UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

ARTHUR SILVA CAVALCANTE FERREIRA

DISTINÇÃO ENTRE IMAGENS SINTÉTICAS DE FACES E IMAGENS DE FACES REAIS

CAMPINA GRANDE - PB

2024

ARTHUR SILVA CAVALCANTE FERREIRA

DISTINÇÃO ENTRE IMAGENS SINTÉTICAS DE FACES E IMAGENS DE FACES REAIS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Eanes Torres Pereira

CAMPINA GRANDE - PB

2024

ARTHUR SILVA CAVALCANTE FERREIRA

DISTINÇÃO ENTRE IMAGENS SINTÉTICAS DE FACES E IMAGENS DE FACES REAIS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Eanes Torres Pereira

Orientador – UASC/CEEI/UFCG

Herman Martins Gomes

Examinador – UASC/CEEI/UFCG

Francisco Vilar Brasileiro

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 16 de MAIO de 2024.

CAMPINA GRANDE - PB

RESUMO

As Redes Adversárias Generativas (GAN's) têm aplicações amplas, desde a criação de imagens e vídeos até a geração de texto e design de produtos. No contexto deste estudo, serão avaliadas imagens de faces sintéticas geradas por GAN's. Há benefícios neste uso de GAN's como pesquisas voltadas a entender a complexidade e nuances de imagens de faces e formação de bases de dados anônimas para treinamento de redes neurais com imagens de faces. Entretanto, faces sintéticas podem ser usadas para criar identidades falsas, podendo levar a crimes como fraude de identidade e *phishing*, onde faces sintéticas são usadas para enganar sistemas de segurança baseados em reconhecimento facial. Além disso, também podem ser usadas para criar vídeos e imagens falsos com intenções maliciosas, como difamação, desinformação ou propaganda política. Neste trabalho, foi treinada uma Rede Neural Convolucional Profunda baseada na arquitetura EfficientViT utilizando um conjunto de dados composto por bases de dados disponíveis publicamente e imagens sintéticas geradas pela rede StyleGAN3. Os resultados obtidos indicam uma taxa de acurácia de 99%, semelhante a outros métodos na literatura, porém as bases de dados utilizadas para treinamento e avaliação diferem além da quantidade de imagens utilizadas na avaliação. Ademais, houve uma procura de bases de dados diversificadas a fim de mitigar viés e justiça do modelo em relação à idade/etnia, porém uma análise à parte seria necessária para avaliar o impacto dessa escolha das bases de dados em comparação com outros modelos já treinados na literatura.

DISTINCTION BETWEEN SYNTHETIC IMAGES OF FACES AND IMAGES OF REAL FACES

ABSTRACT

Generative Adversarial Networks (GANs) have broad applications, ranging from image and video creation to text generation and product design. In the context of this study, synthetic face images generated by GANs will be evaluated. There are benefits to using GANs, such as external research to understand the complexity and nuances of facial images and the creation of anonymous databases for training neural networks with facial images. However, synthetic faces can be used to create false identities, leading to crimes such as identity theft and phishing, where synthetic faces are used to deceive facial recognition-based security systems. Additionally, they can also be used to create videos and fake images with malicious intent, such as defamation, misinformation, or political propaganda. In this work, a Deep Convolutional Neural Network based on the EfficientViT architecture was trained using a dataset composed of publicly available databases and synthetic images generated by the StyleGAN3 network. The results obtained indicate an accuracy rate of 99%, similar to other methods in the literature, but the databases used for training and evaluation vary beyond the number of images used in the evaluation. Furthermore, there was a search for diversified databases to mitigate bias and model fairness regarding age/ethnicity, but a separate analysis would be necessary to assess the impact of this choice of databases compared to other models already available in the literature.

Distinção entre imagens sintéticas de faces e imagens de faces reais

Arthur Silva Cavalcante Ferreira
Universidade Federal de Campina Grade
arthur.ferreira@ccc.ufcg.edu.br

Eanes Torres Pereira
Universidade Federal de Campina Grade
eanes@computacao.ufcg.edu.br

RESUMO

As Redes Adversárias Generativas (GAN's) têm aplicações amplas, desde a criação de imagens e vídeos até a geração de texto e design de produtos. No contexto deste estudo, serão avaliadas imagens de faces sintéticas geradas por GAN's. Há benefícios neste uso de GAN's como pesquisas voltadas a entender a complexidade e nuances de imagens de faces e formação de bases de dados anônimas para treinamento de redes neurais com imagens de faces. Entretanto, faces sintéticas podem ser usadas para criar identidades falsas, podendo levar a crimes como fraude de identidade e *phishing*, onde faces sintéticas são usadas para enganar sistemas de segurança baseados em reconhecimento facial. Além disso, também podem ser usadas para criar vídeos e imagens falsos com intenções maliciosas, como difamação, desinformação ou propaganda política. Neste trabalho, foi treinada uma Rede Neural Convolucional Profunda baseada na arquitetura EfficientViT utilizando um conjunto de dados composto por bases de dados disponíveis publicamente e imagens sintéticas geradas pela rede StyleGAN3. Os resultados obtidos indicam uma taxa de acurácia de 99%, semelhante a outros métodos na literatura, porém as bases de dados utilizadas para treinamento e avaliação diferem além da quantidade de imagens utilizadas na avaliação. Ademais, houve uma procura de bases de dados diversificadas a fim de mitigar viés e justiça do modelo em relação à idade/etnia, porém uma análise à parte seria necessária para avaliar o impacto dessa escolha das bases de dados em comparação com outros modelos já treinados na literatura.

Palavras-chave

Imagens sintéticas de faces; Imagens de faces reais; Redes Neurais Convolucionais Profundas; Redes GAN.

1. INTRODUÇÃO

As Redes Adversárias Generativas (GAN's) foram introduzidas no campo da aprendizagem profunda por Goodfellow et al. (2014). Uma rede GAN, uma forma de modelos generativos, é treinada em uma rede neural profunda de configuração adversária. Mais especificamente, a GAN aprende o modelo generativo de distribuição de dados por meio de métodos adversários. As GAN são o modelo generativo de maior sucesso desenvolvido nos últimos anos e são uma das áreas de pesquisa mais recentes no campo da inteligência artificial [2].

Um exemplo de aplicação das GAN's é a geração de faces. Sites como thispersondoesnotexist.com mostram fotos de pessoas geradas pelas GAN's [3]. Além disso, o modelo GFP-GAN para Restauração de Face Cega retém a qualidade da imagem e restaura as características faciais presentes na imagem melhor do que os modelos tradicionalmente usados [4].

Figura 1. Exemplo de face sintética adquirida no site thispersondoesnotexist.com, que utiliza imagens geradas pelo algoritmo StyleGAN2 [6].

As GAN's levaram à geração de imagens de face realistas, como a Figura 1, que foram usadas em contas falsas de mídia social e outros assuntos de desinformação que podem gerar impactos profundos. Isso levanta preocupações sobre a disseminação de informações falsas e a manipulação de percepções públicas. Por exemplo, um estudante do ensino médio criou um candidato falso usando uma face gerada por uma GAN em um evento de votação que enganou o Twitter para obter uma cobiçada marca azul, verificando assim a autenticidade da falsa candidatura. Este candidato falso, passando pela verificação, poderia configurar canais de doação para absorver fundos públicos, o que não só danifica as leis relacionadas à propriedade, mas também diminui a integridade da eleição. Portanto, as técnicas correspondentes de detecção de faces geradas pelas GAN's estão em desenvolvimento ativo para expor essas faces sintéticas [5].

O principal objetivo deste artigo é apresentar um modelo de rede neural convolucional com alta taxa de acerto com o uso de bases diversificadas, a fim de garantir a diversidade em relação à etnia/idade dos indivíduos, na classificação entre imagens de faces sintéticas geradas por GAN's e imagens de faces reais. Para tanto, a metodologia qualitativa foi adotada, com experimento utilizando bases de imagens sintéticas de faces geradas e imagens de faces reais disponibilizadas publicamente, previamente rotuladas quanto à autenticidade da face.

2. REVISÃO DA LITERATURA

Esta seção apresenta os trabalhos encontrados e relacionados ao objeto de estudo deste artigo, sendo subdividido nas seguintes subseções: 2.1 contendo trabalhos que usaram técnicas como características visuais e Aprendizagem de Máquina na tarefa de distinção de imagens sintéticas de faces e imagens de faces reais; 2.2 contendo informações sobre a arquitetura utilizada como modelo proposto por este trabalho.

2.1 Abordagens de distinção de imagens sintéticas e imagens de faces reais já exploradas

Gupta et al. (2013) investigaram a disseminação de imagens falsas no Twitter durante o furacão Sandy. O estudo coletou um conjunto de dados de imagens do Twitter relacionadas ao evento e as classificou manualmente como reais ou falsas. Em seguida, características visuais e sociais das imagens foram extraídas e utilizadas para treinar um modelo de aprendizado de máquina para identificar automaticamente imagens falsas. O estudo identificou três tipos principais de imagens falsas:

- Imagens manipuladas: imagens reais alteradas para adicionar ou remover elementos.
- Imagens reencenadas: imagens que representam eventos falsos ou enganosos.
- Imagens de contexto falso: imagens reais usadas com legendas falsas ou enganosas.

Gupta et al. (2013) concluem que o método proposto pode ser utilizado para identificar automaticamente imagens falsas em outros contextos. O estudo apresenta um método robusto para identificar imagens falsas no Twitter. Esse método pode ser adaptado e aplicado em pesquisas que visam analisar a desinformação visual em outras plataformas de mídia social ou em diferentes contextos.

Minyoung et al. (2018) propõem um método para detecção de manipulação de imagens, especificamente manipulação conhecida como "splicing", aonde partes de diferentes imagens são mescladas para criar uma falsa representação. O método proposto baseia-se na autoconsistência aprendida para identificar regiões suspeitas em uma imagem. Minyoung et al. (2018) utilizam uma abordagem baseada em rede neural para aprender padrões de autoconsistência em imagens autênticas e, em seguida, aplicam esses padrões para detectar áreas suspeitas em imagens alvo. Os resultados experimentais indicam que o método proposto supera outras abordagens existentes na detecção de splicing, fornecendo uma contribuição significativa para combater a disseminação de notícias falsas através da manipulação de imagens [7].

Dang et al. (2020) identificaram padrões sutis e artefatos característicos de manipulações digitais em imagens de faces. Esses incluem discrepâncias na textura da pele, inconsistências na iluminação e sombreamento, distorções nas proporções faciais, e a presença de bordas irregulares ou borradas ao redor da face. Além disso, o sistema desenvolvido pode detectar distorções em áreas específicas, como olhos e boca, os quais são comumente manipuladas em imagens sintéticas. Ao analisar esses sinais visuais distintivos, a rede neural convolucional consegue diferenciar eficazmente entre imagens reais e manipuladas, proporcionando uma detecção precisa e confiável de fraudes visuais em faces digitais [15].

Wang et al. (2019) destacam características distintivas e padrões visuais específicos utilizados para detectar faces sintéticas geradas por inteligência artificial. Essas características incluem inconsistências na textura da pele, falta de detalhes faciais realistas, artefatos de suavização excessiva e distorções nas proporções faciais. Além disso, o FakeSpotter identifica padrões de pixelização anômalos e irregularidades nas regiões dos olhos, boca e cabelo, comuns em imagens de faces falsas. Os resultados apresentados no estudo demonstram a eficácia do FakeSpotter na detecção de faces sintéticas, alcançando altas taxas de precisão e recall na distinção entre imagens sintéticas e reais. O sistema foi testado em conjuntos de dados abrangentes e desafiadores, revelando uma capacidade robusta e confiável na identificação de manipulações digitais. A abordagem simplificada do FakeSpotter, combinada com seu desempenho excepcional na detecção de faces falsas, destaca sua relevância como uma linha de base essencial para a detecção de conteúdo visual gerado artificialmente [16].

Yang et al. (2019) abordam a distinção entre imagens sintéticas de faces geradas por redes generativas adversariais (GANs) e imagens de faces reais. O objetivo do estudo é desenvolver um método eficiente para detectar imagens sintéticas de faces geradas por GANs, com base nas localizações dos landmarks faciais. Os landmarks faciais são pontos-chave nas faces, como os cantos dos olhos, nariz e boca, que podem fornecer informações cruciais sobre a autenticidade de uma imagem facial. O artigo propõe um algoritmo que utiliza técnicas de processamento de imagens e aprendizado de máquina para extrair automaticamente os landmarks faciais nas imagens. Em seguida, são aplicadas métricas de distância e análise de agrupamento para distinguir entre imagens sintéticas e reais. Os resultados experimentais demonstram que o método proposto pode identificar com precisão as imagens sintéticas geradas por GANs. Essa abordagem pode ter implicações significativas na detecção de imagens falsas e na preservação da autenticidade na era das deepfakes [17].

He et al. (2019) abordam a distinção entre imagens sintéticas de faces geradas artificialmente e imagens de faces reais, utilizando um método baseado em conjuntos de representações profundas de diferentes espaços de cores. O estudo desenvolve uma abordagem eficaz para detectar imagens falsas, considerando a diversidade de cores e características presentes nas imagens. O estudo propõe um sistema de detecção que utiliza uma combinação de representações profundas extraídas de múltiplos espaços de cores, como RGB, HSV e LAB. Essas representações são alimentadas em um modelo de aprendizado de máquina, treinado para distinguir entre imagens reais e falsas com base em suas características visuais. O método proposto demonstrou excelentes resultados na detecção de imagens falsas, superando abordagens anteriores. A análise dos resultados mostrou que a combinação de representações profundas de múltiplos espaços de cores permite uma melhor discriminação entre imagens reais e falsas, considerando as características sutis presentes em cada espaço de cor. O artigo conclui que a detecção de imagens falsas é um desafio importante no contexto da segurança digital, e que o uso de conjuntos de representações profundas de diferentes espaços de cores pode ser uma abordagem promissora para melhorar a precisão e a eficácia dos sistemas de detecção [18].

2.2 EfficientViT

O modelo EfficientViT é uma arquitetura de rede neural convolucional baseada em Transformers, projetada para tarefas de classificação de imagens. Essa arquitetura é uma variante do Vision Transformer (ViT), que foi introduzido como uma

abordagem inovadora para o processamento de imagens utilizando a técnica de auto atentividade dos Transformers [20].

O funcionamento do EfficientViT pode ser dividido em três etapas principais:

- Extração de características: A primeira etapa do modelo envolve a extração de características das imagens de entrada. Isso é feito por meio de uma camada de convolução inicial, que ajuda a capturar informações de baixo nível, como bordas e texturas. Em seguida, as características extraídas são divididas em patches, os quais são vetores bidimensionais de características. Esses patches são então achatados e linearizados antes de serem passados para a etapa de transformação.

- Transformação: A etapa de transformação é onde ocorre a principal operação do modelo. Ela é composta por vários blocos de transformação, os quais são unidades repetitivas de processamento. Cada bloco de transformação consiste em uma camada de atenção multi-cabeça e uma camada totalmente conectada. A camada de atenção permite que o modelo capture as relações entre os patches de características, enquanto a camada totalmente conectada ajuda a agregar as informações contextuais.

- Classificação: Após a etapa de transformação, as informações contextuais agregadas são passadas para um classificador linear, responsável por atribuir uma classe às imagens. Esse classificador é treinado com base em um conjunto de dados rotulados e pode aprender a mapear as características transformadas para as classes corretas. Ao final do processo, o modelo gera uma distribuição de probabilidade sobre as classes, indicando a probabilidade de cada classe para a imagem de entrada.

A estratégia principal usada pelo EfficientViT para alcançar eficiência computacional é a compactação da rede neural, uma vez que o mesmo emprega uma rede neural mais compacta em comparação com outras arquiteturas convolucionais. Isso é alcançado através do uso de blocos de transformação eficientes que reduzem o número de parâmetros da rede. Essa compactação resulta em uma arquitetura mais leve, exigindo menos recursos computacionais durante a inferência. No entanto, essa compactação não compromete significativamente o desempenho do modelo, permitindo que o EfficientViT mantenha uma boa capacidade de representação e classificação de imagens. A variação utilizada neste trabalho foi a EfficientViT-B2 onde a entrada possui as dimensões de 256 x 256, resultando em um modelo com cerca de 24 milhões de parâmetros [20].

Uma métrica importante a ser considerada ao avaliar o desempenho do EfficientViT é a latência, que se refere ao tempo necessário para processar uma imagem e gerar uma predição de classe. O EfficientViT é conhecido por sua baixa latência em comparação com outras arquiteturas de redes neurais convolucionais, tornando-o adequado para aplicações em tempo real [20].

Além disso, a interpretabilidade do modelo é um aspecto relevante a ser mencionado. Devido à sua arquitetura baseada em Transformers, o EfficientViT consegue capturar informações contextuais em diferentes partes da imagem, possibilitando uma melhor interpretação das decisões tomadas pelo modelo em relação à classificação das imagens [20].

3. ABORDAGEM PROPOSTA

Esta seção apresenta a abordagem proposta neste trabalho sendo dividida nas seguintes subseções: 3.1 para a exposição das bases de dados utilizadas e os procedimentos realizados nelas; 3.2 para a contextualização do pré-processamento adotado nas imagens para o treinamento e avaliação da arquitetura; 3.3 para apresentar o método adotado para a divisão do conjunto de dados para treino, validação e teste.

3.1 Bases de dados

3.1.1 Imagens sintéticas

Para a construção do conjunto de imagens sintéticas usadas para treinamento e avaliação foi utilizada uma rede GAN para a geração de imagens, sendo a StyleGAN3 [19] utilizada neste trabalho. Cerca de 206000 imagens foram geradas nas dimensões 256x256 com o modelo pré-treinado distribuído pela NVIDIA.

3.1.2 CASIA-Face-Africa

As imagens do banco de dados foram capturadas em vários locais da Nigéria, na África. Cerca de 1.150 voluntários participaram, onde para cada indivíduo as imagens foram capturadas simultaneamente por meio de 3 câmeras. Duas câmeras de comprimento de onda visível (VW) e uma câmera de infravermelho próximo (NIR). A captura foi feita em diversas sessões durante um período de 3 meses. Além disso, alguns sujeitos foram solicitados a usar acessórios faciais, como óculos, para capturas múltiplas. A base de dados organizada compreende um total de 38.546 imagens de 1.183 sujeitos. Especificamente, 12.063 imagens capturadas pela câmera VW 1 na resolução de 1332×1080, 13.232 imagens são capturadas pela câmera VW 2 na resolução de 787 × 962 e 13.251 imagens são capturadas pela câmera NIR na resolução de 983 × 877. Apenas as imagens capturadas pela câmera VW 1 foram utilizadas, por apresentarem menor ruído em comparação com o outro subconjunto em relação à luminosidade, por exemplo, o subconjunto capturado por infravermelho não se aplica no escopo deste trabalho [10].

3.1.3 CASIA-FaceV5

CASIA Face Image Database Versão 5.0 (ou CASIA-FaceV5) contém 2.500 imagens faciais coloridas de 500 indivíduos. As imagens desta base de dados são capturadas usando uma câmera USB Logitech em uma sessão. Todas as imagens de rosto são arquivos BMP de 16 bits e a resolução da imagem é de 640 * 480. As variações intra-classes típicas incluem iluminação, pose, expressão, óculos, distância de imagem, entre outras [12].

3.1.4 BUPT-CBFace

O conjunto de treinamento de reconhecimento facial de média escala BUPT-CBFace é construído explorando a estrutura de dados ideal de dados massivos coletados na internet, a fim de proporcionar um treinamento com melhor desempenho de reconhecimento facial a partir do balanceamento de imagens por indivíduo presente na base de dados. Nas tarefas de reconhecimento, o BUPT-CBFace não apenas apresenta um equilíbrio entre o reconhecimento variando idade e pose, mas também reduz o viés de reconhecimento em relação à etnia racial.

Duas versões do BUPT-CBFace estão publicamente disponíveis:

- BUPT-CBFace-50: Conjunto de dados com 10.000 indivíduos e 50 imagens por indivíduos, resultando em 500.000 imagens.

- BUPT-CBFace-12: Conjunto de dados com 41.667 indivíduos e 12 imagens por indivíduo, resultando em 500.004 imagens.

Para evitar o aprendizado de características em relação a uma face específica, optou-se por utilizar o BUPT-CBFace-12 [11].

Analisando as imagens da base de dados, foi visto não haver padronização quanto à resolução das imagens, o que permitia resultados indesejados na etapa de pré-processamento. Portanto, todas as imagens foram previamente filtradas, onde apenas aquelas com resolução de pelo menos 200 x 200 foram mantidas para treinamento, validação e teste.

3.2 Processamento

Todas as bases de dados passaram pelo seguinte pipeline de pré-processamento:

1. Utilizando o framework público DeepFace [9] em Python, as imagens passaram pelo detector de faces Retinaface, uma vez que apresenta a melhor precisão média quando comparado com outros modelos de estado da arte desenvolvidos para a tarefa de detecção [8].
2. Caso o detector indique haver mais de uma face na imagem, a mesma é descartada, uma vez que a presença de mais de uma face na imagem pode introduzir ruído para o modelo.
3. Após a filtragem, a imagem é cortada nas dimensões indicadas pelo detector. Como o recorte pode gerar uma nova imagem na qual a proporção não é quadrada, é aplicado um padding de pixels pretos na dimensão de menor tamanho. Foi escolhido o pixel de cor preta, pois o mesmo apresenta o valor (0, 0, 0) no sistema RGB, evitando assim a ativação de neurônios. Por fim, a imagem é redimensionada para 256x256.

A Figura 1 mostra o fluxograma do pipeline do processamento.

As Figuras (2, 3, 4 e 5) demonstram uma imagem de exemplo de cada base de dados antes e pós o processamento adotado. O lado esquerdo apresenta a imagem original, onde a borda cinza é apenas para identificar a proporção original da imagem, o lado direito apresenta a imagem processada.

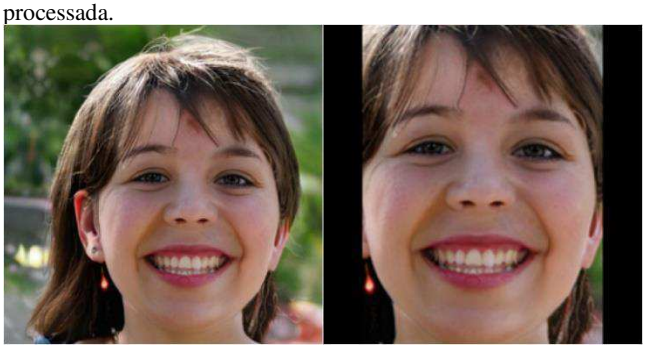

Figura 2. Imagens da StyleGAN3. Esquerda: imagem original. Direita: imagem processada.

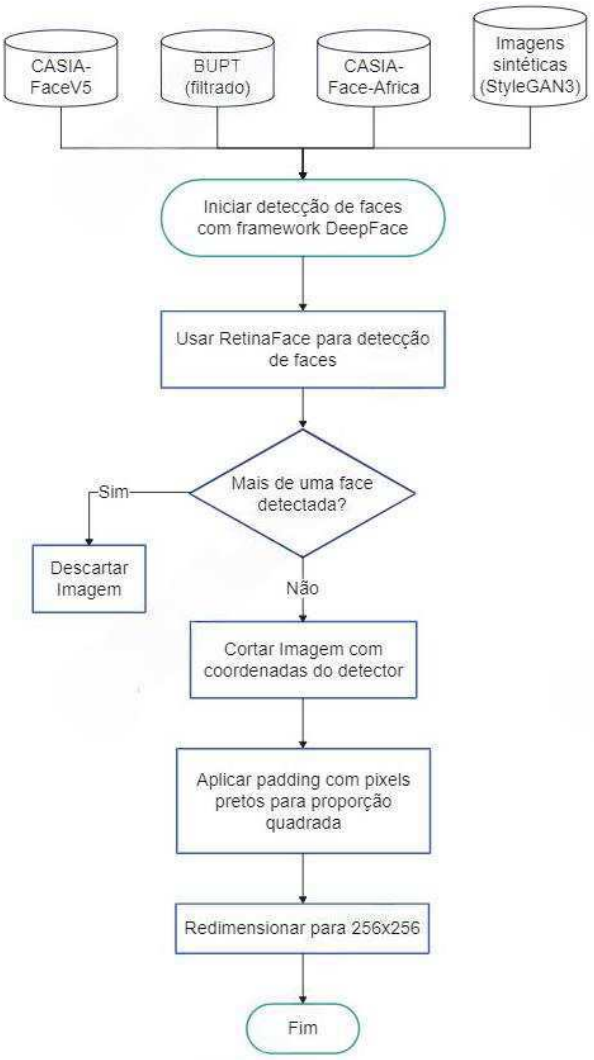

Figura 1. Fluxograma do processamento do conjunto de dados adotado.

Figura 3. Imagens da CASIA-Face-Africa. Esquerda: imagem original. Direita: imagem processada.

Figura 4. Imagens da CASIA-FaceV5. Esquerda: imagem original. Direita: imagem processada.

Figura 5. Imagens da BUPT-CBFace. Esquerda: imagem original. Direita: imagem processada.

3.3 Conjunto de dados

Utilizando as bases de dados mencionadas (StyleGAN3, CASIA-Face-Africa, CASIA-FaceV5, e BUPT-CBFace), o conjunto de dados foi dividido da seguinte forma: 10% dos dados foram reservados para validação, 10% para teste e 80% para treinamento do modelo de reconhecimento facial.

A distribuição exata das imagens de cada base de dados (CASIA-Face-Africa, CASIA-FaceV5, e BUPT-CBFace) apresentada na Tabela 1 nos conjuntos de treinamento, validação e teste foi feita de maneira proporcional, mantendo a distribuição adequada para garantir a representatividade e diversidade do conjunto de dados.

Tabela 1. Quantidade exata de imagens de cada base de dados em cada conjunto.

Base de dados	Treinamento	Validação	Teste
CASIA-Face-Africa	10.575	1.321	1.321
CASIA-FaceV5	1.997	249	249
BUPT-CBFace	130.486	16.310	16.310
Sintéticas (StyleGAN3)	164.827	20.603	20.603
Total	**307.885**	**38.483**	**38.483**

4. RESULTADOS

Esta seção apresenta os resultados obtidos nos experimentos realizados sendo dividida nas seguintes subseções: 4.1 para os experimentos realizados com diferentes arquiteturas de redes; 4.2 para as discussões acerca dos resultados encontrados.

4.1 Resultados da rede

Tabela 2. Métricas do modelo.

Acurácia	Precisão	Revocação	Tempo de inferência (Android)
99%	99%	99%	149072 ms

A partir da Tabela 2, observa-se que o modelo atingiu uma acurácia de 99%, indicando que pôde classificar corretamente a grande maioria das imagens sintéticas e reais. A alta acurácia demonstra a eficácia do modelo na tarefa de distinção entre os dois tipos de imagens. A precisão e revocação também foram medidas em 99%, indicando que o modelo teve uma alta taxa de verdadeiros positivos e baixa taxa de falsos positivos. Isso significa que o modelo foi preciso em suas classificações e recuperou a maioria dos exemplos positivos.

Essa alta acurácia apresentada pelo modelo é consistente com estudos anteriores, como o trabalho de Diego et al. (2021), que também obteve resultados semelhantes (acurácia de até 99%), mas as imagens sintéticas de faces utilizadas foram provenientes de redes GAN's anteriores, como a StyleGAN2, que apresenta uma geração de imagens sintéticas menos realistas quando comparadas com aquelas geradas pela StyleGAN3, logo, de maior facilidade de detecção por modelos como imagem sintética.

O tempo de inferência foi medido em 149072 ms em um smartphone Galaxy S22 com CPU Qualcomm Snapdragon 8Gen1 com a ferramenta de benchmarking mobile oficial do PyTorch [13]. Esse tempo refere-se ao tempo necessário para o modelo realizar a classificação de uma imagem no dispositivo móvel. O tempo de inferência é importante para a aplicação prática do modelo em dispositivos móveis, e o resultado obtido pode indicar a viabilidade de implementação em tempo real. Comparativamente, não foram encontrados estudos que tivessem extraído a métrica de inferência com modelos de tarefa similar.

4.2 Discussão

Os resultados obtidos demonstram que o modelo EfficientViT é altamente eficaz na tarefa de distinguir entre imagens sintéticas de faces e imagens de faces reais, com altas taxas de acurácia, precisão e revocação. A utilização do pipeline de pré-processamento, incluindo a detecção de faces com Retinaface e o redimensionamento das imagens para 256x256, contribuiu para a qualidade dos resultados obtidos. A aplicação do modelo em um dispositivo móvel de alto desempenho demonstrou que ele pode ser utilizado em cenários práticos, como aplicações de reconhecimento facial em tempo real. Esses resultados destacam a eficácia do modelo EfficientViT na distinção entre imagens sintéticas e reais, com potencial para aplicações práticas em segurança, autenticação e reconhecimento facial.

5. CONSIDERAÇÕES FINAIS

Neste estudo, foi explorada a classificação entre imagens sintéticas de faces geradas por GANs e imagens de faces reais, utilizando o modelo EfficientViT e um pipeline de pré-processamento que incluiu a detecção de faces com Retinaface e redimensionamento das imagens para 256x256. Os resultados obtidos demonstraram uma acurácia de 99%, bem como altas

taxas de precisão e revocação, indicando a eficácia do modelo na distinção entre os tipos de imagens.

A utilização do modelo EfficientViT revelou-se promissora para a tarefa em questão, proporcionando resultados consistentes e confiáveis. O pipeline de pré-processamento adotado contribuiu para a qualidade dos resultados, permitindo a padronização e correto dimensionamento das imagens, o que impactou diretamente na capacidade de classificação do modelo.

O tempo de inferência medido no dispositivo móvel Galaxy S22 demonstrou a viabilidade de implementação do modelo em cenários práticos, com uma resposta rápida que permite a utilização em tempo real. Isso é essencial para aplicações de segurança, autenticação e reconhecimento facial, onde a velocidade de processamento é crucial.

Além disso, a diversidade nas bases de dados utilizadas para treinamento e avaliação do modelo também é um aspecto crucial a ser considerado. A busca por bases de dados representativas e diversificadas em relação à idade, etnia e características faciais pode ajudar a evitar viés e garantir a justiça e equidade do modelo em diferentes contextos. Porém, é de extrema importância destacar a necessidade de investigação utilizando ferramentas apropriadas para identificar viés e garantir a justiça do modelo. A detecção de viés é fundamental para evitar a discriminação e assegurar a equidade nos resultados. Diversas ferramentas se propõem para essa finalidade, tais como o estudo de Buolamwini e Gebru (2018) que introduziu o conceito de "FairFace" como uma métrica para avaliar a justiça em sistemas de reconhecimento facial, considerando a representação adequada de diferentes grupos demográficos (Buolamwini & Gebru, 2018). Além disso, a biblioteca de código aberto "AI Fairness 360" (IBM Research, 2020) oferece uma coleção de algoritmos e métricas para avaliar e mitigar viés em modelos de aprendizado de máquina.

A interpretabilidade do modelo também é um estudo futuro a ser considerado. A capacidade de entender e explicar as decisões tomadas pelo algoritmo é essencial para garantir a transparência do sistema e identificar possíveis erros ou vieses ocultos. Diversas ferramentas se desenvolveram para auxiliar nesse aspecto. Por exemplo, a técnica Grad-CAM (Gradient-weighted Class Activation Mapping) proposta por Selvaraju et al. (2017) gera mapas de ativação para visualizar quais áreas da imagem foram mais relevantes para a decisão do modelo (Selvaraju et al., 2017).

Investigar e abordar questões de viés e interpretabilidade são passos essenciais para garantir a confiabilidade e ética dos sistemas de classificação de faces como sintéticas ou reais. Ao utilizar abordagens como as mencionadas, os pesquisadores e desenvolvedores podem ter uma compreensão mais profunda do funcionamento do modelo e tomar medidas para melhorar sua justiça e interpretabilidade.

Apesar dos resultados promissores expostos neste trabalho, é importante ressaltar a necessidade de considerar questões éticas e de privacidade relacionadas ao uso de tecnologias de reconhecimento facial e geração de imagens falsas. A criação e disseminação de deepfakes e identidades falsas podem ter impactos negativos significativos, reforçando a importância do desenvolvimento de métodos para detectar e combater essas práticas.

Em conclusão, o trabalho contribui para a pesquisa em distinção de imagens sintéticas de faces e imagens de faces reais, segurança cibernética e ética na inteligência artificial, fornecendo percepções importantes para aplicações futuras e o desenvolvimento de estratégias de mitigação de riscos associados ao uso de tecnologias de imagem sintética.

6. REFERÊNCIAS

[1] Ian Goodfellow, et al. 2014. Generative adversarial nets. Advances in neural information processing systems, vol. 27.

[2] Lan Lan, et al. 2020. Generative Adversarial Networks and Its Applications in Biomedical Informatics. Frontiers in public health, vol. 8, 164.

[3] Hadi Mansourifar, e Weidong Shi. 2020. One-shot gan generated fake face detection. arXiv preprint arXiv:2003.12244.

[4] Xintao Wang, et al. 2021. Towards Real-World Blind Face Restoration with Generative Facial Prior. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9164–9174.

[5] Xin Wang, et al. 2022. Gan-generated faces detection: A survey and new perspectives. arXiv preprint arXiv:2202.07145.

[6] Aditi Gupta, et al. 2013. Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy. In Proceedings of the 22nd International Conference on World Wide Web.

[7] Minyoung Huh, et al. 2018. Fighting fake news: Image splice detection via learned self-consistency. In Proceedings of the European conference on computer vision (ECCV).

[8] J. Deng, J. Guo, E. Ververas, I. Kotsia, e S. Zafeiriou. 2020. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5202–5211.

[9] Sefik Ilkin Serengil, e Alper Ozpinar. 2021. HyperExtended LightFace: A Facial Attribute Analysis Framework. In 2021 International Conference on Engineering and Emerging Technologies (ICEET), 1–4.

[10] Muhammad Jawad, et al. 2021. CASIA-Face-Africa: A Large-scale African Face Image Database. IEEE Transactions on Information Forensics and Security (TIFS), vol. 16, 3634–3646.

[11] Yaobin Zhang, e Weihong Deng. 2020. Class-balanced training for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 824–825.

[12] Instituto de Automática da Academia Chinesa de Ciências. CASIA Face Image Database Version 5.0 [conjunto de dados]. Disponível em: http://www.cbsr.ia.ac.cn/english/CASIA-WebFace-Database.html. Acesso em: 7 de maio de 2024.

[13] PyTorch. 2024. Android - Benchmarking Setup. In: PyTorch Tutorials 2.3.0+cu121 documentation. Disponível em: https://pytorch.org/tutorials/recipes/mobile_perf.html. Acessado em: 7 de maio de 2024.

[14] Ziwei Liu, et al. 2015. Deep Learning Face Attributes in the Wild. In Proceedings of International Conference on Computer Vision (ICCV).

[15] H. Dang, et al. 2020. On the Detection of Digital Face Manipulation. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5780–5789.

[16] R. Wang, et al. 2021. FakeSpotter: A simple yet robust baseline for spotting AI-synthesized fake faces. In Proceedings of

the 29th International Joint Conference on Artificial Intelligence, IJCAI 2020, 3444–3451.

[17] Xin Yang, et al. 2019. Exposing GAN-synthesized Faces Using Landmark Locations. In Proceedings of the ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec'19), 113–118.

[18] P. He, et al. 2019. Detection of Fake Images Via The Ensemble of Deep Representations from Multi Color Spaces. In 2019 IEEE International Conference on Image Processing (ICIP), 2299–2303.

[19] Tero Karras, et al. 2021. Alias-Free Generative Adversarial Networks. Proc. NeurIPS.

[20] Han Cai, Chuang Gan, e Song Han. 2022. Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. arXiv preprint arXiv:2205.14756.

[21] Diego Gragnaniello, et al. 2021. Are GAN Generated Images Easy to Detect? A Critical Analysis of the State-Of-The-Art. In 2021 IEEE International Conference on Multimedia and Expo (ICME), 1–6.

[22] Joy Buolamwini e Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 77–91.

[23] IBM Research. 2020. AI Fairness 360. Recuperado de https://aif360.mybluemix.net/.

[24] Ramprasaath R. Selvaraju, et al. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, 618–626.