UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Otávio Malta Borges

Detecção de Imagens Geradas por Inteligência Artificial: Um Estudo sobre Técnicas e Desafios

Uberlândia, Brasil

2025

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Otávio Malta Borges

Detecção de Imagens Geradas por Inteligência Artificial: Um Estudo sobre Técnicas e Desafios

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Sistemas de Informação.

Orientador: David Pereira de Araújo

Coorientador: Henrique Coelho Fernandes

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2025

Otávio Malta Borges

Detecção de Imagens Geradas por Inteligência Artificial: Um Estudo sobre Técnicas e Desafios

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Sistemas de Informação.

David Pereira de Araújo
Orientador

Professor

Professor

Uberlândia, Brasil
2025

Dedico este trabalho à minha família e aos meus amigos, pelo apoio, incentivo e presença em todos os momentos desta caminhada.

Per aspera ad astra.

Resumo

Palavras-chave: detecção de imagens sintéticas, revisão sistemática da literatura, *deepfake*, inteligência artificial, CNNs.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ACM	Association for Computing Machinery
ADM	Denoising Diffusion Implicit Model
AI	Artificial Intelligence
AP	Average Precision
AUROC	Area Under the Receiver Operating Characteristic Curve
AUC	Area Under the Curve
CASIA	Chinese Academy of Sciences Image Dataset
CelebA	CelebFaces Attributes Dataset
CLIP	Contrastive Language Image Pretraining
CNN	Convolutional Neural Network
COCO	Common Objects in Context
DCGAN	Deep Convolutional Generative Adversarial Network
DDPM	Denoising Diffusion Probabilistic Model
DIRE	Diffusion Reconstruction Error
FFHQ	Flickr-Faces-HQ
GAN	Generative Adversarial Network
IA	Inteligência Artificial
IEEE	Institute of Electrical and Electronics Engineers
LDM	Latent Diffusion Model
LSUN	Large-scale Scene Understanding Dataset
NIST16	National Institute of Standards and Technology 2016 Dataset
PICOC	Population, Intervention, Comparison, Outcomes, Context
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses

RAISE	Raw Images Dataset
RGB	Red, Green, Blue
ROC	Receiver Operating Characteristic
RSL	Revisão Sistemática da Literatura
SIDBench	Synthetic Image Detection Benchmark
SRM	Steganalysis Rich Model
UFU	Universidade Federal de Uberlândia
VAE	Variational Autoencoder
ViT	Vision Transformer
VGG	Visual Geometry Group

Sumário

1 Introdução

A evolução constante das tecnologias de Inteligência Artificial (IA) nos últimos anos tornou possível a geração de imagens extremamente realistas, tornando cada vez mais difícil diferenciá-las daquelas tiradas de aparelhos convencionais ou criadas por seres humanos. O uso dessa funcionalidade tem se expandido em diversas áreas da sociedade, como no entretenimento, moda e design. Porém, também têm sido utilizadas para fins maliciosos, como a disseminação de informações falsas e *deepfakes*, além de comprometer a confiabilidade de provas judiciais.

Segundo (NAKANISHI, 2023), o crescimento dos *deepfakes* e o avanço das tecnologias de geração de imagens estão tornando cada vez mais difícil distinguir um conteúdo autêntico de um artificial, o que aumenta a probabilidade de fraudes, golpes financeiros e perda de reputação. Aliado a isso, a desinformação se espalha pelas redes sociais rapidamente, ampliando a disseminação de informações falsas e comprometendo a credibilidade de veículos de comunicação. Este debate afeta também outras áreas, como no âmbito jurídico, a manipulação de evidências visuais levanta questionamentos sobre a confiabilidade de provas apresentadas em processos judiciais, o que exige o desenvolvimento de métodos robustos de verificação da autenticidade dessas mídias. Na área de segurança digital, a incapacidade de distinguir uma face verdadeira de uma face artificial é uma ameaça para sistemas que utilizam a autenticação com base em biometria facial.

Diversos estudos têm explorado o problema da detecção de imagens geradas por IA nos últimos anos, a fim de aprimorar os métodos e técnicas utilizados nesse processo. É ressaltada por (DAMASCENO, 2024) a importância de desenvolver mecanismos eficientes para a moderação de conteúdos digitais, alertando sobre o uso de tecnologias de IA na propagação de informações falsas e *deepfakes*. Por outro lado, (SALVO, 2020) explora o uso da inteligência artificial no meio jurídico, destacando os obstáculos na validação de evidências visuais e os riscos que isso representa para a integridade dos processos judiciais. Esses estudos deixam claro que a crescente sofisticação das imagens sintéticas impõe desafios significativos, tornando essencial o avanço de métodos que garantam maior confiabilidade no ambiente digital.

O objetivo geral deste trabalho é investigar e analisar as principais técnicas utilizadas para a detecção de imagens geradas por inteligência artificial, avaliando seus desafios e limitações. Para isso, será realizado um estudo comparativo entre diferentes abordagens. Além do enfoque técnico, será apresentada uma breve discussão sobre o impacto social e ético dessas tecnologias, de modo a contextualizar sua influência na sociedade contemporânea.

2 Fundamentação Teórica

2.1 *Deepfake*

A evolução da inteligência artificial nos últimos anos possibilitou avanços significativos na geração e manipulação de imagens. O termo *deepfake* refere-se à criação de imagens e vídeos sintéticos altamente realistas, geralmente utilizando redes neurais profundas (*Deep Neural Networks* – DNNs) para modificar a aparência ou identidade de indivíduos. Segundo (BOUGUEFFA et al., 2024), técnicas baseadas em redes generativas adversariais (*Generative Adversarial Networks* - GANs), *Transformers* e modelos de difusão (*Diffusion Models*) são amplamente utilizadas para criar conteúdos digitais manipulados, impactando setores como entretenimento e segurança digital.

Os primeiros *deepfakes* eram gerados por autoencoders, redes neurais treinadas para comprimir e reconstruir dados. No entanto, a introdução das redes generativas adversariais revolucionou a área, permitindo a criação de imagens muito mais realistas. Entre os principais modelos, destacam-se:

- DCGAN (Deep Convolutional GAN): Foi a primeira arquitetura a utilizar redes convolucionais profundas para melhorar a qualidade das imagens;
- StyleGAN: Foi desenvolvido pela NVIDIA, possibilita maior controle sobre a aparência das imagens geradas, como idade, gênero e expressões faciais;
- StyleGAN2 e StyleGAN3: Melhoraram a qualidade das imagens, eliminando anomalias e inconsistências;
- GigaGAN (2023): É um dos modelos mais recentes, combinando GANs e *Transformers* para gerar imagens de alta fidelidade (KANG et al., 2023);

Além das GANs, os modelos de difusão surgiram como uma alternativa para a geração de imagens. Modelos como DALL · E e MidJourney superaram as GANs em diversas aplicações, gerando imagens de qualidade. De acordo com Cantanhede (2024), esses modelos de difusão "produzem amostras de maior qualidade e são mais fáceis de dimensionar e controlar".

As aplicações dos deepfakes abrangem desde o entretenimento e a criação artística até investigações em ambientes clínicos. No contexto médico, esse tipo de geração sintética é empregado para ampliar bases de dados limitadas, preservar a privacidade dos pacientes e simular condições raras ou pouco representadas, de modo a treinar modelos de IA

com maior variedade e robustez. Por exemplo, estudos mostram que imagens artificiais geradas por redes generativas podem servir como ferramenta de anonimização e aumentar a diversidade de dados sem comprometer a identidade real dos pacientes (**????**). No entanto, esse mesmo poder técnico também abre caminho para usos maliciosos, como a disseminação de informações falsas, fraudes digitais e ataques de identidade.

2.2 Redes Neurais Convolucionais

As redes neurais convolucionais (*Convolutional Neural Networks* – CNNs) são um dos principais modelos utilizados para a análise e classificação de imagens. Foram baseadas no córtex visual dos mamíferos e processam dados visuais, extraindo características hierárquicas, como bordas e texturas. Segundo (CALIS, 2018), as CNNs demonstram grande eficácia na extração de características visuais complexas, permitindo seu uso em sistemas de reconhecimento de padrões .

O conceito de CNNs apareceu na década de 1980 com os estudos de Fukushima. Em 1998, LeCun desenvolveu o LeNet-5, uma das primeiras arquiteturas de CNN que obteve sucesso. Em 2012, o modelo AlexNet demonstrou a efetividade das CNNs em atividades de reconhecimento de imagens. Desde então, estruturas como VGG, ResNet e EfficientNet forneceram avanços significativos, diminuindo a complexidade computacional e elevando a precisão.

Nos últimos anos, os Vision Transformers (ViTs) têm se consolidado como uma alternativa promissora às CNNs na análise de imagens. Como aponta (HAN et al., 2022), esses modelos apresentam desempenho comparável ou superior em diversos *benchmarks* visuais, chamando a atenção da comunidade científica.

Todavia, as CNNs continuam sendo uma função essencial na identificação de *deepfakes* pois conseguem identificar características sutis deixadas por GANs, como inconsistências em bordas, sombras e texturas. Avanços recentes incluem a combinação de CNNs com ViTs e modelos híbridos. Atualmente, as CNNs continuam sendo fundamentais para muitas aplicações de visão computacional, mas estão sendo superadas pelos Transformers em algumas tarefas, como reconhecimento de ações.

As CNNs são amplamente utilizadas em aplicações como reconhecimento facial, diagnóstico médico e detecção de objetos em imagens. Em relação às *deepfakes*, CNNs são usadas para identificar padrões artificiais em imagens geradas por IA, auxiliando na segurança digital e na prevenção de fraudes.

2.3 *Transformers*

Os *Transformers* constituem uma arquitetura de aprendizado profundo originalmente proposta para o processamento de linguagem natural, mas que rapidamente passou a ser aplicada também em visão computacional. Sua principal inovação é o mecanismo de *self-attention,* que permite ao modelo capturar relações de longo alcance entre diferentes partes de uma sequência ou de uma imagem.

Na área de visão computacional, os *Vision Transformers* dividem a imagem em pequenas regiões denominadas *patches*, que são tratadas como sequências, de forma semelhante às palavras em um texto. Isso possibilita a identificação de padrões globais e dependências contextuais que, muitas vezes, não são capturados de forma eficiente por redes convolucionais tradicionais.

2.4 Métricas de Avaliação e Parâmetros de Treinamento

As métricas de avaliação são fundamentais para medir a eficiência e a confiabilidade dos modelos de aprendizado de máquina, permitindo a comparação entre diferentes abordagens e garantindo que o modelo selecionado seja adequado ao problema em questão.

Conforme (PINHEIRO; GADOTTI; BERNARDY, 2021), a escolha das métricas de avaliação é essencial para assegurar um desempenho confiável, especialmente em dados desbalanceados ou problemas complexos, como classificação médica e detecção de fraudes.

A avaliação do desempenho dos detectores geralmente começa com a construção da matriz de confusão, que organiza os resultados das previsões do modelo em quatro categorias:

- VP (Verdadeiros Positivos): casos corretamente classificados como positivos;
- VN (Verdadeiros Negativos): casos corretamente classificados como negativos;
- FP (Falsos Positivos): casos negativos incorretamente classificados como positivos;
- FN (Falsos Negativos): casos positivos incorretamente classificados como negativos.

Tabela 1 – Matriz de confusão

	Classe Predita Positiva	**Classe Predita Negativa**
Classe Real Positiva	VP (Verdadeiro Positivo)	FN (Falso Negativo)
Classe Real Negativa	FP (Falso Positivo)	VN (Verdadeiro Negativo)

Fonte: Adaptado de (BIOINFO, 2023).

A partir dessa matriz, derivam-se diversas métricas de desempenho.

- Acurácia – mede a proporção de previsões corretas em relação ao total de previsões:

$$acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.1}$$

- Precisão – indica a proporção de previsões positivas que são realmente corretas:

$$precisão = \frac{VP}{VP + FP} \tag{2.2}$$

Revocação (Sensibilidade) – mede a capacidade do modelo de identificar corretamente os exemplos positivos:

$$sensibilidade = \frac{VP}{VP + FN} \tag{2.3}$$

- F1-Score – combina precisão e revocação em uma única métrica harmônica, útil em conjuntos de dados desbalanceados:

$$F1 = 2 \times \frac{precisão \times sensibilidade}{precisão + sensibilidade} \tag{2.4}$$

- Área Sob a Curva ROC (AUROC) – corresponde à área sob a curva que relaciona a taxa de verdadeiros positivos (TPR) e a taxa de falsos positivos (FPR). Quanto maior a AUROC, melhor a capacidade do modelo de separar corretamente as classes positivas e negativas.

$$TPR = \frac{VP}{VP + FN}, FPR = \frac{FP}{FP + VN}, AUROC = \int_0^1 TPR(FPR)\, d(FPR) \tag{2.5}$$

- Função de Perda (*Loss*) – não é uma métrica de avaliação, mas uma medida interna utilizada durante o treinamento. A função de perda quantifica o erro entre as previsões do modelo e os valores reais, guiando o ajuste dos parâmetros. Valores menores indicam melhor aprendizagem e convergência, enquanto valores altos sugerem maior discrepância entre previsão e verdade.

Durante a revisão, observou-se que diversos estudos utilizam métricas adicionais além das tradicionais. Em alguns trabalhos, os autores empregam métricas já consolidadas na literatura, como precisão média, acurácia balanceada ou similaridade cosseno. Entretanto, alguns trabalhos adotam métricas definidas especificamente para seus cenários experimentais , como taxa de ataques bem-sucedidos, acurácia GAN, acurácia entre detecções cruzadas, pontuação média de confiança ou velocidade de detecção. Essa variedade evidencia que diferentes estudos priorizam aspectos distintos, como robustez, qualidade visual, resistência a ataques ou desempenho em tempo real, o que explica a diversidade de indicadores encontrados.

As métricas de avaliação são amplamente aplicadas em visão computacional, incluindo detecção de fraudes bancárias, diagnóstico médico e verificação de autenticidade de imagens. No contexto dos *deepfakes*, essas métricas são essenciais para medir a eficácia dos detectores na identificação de imagens manipuladas, garantindo maior segurança e confiabilidade nos resultados.

3 Trabalhos Relacionados

Um dos trabalhos mais relevantes dentro da temática desta pesquisa é o estudo *FaceForensics++: Learning to Detect Manipulated Facial Images* (RöSSLER et al., 2019), com o objetivo de investigar métodos eficazes para detectar alterações faciais em vídeos manipulados, com foco em técnicas como Face2Face, DeepFakes, FaceSwap e NeuralTextures. Para isso, os pesquisadores construíram uma extensa base de dados, composta por milhares de vídeos adulterados em diferentes níveis de compressão, aspecto que se revelou fundamental para o desempenho dos modelos testados. Entre os algoritmos avaliados, o XceptionNet apresentou os melhores resultados em vídeos de alta qualidade, mas teve desempenho significativamente inferior em conteúdos com compressão agressiva. Esse contraste evidencia que, na prática, a eficácia das ferramentas de detecção ainda depende fortemente das condições do material analisado. Assim, o estudo de Rössler et al. (2019) é particularmente relevante para este trabalho, pois demonstra as limitações dos detectores atuais e reforça a importância de desenvolver soluções mais robustas para ambientes reais.

Complementando essa perspectiva, o artigo *Media Forensics and DeepFakes: An Overview* (VERDOLIVA, 2020) oferece uma análise crítica e abrangente sobre o campo da forense digital, especialmente no que diz respeito à detecção de *deepfakes*. A autora inicia contextualizando o crescimento acelerado das tecnologias de geração de mídia sintética e os impactos associados a esse avanço, abrangendo aspectos sociais, políticos, éticos e tecnológicos. Em termos metodológicos, Verdoliva (2020) compara duas abordagens distintas: as tradicionais, que exploram inconsistências físicas na imagem (como falhas de iluminação e artefatos de compressão), e as modernas, baseadas em aprendizado profundo. Um dos principais pontos levantados é a crescente sofisticação dos métodos de manipulação, que torna sua detecção cada vez mais desafiadora. Dessa forma, o artigo contribui de maneira significativa para esta pesquisa ao apresentar um panorama atualizado do estado da arte e ao apontar lacunas e direções promissoras para investigações futuras no campo.

Mais recentemente, o artigo *Deepfake Generation and Detection: Case Study and Challenges* (PATEL et al., 2023), apresenta uma revisão abrangente das técnicas de geração e detecção de *deepfakes*, com foco em abordagens multimodais que abrangem imagem, vídeo e áudio. Os autores descrevem de forma detalhada os principais modelos utilizados na criação de *deepfakes*, como autoencoders, GANs e suas variantes (StyleGAN, CycleGAN, StarGAN, entre outros), e exploram diferentes ferramentas populares de geração de conteúdo sintético. No campo da detecção, o estudo organiza soluções baseadas em características físicas, sinal digital e modelos baseados em aprendizado profundo, destacando arquiteturas como XceptionNet, FaceForensics++ e FakeSpotter. Um diferencial do trabalho é o estudo de caso IBMM, que analisa inconsistências entre diferentes tipos

de dados em sistemas de detecção. A pesquisa também aponta desafios enfrentados na implementação de soluções robustas, como a necessidade de generalização dos modelos. Por sua abrangência e atualidade, o estudo é altamente relevante para este projeto.

Por fim, o trabalho *Learning Rich Features for Image Manipulation Detection* (ZHOU et al., 2018) propõe uma abordagem específica para a detecção de manipulações em imagens digitais, como splicing, copy-move e removal. O método desenvolvido pelos autores é uma arquitetura de rede convolucional de duas vias, inspirada no modelo Faster R-CNN, que combina características extraídas da imagem RGB com informações de ruído geradas a partir de filtros do tipo steganalysis rich model (SRM). Essa combinação permite que o modelo identifique tanto alterações visíveis quanto inconsistências mais sutis. A eficácia do método foi comprovada por meio de testes em quatro bases de dados amplamente utilizadas: CASIA, Columbia, NIST16 e COVER, mesmo sob condições adversas como compressão e redimensionamento da imagem. Embora não tenha sido desenvolvido com foco direto em *deepfakes*, o estudo de Zhou et al. (2018) é relevante para este projeto, pois demonstra que abordagens que integram sinais de baixa e alta complexidade têm potencial para detectar manipulações visuais refinadas, como as produzidas por IAs.

4 Métodos

Este capítulo apresenta a metodologia adotada para a realização deste trabalho, que consiste em uma revisão sistemática da literatura (RSL) sobre técnicas e desafios na detecção de imagens geradas por inteligência artificial. A RSL foi conduzida seguindo as diretrizes de (KITCHENHAM; CHARTERS et al., 2007) e o protocolo PRISMA (PAGE et al., 2021), garantindo rigor metodológico e transparência em todas as etapas.

4.1 Revisão Sistemática da Literatura

A revisão sistemática da literatura é um método de pesquisa que visa identificar, avaliar e interpretar todas as pesquisas relevantes disponíveis sobre uma questão específica, área temática ou fenômeno de interesse (KITCHENHAM; CHARTERS et al., 2007). Diferente de revisões narrativas tradicionais, a RSL segue um protocolo rigoroso, minimizando vieses e garantindo a reprodutibilidade do processo. A escolha da RSL como metodologia para este trabalho justifica-se pela necessidade de uma análise abrangente e sistemática do estado da arte em detecção de imagens geradas por IA, um campo em rápida evolução. A RSL permite não apenas mapear as técnicas existentes, mas também identificar tendências emergentes, lacunas de pesquisa e desafios persistentes.

4.2 Questões de Pesquisa

O primeiro passo na condução da RSL foi a definição das questões de pesquisa (RQs), que orientaram todo o processo subsequente. As questões foram formuladas seguindo o modelo PICOC (População, Intervenção, Comparação, Resultados e Contexto), conforme recomendado por (KITCHENHAM; CHARTERS et al., 2007):

- RQ1: Quais são as principais técnicas utilizadas para detecção de imagens geradas por IA no período 2022-2025?
- RQ2: Quais modelos geradores de imagem são mais frequentemente alvo dos métodos de detecção?
- RQ3: Quais *datasets* (conjunto de dados) são utilizados para treinamento e avaliação dos métodos de detecção?
- RQ4: Quais métricas são empregadas para avaliar a eficácia dos métodos de detecção?

- RQ5: Quais são os principais desafios e limitações reportados na literatura recente sobre detecção de imagens geradas por IA?

Estas questões foram elaboradas para cobrir aspectos técnicos (RQ1, RQ2), metodológicos (RQ3, RQ4) e desafios (RQ5) relacionados à detecção de imagens geradas por IA, fornecendo uma visão abrangente do estado da arte no período analisado.

4.3 Protocolo da Revisão

O protocolo da RSL foi desenvolvido seguindo as diretrizes de (KITCHENHAM; CHARTERS et al., 2007) e documentado conforme o checklist PRISMA (PAGE et al., 2021). O protocolo especifica os métodos utilizados para identificar, selecionar e avaliar estudos relevantes, bem como para extrair e sintetizar dados.

4.3.1 Fontes de Dados

Para garantir uma cobertura abrangente da literatura relevante, foram selecionadas três bases de dados científicas reconhecidas na área de computação e processamento de imagens:

- IEEE Xplore: Base de dados da IEEE (Institute of Electrical and Electronics Engineers), que cobre publicações em engenharia elétrica, ciência da computação e eletrônica.
- ACM Digital Library: Biblioteca digital da ACM (Association for Computing Machinery), com foco em publicações de ciência da computação.
- ScienceDirect: Base de dados que reúne artigos científicos revisados por pares em diversas áreas do conhecimento, com destaque para ciências exatas, biológicas, da saúde e engenharias.

As strings de busca para cada uma das bases de dados foram:

- IEEE Xplore: ((((("Document Title":"AI-generated image"OR "Document Title":"synthetic image"OR "Document Title":"GAN image"OR "Document Title":"diffusion model image"OR "Document Title":"generated image")) AND (("Document Title":"detection"OR "Document Title":"forensics"OR "Document Title":"authentication"OR "Document Title":"identification"OR "Document Title":"method"))))
- ACM Digital Library:AllField:((Title: "ai-generated image"OR Title: "synthetic image"OR Title: "gan image"OR Title: "diffusion model image"OR Title: "generated image")

AND (Title: "detection"OR Title: "forensics"OR Title: "authentication"OR Title: "identification"))

- ScienceDirect: ("AI-generated image"OR "synthetic image") AND ("method"OR "detection"OR "forensics"OR "authentication"OR "identification") e ("GAN image"OR "diffusion model image"OR "generated image") AND ("method"OR "detection"OR "forensics"OR "authentication"OR "identification")

Estas bases foram escolhidas por sua relevância e qualidade das publicações indexadas nas áreas de visão computacional, processamento de imagens e inteligência artificial.

4.3.2 Estratégia de Busca

A estratégia de busca foi desenvolvida para maximizar a sensibilidade e a especificidade, ou seja, a capacidade de recuperar estudos relevantes e de excluir irrelevantes. Foram utilizadas strings de busca específicas para cada base de dados, adaptadas à sua sintaxe particular, mas mantendo a equivalência semântica entre elas. As strings de busca foram construídas combinando termos relacionados a:

1. Imagens geradas por IA (ex: "AI-generated image", "synthetic image", "GAN image")
2. Detecção e análise forense (ex: "detection", "forensics", "authentication")

A busca foi restrita a publicações do período de janeiro de 2022 a julho de 2025, escritas em inglês, e incluindo apenas artigos de periódicos, conferências e revisões.

4.3.3 Critérios de Inclusão e Exclusão

Os critérios de inclusão e exclusão foram definidos para garantir que apenas estudos relevantes para as questões de pesquisa fossem incluídos na revisão.

Critérios de inclusão (CI):

- CI1: O estudo foi publicado entre janeiro de 2022 e junho de 2025
- CI2: O estudo está escrito em inglês.
- CI3: O estudo é uma publicação primária revisada por pares (artigo de periódico ou conferência completa) ou uma revisão sistemática relevante que aborde diretamente as RQs.
- CI4: O estudo foca primariamente na detecção ou análise forense de imagens geradas ou manipuladas por IA.

- CI5: O estudo apresenta ou avalia técnicas, metodologias, *datasets*, métricas ou discute desafios/limitações relacionados à detecção.

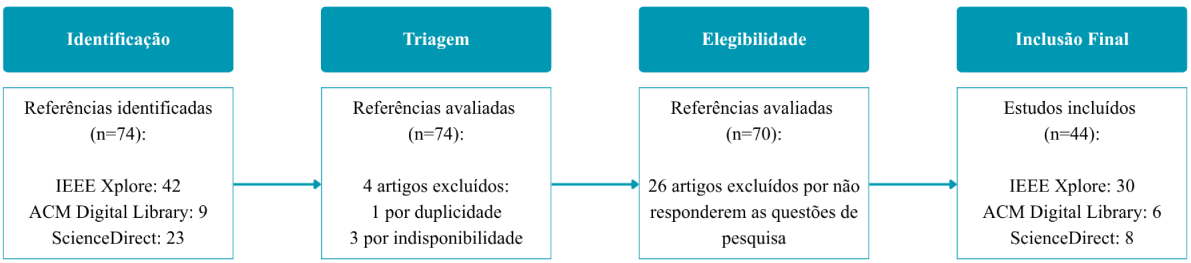
Critérios de exclusão (CE):

- CE1: Foco principal na geração de imagens e não na detecção.
- CE2: Foco exclusivo em vídeo ou áudio *deepfake*, sem abordagem significativa em imagens estáticas.
- CE3: Foco em manipulações que não envolvem IA (ex: edições manuais simples sem uso de redes neurais).
- CE4: Artigos duplicados (a versão mais completa ou final será mantida).
- CE5: Artigo indisponível em texto completo após esforços razoáveis para obtê-lo.

4.3.4 Processo de Seleção dos Estudos

A seleção dos estudos seguiu as diretrizes da revisão sistemática da literatura, conforme o protocolo metodológico definido, baseado nas recomendações de (KITCHENHAM; CHARTERS et al., 2007). Para garantir transparência e reprodutibilidade no processo de filtragem, adotou-se também o modelo PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), que estabelece um conjunto de princípios para relatar de forma clara as etapas de identificação, triagem, avaliação da elegibilidade e inclusão final dos estudos. Essas etapas estão representadas na Figura 1, elaborada com base no modelo proposto por (PAGE et al., 2021).

Figura 1 – Fluxograma do Processo de Identificação e Seleção dos Estudos segundo o modelo PRISMA.



Fonte: Elaboração própria (2025).

Foram inicialmente identificados 74 registros por meio de buscas nas bases *IEEE Xplore*, *ACM Digital Library* e *ScienceDirect*. Após a remoção de 1 duplicado e a

exclusão de 3 artigos devido à indisponibilidade do texto completo, restaram 70 estudos para leitura na íntegra.

A triagem foi realizada com base nos títulos, resumos e posteriormente nos textos completos. Foram excluídos os artigos que não abordavam diretamente a detecção de imagens geradas por inteligência artificial ou que não apresentavam contribuições metodológicas relevantes. Ao final do processo, 44 artigos foram incluídos na análise final, listados na tabela a seguir.

Tabela 2 – Lista de Artigos Incluídos

Artigo	**Título**
(SONG et al., 2024)	ACM Multimedia 2024 Grand Challenge Report for Artificial Intelligence Generated Image Detection
(HERUR et al., 2025)	Addressing Diffusion Model Based Counter-Forensic Image Manipulation for Synthetic Image Detection
(HOSSAIN; ZAMAN; ISLAM, 2023)	Advancing AI-Generated Image Detection: Enhanced Accuracy through CNN and Vision Transformer Models with Explainable AI Insights
(XI et al., 2023)	AI-Generated Image Detection using a Cross-Attention Enhanced Dual-Stream Network
(CHEN et al., 2025)	AI-Generated Image Detection Using Semantic Feature
(LYU et al., 2024)	AI-Generated Image Detection With Wasserstein Distance Compression and Dynamic Aggregation
(MENG et al., 2024)	Artifact feature purification for cross-domain detection of AI-generated image
(RAHMAN et al., 2023)	Artifact: A Large-Scale Dataset With Artificial And Factual Images For Generalizable And Robust Synthetic Image Detection
(XUE; JI; LI, 2024)	Computer-Generated Image Detection Based on Multi-Scale Feature Fusion Attention Module
(TAN et al., 2025)	DC-BiNet: Towards interpretable generated image detection with dark channel prior
(SINITSA; FRIED, 2024)	Deep Image Fingerprint: Towards Low Budget Synthetic Image Detection and Model Lineage Analysis
(BYEON et al., 2024)	Deep learning model to detect deceptive generative adversarial network generated images using multimedia forensic
(JAGADISH; JASMINE, 2024)	Detection of AI-Generated Image Content in News and Journalism
(LIU et al., 2023)	Detection of GAN generated image using color gradient representation
(WANG et al., 2023b)	DIRE for Diffusion-Generated Image Detection
(ROUT; MISHRA, 2025)	Enhanced CNN Architecture with Residual Blocks and Regularization for AI-Generated Image Detection
(WEIR et al., 2024)	Enhancing AI-Generated Image Detection with a Novel Approach and Comparative Analysis
(LIN; SHANG; GAO, 2023)	Enhancing Interpretability in AI-Generated Image Detection with Genetic Programming
(JAVAHERI; MOTAMEDNIA; MAHMOUDI-AZANVEH, 2024)	Enhancing the Generalization of Synthetic Image Detection Models through the Exploration of Features in Deep Detection Models
(CHAN et al., 2024)	Evasion on general GAN-generated image detection by disentangled representation
(ROSA et al., 2024)	Exploring the Adversarial Robustness of CLIP for AI-generated Image Detection
(XU et al., 2025)	FAMSeC: A Few-Shot-Sample-Based General AI-Generated Image Detection Method
(LIU et al., 2024)	Forgery-aware Adaptive Transformer for Generalizable Synthetic Image Detection
(WANG et al., 2023a)	General GAN-generated Image Detection by Data Augmentation in Fingerprint Domain
(RAJ; MATHEW; MONDAL, 2024)	Generalized and robust model for GAN-generated image detection
(ZHANG et al., 2022)	Improving GAN-Generated Image Detection Generalization Using Unsupervised Domain Adaptation
(LI et al., 2025)	Improving Synthetic Image Detection Towards Generalization: An Image Transformation Perspective
(DOGOULIS et al., 2023)	Improving Synthetically Generated Image Detection in Cross-Concept Setting

Tabela 2 – Lista de Artigos Incluídos (continuação)

Artigo	**Título**
(LUO et al., 2024)	LaRE2: Latent Reconstruction Error Based Method for Diffusion-Generated Image Detection
(GHOSH; NASKAR, 2023)	Leveraging Image Gradients for Robust GAN-Generated Image Detection in OSN context
(XU et al., 2024)	MDTL-NET: Computer-generated image detection based on multi-scale deep texture learning
(GUAN et al., 2025)	Noise-Informed Diffusion-Generated Image Detection With Anomaly Attention
(PARK; NA; CHOI, 2024)	Performance Comparison and Visualization of AI-Generated-Image Detection Methods
(S; R, 2023)	Performance Comparison of Deep Learning Models for Computer Generated Image Detection
(WU; LI, 2024)	Photo Response Non-Uniformity Based AI-Generated Image Detection
(COZZOLINO et al., 2024)	Raising the Bar of AI-generated Image Detection with CLIP
(GYE et al., 2025)	Reducing the Content Bias for AI-generated Image Detection
(JEONG et al., 2024)	Self-supervised scheme for generalizing GAN image detection
(YU et al., 2024)	SemGIR: Semantic-Guided Image Regeneration Based Method for AI-generated Image Detection and Attribution
(SCHINAS; PAPADOPOULOS, 2024)	SIDBench: A Python framework for reliably assessing synthetic image detection method
(COZZOLINO et al., 2023)	Synthetic Image Detection: Highlights from the IEEE Video and Image Processing Cup 2022 Student Competition [SP Competitions]
(KONSTANTINIDOU; KOUTLIS; PAPADOPOULOS, 2025)	TextureCrop: Enhancing Synthetic Image Detection Through Texture-Based Cropping
(GANGAN; KADAN; L., 2025)	Toward Exploring Fairness in Visual Transformer Based Natural and GAN Image Detection Systems
(GUPTA; SHRENETER; SEHGAL, 2024)	Visual Veracity: Advancing AI-Generated Image Detection with Convolutional Neural Networks

4.3.5 Extração e Síntese de Dados:

Para cada estudo incluído, foram extraídas as seguintes informações:

- Título
- Foco Principal do Estudo
- Técnicas de detecção propostas ou avaliadas
- *Datasets* utilizados
- Métricas de avaliação adotadas
- Resultados de desempenho
- Limitações identificadas
- Desafios abordados
- Ferramentas promissoras
- Implicações sociais

Os dados extraídos foram organizados em uma matriz de síntese, permitindo comparações sistemáticas entre diferentes abordagens e identificação de padrões, tendências e lacunas na literatura.

Para a contagem de frequência de arquiteturas base, foram agrupadas variantes de uma mesma família sob uma única categoria. Por exemplo, todas as versões da arquitetura ResNet (como ResNet18, ResNet34, ResNet50) foram contabilizadas como "ResNet". O mesmo critério foi adotado para outras famílias de arquiteturas, como EfficientNet, VGG e Inception.

5 Resultados

Este capítulo apresenta os resultados desta RSL. A análise integra dados quantitativos e qualitativos provenientes dos 44 artigos selecionados, conforme a metodologia detalhada no Capítulo 4, visando identificar tendências, lacunas e contribuições significativas neste campo em rápida evolução.

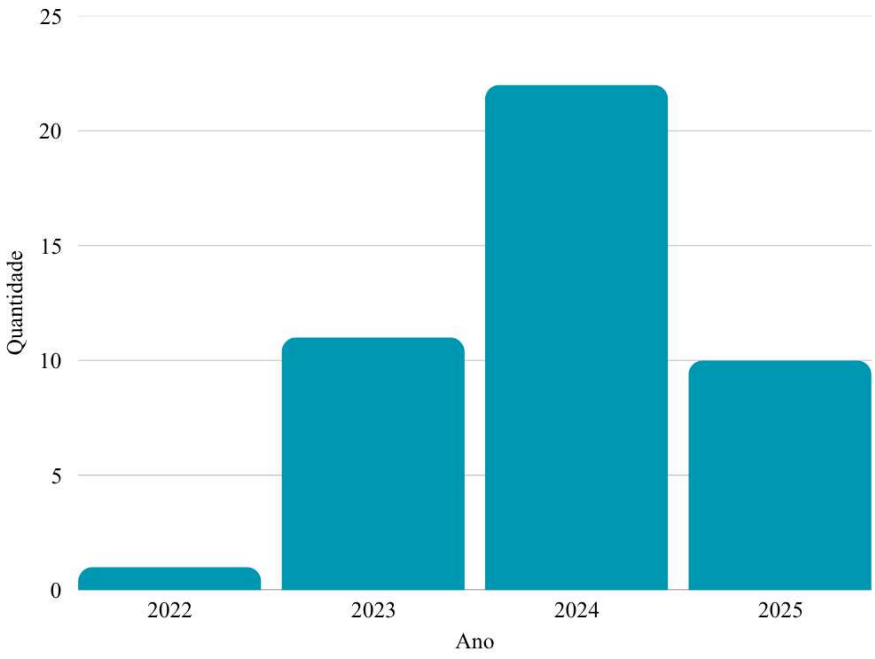
5.1 Análise Quantitativa

A análise quantitativa oferece uma perspectiva estruturada sobre a evolução e as tendências da pesquisa em detecção de imagens geradas por IA. Os dados foram categorizados para identificar padrões na distribuição temporal dos estudos e nas técnicas, modelos generativos e *datasets* mais citados.

5.1.1 Distribuição Temporal dos Estudos

O período abrangido por este estudo, de janeiro de 2022 a junho de 2025 , demonstra uma evolução notável na quantidade de publicações, refletindo o interesse e a urgência em abordar os desafios impostos pelos avanços nos modelos generativos. A Figura 2 ilustra a distribuição dos artigos ao longo dos anos, com base nos dados coletados:

Figura 2 – Distribuição temporal dos artigos por ano de publicação.



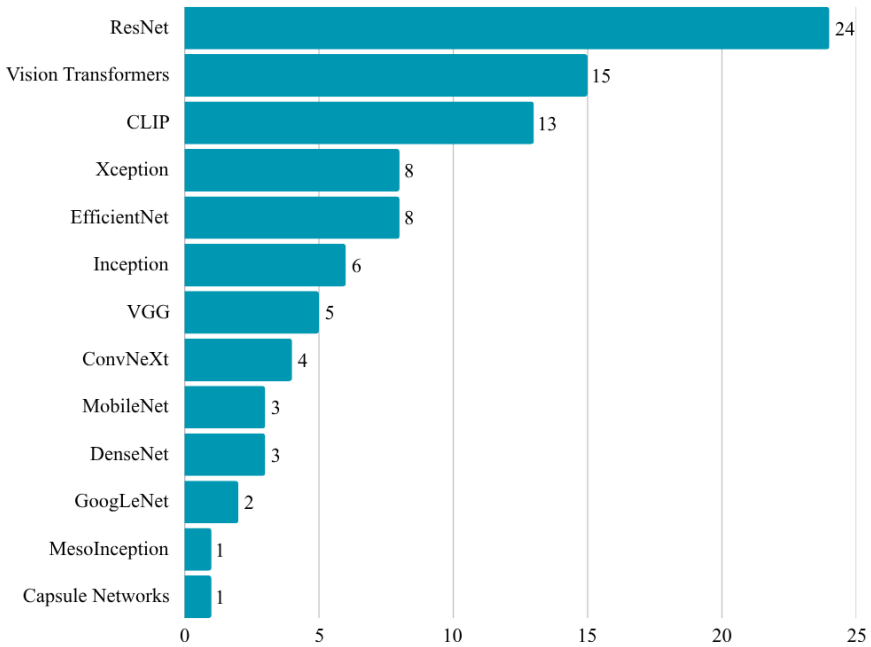
Fonte: Elaboração própria (2025).

Observa-se um aumento expressivo no número de publicações a partir de 2023, atingindo seu pico em 2024. Este crescimento acentuado sugere que a comunidade científica intensificou seus esforços na detecção de imagens geradas por IA, provavelmente em resposta à crescente sofisticação e acessibilidade de modelos generativos como o *Stable Diffusion* e o *MidJourney*, que se popularizaram amplamente a partir de 2022. O leve decréscimo observado em 2025 pode ser explicado pela limitação temporal da coleta, realizada ainda no primeiro semestre do ano. Assim, é possível que parte das publicações mais recentes ainda não tenha sido indexada nas bases consultadas no momento da análise. Essa tendência reflete a relevância e a dinamicidade da área, impulsionada pela necessidade de desenvolver métodos robustos e generalizáveis.

5.1.2 Arquiteturas Base

A análise das arquiteturas base (ou *backbones*) utilizadas nos métodos de detecção revela um predomínio de redes neurais convolucionais clássicas e modernas, assim como o crescimento do uso de Vision Transformers. Em termos gerais, a arquitetura base corresponde ao modelo fundamental de rede utilizado para extrair características da imagem, sobre o qual são construídas as etapas posteriores de detecção e classificação. A Figura 3 ilustra a distribuição das arquiteturas mais frequentemente empregadas nos artigos analisados.

Figura 3 – Distribuição das arquiteturas base dos métodos de detecção.



Fonte: Elaboração própria (2025).

A análise dos dados revela que as arquiteturas ResNet são as mais predominantes,

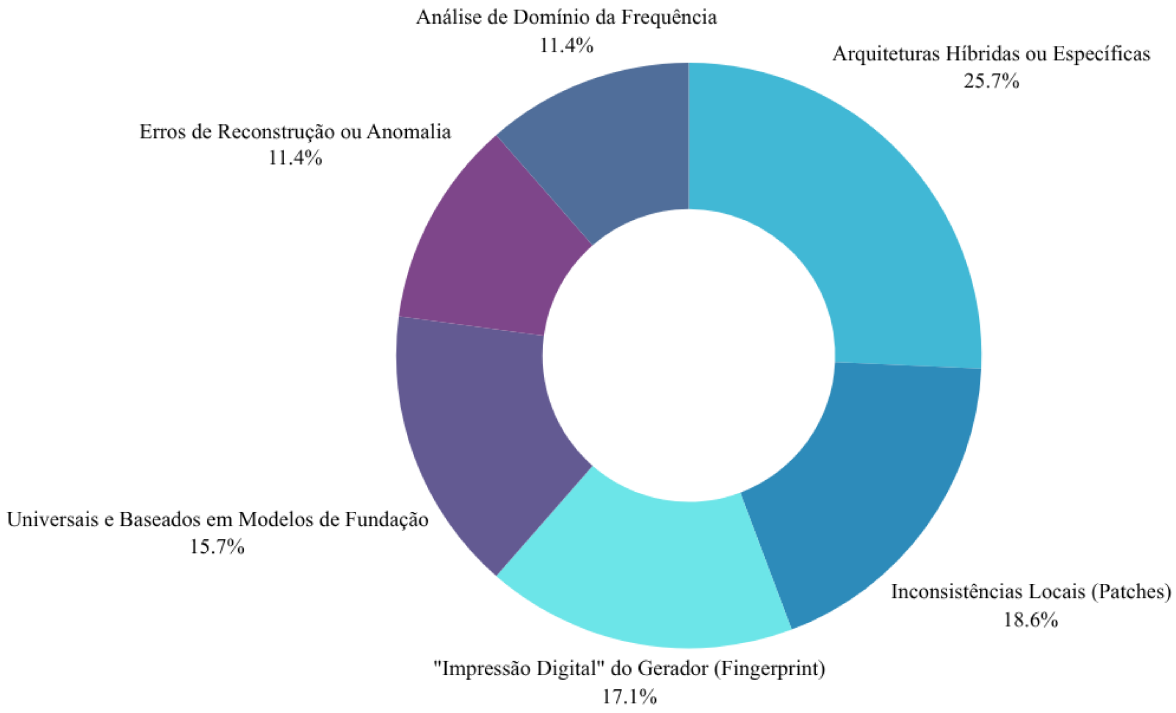
com 24 ocorrências. Sua robustez e desempenho comprovado em diversas tarefas de visão computacional as tornam uma escolha popular como *backbone* para detectores de imagens sintéticas. Modelos baseados em Transformers também demonstram uma presença significativa (15 ocorrências), refletindo a ascensão dessas arquiteturas em tarefas de processamento de imagem, especialmente pela sua capacidade de capturar dependências de longo alcance e padrões globais.

O CLIP, com 13 ocorrências, destaca-se como uma arquitetura que integra visão e linguagem, sendo cada vez mais utilizada em cenários de detecção por sua capacidade de generalização e compreensão contextual. Xception e EfficientNet, ambas com 8 ocorrências, são valorizadas por sua eficiência computacional e capacidade de extração de características detalhadas. Outras arquiteturas notáveis incluem Inception, VGG, ConvNeXt, MobileNet e DenseNet, que, embora com menor frequência, ainda contribuem para a diversidade metodológica na área. A variedade de arquiteturas empregadas sublinha a complexidade da tarefa de detecção e a busca contínua por modelos que possam efetivamente diferenciar entre imagens reais e sintéticas, adaptando-se às nuances introduzidas pelos diferentes modelos geradores de IA.

5.1.3 Modelos de Detecção

Foram identificados 36 modelos de detecção distintos, os quais foram classificados em seis categorias principais, conforme sua abordagem técnica predominante. A Figura 4 apresenta a distribuição percentual dessas categorias.

Figura 4 – Distribuição das categorias dos modelos de detecção.



Fonte: Elaboração própria (2025).

A categoria mais representativa é a de arquiteturas híbridas ou específicas, com 25,7% dos detectores analisados. Esses métodos geralmente combinam múltiplas estratégias ou redes neurais customizadas para capturar características profundas das imagens sintéticas. Detectores como LGrad, fusing e NPR se destacam nesse grupo por integrarem abordagens multiescalares, atencionais ou baseadas em aprendizado de explicabilidade.

Em seguida, estão os detectores focados em inconsistências locais (*patches*)) (18,6%), como o GramNet, que explora descontinuidades sutis ou artefatos em regiões específicas da imagem, refletindo variações regionais geradas durante o processo de síntese.

A categoria de "impressão digital" do gerador (*fingerprint*), com 17,1%, abrange métodos como CNNSpot e CNNDetect, os quais buscam identificar padrões recorrentes introduzidos por modelos generativos específicos, funcionando como assinaturas digitais.

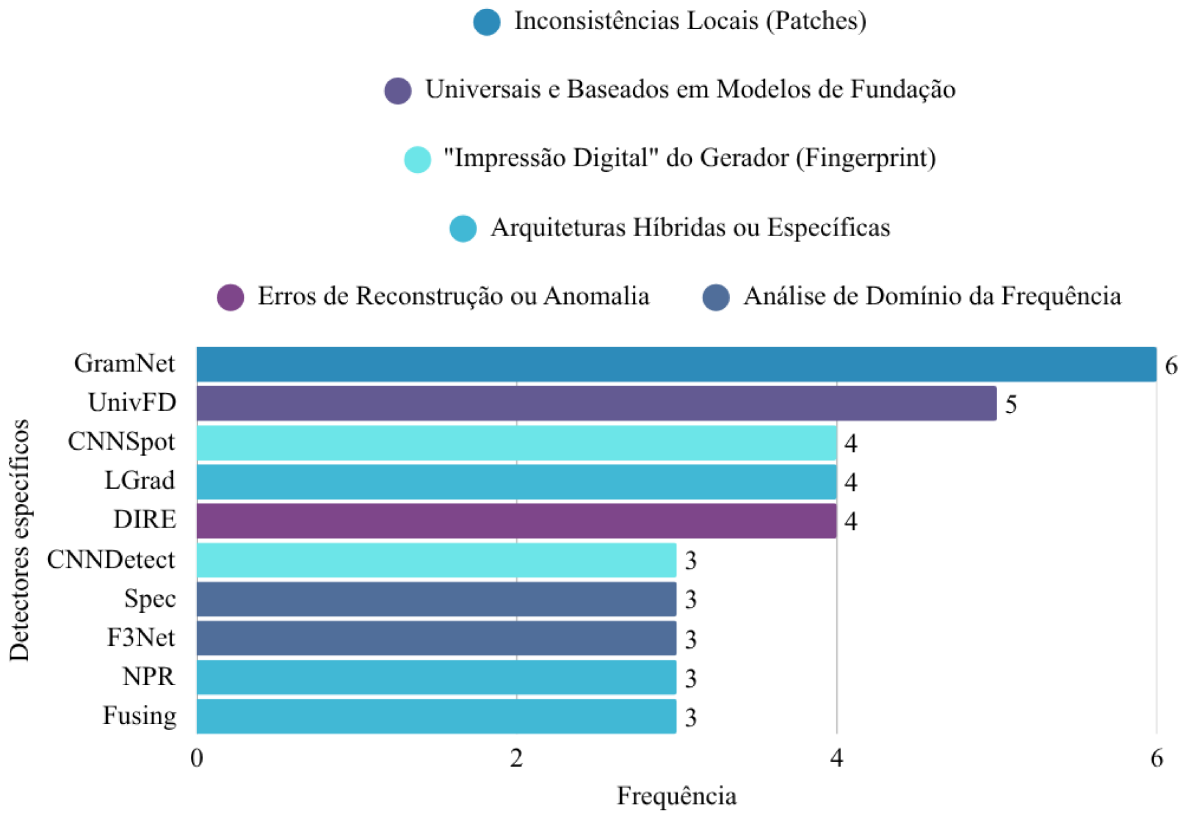
Detectores classificados como universais e baseados em modelos de fundação (15,7%) incluem propostas como o UnivFD, que visam identificar imagens sintéticas de forma generalizável, sem depender de um gerador específico. Tais abordagens são especialmente valiosas diante do crescimento e diversificação dos modelos de geração.

As categorias de erros de reconstrução ou anomalia e análise de domínio da frequência concentram 11,4% dos detectores cada. O primeiro grupo inclui métodos como DIRE, que buscam inconsistências geradas por falhas no processo de reconstrução da imagem

sintética. Já o segundo engloba detectores como F3Net (QIAN et al., 2020), que atuam sobre espectros de frequência para identificar padrões não naturais.

A Figura 5 detalha os dez detectores mais frequentes identificados nesta revisão.

Figura 5 – Os 10 detectores mais citados.

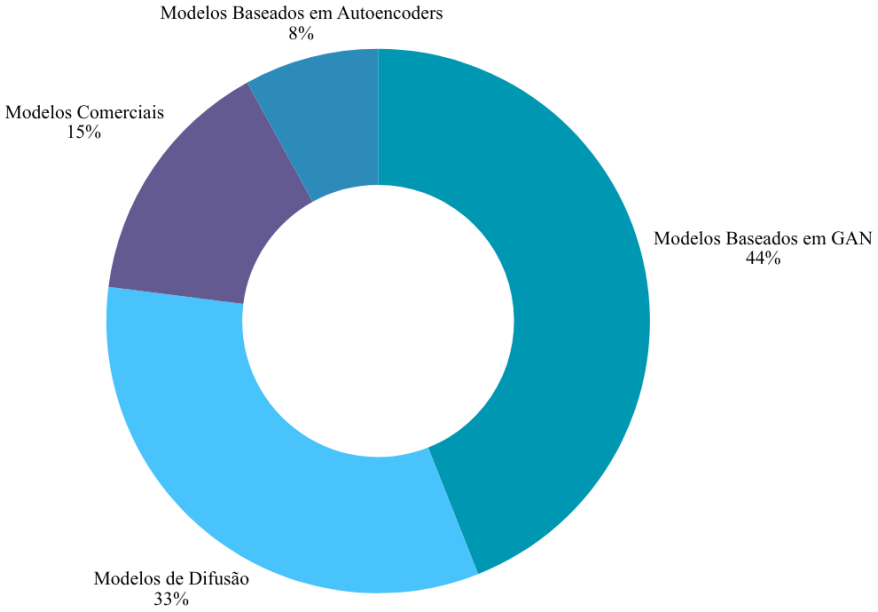


Fonte: Elaboração própria (2025).

5.1.4 Modelos Geradores de Imagem

Foram identificados 41 diferentes modelos geradores de imagens utilizados nos estudos analisados, os quais foram classificados em quatro grandes categorias, conforme a abordagem arquitetural: modelos baseados em GAN, modelos de difusão, modelos baseados em *autoencoders* e modelos comerciais. A Figura 6 apresenta a distribuição percentual dessas categorias.

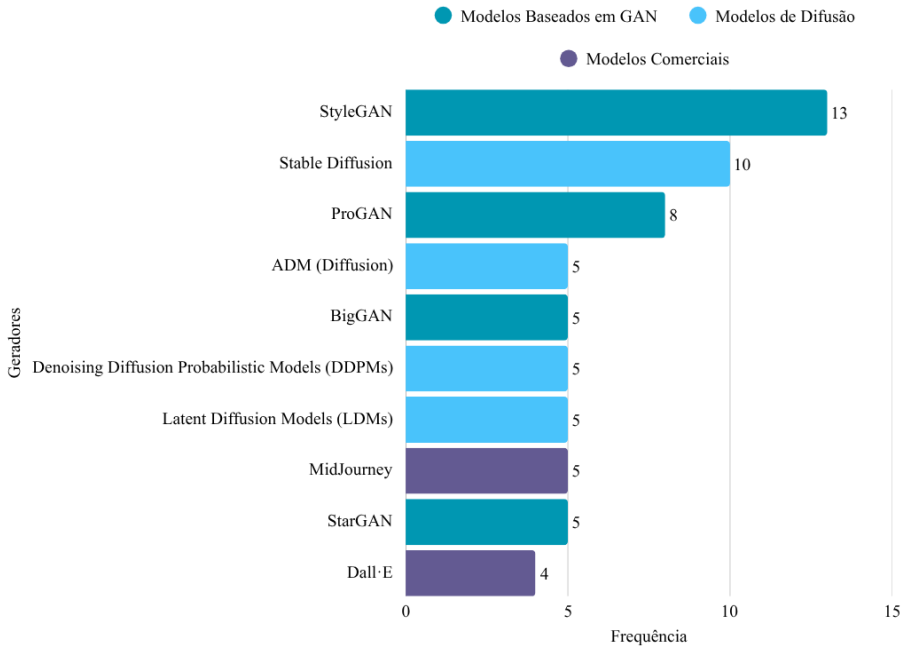
Figura 6 – Distribuição das categorias dos modelos geradores de imagem.



Fonte: Elaboração própria (2025).

A Figura 7 detalha os dez geradores mais citados, reforçando a liderança do StyleGAN e a presença crescente dos modelos de difusão no cenário atual. A diversidade de geradores abordados nos artigos analisados evidencia a necessidade de métodos de detecção capazes de lidar com múltiplas arquiteturas, especialmente considerando a evolução constante e a acessibilidade crescente de ferramentas de geração de imagens por IA.

Figura 7 – Os 10 modelos geradores de imagem mais citados.



Fonte: Elaboração própria (2025).

Observa-se o predomínio de modelos baseados em GAN, responsáveis por 44% das ocorrências. Essa predominância é justificada pela ampla adoção desta arquitetura na geração de imagens realistas. Destaque para o StyleGAN, que lidera individualmente com 13 ocorrências, seguido por ProGAN (8), BigGAN (5) e StarGAN (5). Esses modelos têm sido amplamente utilizados como referência em *benchmarks* de detecção, em virtude da alta qualidade visual das imagens que produzem.

Os modelos de difusão aparecem como a segunda categoria mais representativa, com 33% das ocorrências. Essa família de modelos ganhou notoriedade nos anos mais recentes devido à sua capacidade de gerar imagens altamente detalhadas. Modelos como Stable Diffusion (10 ocorrências), DDPMs, ADM e LDMs (todos com 5 ocorrências) são amplamente empregados nas pesquisas recentes, refletindo a ascensão dessa abordagem como paradigma dominante na geração de imagens sintéticas.

A categoria de modelos comerciais, que representa 15% dos casos, inclui geradores amplamente utilizados por usuários finais e que têm grande impacto social e midiático. Modelos como MidJourney (5 ocorrências) e DALL · E (4) foram alvos frequentes nos estudos por sua capacidade de gerar imagens visualmente impressionantes e por estarem disponíveis em larga escala para o público geral.

Por fim, os modelos baseados em autoencoders corresponderam a 8% das ocorrências. Essa categoria inclui abordagens como VAE (Variational Autoencoders) e outros modelos menos recorrentes, mas que ainda desempenham papel relevante em tarefas es-

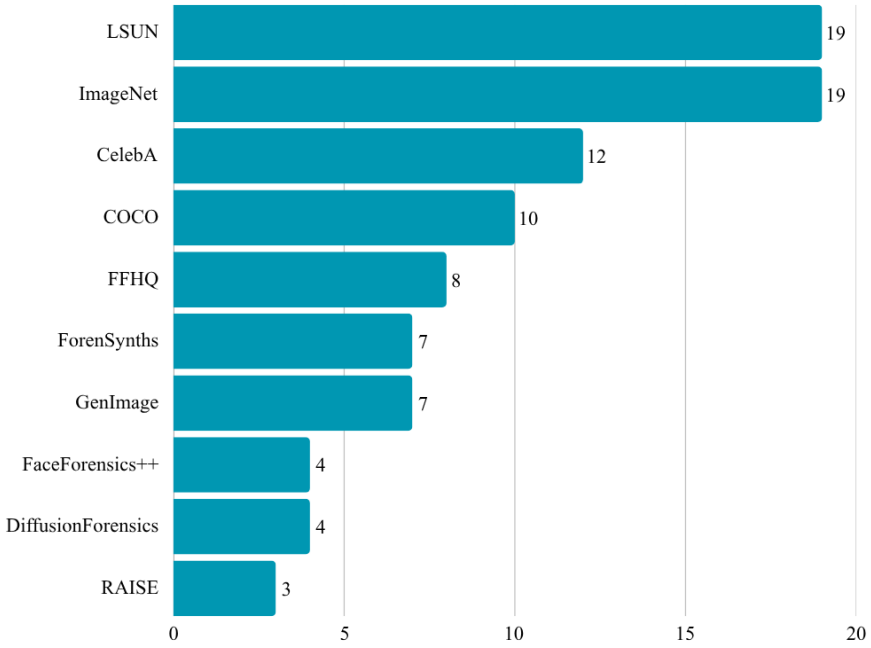
pecíficas de síntese ou reconstrução de imagem.

5.1.5 *Datasets*

A análise dos *datasets* utilizados para o treinamento e/ou avaliação dos detectores revelou uma ampla diversidade de bases de dados, refletindo os múltiplos contextos e abordagens empregados na detecção de imagens geradas por IA. Ao todo, foram identificados 34 *datasets* distintos, com diferentes propósitos e formatos, desde imagens de cenas gerais até rostos sintéticos e imagens manipuladas digitalmente.

A Figura 8 apresenta os dez *datasets* mais frequentemente utilizados. LSUN e ImageNet lideram o ranking com 19 ocorrências cada, o que evidencia sua ampla utilização como fontes de imagens reais de alta variedade visual, frequentemente empregadas tanto para treinamento quanto como base de comparação para imagens sintéticas. Na sequência, destaca-se o CelebA com 12 citações, voltado para imagens faciais, e o COCO com 10, amplamente usado pela sua diversidade de objetos e cenários.

Figura 8 – Os 10 *Datasets* mais citados.



Fonte: Elaboração própria (2025).

Bases mais específicas como FFHQ (8 ocorrências), ForenSynths e GenImage (ambos com 7), apresentam-se como alternativas mais recentes e direcionadas à detecção de imagens geradas por modelos modernos, como GANs e modelos de difusão. Esses *datasets* geralmente incluem pares *real-fake* ou composições sintéticas com características desafiadoras, contribuindo para o desenvolvimento de métodos mais robustos.

Outros *datasets* relevantes, ainda que com menor frequência, incluem o FaceForensics++ e o DiffusionForensics (4 ocorrências cada), ambos voltados para avaliação de imagens manipuladas ou geradas, e o RAISE (3 ocorrências), utilizado tradicionalmente em análises forenses de imagens reais.

Além da utilização de *datasets* públicos amplamente difundidos, observou-se que diversos artigos revisados optaram por construir seus próprios conjuntos de dados personalizados. Essa construção, em muitos casos, foi realizada por meio da mesclagem de diferentes bases públicas, com imagens sintéticas geradas a partir de modelos como StyleGAN, ProGAN, BigGAN, DDPM, ADM e Stable Diffusion. Essa abordagem visa criar pares real-fake com maior controle experimental, além de permitir a simulação de cenários específicos. Embora essa prática contribua para a diversificação dos dados de treinamento e teste, ela também introduz variações metodológicas que podem dificultar a comparação direta entre os diferentes detectores.

5.2 Análise Qualitativa

5.2.1 Tendências Tecnológicas Emergentes

A análise qualitativa dos artigos revela um cenário de rápida evolução e diversificação das estratégias para detecção de imagens geradas por IA. Um dos avanços mais recorrentes está no desenvolvimento de detectores com maior capacidade de generalização, capazes de atuar de forma eficaz sobre múltiplos modelos geradores e em cenários distintos. Diversas propostas demonstram essa preocupação ao empregar técnicas de domain generalization, transferência de aprendizado e estruturas adaptativas, que permitem a detecção mesmo diante de conteúdos sintéticos não vistos previamente durante o treinamento. Um exemplo é o trabalho de (MENG et al., 2024), que realiza a purificação explícita e implícita das features (características extraídas da imagem) para melhorar o desempenho entre domínios (cross-domain). Outro caso é o estudo de (CHEN et al., 2025), que utiliza features semânticas (representações vetoriais de significado visual) extraídas pelo CLIP ViT para aumentar a robustez frente a geradores inéditos.

Além disso, surgem estratégias complementares que buscam mitigar fragilidades específicas. Entre elas, destaca-se a compressão informacional em espaço de Wasserstein com agregação dinâmica, proposta por (LYU et al., 2024), que visa reduzir redundâncias no conjunto de treinamento e aumentar a robustez em bases desbalanceadas. Outro exemplo é o FatFormer, uma arquitetura transformadora adaptativa que integra pistas espaciais e de frequência para melhorar a detecção em geradores não vistos (LIU et al., 2024). Entre as propostas complementares, destaca-se ainda o PatchShuffle, explorado no SFLD, que embaralha regiões da imagem para reduzir o viés de conteúdo e reforçar a análise de texturas (GYE et al., 2025).

Outro avanço relevante diz respeito ao uso crescente de arquiteturas híbridas, nas quais diferentes paradigmas de rede neural são combinados para explorar simultaneamente padrões globais e inconsistências locais. Esse princípio está presente em diversos trabalhos analisados, como redes *dual-stream* com mecanismos de *cross-attention*, modelos que integram fusão multiescalar e atenção espacial/de canal, arquiteturas CNN–Transformer e sistemas de múltiplos ramos com atenção multi-cabeças, todos voltados a potencializar a detecção de artefatos visuais sutis (XI et al., 2023; XUE; JI; LI, 2024; XU et al., 2024; RAJ; MATHEW; MONDAL, 2024). Essas técnicas ampliam a capacidade de foco do modelo sobre regiões críticas da imagem, especialmente em cenários de alta complexidade visual.

O protagonismo crescente de arquiteturas baseadas em Transformers é outro aspecto notável. Sua capacidade de capturar dependências de longo alcance e modelar relações contextuais entre regiões da imagem tem-se mostrado eficaz na identificação de padrões sintéticos, como evidenciado em propostas que adaptam Transformers para detecção de imagens sintéticas, a exemplo do FatFormer, que combina pistas espaciais e de frequência (LIU et al., 2024), e de modelos híbridos CNN–ViT com mecanismos adicionais de atenção (WEIR et al., 2024). Em muitos casos, essas arquiteturas são utilizadas em conjunto com redes convolucionais, formando sistemas que equilibram desempenho e capacidade de generalização. Nesse cenário, ganha destaque a incorporação de modelos multimodais como o CLIP, explorada em diferentes estudos pela sua habilidade de integrar informações visuais e linguísticas, o que potencializa a compreensão e detecção de conteúdos gerados com alto grau de semântica visual (CHEN et al., 2025; COZZOLINO et al., 2024; ROSA et al., 2024).

A literatura também aponta para uma valorização de abordagens que atuam em domínios alternativos à visualização direta da imagem. Técnicas que operam no domínio da frequência ou em espaços latentes têm ganhado espaço por sua habilidade de detectar padrões recorrentes e não naturais inseridos por algoritmos generativos. O LaRE² proposto por (LUO et al., 2024), por exemplo, explora discrepâncias no espaço latente entre uma imagem e sua reconstrução, revelando inconsistências que se mostraram especialmente eficazes contra imagens de difusão. Já o método de (HERUR et al., 2025) foca em representações no domínio da frequência, utilizando transformadas (DCT) para capturar artefatos sutis introduzidos por autoencoders de difusão, alcançando bons resultados ao evidenciar assinaturas espectrais ocultas no processo de síntese.

Outra tendência observada envolve a adoção de estratégias de purificação ou reconstrução de imagem, nas quais módulos de *autoencoder* são empregados para realçar discrepâncias entre a imagem original e sua reconstrução. Trabalhos como o de (WANG et al., 2023a), que utilizam *autoencoder* para reconstruir e comparar imagens explorando diferenças sutis, e o de (WANG et al., 2023b), que mede erros de reconstrução para de-

tecção, têm demonstrado elevado desempenho na exposição de falhas características de conteúdos sintéticos.

No que se refere aos dados utilizados, embora muitos trabalhos recorram a bases consagradas como FFHQ, CelebA, COCO e ImageNet, há uma forte presença de *datasets* personalizados, construídos por meio da mesclagem de múltiplas fontes reais com imagens geradas por modelos específicos, como StyleGAN ou Stable Diffusion. Essa prática, observada em estudos como o de (RAHMAN et al., 2023), que propõe o *Artifact*, um grande conjunto de dados integrando múltiplos domínios e geradores, e o de (GUPTA; SHRENETER; SEHGAL, 2024), que utiliza bases como CIFAKE e ImageNet para simular cenários diversos, busca criar ambientes mais realistas e desafiadores, permitindo testes mais próximos dos contextos práticos. No entanto, também evidencia a carência de *benchmarks* padronizados que atendam à crescente diversidade de modelos generativos.

Essa evolução reflete uma mudança de foco na comunidade de pesquisa, que, historicamente, se concentrava na detecção de artefatos visuais mais óbvios gerados por GANs. Com o advento dos modelos de difusão, o desafio passou a ser a identificação de inconsistências mais sutis e de alta fidelidade, exigindo métodos mais sofisticados, interpretáveis e generalizáveis. As tendências observadas, portanto, não representam apenas avanços técnicos, mas uma resposta direta à dinâmica acelerada e à crescente sofisticação do ecossistema de geração de imagens por IA.

5.2.2 Desempenho e Robustez dos Detectores

A literatura mostra avanços relevantes na generalização e na robustez dos detectores, mas também evidencia limites claros quando os métodos são avaliados em domínios abertos. Em cenários controlados, muitos modelos atingem acurácias superiores a 95%; porém, avaliações entre geradores (*cross-generator*) revelam quedas substanciais. Nesses testes, o treinamento é realizado em imagens de determinados geradores e a validação em imagens produzidas por outros, não vistos previamente. Nessas condições, o CLIP ViT obteve média de 72,9% de acurácia ao ser avaliado em modelos de difusão e GANs distintos, superando arquiteturas tradicionais como ResNet-50 e Swin-T (CHEN et al., 2025). Esses resultados confirmam o potencial do uso de características semânticas para reduzir a dependência de artefatos específicos de cada gerador e aumentar a capacidade de generalização.

Tabela 3 – Desempenho comparativo de detectores em avaliação cross-generator. Valores em acurácia (%) de diferentes métodos quando treinados em um gerador e testados em outro. O CLIP ViT apresentou média superior, alcançando 72,9%.

Method	ADM	BigGAN	GLIDE	Midjourney	SD V1.4	SD V1.5	VQDM	Wukong	Avg Acc. (%)
ResNet-50	59.7	66.6	73.1	59.0	72.3	72.4	60.9	71.4	66.9
DeiT-S	59.5	66.3	71.1	60.7	74.2	74.2	61.7	73.1	67.6
Swin-T	61.3	**69.5**	76.9	**61.7**	**76.0**	**76.1**	65.8	**75.1**	70.3
CNNSpot	57.0	56.6	57.1	58.2	70.3	70.2	56.7	67.7	61.7
Spec	57.9	64.3	65.4	56.7	72.4	72.3	61.7	70.3	65.1
F3Net	66.5	56.5	57.8	55.1	73.1	73.1	62.1	72.3	64.6
GramNet	58.7	61.2	65.3	58.1	72.8	72.7	57.8	71.3	64.7
CLIP	**80.6**	69.3	**82.7**	59.7	71.5	72.0	**75.1**	72.7	**72.9**

Fonte: Adaptado de (CHEN et al., 2025).

Detectores desenvolvidos especificamente para modelos de difusão, como o DIRE, reportaram resultados excepcionais em *benchmarks* dedicados, mostrando que abordagens baseadas em erro de reconstrução latente podem suprir fragilidades de métodos tradicionais (WANG et al., 2023b). Ao mesmo tempo, estudos de evasão demonstram que detectores dependentes de artefatos específicos podem ser burlados por técnicas de disentanglement, reduzindo drasticamente as taxas de detecção (CHAN et al., 2024). Esses achados reforçam a necessidade de protocolos de avaliação que incluam cenários adversariais, uso de *benchmarks* realistas, como ArtiFac, DiffusionForensics e SIDBench, e métricas robustas além da acurácia.

Em contraste, detectores baseados exclusivamente em GANs demonstraram baixa capacidade de generalização, falhando frente a imagens de difusão (XU et al., 2024). A resistência a perturbações também se mostrou variável, mostrando diferenças relevantes entre os métodos: enquanto abordagens convencionais sofrem degradação considerável, o FatFormer apresentou maior robustez, mantendo desempenho mais estável (LIU et al., 2024). Já métodos que dependem de artefatos espectrais, como F3Net (QIAN et al., 2020), foram severamente impactados, enquanto em ataques adversariais direcionados observou-se queda para menos de 60% de acurácia (CHAN et al., 2024).

O uso de arquiteturas híbridas, integrando CNNs e Transformers com mecanismos de atenção, demonstrou potencial significativo para enfrentar essas dificuldades, com modelos alcançando até 99,77% de acurácia em determinados cenários de dados não vistos (WEIR et al., 2024). Dentro desse contexto, o NASA-Swin (GUAN et al., 2025) se destacou ao incorporar atenção sensível ao ruído (Noise-Aware Self-Attention) aliados à arquitetura Swin Transformer, apresentando maior resiliência contra perturbações como compressão JPEG, ruído e redimensionamento, além de manter desempenho competitivo em cenários entre domínios (interdomínio). Já o LaRE2 (LUO et al., 2024) avançou ao explorar discrepâncias no espaço latente, evidenciando inconsistências entre a imagem e sua reconstrução, o que se mostrou especialmente eficaz contra imagens de difusão, nas quais o realismo visual dificulta a identificação de padrões artificiais.

Tabela 4 – Desempenho comparativo de diferentes arquiteturas em detecção de imagens geradas por IA. A combinação híbrida CNN/ViT com mecanismos de atenção apresentou os melhores resultados, alcançando 99,77% de acurácia, com perdas reduzidas e maior precisão em comparação a abordagens isoladas..

Modelo	**Acurácia**	**Fução de Perda**	**Precisão**
CNN	94%	23%	96,30%
ViT	93,15%	18%	—
CNN/ViT	99,67%	0,60%	99,52%
CNN/ViT + Attention	99,77%	0,50%	99,80%
CIFAKE (CNN)	92,98%	18%	93,60%

Fonte: Adaptado de (WEIR et al., 2024).

Quanto às métricas de avaliação, a acurácia segue predominando, seguida por indicadores como precisão média (AP), AUC (Area Under the Curve), revocação, precisão e F1-score. Contudo, poucas pesquisas exploraram métricas mais sensíveis à robustez, como a acurácia balanceada, que considera o desempenho de forma equitativa entre classes, ou ainda medidas como taxa de falsos positivos e taxa de falsos negativos, essenciais em cenários adversariais.

Ainda que diversos detectores apresentem resultados notáveis em cenários controlados, a eficácia diante de condições adversas e a capacidade de generalização para domínios abertos continuam sendo obstáculos significativos. A literatura evidencia avanços consistentes, como o uso de arquiteturas híbridas, estratégias em espaço latente e *benchmarks* mais realistas, mas também mostra que tais soluções ainda não garantem resiliência em situações práticas complexas. Nesse sentido, torna-se imprescindível direcionar os esforços futuros não apenas para ganhos de acurácia, mas também para o fortalecimento da robustez, da interpretabilidade e da confiabilidade dos modelos. Essas questões apontam para um conjunto mais amplo de desafios e limitações, que serão discutidos na próxima seção.

5.2.3 Desafios e Limitações

Entre os desafios mais recorrentes identificados na literatura está a dificuldade de generalização dos detectores para domínios não vistos. Muitos modelos, embora apresentem desempenho elevado em cenários controlados, sofrem quedas expressivas de acurácia quando testados em arquiteturas ou conjuntos de dados diferentes daqueles utilizados no treinamento. Essa limitação é ressaltada em trabalhos como o de (MENG et al., 2024), que mostram que detectores tradicionais tendem a superajustar artefatos específicos de geradores, falhando em cenários *cross-domain*. O estudo de (CHEN et al., 2025) também evidencia que detectores baseados em características semânticas, como o CLIP, embora mais robustos que os de baixo nível, ainda enfrentam dificuldades quando expostos a ge-

radores inéditos. Esse desafio é agravado pela rápida evolução dos modelos de geração, especialmente os de difusão, como Stable Diffusion, que produzem imagens cada vez mais realistas e com artefatos imperceptíveis aos métodos convencionais.

Outro desafio central refere-se à vulnerabilidade a manipulações e transformações comuns no mundo real, como compressão JPEG, redimensionamento, corte, adição de ruído e desfoque. Trabalhos como o de (PARK; NA; CHOI, 2024) demonstram que a acurácia de vários detectores cai significativamente quando expostos a imagens comprimidas ou corrompidas, o que os torna frágeis em cenários práticos. Além disso, ataques adversariais exploram essas fragilidades, levando modelos a apresentarem desempenhos próximos ao aleatório, como evidenciado no estudo de (ROSA et al., 2024), que analisou a robustez adversarial de arquiteturas baseadas em CNN e CLIP.

A escassez de *benchmarks* abrangentes e representativos constitui outra limitação crítica. Grande parte dos estudos se apoia em bases artificiais ou restritas, que não refletem a diversidade e complexidade de aplicações reais. Para enfrentar essa lacuna, iniciativas como o ArtiFact dataset (RAHMAN et al., 2023) e o SIDBench framework (SCHINAS; PAPADOPOULOS, 2024) foram propostas, destacando a importância da padronização de métricas e protocolos para comparações justas.

Outro ponto crítico é a especialização excessiva de alguns detectores. Modelos como o Deep Image Fingerprint (SINITSA; FRIED, 2024) apresentam ótimo desempenho para identificar imagens de um gerador específico, mas sua eficácia cai ao lidar com outros geradores ou domínios. Essa limitação reforça a necessidade de detectores mais universais. Além disso,(CHEN et al., 2025) mostra que a utilização de representações semânticas extraídas pelo CLIP ViT contribui para uma melhor generalização, superando modelos tradicionais como o ResNet-50 por não depender apenas de artefatos específicos dos geradores.

Do ponto de vista prático, a complexidade e o custo computacional representam desafios significativos, sobretudo em aplicações em tempo real. Trabalhos como o de (LYU et al., 2024) propõem estratégias de compressão no espaço de *Wasserstein* com agregação dinâmica, visando reduzir os custos de treinamento e melhorar a eficiência sem comprometer o desempenho.

Em síntese, os principais desafios e limitações identificados abrangem a generalização limitada, a vulnerabilidade a manipulações e ataques, a escassez de *benchmarks* representativos, a especialização excessiva, o alto custo computacional e a falta de interpretabilidade. Superá-los exigirá o desenvolvimento de abordagens mais generalizáveis, resistentes e transparentes, baseadas em avaliações realistas e padronizadas, que integrem também considerações éticas e de aplicabilidade prática.

5.2.4 Implicações Éticas e Sociais

O avanço dos modelos generativos tem intensificado preocupações sobre o uso malicioso de imagens sintéticas, sobretudo na produção de *deepfakes* e conteúdos manipulados que podem disseminar desinformação e comprometer processos democráticos (JAGADISH; JASMINE, 2024). Esse tipo de material representa um risco direto para a esfera pública, já que amplia a propagação de notícias falsas e fraudes visuais em ambientes digitais, favorecendo a manipulação de opiniões coletivas.

Esses problemas reverberam em um efeito mais amplo: a erosão da confiança pública em imagens digitais. (TAN et al., 2025) enfatizam que "o surgimento de imagens geradas, visualmente indistinguíveis, representa riscos substanciais para o domínio judicial", instaurando um ambiente em que até mesmo registros autênticos passam a ser questionados.

Além dessas implicações, é fundamental considerar os riscos éticos e de privacidade associados à geração e uso de imagens sintéticas. A criação de rostos realistas de pessoas inexistentes ou a manipulação de identidades verdadeiras levanta questões sobre consentimento, violação de direitos individuais e uso indevido de dados biométricos. Em contextos sensíveis, como segurança, justiça e vigilância, o emprego de imagens sem autorização para treinar ou validar modelos pode comprometer a privacidade e expor indivíduos a riscos adicionais.

Assim, compreender esses riscos vai além do debate técnico: envolve reconhecer que a detecção de imagens geradas por IA desempenha um papel essencial na proteção da privacidade, na preservação da confiança pública e na prevenção de abusos e manipulações. O enfrentamento desses desafios exige diretrizes claras, práticas responsáveis e mecanismos de governança que assegurem a aplicação ética dessas tecnologias.

6 Conclusão

Em síntese, a revisão confirmou as hipóteses iniciais de que a detecção de imagens sintéticas representa um campo promissor, mas ainda marcado por limitações significativas. Os objetivos propostos foram atendidos: foram identificados os métodos mais utilizados, os mais promissores e os principais desafios técnicos, sociais e éticos associados ao tema. As evidências apontam que, embora haja progressos claros em termos de desempenho e inovação tecnológica, a consolidação da detecção de imagens geradas por IA como ferramenta confiável exigirá avanços adicionais que unam robustez técnica, governança responsável e engajamento social. Ressalta-se, por fim, que a relevância do tema transcende a esfera técnica, tendo implicações diretas para a ciência, para a proteção social e para a manutenção da confiança pública em um ecossistema digital cada vez mais permeado por conteúdos sintéticos.

A revisão sistemática realizada evidencia que a detecção de imagens geradas por inteligência artificial é um campo em rápida evolução, no qual avanços técnicos significativos coexistem com desafios persistentes. A análise quantitativa dos artigos finais após os filtros revelou um crescimento expressivo da produção científica entre 2022 e 2025, acompanhado pela diversificação de arquiteturas e metodologias de detecção. Modelos baseados em CNNs e *Transformers*, além de arquiteturas híbridas e multimodais, mostraram-se predominantes, com destaque para detectores como CLIP ViT, DIRE, DC-BiNet e FatFormer, que obtiveram resultados superiores em cenários de generalização e sob condições adversariais.

Do ponto de vista qualitativo, observou-se que o desempenho dos detectores é elevado em ambientes controlados, muitas vezes superando 95% de acurácia, mas apresenta quedas significativas em cenários entre domínios ou diante de modelos de difusão recentes. Essa constatação reforça a dificuldade de generalização, apontada como uma das principais limitações. Estratégias emergentes, como purificação de artefatos, exploração de espaços latentes, compressão no espaço de *Wasserstein*, mecanismos de atenção multi-escala e integração de representações semânticas, têm buscado mitigar essas fragilidades, embora ainda não eliminem por completo a vulnerabilidade a manipulações e ataques adversariais.

6.1 Desafios e Limitações

Os desafios e limitações recorrentes incluem a dependência de *benchmarks* artificiais ou restritos, a especialização excessiva de alguns detectores, o alto custo computacional e a escassez de abordagens explicáveis. Esses pontos indicam que, apesar do avanço técnico,

a maturidade prática da área ainda é incipiente. Em paralelo, a análise das implicações éticas e sociais destacou riscos como desinformação, violação de privacidade, uso malicioso de *deepfakes* e erosão da confiança pública em conteúdos visuais. Também foi evidenciada a ausência de regulamentações específicas e a necessidade de maior transparência e auditabilidade nos sistemas de detecção.

A presente revisão, no entanto, também apresenta limitações. A busca concentrou-se em bases específicas, publicações entre 2022 e 2025 e majoritariamente na língua inglesa, o que pode ter levado à exclusão de estudos relevantes em outros idiomas ou períodos. Além disso, a heterogeneidade metodológica encontrada, com diferentes métricas, *datasets* e protocolos experimentais, dificultou comparações padronizadas entre os detectores.

No que diz respeito às perspectivas futuras, observa-se a necessidade de desenvolvimento de detectores mais robustos e generalizáveis, capazes de lidar com ambientes interdomínio e com a sofisticação crescente dos modelos de difusão. Estratégias multimodais, integrando imagem, texto e metadados, surgem como alternativas promissoras, assim como a criação de *benchmarks* padronizados e realistas, que permitam comparações mais consistentes entre abordagens. Avanços em interpretabilidade e transparência também são fundamentais, especialmente para o uso em contextos críticos, como jornalismo, perícia digital e segurança institucional. Nesse sentido, esforços interdisciplinares envolvendo aspectos técnicos, éticos e regulatórios podem contribuir para um ecossistema mais seguro e confiável.

Referências

BIOINFO. **Métricas de avaliação em Machine Learning: acurácia, sensibilidade, precisão, especificidade e F-Score**. 2023. Disponível em: <https://bioinfo.com.br/metricas-de-avaliacao-em-machine-learning-acuracia-sensibilidade-precisao-especificidade-e-f-score/>. Citado na página 15.

BOUGUEFFA, H.; KEITA, M.; HAMIDOUCHE, W.; TALEB-AHMED, A.; LIZ-LóPEZ, H.; MARTíN, A.; CAMACHO, D.; HADID, A. Advances in ai-generated images and videos. **International Journal of Interactive Multimedia and Artificial Intelligence**, 2024. [Online; accessed 24-Março-2025]. Disponível em: <https://www.ijimai.org/journal/bibcite/reference/3512>. Citado na página 13.

BYEON, H.; SHABAZ, M.; SHRIVASTAVA, K.; JOSHI, A.; KESHTA, I.; OAK, R.; SINGH, P. P.; SONI, M. Deep learning model to detect deceptive generative adversarial network generated images using multimedia forensic. **Computers and Electrical Engineering**, v. 113, p. 109024, 2024. ISSN 0045-7906. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0045790623004482>. Citado na página 24.

CALIS, J. O. G. Trabalho de Conclusão de Curso, **Aplicação de redes neurais convolucionais para reconhecimento automático de placas de veículos**. 2018. <https://tcc.dcce.ibilce.unesp.br/media/2018/12/12/2adebeaedc13491d8ec7e19de050a8ae.pdf>. [Online; accessed 24-Março-2025]. Citado na página 14.

CHAN, P. P.; ZHANG, C.; CHEN, H.; DENG, J.; MENG, X.; YEUNG, D. S. Evasion on general gan-generated image detection by disentangled representation. **Information Sciences**, v. 683, p. 121267, 2024. ISSN 0020-0255. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0020025524011812>. Citado 2 vezes nas páginas 24 e 38.

CHEN, H.-C.; SHEN, Y.-P.; YEH, Y.-H.; CHEN, J.-L. Ai-generated image detection using semantic feature. In: **2025 27th International Conference on Advanced Communications Technology (ICACT)**. [S.l.: s.n.], 2025. p. 459–463. Citado 7 vezes nas páginas 24, 35, 36, 37, 38, 39 e 40.

COZZOLINO, D.; NAGANO, K.; THOMAZ, L.; MAJUMDAR, A.; VERDOLIVA, L. Synthetic image detection: Highlights from the ieee video and image processing cup 2022 student competition [sp competitions]. **IEEE Signal Processing Magazine**, v. 40, n. 7, p. 94–100, 2023. Citado na página 25.

COZZOLINO, D.; POGGI, G.; CORVI, R.; NIEßNER, M.; VERDOLIVA, L. Raising the bar of ai-generated image detection with clip. In: **2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**. [S.l.: s.n.], 2024. p. 4356–4366. Citado 2 vezes nas páginas 25 e 36.

DAMASCENO, G. F. Trabalho de Conclusão de Curso, **A responsabilidade civil dos provedores de aplicações de internet: análise sobre a moderação de conteúdo virtual frente às novas tecnologias de inteligência artificial**.

2024. <https://repositorio.ufu.br/handle/123456789/43716>. [Online; accessed 14-Março-2025]. Citado na página 12.

DOGOULIS, P.; KORDOPATIS-ZILOS, G.; KOMPATSIARIS, I.; PAPADOPOULOS, S. Improving synthetically generated image detection in cross-concept settings. In: **Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation**. Association for Computing Machinery, 2023. p. 28–35. ISBN 9798400701870. Disponível em: <https://doi.org/10.1145/3592572.3592846>. Citado na página 24.

GANGAN, M. P.; KADAN, A.; L., L. V. Toward exploring fairness in visual transformer based natural and gan image detection systems. **IEEE Transactions on Computational Social Systems**, v. 12, n. 3, p. 1068–1079, 2025. Citado na página 25.

GHOSH, T.; NASKAR, R. Leveraging image gradients for robust gan-generated image detection in osn context. In: **2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)**. [S.l.: s.n.], 2023. p. 1–5. Citado na página 25.

GUAN, W.; WANG, W.; PENG, B.; HE, Z.; DONG, J.; CHENG, H. Noise-informed diffusion-generated image detection with anomaly attention. **IEEE Transactions on Information Forensics and Security**, v. 20, p. 5256–5268, 2025. Citado 2 vezes nas páginas 25 e 38.

GUPTA, A. S.; SHRENETER, K. P.; SEHGAL, S. Visual veracity: Advancing ai-generated image detection with convolutional neural networks. In: **2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)**. [S.l.: s.n.], 2024. p. 1–6. Citado 2 vezes nas páginas 25 e 37.

GYE, S.; KO, J.; SHON, H.; KWON, M.; KIM, J. Reducing the content bias for ai-generated image detection. In: **2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)**. [S.l.: s.n.], 2025. p. 399–408. Citado 2 vezes nas páginas 25 e 35.

HAN, K.; WANG, Y.; CHEN, H.; CHEN, X.; GUO, J. A survey on vision transformer. **IEEE Transactions on Neural Networks and Learning Systems**, 2022. [Online; accessed 24-Março-2025]. Disponível em: <https://ieeexplore.ieee.org/abstract/document/9716741/>. Citado na página 14.

HERUR, A. N.; SANTHOSH, V.; SHETTY, N.; SEELAMANTULA, C. S. Addressing diffusion model based counter-forensic image manipulation for synthetic image detection. In: **Proceedings of the Fifteenth Indian Conference on Computer Vision Graphics and Image Processing**. New York, NY, USA: Association for Computing Machinery, 2025. ISBN 9798400710759. Disponível em: <https://doi.org/10.1145/3702250.3702296>. Citado 2 vezes nas páginas 24 e 36.

HOSSAIN, M. Z.; ZAMAN, F. U.; ISLAM, M. R. Advancing ai-generated image detection: Enhanced accuracy through cnn and vision transformer models with explainable ai insights. In: **2023 26th International Conference on Computer and Information Technology (ICCIT)**. [S.l.: s.n.], 2023. p. 1–6. Citado na página 24.

JAGADISH, T.; JASMINE, S. G. Detection of ai-generated image content in news and journalism. In: **2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)**. [S.l.: s.n.], 2024. p. 1–6. Citado 2 vezes nas páginas 24 e 41.

JAVAHERI, A. H.; MOTAMEDNIA, H.; MAHMOUDI-AZANVEH, A. Enhancing the generalization of synthetic image detection models through the exploration of features in deep detection models. In: **2024 13th Iranian/3rd International Machine Vision and Image Processing Conference (MVIP)**. [S.l.: s.n.], 2024. p. 1–6. Citado na página 24.

JEONG, Y.; KIM, D.; KIM, P.; RO, Y.; CHOI, J. Self-supervised scheme for generalizing gan image detection. **Pattern Recognition Letters**, v. 184, p. 219–224, 2024. ISSN 0167-8655. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167865524002009>. Citado na página 25.

KANG, M.; ZHU, J. Y.; ZHANG, R.; PARK, J.; SHECHTMAN, E.; PARIS, S.; PARK, T. Scaling up gans for text-to-image synthesis. **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, p. 10124–10134, 2023. [Online; accessed 24-Março-2025]. Disponível em: <https://openaccess.thecvf.com/content/CVPR2023/html/Kang_Scaling_Up_GANs_for_Text-to-Image_Synthesis_CVPR_2023_paper.html>. Citado na página 13.

KITCHENHAM, B.; CHARTERS, S. et al. **Guidelines for performing systematic literature reviews in software engineering**. [S.l.]: Keele, UK, 2007. Citado 3 vezes nas páginas 20, 21 e 23.

KONSTANTINIDOU, D.; KOUTLIS, C.; PAPADOPOULOS, S. Texturecrop: Enhancing synthetic image detection through texture-based cropping. In: **2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)**. [S.l.: s.n.], 2025. p. 1369–1378. Citado na página 25.

LI, O.; CAI, J.; HAO, Y.; JIANG, X.; HU, Y.; FENG, F. Improving synthetic image detection towards generalization: An image transformation perspective. In: **Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1**. Association for Computing Machinery, 2025. p. 2405–2414. ISBN 9798400712456. Disponível em: <https://doi.org/10.1145/3690624.3709392>. Citado na página 24.

LIN, M.; SHANG, L.; GAO, X. Enhancing interpretability in ai-generated image detection with genetic programming. In: **2023 IEEE International Conference on Data Mining Workshops (ICDMW)**. [S.l.: s.n.], 2023. p. 371–378. Citado na página 24.

LIU, H.; TAN, Z.; TAN, C.; WEI, Y.; WANG, J.; ZHAO, Y. Forgery-aware adaptive transformer for generalizable synthetic image detection. In: **2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2024. p. 10770–10780. Citado 4 vezes nas páginas 24, 35, 36 e 38.

LIU, Y.; WAN, Z.; YIN, X.; YUE, G.; TAN, A.; ZHENG, Z. Detection of gan generated image using color gradient representation. **Journal of Visual Communication and**

Image Representation, v. 95, p. 103876, 2023. ISSN 1047-3203. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1047320323001268>. Citado na página 24.

LUO, Y.; DU, J.; YAN, K.; DING, S. Lare2: Latent reconstruction error based method for diffusion-generated image detection. In: **2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2024. p. 17006–17015. Citado 3 vezes nas páginas 25, 36 e 38.

LYU, Z.; XIAO, J.; ZHANG, C.; LAM, K.-M. Ai-generated image detection with wasserstein distance compression and dynamic aggregation. In: **2024 IEEE International Conference on Image Processing (ICIP)**. [S.l.: s.n.], 2024. p. 3827–3833. Citado 3 vezes nas páginas 24, 35 e 40.

MENG, Z.; PENG, B.; DONG, J.; TAN, T.; CHENG, H. Artifact feature purification for cross-domain detection of ai-generated images. **Computer Vision and Image Understanding**, v. 247, p. 104078, 2024. ISSN 1077-3142. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1077314224001590>. Citado 3 vezes nas páginas 24, 35 e 39.

NAKANISHI, M. F. M. Trabalho de Conclusão de Curso, **A problemática jurídica dos deepfakes: uma análise do uso da inteligência artificial na produção de provas e suas repercussões penais**. 2023. <https://repositorio.uniceub.br/jspui/handle/prefix/17157>. [Online; accessed 14-Março-2025]. Citado na página 12.

PAGE, M. J.; MCKENZIE, J. E.; BOSSUYT, P. M.; BOUTRON, I.; HOFFMANN, T. C.; MULROW, C. D.; SHAMSEER, L.; TETZLAFF, J. M.; AKL, E. A.; BRENNAN, S. E.; CHOU, R.; GLANVILLE, J.; GRIMSHAW, J. M.; HRóBJARTSSON, A.; LALU, M. M.; LI, T.; LODER, E. W.; MAYO-WILSON, E.; MCDONALD, S.; MCGUINNESS, L. A.; STEWART, L. A.; THOMAS, J.; TRICCO, A. C.; WELCH, V. A.; WHITING, P.; MOHER, D. The prisma 2020 statement: An updated guideline for reporting systematic reviews. **The BMJ**, v. 372, 2021. Cited by: 53027; All Open Access, Green Open Access, Hybrid Gold Open Access. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85103451714&doi=10.1136%2fbmj.n71&partnerID=40&md5=be0fc2e995f7f042e89a77c3cd53ab9b>. Citado 3 vezes nas páginas 20, 21 e 23.

PARK, D.; NA, H.; CHOI, D. Performance comparison and visualization of ai-generated-image detection methods. **IEEE Access**, v. 12, p. 62609–62627, 2024. Citado 2 vezes nas páginas 25 e 40.

PATEL, Y.; TANWAR, S.; GUPTA, R.; BHATTACHARYA, P.; DAVIDSON, I. E.; NYAMEKO, R.; ALUVALA, S.; VIMAL, V. Deepfake Generation and Detection: Case Study and Challenges. **IEEE Access**, v. 11, p. 2169–3536, 2023. [Online; accessed 10-Abril-2025]. Disponível em: <https://ieeexplore.ieee.org/document/10354308>. Citado na página 18.

PINHEIRO, R. M.; GADOTTI, G. I.; BERNARDY, R. **Análise do desempenho de técnicas de aprendizado de máquina para classificação de lotes de sementes de soja**. 2021. <https://guaiaca.ufpel.edu.br/handle/prefix/12817>. [Online; accessed 24-Março-2025]. Citado na página 15.

QIAN, Y.; YIN, G.; SHENG, L.; CHEN, Z.; SHAO, J. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: VEDALDI, A.; BISCHOF, H.; BROX, T.; FRAHM, J.-M. (Ed.). **Computer Vision – ECCV 2020**. [S.l.]: Springer International Publishing, 2020. p. 86–103. ISBN 978-3-030-58610-2. Citado 2 vezes nas páginas 31 e 38.

RAHMAN, M. A.; PAUL, B.; SARKER, N. H.; HAKIM, Z. I. A.; FATTAH, S. A. Artifact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection. In: **2023 IEEE International Conference on Image Processing (ICIP)**. [S.l.: s.n.], 2023. Citado 3 vezes nas páginas 24, 37 e 40.

RAJ, S.; MATHEW, J.; MONDAL, A. Generalized and robust model for gan-generated image detection. **Pattern Recognition Letters**, v. 182, p. 104–110, 2024. ISSN 0167-8655. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167865524001193>. Citado 2 vezes nas páginas 24 e 36.

ROSA, V. D.; GUILLARO, F.; POGGI, G.; COZZOLINO, D.; VERDOLIVA, L. Exploring the adversarial robustness of clip for ai-generated image detection. In: **2024 IEEE International Workshop on Information Forensics and Security (WIFS)**. [S.l.: s.n.], 2024. p. 1–6. Citado 3 vezes nas páginas 24, 36 e 40.

ROUT, J.; MISHRA, M. Enhanced cnn architecture with residual blocks and regularization for ai-generated image detection. In: **2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)**. [S.l.: s.n.], 2025. v. 3, p. 1–6. Citado na página 24.

RöSSLER, A.; COZZOLINO, D.; VERDOLIVA, L.; RIESS, C.; THIES, J.; NIEßNER, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In: **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**. [s.n.], 2019. p. 1–14. [Online; accessed 10-Abril-2025]. Disponível em: <https://arxiv.org/abs/1901.08971>. Citado na página 18.

S, S. C.; R, S. Performance comparison of deep learning models for computer generated image detection. In: **2023 International Conference on Control, Communication and Computing (ICCC)**. [S.l.: s.n.], 2023. p. 1–5. Citado na página 25.

SALVO, R. de V. Trabalho de Conclusão de Curso, **Juízes artificiais: Aplicação da Inteligência Artificial no julgamento de processos**. 2020. <https://repositorio.ufu.br/handle/123456789/30070>. [Online; accessed 14-Março-2025]. Citado na página 12.

SCHINAS, M.; PAPADOPOULOS, S. Sidbench: A python framework for reliably assessing synthetic image detection methods. In: **Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation**. Association for Computing Machinery, 2024. p. 55–64. ISBN 9798400705526. Disponível em: <https://doi.org/10.1145/3643491.3660277>. Citado 2 vezes nas páginas 25 e 40.

SINITSA, S.; FRIED, O. Deep image fingerprint: Towards low budget synthetic image detection and model lineage analysis. In: **2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)**. [S.l.: s.n.], 2024. p. 4055–4064. Citado 2 vezes nas páginas 24 e 40.

SONG, S.; YANG, J.; CHEN, J.; QI, H.; XUE, Y.; LAO, Y.; YU, Y. Acm multimedia 2024 grand challenge report for artificial intelligence generated image detection. In: **Proceedings of the 32nd ACM International Conference on Multimedia**. New York, NY, USA: Association for Computing Machinery, 2024. p. 11470–11471. ISBN 9798400706868. Disponível em: <https://doi.org/10.1145/3664647.3689001>. Citado na página 24.

TAN, D.; NIU, C.; YANG, Y.; YANG, D.; TAN, B. Dc-binet: Towards interpretable generated image detection with dark channel prior. **Expert Systems with Applications**, v. 280, p. 127508, 2025. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417425011303>. Citado 2 vezes nas páginas 24 e 41.

VERDOLIVA, L. Media Forensics and DeepFakes: An Overview. **IEEE Journal of Selected Topics in Signal Processing**, v. 14, n. 5, p. 910–932, 2020. [Online; accessed 10-Abril-2025]. Disponível em: <https://arxiv.org/abs/2001.06564>. Citado na página 18.

WANG, H.; FEI, J.; DAI, Y.; LENG, L.; XIA, Z. General gan-generated image detection by data augmentation in fingerprint domain. In: **2023 IEEE International Conference on Multimedia and Expo (ICME)**. [S.l.: s.n.], 2023. p. 1187–1192. Citado 2 vezes nas páginas 24 e 36.

WANG, Z.; BAO, J.; ZHOU, W.; WANG, W.; HU, H.; CHEN, H.; LI, H. Dire for diffusion-generated image detection. In: **2023 IEEE/CVF International Conference on Computer Vision (ICCV)**. [S.l.: s.n.], 2023. Citado 3 vezes nas páginas 24, 36 e 38.

WEIR, S.; KHAN, M. S.; MORADPOOR, N.; AHMAD, J. Enhancing ai-generated image detection with a novel approach and comparative analysis. In: **2024 17th International Conference on Security of Information and Networks (SIN)**. [S.l.: s.n.], 2024. p. 1–7. Citado 4 vezes nas páginas 24, 36, 38 e 39.

WU, K.; LI, X. Photo response non-uniformity based ai-generated image detection. In: **2024 IEEE 12th International Conference on Computer Science and Network Technology (ICCSNT)**. [S.l.: s.n.], 2024. p. 93–97. Citado na página 25.

XI, Z.; HUANG, W.; WEI, K.; LUO, W.; ZHENG, P. Ai-generated image detection using a cross-attention enhanced dual-stream network. In: **2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)**. [S.l.: s.n.], 2023. Citado 2 vezes nas páginas 24 e 36.

XU, J.; YANG, Y.; FANG, H.; LIU, H.; ZHANG, W. Famsec: A few-shot-sample-based general ai-generated image detection method. **IEEE Signal Processing Letters**, v. 32, p. 226–230, 2025. Citado na página 24.

XU, Q.; JIA, S.; JIANG, X.; SUN, T.; WANG, Z.; YAN, H. Mdtl-net: Computer-generated image detection based on multi-scale deep texture learning. **Expert Systems with Applications**, v. 248, p. 123368, 2024. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417424002331>. Citado 3 vezes nas páginas 25, 36 e 38.

XUE, Y.; JI, L.; LI, S. Computer-generated image detection based on multi-scale feature fusion attention module. In: **2024 IEEE 2nd International Conference on Control, Electronics and Computer Technology (ICCECT)**. [S.l.: s.n.], 2024. p. 273–277. Citado 2 vezes nas páginas 24 e 36.

YU, X.; CHEN, K.; ZENG, K.; FANG, H.; YANG, Z.; SHANG, X.; QI, Y.; ZHANG, W.; YU, N. Semgir: Semantic-guided image regeneration based method for ai-generated image detection and attribution. In: **Proceedings of the 32nd ACM International Conference on Multimedia**. Association for Computing Machinery, 2024. p. 8480–8488. ISBN 9798400706868. Disponível em: <https://doi.org/10.1145/3664647.3680776>. Citado na página 25.

ZHANG, M.; WANG, H.; HE, P.; MALIK, A.; LIU, H. Improving gan-generated image detection generalization using unsupervised domain adaptation. In: **2022 IEEE International Conference on Multimedia and Expo (ICME)**. [S.l.: s.n.], 2022. p. 1–6. Citado na página 24.

ZHOU, P.; HAN, X.; MORARIU, V. I.; DAVIS, L. S. Learning Rich Features for Image Manipulation Detection. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [s.n.], 2018. p. 1053–1061. [Online; accessed 10-Abril-2025]. Disponível em: <https://arxiv.org/abs/1805.04953>. Citado na página 19.