VIII Simpósio de Iniciação Científica 2025

"Do conhecimento acadêmico à transformação sustentável: inovação com validação científica"

VULNERABILIDADES E CONTRA-MEDIDAS NA UTILIZAÇÃO DE BIOMETRIA UTILIZANDO INTELIGÊNCIA ARTIFICIAL

Guilherme S. Teixeira[1] (IC), Otávio de S. M. Gomes (PQ)[1]
[1]*Universidade Federal de Itajubá (UNIFEI).*

Palavras-chave: Cybersecurity. Deepfake. Face recognition. Liveness detection. Spoofing attacks.

Introdução

O desenvolvimento na área da inteligência artificial proporciona a criação e a melhoria de diversas ferramentas e aplicações, as quais influenciam os comportamentos das pessoas na sociedade dentro e fora da Internet. Dentro desse desenvolvimento, um grande progresso foi feito na Inteligência Artificial Generativa (IAG), que é a capacidade de modelos de aprendizado de máquina aprenderem distribuições de dados e gerar novas instâncias que se assemelhem à essas distribuições dos dados de treinamento **[4]**.

Entre as diversas possibilidades da IAG, está sendo obtido um grande avanço tecnológico nas técnicas de geração facial, as quais já alcançam um nível de realismo suficiente para enganar a percepção humana. Esse aprimoramento é capaz de causar grandes impactos positivos em algumas indústrias, como a do cinema e a da educação. Contudo, também existem impactos negativos, os quais levantam discussões éticas e sociais, principalmente com relação à segurança. A facilidade de troca de rostos abre portas para diversas fraudes e a disseminação de informações falsas, ou fake news, em grande escala. Para ser possível evitar os impactos negativos da técnica de geração facial e continuar tendo acesso às aplicações positivas, é necessário a compreensão e o desenvolvimento dessa tecnologia de forma responsável e ética.

Nesse contexto, o problema investigado neste trabalho é a crescente sofisticação das técnicas de geração facial baseadas em GANs, como os deepfakes, que ampliam riscos de fraude, manipulação e desinformação. O objetivo é apresentar uma análise das principais arquiteturas utilizadas nesse processo, como StyleGAN, CycleGAN e Face Swap, destacando seus fundamentos, aplicações e perigos associados. A contribuição consiste em reunir e discutir essas abordagens sob a perspectiva da cibersegurança, fornecendo uma visão crítica dos riscos técnicos e éticos envolvidos na utilização da IAG.

Metodologia

Este trabalho foi desenvolvido por meio de uma pesquisa exploratória e descritiva, baseada em revisão bibliográfica e análise comparativa de técnicas de Inteligência Artificial Generativa aplicadas à síntese de imagens faciais.

Inicialmente, foram estudados os fundamentos das Redes Adversariais Generativas (GANs), com ênfase em seu funcionamento e limitações. Em seguida, investigaram-se variações relevantes, como o StyleGAN, voltado para geração de imagens de alta fidelidade e controle de atributos, e o CycleGAN, aplicado a tarefas de tradução de imagens entre domínios sem a necessidade de pares de dados. Por fim, foi analisada a técnica de face swap, destacando sua aplicação prática por meio de ferramentas disponíveis em código aberto, como o Deep Live Cam. Essa abordagem permitiu discutir não apenas os mecanismos de geração e manipulação de imagens, mas também seus potenciais impactos em segurança, privacidade e disseminação de conteúdos falsos. A metodologia adotada buscou, portanto, sistematizar conceitos e aplicações dessas arquiteturas, com foco na identificação de vulnerabilidades e desafios de cibersegurança associados ao uso de deepfakes.

Resultados e discussão

A Inteligência Artificial Generativa (IAG) é fundamentada em modelos de aprendizado profundo capazes de aprender representações complexas e gerar novos dados semelhantes aos de treinamento. Entre suas arquiteturas, destacam-se as redes convolucionais (CNNs), aplicadas em imagens, e as recorrentes (RNNs), voltadas a dados sequenciais **[2]**. Diferente dos modelos discriminativos, que classificam, os modelos generativos criam novos exemplos, possibilitando tarefas como geração de imagens, vídeos, textos e áudios. Técnicas como Redes Adversariais Generativas

(GANs) **[4]** e Autoencoders Variacionais (VAEs) **[6]** foram marcos nesse avanço, permitindo sínteses altamente realistas e cada vez mais difíceis de distinguir da realidade.

Dentro desse contexto, surgem os *deepfakes*, mídias sintéticas em que rostos ou vozes são substituídos por meio de aprendizado profundo **[7]**. Embora ofereçam benefícios em áreas como cinema, treinamento e acessibilidade, também representam riscos significativos à cibersegurança.

A criação e disseminação facilitada desses conteúdos ampliam ameaças de fraudes, manipulação de informações e ataques contra sistemas biométricos. Softwares de código aberto, como o *FaceSwap* **[8]**, ilustram a acessibilidade dessas ferramentas, que podem ser usadas tanto em aplicações legítimas quanto em atividades maliciosas, tornando essencial compreender seus impactos e limitações.

Desde que foi introduzido por Goodfellow et al. (2014), o conceito de redes generativas adversariais (GANs) consolidou-se como um marco no campo da geração de dados sintéticos.

A arquitetura é composta por dois componentes principais: o gerador (G) e o discriminador (D), que são treinados de maneira adversarial. Nesse processo, o gerador busca produzir amostras cada vez mais realistas com o objetivo de enganar o discriminador, enquanto este se aperfeiçoa continuamente para distinguir com precisão entre amostras autênticas e geradas.

Apesar de sua relevância, o treinamento das GANs apresenta desafios como *mode collapse*, *vanishing gradients* e instabilidade de convergência, o que levou ao surgimento de variantes mais robustas, como a WGAN. Esses avanços ampliaram suas aplicações em geração de rostos, restauração de imagens e síntese audiovisual **[1][3]**, criando a base para arquiteturas mais sofisticadas, entre elas o StyleGAN, que trouxe maior controle sobre os atributos visuais das imagens.

Proposto por Karras et al. (2019), o StyleGAN representa uma evolução das GANs tradicionais, com foco na geração de imagens fotorrealistas e no controle refinado de atributos visuais. A principal inovação está no mapeamento do espaço latente z para um espaço intermediário w, mais disentangled, cujos vetores modulam o gerador por meio da técnica Adaptive Instance Normalization (AdaIN). **[5]** Esse mecanismo permite controlar desde estruturas globais até detalhes finos, enquanto ruídos estocásticos acrescentam realismo não determinístico. O treinamento é realizado de forma progressiva, aumentando a resolução de 4×4 até 1024×1024 pixels, com estabilidade garantida por técnicas como *equalized learning rate* e *fused leaky ReLU*. A implementação oficial foi disponibilizada em código aberto pela NVIDIA e, desde então, vem sendo amplamente aplicada em datasets como FFHQ (Flickr-Faces-HQ) e CelebA-HQ **[4][5]**. O StyleGAN se destaca por separar conteúdo e estilo, permitindo edições faciais precisas e controle semântico detalhado. Essas características favoreceram aplicações como deepfakes, reconstrução facial e manipulação por atributos. As versões subsequentes, StyleGAN2 e 3, aprimoraram a remoção de artefatos, a estabilidade do treinamento e a consistência espacial, mantendo a modulação de estilo como base da geração facial atual.

O CycleGAN,apresentado por Zhu et al. (2017), trouxe uma proposta diferente para a tradução de imagens entre domínios visuais distintos, dispensando o uso de dados pareados. Isso marcou um avanço importante em relação a modelos como o Pix2Pix, que dependem de imagens de entrada e saída perfeitamente alinhadas. **[10]** A arquitetura do CycleGAN é composta por dois geradores (G e F) e dois dis criminadores (DA e DB). O gerador G é responsável por mapear imagens do domínio A para o domínio B ($G : A \rightarrow B$), enquanto F realiza o caminho inverso ($F : B \rightarrow A$). Os discriminadores DA e DB avaliam se as imagens geradas pertencem aos respectivos domínios, treinando os geradores de forma adversarial.

O principal diferencial do CycleGAN está na perda de ciclo-consistência, que impõe que, ao transformar uma imagem de A para B e depois reconvertê-la de volta para A, o resultado seja similar à imagem original.

A técnica de Face Swap tem como objetivo substituir o rosto de uma pessoa pelo de outra, mantendo elementos como a expressão facial, a pose e a iluminação da imagem original. Essa abordagem ganhou bastante evidência com a popularização das deepfakes, sendo também aplicada em efeitos de vídeo e em ferramentas de edição facial automatizada. **[6]**

Modelos modernos de face swap, como o FaceController ou SIMswap, operam a partir de três pilares principais: extração da identidade do rosto-fonte, preservação da estrutura do rosto-alvo e recomposição refinada da imagem.**[7]** Essa tarefa pode ser formali zada da seguinte forma:

Onde:

- Is é a imagem-fonte (cujo rosto será inserido);

- It é a imagem-alvo (que manterá pose, expressão e iluminação);

- Es e Et são os extratores de características de identidade e estrutura, respectiva mente;
- G é ogerador responsável por compor a nova face combinando os atributos de Is com a geometria de It.

A eficácia do face swap depende do equilíbrio entre identidade e fidelidade estrutural. Assim, é comum utilizar uma função de perda composta, como:

$$Ltotal = \lambda_{id} L_{id} + \lambda_{rec} L_{rec} + \lambda_{adv} L_{adv}$$

Onde:

- Lid é a perda de identidade, que penaliza desvios do rosto-fonte;
- Lrec é a perda de reconstrução (como L1), garantindo fidelidade estrutural;
- Ladv é a perda adversarial, forçando realismo;
- Os hiperparâmetros λid, λrec e λadv ponderam o impacto de cada termo.

É uma técnica robusta, capaz de lidar com variações de iluminação e poses distintas, mantendo a integridade da imagem de fundo. Com módulos especializados de forma e textura, alcança resultados de alta fidelidade em tempo real.

Durante o minicurso ministrado pelo autor, foi apresentada a aplicação **Deep Live Cam**, ferramenta de código aberto para troca de rostos em tempo real. Sua simplicidade evidencia como técnicas complexas de face swap já estão acessíveis ao público, ampliando riscos de uso malicioso, como fraudes em entrevistas virtuais e reuniões online, o que reforça a importância de discutir seus impactos na cibersegurança.

Para fins exclusivamente éticos e educacionais, foi utilizado um rosto sintético gerado por IA (via *thispersondoesnotexist.com*), evitando-se o uso de imagens de pessoas reais ou figuras públicas. A demonstração evidenciou tanto o potencial quanto os riscos das técnicas de *face swapping*, especialmente em cenários de engenharia social, manipulação de identidade e desinformação audiovisual. A aplicação *Deep Live Cam* está disponível publicamente no GitHub, com código-fonte aberto e documentação acessível em:

https://github.com/ hacksider/Deep-Live-Cam

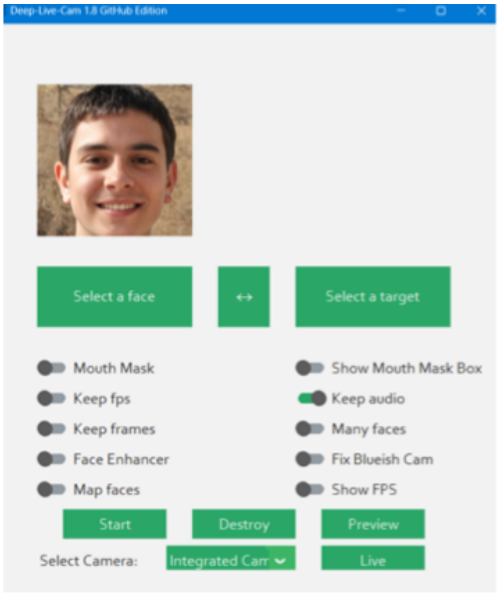

Figura 1 - Interface da aplicação Deep Live Cam. A imagem-fonte utilizada foi gerada por uma redegenerativa de rostos sintéticos. A ferramenta permite alternar modos de mapeamento, mascaramento labial, preservação de FPS e renderização em tempo real. Fonte: elaborado pelo autor.

Nos últimos anos, os deepfakes deixaram de ser apenas uma curiosidade tecnológica e passaram a ser usados em esquemas de fraude sofisticados. Um dos primeiros casos amplamente divulgados ocorreu em 2019, quando criminosos clonaram a voz de um CEO para induzir uma transferência de US$ 243 mil **[12]**. Desde então, os golpes evoluíram: em Hong Kong, por exemplo, uma funcionária foi enganada por uma videochamada em que todos os participantes eram executivos falsos gerados por IA, resultando em perdas de cerca de 200 milhões de dólares de Hong Kong **[13]**.

Além das fraudes financeiras, o uso malicioso também chegou ao mercado de trabalho. O FBI alertou sobre entrevistas de emprego realizadas por pessoas que utilizavam deepfakes para ocultar sua identidade, prática já documentada em diferentes setores **[14][15]**. Esses episódios demonstram como a tecnologia, antes restrita a ambientes de pesquisa, já é explorada em crimes concretos, ampliando os desafios para a cibersegurança.

Apesar dos riscos apontados, as técnicas de geração facial também apresentam potenciais benefícios quando aplicadas de forma ética. Tais recursos podem contribuir para avanços na educação, na acessibilidade e em produções criativas, permitindo novas formas de

interação digital e suporte a diferentes públicos. Esse cenário reforça a necessidade de equilíbrio entre inovação tecnológica e responsabilidade social, de modo que as aplicações positivas não sejam ofuscadas pelos riscos de uso malicioso.

Conclusões

A Inteligência Artificial Generativa (IAG) aplicada à síntese e manipulação de rostos digitais apresenta avanços significativos, com potencial para aplicações positivas em áreas como cinema, acessibilidade e educação. Entretanto, o mesmo realismo que torna essas técnicas úteis também amplia riscos de fraudes, manipulação de identidade e disseminação de desinformação, exigindo atenção da comunidade de cibersegurança. Este trabalho destacou a evolução de arquiteturas como GANs, StyleGAN, CycleGAN e Face Swap, evidenciando seus fundamentos, capacidades e implicações.

Conclui-se que o desenvolvimento e a utilização dessas tecnologias devem ser orientados por princípios éticos e acompanhados de investimentos em métodos de detecção de deepfakes, de modo a mitigar ameaças e assegurar que seus benefícios sejam aproveitados de forma segura e responsável.

Agradecimentos

Gostaria de expressar minha profunda gratidão às pessoas e instituições que contribuíram de maneira essencial para o desenvolvimento deste trabalho. Agradeço imensamente ao professor Otávio de S. M. Gomes, meu orientador, pela confiança em me conceder esta grande oportunidade, pela orientação dedicada, pelos valiosos conselhos e pelo apoio constante ao longo de todo o projeto.

Ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), pelo financiamento e suporte que tornaram este trabalho possível. À Universidade Federal de Itajubá (UNIFEI), pela infraestrutura, ambiente acadêmico e recursos disponibilizados para a execução deste estudo.

Referências

[1] BENGESI, Staphord; EL-SAYED, Hoda; SARKER, Md Kamruzzaman; HOUKPATI, Yao; IRUNGU, John; OLADUNNI, Timothy. Advancements in generative AI: a comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers. *arXiv preprint*, arXiv:2311.10242, 2023.

[2] BENGIO, Yoshua; SIMARD, Patrice; FRASCONI, Paolo. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, v. 5, n. 2, p. 157-166, 1994. *IEEE:* 10.1109/ICNN.1993.298725

[3] CHAKRABORTY, Tanujit; REDDY, Ujjwal K. S.; NAIK, Shraddha M.; PANJA, Madhurima; MANVITHA, Bayapureddy. Ten years of generative adversarial nets (GANs). *arXiv preprint*, arXiv:2308.16316, 2023.

[4] GOODFELLOW, Ian; POUGET-ABADIE, Jean; MIRZA, Mehdi; XU, Bing; WARDE-FARLEY, David; OZAIR, Sherjil; COURVILLE, Aaron; BENGIO, Yoshua. Generative adversarial nets. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S. l.: s. n.], arXiv: 1406.2661

[5] KARRAS, Tero; LAINE, Samuli; AILA, Timo. A style-based generator architecture for generative adversarial networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S. l.: s. n.], 2019. p. 4401-4410. doi: 10.1109/CVPR.2019.00453

[6] KINGMA, Diederik P.; WELLING, Max. Auto-encoding variational bayes. *arXiv preprint*, arXiv:1312.6114, 2013.

[7] LI, Yanan; YUAN, Yiming; LIU, Xinyu; WANG, Yajing; ZHANG, Wei; WANG, Yizhou. Fine-grained face swapping via editing with regional GAN inversion. *arXiv preprint*, arXiv:2310.15081, 2023.

[8] NIRKIN, Yuval; KELLER, Yosi; HASSNER, Tal. FSGAN: subject agnostic face swapping and reenactment. *arXiv preprint*, arXiv:1908.05932, 2019.

[9] RÖSSLER, Andreas; COZZOLINO, Davide; VERDOLIVA, Luisa; RIESS, Christian; THIES, Justus; NIEßNER, Matthias. FaceForensics++: learning to detect manipulated facial images. *arXiv preprint*, arXiv:1901.08971, 2019.

[10] ZHU, Jun-Yan; PARK, Taesung; ISOLA, Phillip; EFROS, Alexei A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. [S. l.: s. n.], 2017. p. 2223-2232. arXiv: 1703.10593

[11] ZHANG, Xiaoyu; WANG, Yifan; LIU, Yifan; LI, Yanan; WANG, Yizhou. High-fidelity face swapping with style blending. *arXiv preprint*, arXiv:2312.10843, 2023.

[12] FORBES. Criminals used AI to mimic CEO's voice in $243,000 fraud. Forbes, 2019. Disponível em: https://www.forbes.com/sites/thomasbrewster/2019/09/03/criminals-used-ai-to-mimic-ceos-voice-in-243000-fraud/

[13] THE GUARDIAN. Hong Kong woman loses HK$200m in deepfake video call scam. The Guardian, 2024. Disponível em:https://www.theguardian.com/world/2024/feb/04/hong-kong-woman-loses-hk200m-in-deepfake-video-call-scam

[14] FBI. FBI warns of deepfake job applicants using stolen personal information. Federal Bureau of Investigation, 2022. Disponível; https://www.ic3.gov/Media/Y2022/PSA220628.

[15] BUSINESS INSIDER. Cybercriminals are using deepfake job interviews to steal remote work. Business Insider, 2022. Disponível em: https://www.businessinsider.com/deepfake-job-interviews-remote-work-scam-2022-7