconsideração dessas amostras para

o cálculo de deformação. A partir daí, analisamos o comportamento de falhas de detecção de faces do MediaPipe pela ótica do *Group Fairness*. Paralelamente, investigamos os resultados qualitativos relativos aos valores extremos da métrica IFQA, os quais evidenciaram um comportamento enviesado da métrica, que será explicado nesta seção.

5.3.1 Geração de Malhas Faciais

Ao quantificar os casos em que o MediaPipe não consegue identificar rostos nas imagens, observamos que 91,12% dessas falhas são provenientes de imagens anonimizadas. Ou seja, a anonimização compromete a performance do algoritmo de identificação de faces do MediaPipe, o que pode ser interpretado como uma deficiência das técnicas em gerar imagens realistas.

Analisando por algoritmos, vemos que o que provocou a maior quantidade de falhas no MediaPipe foi o GANonymization, com 44,71% dos seus dados sem malhas geradas. Em segundo lugar, vem o DeepPrivacy2, com 35,33% de falhas, seguido pelo Pixelation, com 32,06% de falhas. O que teve a melhor taxa de reconhecimento pelo MediaPipe foi o FADM, com apenas 13,65% de falhas, o que indica que suas imagens anonimizadas podem ser mais realistas.

Investigamos também as taxas de falha do MediaPipe por gênero e raça, com o objetivo de detectar possíveis padrões. Na Tabela 7, vemos que o gênero masculino é pior detectado pelo *framework* em 11 dos 12 experimentos. A taxa é calculada pela porcentagem dos dados correspondentes ao total daquela classe que não foram detectados pelo MediaPipe. Analisamos qualitativamente 120 imagens anonimizadas do gênero masculino que geraram falhas de detecção e encontramos alguns padrões qualitativos: rostos com ângulos na lateral, idades extremas (idosos e crianças) e faces distorcidas de forma irreal, como mostrado na Figura 14.

Tabela 7 – Taxa de falha do MediaPipe por gênero.

	FADM		**DeepPrivacy2**		**GANonymization**		**Pixelation**	
	Male	Female	Male	Female	Male	Female	Male	Female
FairFace	18%	12%	41%	34%	6%	4%	39%	32%
UTKFace	1%	0,8%	11%	8%	1%	1%	21%	14%
FEI	3%	0,8%	3%	2%	1%	0,7%	11%	10%

Já com relação à análise por raça, vemos na Tabela 8 que a anonimização do UTKFace pelo Pixelation, apesar de não possuir a maior taxa de falha absoluta, possui a maior discrepância

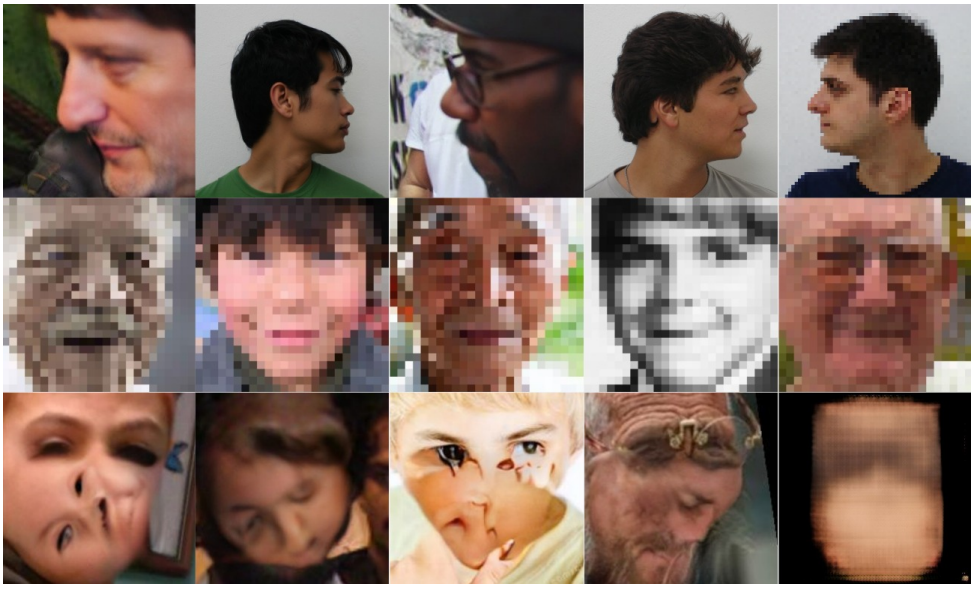


Figura 14 – Exemplos de faces geradas da categoria *Male* que não foram detectadas pelo MediaPipe.

entre classes, contendo a categoria *Black* a maior porcentagem de falhas do experimento, com 30% de erro, e a categoria *Indian* em segundo lugar, com 17% de erro. Outros dois experimentos mostram a categoria *Black* com a maior porcentagem de falhas, enquanto um terceiro mostra a categoria *White* no topo. Um trabalho anterior (MÜLLER et al., 2022) mostra que tons de pele mais escuros podem comprometer o desempenho do MediaPipe em uma aplicação de rastreamento de mãos em contexto cirúrgico. Sendo assim, a presença desse viés também na detecção de faces, como indicado nos resultados anteriores, compromete a imparcialidade da métrica em fazer avaliações equitativas. Na Figura 15, vemos uma amostra dos dados anonimizados pelo Pixelation com falhas de detecção no MediaPipe para a raça *Black*.

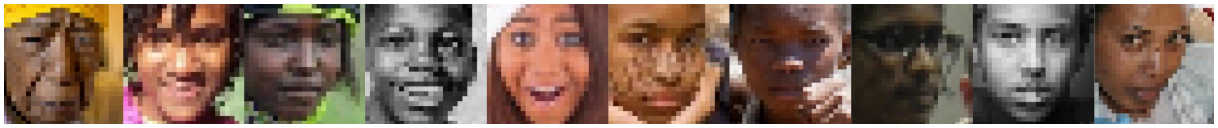


Figura 15 – Exemplos de imagens de pessoas negras anonimizadas pelo Pixelation que o MediaPipe não conseguiu detectar.

Considerando esses desdobramentos, nós reavaliamos os resultados da Tabela 5 relacionados à comparação de deformação facial entre raças. Eles mostram que todos os experimentos tiveram a raça *Black* com maior deformação, o que inicialmente foi interpretado como um padrão das técnicas de anonimização. Porém, tendo em vista que o MediaPipe teve uma taxa de falha maior para imagens da raça *Black*, precisamos considerar a possibilidade de a deformação mais alta ser proveniente da dificuldade do MediaPipe em detectar com precisão as

faces desse grupo.

Tabela 8 – Taxa de falha do MediaPipe por raça.

		FairFace	UTKFace
FADM	Asian	12%	1%
	White	18%	0,9%
	Black	15%	1%
	Indian	14%	1%
	Other	16%	0,5%
DeepPrivacy2	Asian	37%	8%
	White	41%	10%
	Black	41%	12%
	Indian	33%	8%
	Other	37%	7%
GANonymization	Asian	5%	1%
	White	5%	0,09%
	Black	4%	2%
	Indian	4%	2%
	Other	5%	1%
Pixelation	Asian	30%	10%
	White	40%	16%
	Black	42%	30%
	Indian	34%	17%
	Other	35%	9%

5.3.2 Análise de Qualidade Perceptível

Além da análise de taxa de falha do MediaPipe, fizemos uma análise qualitativa para os melhores e piores *scores* de IFQA, com o intuito de identificar padrões visuais que podem influenciar uma melhor avaliação de qualidade perceptível pela métrica. Para isso, selecionamos as 10 imagens com maiores *scores* de cada experimento, assim como as 10 imagens com menores *scores* de cada um, totalizando 240 imagens. Algumas amostras dos melhores resultados podem ser vistas na Figura 16, enquanto algumas das piores podem ser vistas na Figura 17.

Ao se analisar a amostra de 240 imagens, identificou-se um padrão que sugere a presença de um viés na métrica IFQA. Ao comparar os experimentos do DeepPrivacy2 e do FADM

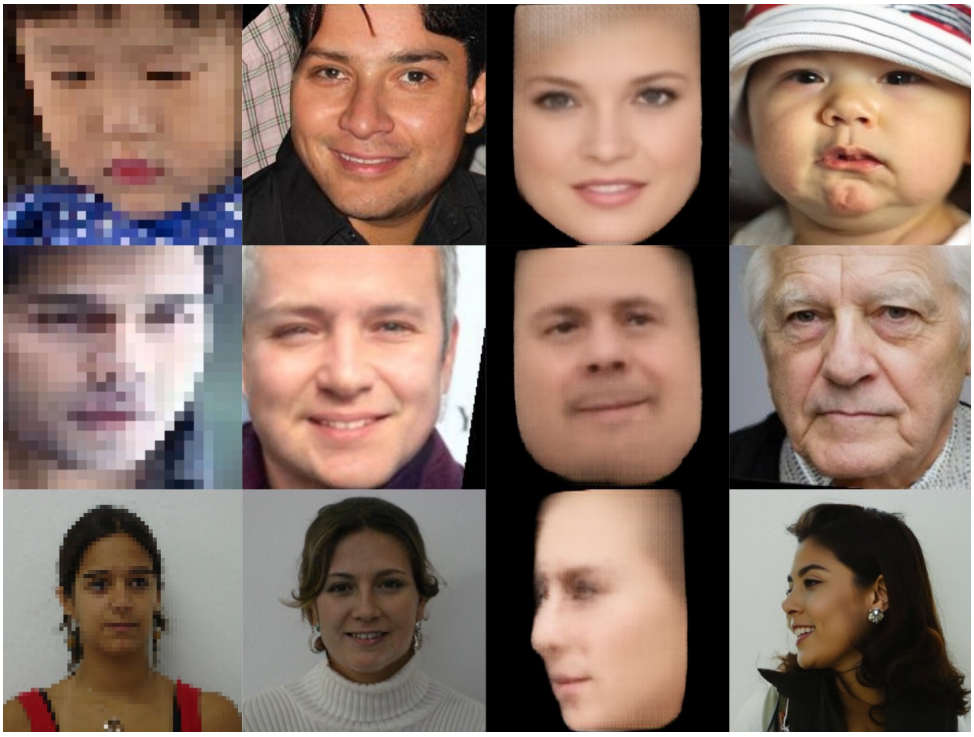


Figura 16 – Amostras de imagens dos melhores *scores* de qualidade perceptível. Cada coluna representa um algoritmo, sendo eles Pixelation, DeepPrivacy2, GANonymization e FADM, enquanto cada linha representa um conjunto de dados, sendo eles FairFace, UTKFace e FEI, respectivamente.

com o conjunto FairFace, percebe-se que a maioria das imagens com os melhores *scores* do DeepPrivacy2 têm faces altamente distorcidas e não naturais, o que não é esperado. Simultaneamente, as faces com menores *scores* da técnica FADM são melhor reconstruídas e mais naturais, em comparação com as do DeepPrivacy2. Na Figura 18, vemos os dois conjuntos e observamos que, além da qualidade perceptível da reconstrução das faces, a principal diferença entre as imagens é o nível de borramento. Enquanto as imagens geradas pela FADM são mais borradas, as que são geradas pela DeepPrivacy2 são mais nítidas.

Considerando que o treinamento do gerador do IFQA é baseado na restauração de imagens corrompidas por ruídos e outros fatores, levantamos a hipótese de que esses ruídos podem estar influenciando mais o cálculo do *score* do que, de fato, a qualidade perceptível da face gerada. Com o intuito de validar essa hipótese, selecionamos as 120 imagens com os melhores *scores* dentre as 240 analisadas anteriormente, aplicamos nelas um filtro gaussiano de desfocagem progressiva e medimos novamente os valores dos *scores* para cada nível de desfoque. Os quatro níveis da máscara do filtro podem ser vistos na Figura 19.

Calculamos a correlação de Pearson entre o nível do desfoque e o *score* da métrica como igual a -0,74. Esse valor indica uma relação moderadamente forte entre as duas variáveis, o que significa que a métrica IFQA tem uma forte dependência da nitidez da imagem para a

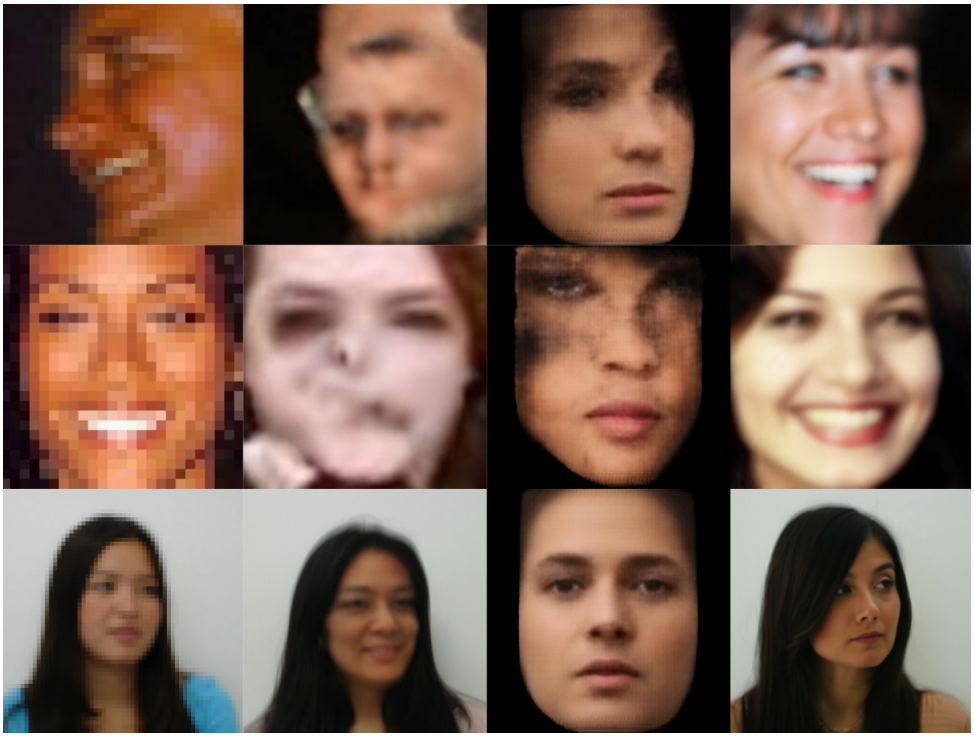


Figura 17 – Amostras de imagens com piores *scores* de qualidade perceptível.

Figura 18 – Comparação entre os melhores *scores* do DeepPrivacy2 com o FairFace (primeira linha) e os piores *scores* do FADM com o FairFace (segunda linha).

Figura 19 – Da esquerda para a direita, vemos uma imagem anonimizada pelo FADM sem borramento e com filtro gaussiano aplicado em 4 níveis de desfoque crescente. Seus *scores* são, respectivamente: 0,64, 0,63, 0,60, 0,31 e 0,02.

avaliação da qualidade perceptível. O valor de p = 1,23e-83 ($p < \alpha$) indica que a relação é estatisticamente significativa, podendo rejeitar a hipótese de que não há relação.

Tendo em vista essa tendência da IFQA, encontramos uma importante limitação da métrica, visto que ela pode favorecer técnicas que priorizam a nitidez em detrimento da autenticidade da reconstrução. O fato do Pixelation ser o algoritmo com os maiores *scores* médios calculados, mesmo sendo a técnica mais elementar, pode ser explicado por essa limitação, de forma que

a IFQA não capta o objetivo real da avaliação. Além disso, a correlação baixa entre os valores de IFQA e EPE calculada na Seção 5.1 pode ser um reflexo deste viés, de forma que a deformação alta da face não impacta tanto a qualidade calculada quanto o filtro gaussiano. Embora essa métrica fosse considerada promissora, ela se revela mais uma alternativa com limitações, deixando em aberto a busca por uma métrica verdadeiramente ideal para avaliar a reconstrução de faces. A ausência de uma ferramenta definitiva continua sendo um desafio a ser superado na área.

6 CONCLUSÃO

Nosso objetivo neste trabalho foi avaliar a justiça de algoritmos de anonimização de faces, relativa ao grau de deformação dos rostos e à qualidade perceptível das faces geradas artificialmente, no contexto de raça e gênero. Através de experimentos e da análise de dados, identificamos um algoritmo que mostrou uma maior discrepância de resultados entre os grupos, prejudicando o gênero feminino e a categoria de pessoas negras. A partir dos resultados preliminares, novas análises mostraram que, além dos algoritmos, as próprias métricas avaliadoras possuem vieses, de forma que elas podem mascarar desigualdades presentes nos algoritmos.

A falta de *benchmarks* na área de anonimização de faces foi uma limitação na pesquisa, de forma que muitos algoritmos utilizam métricas dispersas para se comparar com a literatura existente, dificultando a nossa escolha de métricas. A falta de referências sobre análise de justiça em modelos generativos também constituiu um desafio, de forma que não são modelos diretos de fazer esse tipo de análise, necessitando de maior conhecimento de domínio no problema. Uma outra limitação da nossa pesquisa é a falta de investigação sobre o nível de privacidade dos algoritmos de anonimização, para detectar quanto eles de fato protegem a face da pessoa. Sabemos que uma alta deformação pode inutilizar uma face para tarefas de visão computacional, mas não sabemos se uma face muito pouco deformada tem mais chances de ser reidentificada.

Esses resultados alertam para uma necessidade de desenvolver algoritmos de anonimização mais preocupados com justiça, como também evidenciam a necessidade de melhoria das métricas de avaliação de faces, caminhos que podem ser explorados em trabalhos futuros. Essas métricas devem funcionar de forma ponderada entre a detecção de ruídos na imagem e a medição da deformação da face, com o objetivo de se aproximar cada vez mais da percepção humana de qualidade visual. A criação de *benchmarks* também é uma possibilidade futura de investigação, tendo um potencial de grande impacto para orientar futuros pesquisadores.

REFERÊNCIAS

BABNIK, Ž.; PEER, P.; ŠTRUC, V. Faceqan: Face image quality assessment through adversarial noise exploration. In: IEEE. *2022 26th International Conference on Pattern Recognition (ICPR)*. [S.l.], 2022. p. 748–754.

BARATTIN, S.; TZELEPIS, C.; PATRAS, I.; SEBE, N. Attribute-preserving face dataset anonymization via latent code optimization. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2023. p. 8001–8010.

BAROCAS, S.; SELBST, A. D. Big data's disparate impact. *Calif. L. Rev.*, HeinOnline, v. 104, p. 671, 2016.

BARRETT, T.; CHEN, Q.; ZHANG, A. Skin deep: Investigating subjectivity in skin tone annotations for computer vision benchmark datasets. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. [S.l.: s.n.], 2023. p. 1757–1771.

BRANT, P. K. *Gaze Preservation on Artificially Generated Faces for Privacy Compliance*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2024.

CHENG, M.; DE-ARTEAGA, M.; MACKEY, L.; KALAI, A. T. Social norm bias: residual harms of fairness-aware algorithms. *Data Mining and Knowledge Discovery*, Springer, v. 37, n. 5, p. 1858–1884, 2023.

CHEVRIER, R.; FOUFI, V.; GAUDET-BLAVIGNAC, C.; ROBERT, A.; LOVIS, C. Use and understanding of anonymization and de-identification in the biomedical literature: scoping review. *Journal of medical Internet research*, JMIR Publications Toronto, Canada, v. 21, n. 5, p. e13484, 2019.

CSÁNYI, G. M.; NAGY, D.; VÁGI, R.; VADÁSZ, J. P.; OROSZ, T. Challenges and open problems of legal document anonymization. *Symmetry*, MDPI, v. 13, n. 8, p. 1490, 2021.

FAN, L. Image pixelization with differential privacy. In: SPRINGER. *Data and Applications Security and Privacy XXXII: 32nd Annual IFIP WG 11.3 Conference, DBSec 2018, Bergamo, Italy, July 16–18, 2018, Proceedings 32*. [S.l.], 2018. p. 148–162.

GOLDFARB-TARRANT, S.; MARCHANT, R.; SÁNCHEZ, R. M.; PANDYA, M.; LOPEZ, A. Intrinsic bias metrics do not correlate with application bias. *arXiv preprint arXiv:2012.15859*, 2020.

HELLMANN, F.; MERTES, S.; BENOUIS, M.; HUSTINX, A.; HSIEH, T.-C.; CONATI, C.; KRAWITZ, P.; ANDRÉ, E. Ganonymization: A gan-based face anonymization framework for preserving emotional expressions. *ACM Transactions on Multimedia Computing, Communications and Applications*, ACM New York, NY, 2024.

HUKKELÅS, H.; LINDSETH, F. Deepprivacy2: Towards realistic full-body anonymization. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. [S.l.: s.n.], 2023. p. 1329–1338.

HUKKELÅS, H.; MESTER, R.; LINDSETH, F. Deepprivacy: A generative adversarial network for face anonymization. In: SPRINGER. *International symposium on visual computing*. [S.l.], 2019. p. 565–578.

HUKKELÅS, H.; SMEBYE, M.; MESTER, R.; LINDSETH, F. Realistic full-body anonymization with surface-guided gans. In: *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision*. [S.l.: s.n.], 2023. p. 1430–1440.

HUYNH-THU, Q.; GHANBARI, M. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, IET, v. 44, n. 13, p. 800–801, 2008.

JO, B.; CHO, D.; PARK, I. K.; HONG, S. Ifqa: interpretable face quality assessment. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. [S.l.: s.n.], 2023. p. 3444–3453.

JUNIOR, L. L. d. O.; THOMAZ, C. E. *Captura e alinhamento de imagens: Um banco de faces brasileiro*. [S.l.], 2006.

KARKKAINEN, K.; JOO, J. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. [S.l.: s.n.], 2021. p. 1548–1558.

KARRAS, T.; LAINE, S.; AITTALA, M.; HELLSTEN, J.; LEHTINEN, J.; AILA, T. Analyzing and improving the image quality of stylegan. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2020. p. 8110–8119.

KLEMP, M.; RÖSCH, K.; WAGNER, R.; QUEHL, J.; LAUER, M. Ldfa: Latent diffusion face anonymization for self-driving applications. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2023. p. 3199–3205.

LUGARESI, C.; TANG, J.; NASH, H.; MCCLANAHAN, C.; UBOWEJA, E.; HAYS, M.; ZHANG, F.; CHANG, C.-L.; YONG, M. G.; LEE, J. et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

MALM, S.; RÖNNBÄCK, V.; HÅKANSSON, A.; LE, M.-h.; WOJTULEWICZ, K.; CARLSSON, N. Rad: Realistic anonymization of images using stable diffusion. In: *Proceedings of the 23rd Workshop on Privacy in the Electronic Society*. [S.l.: s.n.], 2023. p. 193–211.

MAXIMOV, M.; ELEZI, I.; LEAL-TAIXÉ, L. Ciagan: Conditional identity anonymization generative adversarial networks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2020. p. 5447–5456.

MITTAL, A.; MOORTHY, A. K.; BOVIK, A. C. Blind/referenceless image spatial quality evaluator. In: IEEE. *2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*. [S.l.], 2011. p. 723–727.

MÜLLER, L.-R.; PETERSEN, J.; YAMLAHI, A.; WISE, P.; ADLER, T. J.; SEITEL, A.; KOWALEWSKI, K.-F.; MÜLLER, B.; KENNGOTT, H.; NICKEL, F. et al. Robust hand tracking for surgical telestration. *International Journal of Computer Assisted Radiology and Surgery*, Springer, v. 17, n. 8, p. 1477–1486, 2022.

NOIRET, S.; LUMETZBERGER, J.; KAMPEL, M. Bias and fairness in computer vision applications of the criminal justice system. In: IEEE. *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. [S.l.], 2021. p. 1–8.

PARLIAMENT, E. *Artificial Intelligence Act: Provisional Agreement Resulting from Interinstitutional Negotiations*. 2024. Disponível em: <https://web.archive.org/web/20240310112041/https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/CJ40/AG/2024/02-13/1296003EN.pdf>.

PERERA, A.; ALETI, A.; TANTITHAMTHAVORN, C.; JIARPAKDEE, J.; TURHAN, B.; KUHN, L.; WALKER, K. Search-based fairness testing for regression-based machine learning systems. *Empirical Software Engineering*, Springer, v. 27, n. 3, p. 79, 2022.

PHILLIPS, M.; DOVE, E. S.; KNOPPERS, B. M. Criminal prohibition of wrongful re-identification: Legal solution or minefield for big data? *Journal of bioethical inquiry*, Springer, v. 14, p. 527–539, 2017.

PINSON, M. H. Why no reference metrics for image and video quality lack accuracy and reproducibility. *IEEE Transactions on Broadcasting*, IEEE, v. 69, n. 1, p. 97–117, 2022.

ROTHE, R.; TIMOFTE, R.; GOOL, L. V. Dex: Deep expectation of apparent age from a single image. In: *Proceedings of the IEEE international conference on computer vision workshops*. [S.l.: s.n.], 2015. p. 10–15.

SELBST, A. D.; BOYD, D.; FRIEDLER, S. A.; VENKATASUBRAMANIAN, S.; VERTESI, J. Fairness and abstraction in sociotechnical systems. In: *Proceedings of the conference on fairness, accountability, and transparency*. [S.l.: s.n.], 2019. p. 59–68.

SEO, J.; HAN, S.; LEE, S.; KIM, H. Computer vision techniques for construction safety and health monitoring. *Advanced Engineering Informatics*, Elsevier, v. 29, n. 2, p. 239–251, 2015.

SMITH, M.; MILLER, S. The ethical application of biometric facial recognition technology. *Ai & Society*, Springer, v. 37, n. 1, p. 167–175, 2022.

STARKE, C.; BALEIS, J.; KELLER, B.; MARCINKOWSKI, F. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, SAGE Publications Sage UK: London, England, v. 9, n. 2, p. 20539517221115189, 2022.

THOMEE, B.; SHAMMA, D. A.; FRIEDLAND, G.; ELIZALDE, B.; NI, K.; POLAND, D.; BORTH, D.; LI, L.-J. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, ACM New York, NY, USA, v. 59, n. 2, p. 64–73, 2016.

WANG, Z.; BOVIK, A. C.; SHEIKH, H. R.; SIMONCELLI, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, IEEE, v. 13, n. 4, p. 600–612, 2004.

ZHANG, R.; ISOLA, P.; EFROS, A. A.; SHECHTMAN, E.; WANG, O. The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2018. p. 586–595.

ZHANG, Z.; SONG, Y.; QI, H. Age progression/regression by conditional adversarial autoencoder. In: IEEE. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.], 2017.

ZWICK, P.; ROESCH, K.; KLEMP, M.; BRINGMANN, O. Context-aware full body anonymization using text-to-image diffusion models. *arXiv preprint arXiv:2410.08551*, 2024.