Ataques adversariais como estratégia de proteção de imagens femininas contra *deepfakes*: um desenho de pesquisa

Cléo Cunha Peixoto[1], **Claudia Lage Rebello da Motta**[1], **Pedro Nuno de Souza Moura**[2]

[1] Programa de Pós-Graduação em Informática (PPGI)
Universidade Federal do Rio de Janeiro (UFRJ)

[2]Programa de Pós-Graduação em Informática (PPGI)
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)

cleocpeixoto@gmail.com, claudiam@nce.ufrj.br, pedro.moura@uniriotec.br

Abstract. *The popularization of generative Artificial Intelligence models has transformed everyday life, enabling the production of deepfake pornography through the non-consensual appropriation of personal images. Predominant narratives place responsibility on individuals for the improper use of technology. Grounded in critical theorization that asserts that technology is not neutral, this work understands deepfakes as expected outcomes of a structure from which large corporations profit. This research investigates the use of adversarial attacks to introduce perturbations into female facial images, rendering them unusable by generative models and broadening the debate on protection in collaborative systems.*

Resumo. *A popularização de modelos generativos de Inteligência Artificial tem transformado o cotidiano das pessoas, viabilizando a produção de pornografia deepfake, a partir da apropriação de imagens pessoais sem consentimento. A narrativa predominante responsabiliza o indivíduo pelo uso inadequado da tecnologia. Este trabalho parte da teorização crítica de que a tecnologia não é neutra e, portanto, compreende os deepfakes como resultados esperados, em que grandes corporações se beneficiam com o lucro desta prática. Esta pesquisa investiga o uso de ataques adversariais para gerar perturbações em imagens faciais femininas a fim de inutilizá-las por modelos generativos, ampliando o debate sobre proteção em sistemas colaborativos.*

1. Introdução

Há uma roupagem elegante na associação entre tecnologia e progresso, que sustenta uma naturalização de que consequências sociais prejudiciais decorrentes do seu uso seriam inevitáveis - uma espécie de "preço do avanço". Para romper com este fatalismo tecnológico, é necessário compreender que tecnologias, quando desenvolvidas em sociedades capitalistas, potencializam opressões ao invés de romper com padrões estruturais de dominação, contrastando as expectativas frequentemente sustentadas pelo senso comum [Campante 2024].

Neste contexto, a popularização da Inteligência Artificial generativa viabiliza possibilidades irrestritas de utilização, desde usos responsáveis como auxílio no diagnóstico

precoce de câncer ao gerar análises de imagens médicas, por exemplo, até o uso exploratório como geração de desinformação (*fake news*) e pornografia artificial, que é o objeto deste estudo. A indústria pornográfica, portanto, encontra nesses modelos um novo mecanismo de lucro e exploração de corpos femininos, cuja eficiência reside na capacidade de produzir em larga escala, com extrema facilidade, rapidez e verosimilhança, conteúdos pornográficos a partir de imagens pessoais de qualquer mulher sem consentimento, que correspondem ao chamado *deepfake*. A análise marxista da falta de neutralidade tecnológica indica que *deepfakes* pornográficos não constituem desvios acidentais, mas resultados coerentes dentro da lógica da maximização de lucro, uma vez que servem a indústrias bilionárias, como a pornográfica - consolidada como uma das mais lucrativas do capitalismo contemporâneo [Dines 2022].

A proposta técnica aqui explorada, ao reconhecer este cenário, posiciona-se como uma resistência política e uma possibilidade de proteção de meninas e mulheres. Não se trata de buscar impedir a ocorrência da criação artificial de imagens si, mas de dificultar a exposição de mulheres sem o devido consentimento. A partir desta perspectiva, esta pesquisa propõe investigar o potencial de aplicação de técnicas adversariais em imagens faciais de mulheres com o objetivo de inviabilizar sua utilização por modelos generativos.

2. Problema de pesquisa

A pornografia *deepfake* trouxe um novo desafio à emancipação feminina ao reconfigurar as dinâmicas de violência sexual contra as mulheres no ambiente digital [Akter and Ahmed 2025]. Para discutir este problema, é interessante observar que as consequências da institucionalização da pornografia em nossa sociedade foram a total desumanização e objetificação sistemática dos corpos das mulheres, alienando-as de suas próprias emoções e desejos [Marx 2011].

Esse processo é corroborado quando Dines, pesquisadora crítica da indústria pornográfica, descreve sua experiência na feira de negócios *Adult Entertainment Expo*, onde observa que o interesse dos homens não está relacionado à sexualidade, mas às estratégias de maximização do lucro que a indústria pode oferecer, anulando qualquer vestígio da presença humana (a mulher) na produção pornográfica [Dines 2022].

Diante da extrema desumanização feminina promovida pela pornografia *deepfake*, a estratégia de enfrentamento proposta por esta pesquisa consiste na investigação da aplicação de técnicas de ataques adversariais, oriundas do campo do Aprendizado Profundo, como mecanismo de proteção contra a exploração feminina. Ataques adversariais consistem na adição de pequenas perturbações aos dados de entrada capazes de induzir redes neurais profundas a produzirem saídas incorretas ou inconsistentes, mesmo com alto grau de confiança [Liang et al. 2022]. Essas perturbações exploram vulnerabilidades inerentes aos modelos, tornando-os incapazes de gerar conteúdos artificiais coerentes a partir das imagens originais.

Embora tradicionalmente investigadas com o objetivo de avaliar ou aumentar a eficácia e a robustez dos modelos, estudos recentes já reconhecem o potencial dessas técnicas como instrumentos de defesa [Guo et al. 2025]. Este trabalho posiciona-se, portanto, neste âmbito menos explorado ao deslocar a utilização dos ataques de uma lógica de fortalecimento à estrutura para uma lógica de defesa, justificando a relevância e atualidade da pesquisa. Desta maneira, o problema de pesquisa que orienta este estudo consiste

em investigar se a própria vulnerabilidade técnica dos modelos pode ser explorada como mecanismo de proteção.

3. Desenho metodológico da pesquisa

Esta pesquisa adota uma abordagem de experimentação computacional aplicada, estruturada a partir da aplicação de técnicas de ataques adversariais sobre as imagens selecionadas, com o objetivo de avaliar sua capacidade de comprometer ou inviabilizar o uso dessas imagens pelos modelos generativos especificados a seguir. O desenho metodológico compreende quatro etapas:

- **Seleção e preparação de um conjunto de imagens faciais** a partir da base de dados pública FairFace, desenvolvido por Karkkainen e Joo [Karkkainen and Joo 2021], que contém em torno de 100 mil imagens de rostos humanos;
- **Aplicação de perturbação adversarial** utilizando três técnicas amplamente reconhecidas na literatura: *Fast Gradient Sign Method* (FGSM), proposta por [Goodfellow et al. 2014]; *Basic Iterative Method* (BIM), proposta por [Kurakin et al. 2016]; *Projected Gradient Descent* (PGD), proposta por [Madry et al. 2017]. Os experimentos serão realizados múltiplas vezes, em número a ser definido, a fim de contemplar o caráter estocástico dos ataques e dos modelos generativos, e algumas medidas serão coletadas;
- **Submissão das imagens originais e perturbadas aos modelos generativos** StyleGAN2, proposto por [Karras et al. 2019], que obteve resultados relevantes em rostos humanos; e Stable Diffusion, modelo de difusão proposto por [Rombach et al. 2022], reconhecido pela geração de imagens de alta qualidade. Ambos reconhecidos pelo alto realismo das imagens geradas, o que os torna particularmente relevantes para avaliação de ataques adversariais.
- **Análise comparativa das saídas obtidas**, através da métrica de distância *Fréchet Inception Distance* (FID), proposta por [Heusel et al. 2017], empregada para avaliação da qualidade das imagens geradas, de modo a garantir a reprodutibilidade e comparabilidade dos resultados.

Considerando o caráter público e anonimizado dos dados, não há necessidade de submissão ao comitê de ética.

4. Avaliação planejada

A avaliação da proposta tem o objetivo de analisar a eficácia das perturbações aplicadas às imagens. Para isso, serão comparados os comportamentos dos modelos quando submetidos às imagens originais e às suas versões perturbadas pelos ataques adversariais. As definições das três categorias principais de avaliação são:

- **Ataque bem-sucedido:** falha completa na geração de imagens; indica que o modelo generativo foi incapaz de produzir uma imagem coerente a partir da entrada perturbada;
- **Ataque parcialmente bem-sucedido:** geração de imagem com distorções perceptíveis; quando há falhas visíveis significativas na imagem gerada, ainda que parcialmente reconhecível;

- **Ataque malsucedido:** geração de uma imagem funcional e realista; o modelo foi capaz de gerar uma imagem coerente com uma aparência natural, mesmo a partir da entrada adversarial.

Essa categorização será combinada com a métrica *Fréchet Inception Distance* (FID), permitindo comparar quantitativamente a divergência entre as distribuições das imagens geradas a partir de entradas originais e perturbadas. Serão também aplicados testes de hipótese para verificar se as diferenças obtidas são estatisticamente significativas. A análise conjunta busca fornecer evidências empíricas sobre a capacidade dos ataques adversariais de inviabilizar a utilização das imagens por modelos generativos.

5. Contribuições, limitações e direções futuras

Do ponto de vista técnico, esta pesquisa contribui ao investigar o uso de ataques adversariais em um contexto ainda pouco explorado na literatura: a proteção de imagens femininas contra a exploração para a geração de *deepfakes* pornográficos. Ao reposicionar técnicas tradicionalmente empregadas para aumentar a robustez de modelos generativos como mecanismo de proteção, o trabalho amplia o escopo de aplicação dos ataques e propõe um uso contra-hegemônico dessas técnicas. Em termos de metodologia, este estudo se baseia em uma experimentação computacional e avaliação por métricas reprodutíveis, possibilitando uma melhor qualidade e consistência da análise da eficácia das perturbações.

No âmbito sociotécnico, a pesquisa contribui para o debate crítico ao propor a associação entre a utilização de técnicas de ataques adversariais e a discussão sobre o uso exploratório de imagens femininas em sistemas colaborativos digitais, compreendendo a técnica como espaço de disputa e resistência.

Entretanto, é fundamental reconhecer as limitações sociais e técnicas desta pesquisa. Por razões éticas e metodológicas, este estudo não envolve imagens de caráter sexual ou nudez e delimita-se à utilização de imagens faciais estáticas, retiradas de uma base de dados pública, anonimizadas e obtidas com condições relativamente controladas de iluminação, enquadramento e qualidade visual, ideais para testes. Este cenário se distancia das condições reais de conteúdos pornográficos que envolvem manipulação de corpos inteiros e também vídeos, bem como a extração das imagens a partir de redes sociais por meio de capturas de tela, sujeitas à ruídos maiores, recortes e outras degradações visuais recorrentes.

Dessa forma, o desenho planejado representa uma etapa inicial, com o objetivo de validar o protocolo metodológico, verificar a viabilidade técnica da abordagem escolhida e fornecer subsídios para experimentos mais robustos e com condições mais abrangentes. Como direções futuras, pode-se refletir a inclusão de operações de redimensionamento, compressão em diferentes níveis de qualidade, adição de ruído e inserção de elementos gráficos típicos de interfaces digitais, entre outros, permitindo avaliar a robustez das técnicas frente a degradações mais próximas de cenários reais de uso. Técnicas de interpretabilidade também serão exploradas.

Apesar de a pesquisa se inserir no âmbito da prevenção, é importante reforçar que os resultados obtidos devem ser compreendidos como uma contribuição para a mitigação do problema, devendo ser articulados com por exemplo, a regulamentação dos ambientes digitais que, conforme argumenta [Meira 2024], são frequentemente marcados pela crescente assimetria no uso de dados.

Uso de Inteligência Artificial

Foram utilizadas ferramentas de Inteligência Artificial Generativa exclusivamente para apoio na tradução do abstract para o idioma inglês e na revisão gramatical e estilística do manuscrito. Não foram empregadas ferramentas de IA para geração de conteúdo, argumentação, análise de dados ou produção de resultados, sendo esta responsabilidade integral acadêmica do(a) autor(a).

Referências

Akter, S. and Ahmed, P. (2025). The emergence of ai-generated deepfakes as a new tool for gender-based violence against women: A brief narrative review of evidence and the implications of the techno-feminist perspective. 13:1–17.

Campante, R. G. (2024). Marx, allende e a rejeição do fatalismo tecnológico.

Dines, G. (2022). *Pornland: como a indústria do sexo sequestrou nossa sexualidade*. Caqui Livros, São Paulo. Tradução de Bruna Della Torre.

Goodfellow, I., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv 1412.6572*.

Guo, Z., Qian, Y., Li, Y., Li, W., Lei, C. T., Zhao, S., Fang, L., Arandjelović, O., and Lau, C. P. (2025). Beyond vulnerabilities: A survey of adversarial attacks as both threats and defenses in computer vision systems.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium*.

Karkkainen, K. and Joo, J. (2021). 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*, pages 1548–1558.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2019). Analyzing and improving the image quality of stylegan.

Kurakin, A., Goodfellow, I., and Bengio, S. (2016). Adversarial examples in the physical world.

Liang, H., He, E., Zhao, Y., Jia, Z., and Hao, l. (2022). Adversarial attack and defense: A survey. *Electronics*, 11:1283.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks.

Marx, K. (2011). Fragmento sobre as máquinas. In *Grundrisse: manuscritos de 1857-1858*, pages aprox. 585–615. Boitempo, São Paulo. Parte dos manuscritos de 1857-1858.

Meira, S. (2024). Ia sabe demais: dados, algoritmos e o futuro da privacidade. Poder360. Acesso em: 10 fev. 2026.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *High-Resolution Image Synthesis with Latent Diffusion Models*, pages 10674–10685.