UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Faculdade de Ciências
Câmpus de Bauru

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FABRÍCIO STEINLE AMOROSO

DETECÇÃO DE DEEPFAKES COM MODELOS DE ESPAÇO DE ESTADO ESTRUTURADO: AVALIAÇÃO DA ARQUITETURA MAMBA

BAURU
2025

FABRÍCIO STEINLE AMOROSO

Detecção de Deepfakes com Modelos de Espaço de Estado Estruturado: Avaliação da Arquitetura Mamba

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Câmpus de Bauru.

Orientador: Prof. Dr. Kelton Augusto Pontara da Costa

BAURU
2025

A524d

Amoroso, Fabrício Steinle
Detecção de Deepfakes com Modelos de Espaço de Estado Estruturado : Avaliação da Arquitetura Mamba / Fabrício Steinle Amoroso. -- Bauru, 2025
49 p. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências, Bauru
Orientador: Kelton Augusto Pontara da Costa

1. Detecção de Deepfakes. 2. Eficiência. 3. Mamba. 4. SSM. 5. Transformers. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE MESTRADO DE FABRÍCIO STEINLE AMOROSO, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO, DA FACULDADE DE CIÊNCIAS - CÂMPUS DE BAURU.

Aos 20 dias do mês de agosto do ano de 2025, às 10h, por meio de Videoconferência, realizou-se a defesa de DISSERTAÇÃO DE MESTRADO de FABRÍCIO STEINLE AMOROSO, intitulada **Detecção de Deepfakes com Modelos de Espaço de Estado Estruturado: Avaliação da Arquitetura Mamba** . A Comissão Examinadora foi constituída pelos seguintes membros: Prof. Dr. KELTON AUGUSTO PONTARA DA COSTA (Orientador(a) - Participação Virtual) do(a) FC / UNESPBauru SP, Prof. Dr. LEANDRO APARECIDO PASSOS JUNIOR (Participação Virtual) do(a) Pós-doutorando do Depto. de Computação / FC/Bauru - Unesp, Prof. Dr. ANDERSON RAYMUNDO ÁVILA (Participação Virtual) do(a) Institut national de la recherche scientifique. Centre Energie, Materiaux, Telecommunications. UMR INRS- UQO. Após a exposição pelo mestrando e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, o discente recebeu o conceito final APROVADO. Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

Documento assinado digitalmente
gov.br
KELTON AUGUSTO PONTARA DA COSTA
Data: 20/08/2025 10:29:42-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. KELTON AUGUSTO PONTARA DA COSTA

Dedico este trabalho à minha esposa, Beatriz, que trilha todos os caminhos ao meu lado. Espero que este trabalho seja um passo, independente do tamanho, em direção a um mundo mais seguro e justo.

Agradecimentos

Agradeço à minha esposa, Beatriz, por estar sempre ao meu lado e reafirmar constantemente que tudo dará certo em nossas vidas, independente dos desafios à nossa frente. Para mim, sua presença diária traz felicidade e é um enorme privilégio tê-la como minha parceira para a vida.

Agradeço aos meus pais, Irene e João, pelo amor incondicional e apoio contínuo, que foram essenciais em todas as etapas da minha vida. Espero honrá-los com cada escolha.

Agradeço também ao Professor Kelton, pelo apoio durante toda a jornada do mestrado. Sob sua orientação me tornei muito mais competente dentro da jornada acadêmica e fui capaz de expandir meus horizontes de pesquisa. Muitas portas se abriram com a sua ajuda.

Resumo

A chegada dos *deepfakes* transformou profundamente o cenário da mídia digital, oferecendo tanto oportunidades quanto desafios significativos. Essas mídias, geradas por técnicas avançadas de aprendizado profundo, são imagens, áudios e vídeos altamente realistas que podem comprometer a confiança em conteúdos digitais. Seu uso malicioso abrange desde manipulação política e desinformação até extorsão, roubo de identidade e violações de propriedade intelectual. Abordagens da comunidade científica para detecção de *deepfakes* de imagem têm sido baseadas principalmente em Redes Neurais Convolucionais (CNNs) e Transformadores, no entanto, existem limitações relacionadas ao alto custo computacional, especialmente quando se trata do processamento de grandes volumes de dados. Neste contexto, este trabalho introduz e avalia o uso do MambaVision, uma arquitetura baseada em Modelos de Espaço de Estado Estruturado (SSMs), partindo da hipótese de que essa abordagem pode constituir uma alternativa competitiva para a detecção de *deepfakes* de imagem, proporcionando ganhos em eficiência computacional sem comprometer a robustez do desempenho. Os resultados experimentais, obtidos a partir de comparações com modelos baseados em CNNs (Xception) e Transformadores (ViT), demonstram que o MambaVision alcançou o maior *throughput* e os tempos totais de teste mais curtos em todos os cenários avaliados. Tudo isso enquanto mantém desempenho competitivo em métricas de acurácia, alcançando os maiores resultados de AUC em todos os conjuntos de dados. Notavelmente, as AUCs de 99.99% e 95.33%, e acurácias de 99.79% e 92.79% nos conjuntos de dados CelebDFv2 e FaceForensics$++$, respectivamente. Em comparação ao Xception e ViT, destacou-se especialmente nas métricas temporais, com *throughput* aproximadamente 22% superior ao Xception e quase 99% superior ao ViT, além de tempos de processamento até 75 vezes menores que o ViT. Esses resultados evidenciam a viabilidade do uso do MambaVision como uma solução prática e eficiente para detecção de *deepfakes* de imagem, contribuindo para a proteção da integridade das mídias digitais.

Palavras-chave: Deepfake, Detecção, Eficiência, Mamba, SSMs

Abstract

The advent of deepfakes has profoundly reshaped the landscape of digital media, presenting both significant opportunities and challenges. Generated through advanced deep learning techniques, these highly realistic images, audio, and videos can undermine trust in digital content. Malicious applications range from political manipulation and misinformation spread to extortion, identity theft, and intellectual property violations. Image deepfake detection approaches proposed by the scientific community have primarily relied on Convolutional Neural Networks (CNNs) and Transformers; however, these methods face limitations due to their high computational cost, especially when processing large volumes of data. In this context, this work proposes and evaluates the use of MambaVision, an architecture grounded in Structured State Space Models (SSMs), based on the hypothesis that this approach can serve as a competitive alternative for image deepfake detection, providing gains in computational efficiency without compromising performance robustness. Experimental results, obtained through comparisons with CNN-based (Xception) and Transformer-based (ViT) models, demonstrate that MambaVision achieved the highest throughput and the shortest overall test times across all evaluated scenarios, while maintaining competitive accuracy and achieving the highest AUC scores in every dataset. Notably, it attained AUCs of 99.99% and 95.33%, and accuracies of 99.79% and 92.79% on the CelebDFv2 and FaceForensics++ datasets, respectively. Compared to Xception and ViT, MambaVision excelled particularly in temporal metrics, with throughput approximately 22% higher than Xception and nearly 99% higher than ViT, along with processing times up to 75 times shorter than ViT. These findings highlight the viability of MambaVision as a practical and efficient solution for image deepfake detection, contributing to the safeguarding of digital media integrity.

Keywords: Deepfake, Detection, Efficiency, Mamba, SSMs

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AUC	Area Under the Curve
CNN	Convolutional Neural Network
CT	Convolutional Traces
DeiT	Data-efficient image Transformers
DF-VAE	DeepFake Variational AutoEncoder
DFDC	DeepFake Detection Challenge Dataset
DiM	Diffusion Mamba
FF++	FaceForensics++
FN	Falso Negativo
FP	Falso Positivo
GAN	Generative Adversarial Network
LFD	Latent Flow Diffusion
PLN	Processamento de Linguagem Natural
PSO	Particle Swarm Optimization
RNN	Recurrent Neural Network
S4	Structured State Space Sequence Models
S6	Selective Structured State Space Sequence Model with a Scan
SS2D	Selective Scan 2D
SSM	State Space Model
ViM	Vision Mamba
ViT	Vision Transformer
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

VSS	Visual State-Space
ZigMa	Zigzag Mamba
ZOH	Zero-Order Hold

Sumário

1 Introdução

A chegada dos *deepfakes* transformou significativamente o cenário da mídia digital, trazendo tanto oportunidades quanto desafios consigo. *Deepfakes*, que são mídias sintéticas criadas para parecerem reais, são geradas por técnicas de aprendizado profundo e tornaram-se um foco de pesquisa devido ao seu impacto potencial na sociedade.

Deepfakes de imagem são criados principalmente usando modelos de aprendizado profundo que se desenvolvem aprendendo com grandes conjuntos de dados. São utilizadas principalmente redes adversárias generativas, *autoencoders* e, mais recentemente, modelos baseados em difusão para produzir mídias cada vez mais realistas, capazes de enganar até mesmo sistemas de detecção sofisticados.

O estado atual do aprendizado de máquina e do aprendizado profundo no contexto de *deepfakes* está evoluindo rapidamente. Avanços recentes levaram ao desenvolvimento de modelos poderosos que podem gerar mídias sintéticas altamente realistas, como StyleGAN (KARRAS; LAINE; AILA, 2019), os trabalhos apresentados por Rossler et al. (2019), Neves et al. (2020), Jiang et al. (2020), Karras et al. (2020), Li et al. (2020), Beckmann, Hilsmann e Eisert (2023), trabalhos baseados em difusão, como Ho, Jain e Abbeel (2020), Rombach et al. (2022), AV et al. (2024), os modelos Imagen e Veo da empresa Google (GOOGLE, 2025), entre outros. Estes métodos possibilitam a produção de imagens e vídeos de alta fidelidade quase indistinguíveis dos reais.

A capacidade de gerar imagens, áudios e vídeos falsos realistas e convincentes resulta em diversos impactos negativos que se extendem desde disseminação de informações falsas para manipulação social, política, perigos à saúde pública, uso indevido de imagem para criação de pornografia não consensual, chantagem, golpes, até riscos à propriedade intelectual e direitos autorais.

Esse progresso também destaca a urgente necessidade de técnicas de detecção igualmente avançadas. A disputa contínua entre as capacidades de geração e detecção é um testemunho da natureza dinâmica deste campo (KAUR et al., 2024). À medida que a qualidade dos *deepfakes* continua a melhorar, também devem melhorar os métodos disponíveis para identificá-los. Atualmente existem diversos métodos baseados em Redes Neurais Convolucionais e Transformadores para detecção de *deepfakes* de imagem. Notavelmente, o uso da XceptionNet (ROSSLER et al., 2019) e de Transformadores de Visão (*Vision Transformers*), como proposto por Khan e Dai (2021), Zhang et al. (2022), Lin et al. (2023), Coccomini et al. (2022) e outros.

É importante notar que o desenvolvimento dos modelos de detecção baseados em aprendizado profundo geralmente requer recursos computacionais extensivos e grandes conjuntos

de dados para alcançar boas métricas. Em especial os métodos baseados em transformadores que, apesar de seus ótimos resultados, possuem eficiência computacional quadrática devido ao mecanismo de atenção, necessitando de capacidade de processamento notável quando recebem longas sequências de entrada, como grandes quantidades de vídeos e imagens de alta resolução (GU; DAO, 2023), (ZHU et al., 2024), (HATAMIZADEH; KAUTZ, 2024).

Mais recentemente, com avanços nos modelos de espaço de estado (*state space models*), uma nova arquitetura nomeada Mamba (GU; DAO, 2023) foi desenvolvida. Apresentando alternativas mais eficientes do que os transformadores devido à sua complexidade computacional linear, enquanto alcança resultados competitivos. Desta forma, com o grande número de *deepfakes* existentes e a gama de métodos disponíveis para geração de mídias sintéticas cada vez maior, se faz relevante buscar técnicas não só com melhores métricas de detecção, como precisão e acurácia, mas que também sejam mais eficientes na análise de grandes quantidades de dados reais e sintéticos, com menores custos computacionais. Nesta busca, o uso e desenvolvimento de abordagens de detecção de imagens e vídeos manipulados baseadas em Mamba se apresenta como uma opção muito promissora.

A principal contribuição desta pesquisa é a aplicação e avaliação do método MambaVision e sua comparação por meio de métricas de precisão e eficiência ao lado de outros métodos estabelecidos no contexto de detecção de *deepfakes* de imagem. No decorrer da pesquisa, busca-se responder se o uso de métodos baseados em Mamba para processamento de imagem são uma boa alternativa para servir como base para modelos de detecção de *deepfakes* de imagem. Alem disso, investiga-se se houveram ganhos em termos de eficiência computacional, enquanto boas métricas de detecção são mantidas neste contexto. Até onde se sabe, este é um dos primeiros trabalhos a utilizar Mamba para esta tarefa. Espera-se que esta pesquisa contribua significativamente para os esforços contra o uso malicioso de *deepfakes*, melhorando a eficiência dos métodos de detecção e protegendo a integridade das mídias digitais.

1.1 Objetivos

O objetivo principal desta dissertação de mestrado é apresentar um dos primeiros casos de uso da nova arquitetura Mamba no contexto de detecção de *deepfakes* de imagem e analisar os resultados obtidos, com ênfase na investigação dos ganhos em eficiência computacional e tempo de processamento em comparação com outros métodos de detecção descritos na literatura.

A motivação desta pesquisa é combater o uso malicioso de *deepfakes*, auxiliando no desenvolvimento de métodos melhores e mais eficientes de detecção.

Os objetivos específicos são:

- Avaliar empiricamente a viabilidade de abordagens baseadas em Mamba para a detecção

de *deepfakes*, investigando suas capacidades, limitações e o equilíbrio entre performance e eficiência na identificação de imagens sintéticas.

- Desenvolver uma estrutura de detecção de *deepfakes* baseado no método MambaVision, adaptando estratégias de treinamento para viabilizar sua aplicação neste contexto.
- Aplicar uma metodologia de testes padronizada, abrangendo desde a preparação dos dados até a condução das avaliações experimentais.

1.2 Organização da Monografia

Esta monografia é apresentada em capítulos, organizados da seguinte forma:

Capítulo 1 : Introdução ao tema da pesquisa, seus objetivos e motivações;

Capítulo 2: Apresenta a fundamentação teórica dos conceitos e métodos importantes para a compreensão da pesquisa;

Capítulo 3: Apresenta a revisão de literatura, contendo trabalhos relacionados ao tema de pesquisa e suas abordagens para o problema em questão;

Capítulo 4: Detalha a metodologia de pesquisa em termos de conjuntos de dados, métricas, experimentos e métodos utilizados;

Capítulo 5: Apresenta e discute os resultados obtidos em métricas, evidenciando comparativamente os pontos fortes e fracos dos métodos apresentados;

Capítulo 6: Contém a conclusão da pesquisa e cita os possíveis trabalhos futuros derivados do tema deste trabalho.

2 Fundamentação Teórica

Este capítulo introduz conceitos importantes para a compreensão dos temas abordados.

2.1 Deepfake

Deepfakes são dados sintéticos que podem ser manipulados ou completamente gerados, com a intenção de parecerem tão autênticos e realistas quanto possível. Esses dados sintéticos são criados através de técnicas de inteligência artificial, em especial, de aprendizado profundo.

Vídeos de *deepfake* são uma preocupação social cada vez maior. Pode-se usar as imagens, vídeos e áudios falsos para disseminação de informações falsas ameaçando, por exemplo, a segurança, privacidade e política, como evidenciado por Kaur et al. (2024). Alguns dos riscos trazidos por *deepfakes* são o roubo e fraude de identidade, chantagem, manipulação de voz e imagem durante autenticações e manipulação, ou mesmo a criação, de evidências falsas (RAO; VERMA; BHATIA, 2021).

Como apresentado na pesquisa de Kaur et al. (2024), *deepfakes* estiveram entre os cinco maiores tipos de fraude de identidade em 2023. Além disso, é apresentada uma pesquisa desenvolvida por uma *startup* focada no problema de identificação de conteúdo falso, DeepMedia, mostrando que entre o perído de 2022 a 2023, o número de *deepfakes* de vídeo triplicou e o de *fakes* de fala aumentou em oito vezes.

Um fator agravante dos problemas gerados pelo uso mal-intencionado de *deepfakes* é que as informações falsas geradas podem alcançar facilmente milhões de pessoas devido ao acesso à internet (WESTERLUND, 2019). Além disso, *deepfakes* foram feitos para serem usados em redes sociais e plataformas de compartilhamento de conteúdo, onde é comum que rumores, conspirações e desinformação se espalhem com facilidade entre os usuários (MASOOD et al., 2023).

Conteúdos de *deepfakes* podem ser de diferentes tipos com base no tipo de dado que está sendo manipulado. Podem ser de imagens, vídeos, áudios e de texto. Existem ainda modificações multimodais, onde mais de um tipo de dado é utilizado, como a sintetização de vídeos acompanhados por áudio. Na Figura 1 é apresentada uma classificação dos tipos principais de *deepfake* com base no tipo de dado, e alguns dos métodos principais de criação.

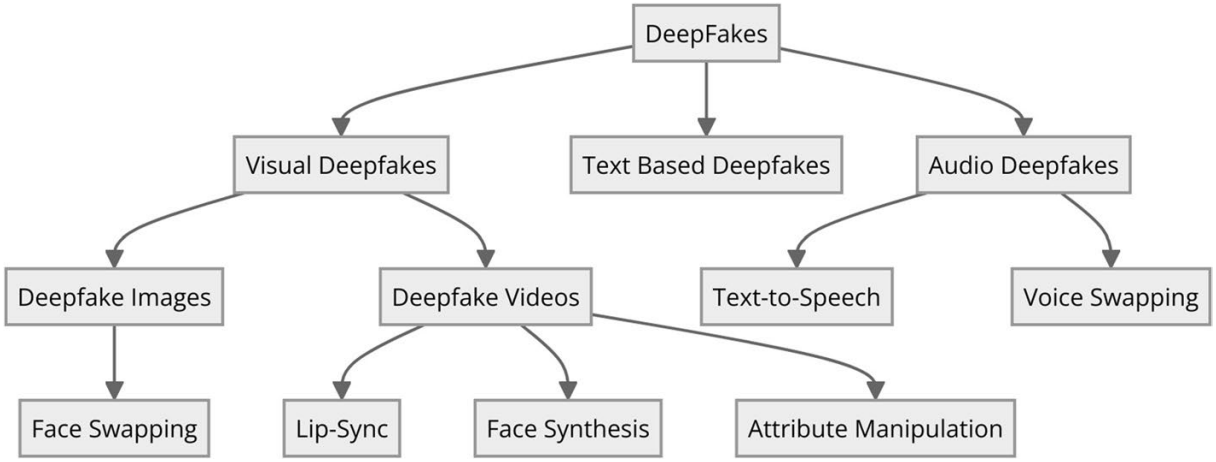

Figura 1 – Classificação hierárquica de deepfakes e métodos de criação de deepfakes visuais e de áudio (KAUR et al., 2024).

Apesar de ser comum ver a tecnologia usada nos *deepfakes* sempre de maneira negativa, vale ressaltar que existem também diversos usos positivos. Como apresentado por Kaur et al. (2024), *deepfakes* podem ser usados em áreas como comunicações digitais, usos nas indústrias de desenvolvimento de conteúdo e comerciais, jogos, filmes, entretenimento e tecnologias para medicina. Alguns exemplos trazidos do uso positivo destas tecnologias:

- Dublagem realista e automatizada de filmes e vídeos em línguas estrangeiras;
- Narração de voz realista e expressões e reações de personagens mais realistas e coordenadas em jogos;
- Sistemas de conferência de vídeo que possibilitem quebrar barreiras linguísticas, fazendo com que todos os integrantes pareçam falar a mesma língua;
- Recuperar digitalmente membros amputados de pacientes ou auxiliar pessoas transgênero a se ver de maneira mais favorável quanto ao seu gênero desejado;
- Auxiliar pessoas que sofrem com doenças de memória, como Alzheimer, a interagir digitalmente com rostos mais novos que eles possam se lembrar.

Infelizmente, apesar do grande potencial para usos positivos, o uso mal-intencionado de *deepfakes* ainda ultrapassa muitos em números o uso com boa intenções (WESTERLUND, 2019). A partir daí surge a necessidade de desenvolver melhores maneiras de combater o uso malicioso dessas mídias sintéticas, auxiliando no desenvolvimento de melhores métodos de mitigação dos seus efeitos negativos.

2.2 SSMs e Mamba

Grande parte dos modelos modernos de detecção e de análise sequencial como processamento de linguagem, áudio, fala, sequências temporais e genoma, entre outros, são baseados predominantemente na arquitetura proposta por Vaswani (2017), o Transformador (ou *Transfomer*) e o seu mecanismo de atenção, que o possibilita modelar dependências presentes em dados complexos dentro de uma janela contextual.

Apesar dos benefícios, o mecanismo de atenção dos transformadores possui desvantagens, como a complexidade quadrática em relação ao tamanho da sequência de entrada, fazendo com que sejam modelos computacionalmente custosos, principalmente quando se trata do contexto de processamento de imagens de alta resolução (TENG et al., 2024).

Para atender a esses desafios computacionais os modelos de espaço de estado foram adaptados para aplicações modernas. SSMs são uma categoria de modelos com o foco em modelagem sequencial, aplicados principalmente em processamentos de sinais e linguagem natural. Modelos SSM mais novos, como o S4 (GU; GOEL; RÉ, 2021), são modelos sequenciais com grande potencial, apresentando capacidades e resultados que se comparam ou superam os transformadores, principalmente quando se trata de grandes sequências de entrada.

O desenvolvimento mais recente em SSMs é o S6, i.e. Mamba (GU; DAO, 2023), que traz um grande salto em comparação ao S4. Mamba apresenta complexidade de tempo linear em relação à entrada e foi projetado para superar ou igualar o desempenho dos transformadores, que apresentam complexidade quadrática.

Sua principal contribuição é um mecanismo inovador de seleção, que permite o processamento eficiente de sequências longas de forma dependente dos dados de entrada. Esse mecanismo possibilita que os parâmetros do modelo, como B e C, sejam ajustados dinamicamente de acordo com as entradas, filtrando informações irrelevantes e adaptando-se melhor à particularidades nos dados.

Mamba foca também em otimização através de operações mais eficientes dentro da GPU e em escalabilidade para processar sequências longas de entrada. Devido a suas características, Mamba se apresenta como uma alternativa promissora para a construção de modelos eficientes e versáteis dentro do domínio dos SSMs.

Apesar dos avanços recentes, o Mamba foi originalmente projetado para lidar com sequências de entrada unidimensionais. Em razão dessa limitação, diversas implementações têm sido desenvolvidas para adaptar o Mamba ao processamento eficiente em tarefas de visão computacional.

Adaptações voltadas para processamento visual, como o ViM (ZHU et al., 2024) e MambaVision (HATAMIZADEH; KAUTZ, 2024), demonstram ótimo desempenho nas referências ImageNet e COCO, oferecendo economias substanciais de memória em comparação ao

DeiT (TOUVRON et al., 2021), por exemplo. Variantes adicionais, como o VMamba (LIU et al., 2024) e o EfficientVMamba (PEI; HUANG; XU, 2024), continuam a refinar a arquitetura para tarefas visuais.

Em específico, MambaVision (HATAMIZADEH; KAUTZ, 2024) é uma arquitetura híbrida que nos primeiros estágios de sua estrutura, usa blocos convolucionais que possibilitam a extração eficiente de características. Já seus estágios finais são compostos por $N/2$ blocos MambaVision e $N/2$ blocos de *self-attention* dos transformadores, dadas N camadas, aprimorando a capacidade de captura de dependências espaciais de longo alcance e, consequentemente, melhores resultados no processamento de características visuais presentes em imagens. Tudo isso enquanto mantém a eficiência computacional inerente do Mamba.

Os blocos MambaVision, por sua vez, são uma adaptação dos blocos Mamba originais, nos quais são substituidas as camadas convolucionais causais por camadas de convolução regulares. Além disso, o bloco é dividido entre dois caminhos simétricos, com a única diferença entre os dois sendo a ausência do uso de SSM em um deles. Para a saída do bloco, as saídas dos dois caminhos são concatenadas e projetadas para uma camada linear final. Essas modificações resultam em uma melhoria na modelagem de contextos globais. Para os objetivos deste trabalho, MambaVision será utilizado como base para as adaptações necessárias para a tarefa de detecção de imagens sintéticas.

2.2.1 Definições Matemáticas

State Space Models foram inicialmente desenvolvidos como um sistema contínuo e invariante a tempo para processar sequências de 1 dimensão, onde um sinal de entrada $x(t) \in \mathbb{R}$ é primeiramente codificado para um vetor de estado latente oculto $h(t) \in \mathbb{R}^N$ e depois decodificado no sinal de saída $y(t) \in \mathbb{R}$ de acordo com a Equação 1:

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t) \\ y(t) &= Ch(t) \end{aligned}, \tag{1}$$

onde h' é a derivada de h e $\mathbf{A} \in \mathbb{R}^{N \times N}$ e $\mathbf{B}, \mathbf{C} \in \mathbb{R}^N$ correspondem à matriz de estado, matriz de entrada e matriz de saída, respectivamente, sendo interpretados como os pesos do SSM.

Mais recentemente, surgiram os S4 (GU; GOEL; RÉ, 2021), que são uma classe de modelos sequenciais para aprendizado profundo inspirados em Redes Neurais Recorrentes, Redes Neurais Convolucionais e no sistema SSM clássico descrito.

Os modelos S4 são definidos por quatro parâmetros (Δ, A, B, C), que vão guiar a transformação da sequência $x(t) \in \mathbb{R} \mapsto y(t) \in \mathbb{R}$.

S4 e Mamba (GU; DAO, 2023) são versões discretizadas do sistema contínuo de SSM descrito em 1 e utilizam Δ como um parâmetro de escala de tempo para fazer a discretização. Nota-se que as entradas do modelo em tarefas de processamento de linguagem natural e de imagens de visão computacional com 2 dimensões, geralmente são valores discretos (TENG et al., 2024), sendo assim a discretização é realizada tendo em vista a eficiência computacional.

Matematicamente, para transformar o sistema de parâmetros contínuos (Δ, A, B) em um sistema de parâmetros discretos ($\bar{A}$, $\bar{B}$), segue-se a seguinte formulação: $\bar{A} = f_A(\Delta, A)$ e $\bar{B} = f_B(\Delta, A, B)$, onde o par ($f_A$, f_B) é chamado de regra de discretização.

A regra de discretização utilizada por Mamba é *zero-order hold*, demonstrada na Equação 2:

$$\begin{aligned} \bar{A} &= \exp(\Delta A) \\ \bar{B} &= (\Delta A)^{-1} \left(\exp(\Delta A) - I\right) \cdot \Delta B \end{aligned}. \tag{2}$$

Após os parâmetros serem discretizados $(\Delta, A, B, C) \mapsto (\bar{A}, \bar{B}, \bar{C})$, o modelo pode ser computado de duas formas, como uma recorrência linear, Equação 3:

$$\begin{aligned} h_t &= \bar{A} h_{t-1} + \bar{B} x_t \\ y_t &= C h_t \end{aligned}, \tag{3}$$

ou como uma convolução global, Equação 4:

$$\begin{aligned} \bar{K} &= \left(C\bar{B},\, C\bar{A}\bar{B},\, \ldots,\, C\bar{A}^k\bar{B}\right) \\ y &= x * \bar{K} \end{aligned}. \tag{4}$$

Com as equações obtidas, o modelo utiliza o modo convolucional para calcular a saída através de paralelização eficiente, onde toda a sequência de entrada pode ser observada, depois alterna para o modo recorrente para inferência autoregressiva, onde as entradas são observadas uma por vez.

2.3 Aprendizado Profundo

Aprendizado profundo (ou *deep learning*) é uma ramificação do aprendizado de máquina e inteligência artificial. Os métodos de aprendizado profundo possuem múltiplas camadas de processamento, possibilitando a representação de dados com vários níveis de abstração.

Deep learning é muito utilizado em tarefas de visão computacional e de reconhecimento de padrões devido à sua grande capacidade de reconhecer padrões em dados com múltiplas

dimensões automaticamente, tarefa que exigiria muito mais esforço com modelos tradicionais de aprendizado de máquina.

Entretanto, para que estes modelos de aprendizado profundo possam aprender a reconhecer consistentemente esses padrões nos dados, eles normalmente precisam ser treinados com grandes conjuntos de dados rotulados. Tanto estes conjuntos de dados quanto o treino dos modelos profundos exige custos computacionais e de tempo relevantes (ZHANG, 2022).

2.4 Transformadores

A principal característica de um transformador é o mecanismo de atenção *multi-head* (VASWANI, 2017). Este mecanismo é o que possibilita ao modelo aprender relações entre os elementos da sequência de entrada. Através de sua forte capacidade de representação e análise em paralelo das entradas, *transformers* se destacam no aprendizado de dependências entre as entradas.

O uso desta arquitetura trouxe ótimos resultados, muitas vezes comparáveis ou melhores do que o estado da arte. Inicialmente o modelo foi proposto para ser utilizado em tarefas de processamento de linguagem natural, específicamente para tradução de textos entre inglês-alemão, substituindo as redes neurais recorrentes e convolucionais, as principais redes que eram utilizadas neste contexto, e trazendo melhores resultados, com melhor paralelização e menos tempo necessário para treinamento. Até hoje são amplamente utilizados em diversas tarefas de PLN, como em aplicações para classificação e segmentação de textos. Existem diversos modelos relevantes baseados em transformadores, entre eles BERT (KENTON; TOUTANOVA, 2019), GPT-3 (BROWN, 2020) e BODE (GARCIA et al., 2024).

O sucesso dos transformadores também foi expandido para outros contextos além de processamento de linguagem, passando também a ser utilizado em diversas abordagens focadas em visão computacional para detecção de objetos, classificação e segmentação de imagens, entre outras aplicações. Especificamente, o ViT (ALEXEY, 2020) utiliza o mecanismo de atenção, dividindo uma imagem em fragmentos que são utilizados como *tokens* na sequência de entrada. Os modelos baseados em transformadores também alcançam desempenho do estado da arte, entretanto o mecanismo de atenção dos transformadores escala de forma quadrática, podendo se tornar uma operação extremamente custosa à medida que o tamanho da sequência de entrada aumenta.

2.5 Redes Neurais Convolucionais

Redes neurais convolucionais (GOODFELLOW et al., 2014) estão entre as arquiteturas de redes neurais profundas mais populares, muito aplicadas para classificação e reconhecimento de imagens e tarefas de visão computacional. CNNs são compostas de camadas convolucionais,

onde é aplicado o processo de convolução, a etapa mais característica dessas redes, bem como camadas de *pooling*, também conhecida como camada de agrupamento, e de camadas totalmente conectadas (ZHANG, 2022).

O diferencial das CNNs reside nas camadas convolucionais, que aplicam filtros e identificam padrões nos dados de entrada, introduzindo não linearidades por meio de funções de ativação. Esse processo é crucial para que a rede aprenda padrões complexos e relevantes em diferentes níveis de abstração, o que a torna especialmente eficaz em tarefas de visão computacional, como detecção de objetos, segmentação de imagens e classificação de cenas.

No contexto de *deepfakes* as CNNs são comuns tanto no processo de geração quanto de detecção de imagens sintéticas, aplicadas para extrair detalhes visuais de rostos e gerar imagens sintéticas a partir das informações obtidas.

3 Revisão de Literatura

Este capítulo apresenta trabalhos recentes e relevantes publicados na área de interesse desta pesquisa. A seção 3.1 apresenta trabalhos que abordam a geração de *deepfakes*; Já a seção 3.2 apresenta trabalhos focados principalmente na detecção de *deepfakes*; Por fim, 3.3 apresenta trabalhos relacionados a Mamba.

3.1 Geração de Deepfakes

São apresentados alguns dos trabalhos mais relevantes e recentes que apresentam metodologias para geração de *deepfakes*. Algumas destas pesquisas resultaram na publicação de conjuntos de dados que podem ser utilizados para complementar o desenvolvimento dos detectores atuais.

Karras, Laine e Aila (2019) desenvolveram uma abordagem chamada StyleGAN que se beneficia de técnicas de transferência de estilo aliadas a uma arquitetura de rede generativa adversarial, possibilitando a extração automática de atributos de alto-nível, como pose, expressão e identidade quando treinados no contexto rostos humanos. Utilizando imagens de alta qualidade no seu treinamento, pode-se obter como resultado desse processo faces sintetizadas realistas a partir da combinação das características de pose, estilo de cabelo, formato de uma face base A e dos traços faciais mais discretos e cores de pele, cabelo e iluminação de outra face B. Posteriormente, os autores melhoraram a arquitetura de seu modelo e revisaram a metodologia de treino, resultando no StyleGAN2 (KARRAS et al., 2020), que corrige algumas imperfeições existentes nas imagens sintetizadas pelo StyleGAN, melhorando ainda mais a qualidade visual.

Rossler et al. (2019) propuseram FaceForensics++, um conjunto de dados com faces falsas e rótulos para auxiliar no desenvolvimento de técnicas de detecção supervisionadas. Para a geração do conjunto foram utilizados 1000 vídeos reais extraídos do YouTube, sobre os quais aplicaram quatro métodos de manipulação de faces: Face2Face, FaceSwap, DeepFakes e NeuralTextures, resultando em mais de 1.8 milhões de imagens falsas. Além disso, é proposto um método de detecção baseado em CNNs aliada a uma técnica de rastreamento de face para extração das áreas de interesse das imagens.

Neves et al. (2020) traz uma nova estratégia para aprimoramento da qualidade de *deepfakes* chamada GANprintR, baseada na remoção de resquícios deixados pelo processo de geração de *fakes* com GANs através uma etapa posterior utilizando Autoencoders treinados em dados reais de alta qualidade. Esta etapa adicional complementa as imagens geradas trazendo traços mais realistas. Através do método proposto foi criado o conjunto de dados iFakeFaceDB.

Pensando na qualidade das falsificações geradas, Jiang et al. (2020) desenvolveram o

conjunto de dados de falsificação de rostos DeeperForensics-1.0, contendo 60.000 vídeos através de uma nova estrutura de troca de rosto, o DeepFake Variational AutoEncoder. O DF-VAE consiste de três módulos principais: módulo de extração da estrutura da face, um segundo módulo responsável pela separação dos atributos de expressão e pose e das características físicas da face, como cor da pele e textura e, por fim, um módulo de fusão que permite utilizar os atributos extraídos de uma face para serem aplicados em outra.

Li et al. (2020) apresentam Celeb-DF, um conjunto contendo 5.639 vídeos *deepfake*, buscando trazer alta qualidade nas falsificações focando no problema de que os outros conjuntos de dados apresentam muitos artefatos visuais. Seu método foi baseado no algoritmo de síntese DeepFake que utiliza *Autoencoders*, mas com melhorias: aumento na resolução das faces geradas de 64×64 ou 128×128 pixels para 256 × 256 pixels através da adição de mais camadas na arquitetura utilizada de Autoencoder; melhorias no problema de incoerência de cor entre as faces originais e as geradas, que foi alcançado através da realização de perturbações aleatórias nas cores das faces de treino, forçando a rede profunda a gerar imagens com o mesmo padrão de cores da imagem de entrada; máscaras para extração da face mais precisas, reduzindo os artefatos encontrados próximos das bordas das faces; e redução da cintilação entre *frames* por meio de uma filtragem utilizando um algoritmo de Kalman para suavização.

Dolhansky et al. (2020) propuseram o DeepFake Detection Challenge Dataset, um conjunto de dados contendo mais de 100.000 vídeos *deepfake* gerados através de diversos métodos populares de *face swapping* com variações entre o nível de realismo das imagens, entre eles: DFAE, MM/NN face swap, NTH (ZAKHAROV et al., 2019), FSGAN (NIRKIN; KELLER; HASSNER, 2019) e StyleGAN (KARRAS; LAINE; AILA, 2019).

Beckmann, Hilsmann e Eisert (2023) apresentam uma arquitetura para *face swapping* utilizando *Autoencoder* em conjunto com uma técnica avançada de combinação de faces, com o objetivo de gerar um conjunto de dados contendo 90 *deepfakes* de alta qualidade. Para verificar o impacto dos *fakes* gerados, detectores do estado da arte foram colocados à prova, resultando em uma queda drástica de performance quando comparados com *fakes* de outros conjuntos de dados. Ainda, foi testado o desempenho realizando o ajuste fino dos detectores com os *fakes* gerados, resultando em uma melhor performance do detector. A partir disso, os autores realçam a importância de que os detectores também complementem seu treinamento com *deepfakes* de alta qualidade e que treinamentos baseados apenas em conjuntos de dados de pesquisa nem sempre refletem os dados encontrados no mundo real.

As abordagens mais recentes de geração de imagens e *deepfakes* são baseadas em modelos de difusão, obtendo representações verossímeis. Ho, Jain e Abbeel (2020) utilizam modelos probabilísticos de difusão inspirados pela termodinâmica de não equilíbrio para síntese de imagens de alta qualidade. No conjunto de dados CIFAR10, o modelo obteve um Inception score de 9,46 e um FID do estado da arte de 3,17.

Pensando nos altos custos de otimização de modelos de difusão, Rombach et al. (2022)

utilizam o espaço latente de *autoencoders* pré-treinados para possibilitar o treinamento com recursos computacionais limitados enquanto mantêm a qualidade das gerações, possibilitando encontrar um melhor balanço entre redução de complexidade e preservação de detalhes. Os resultados desta arquitetura alcançaram um novo estado da arte para sintetização e manipulação de imagens, enquanto reduz os custos computacionais em relação a métodos anteriores de difusão.

É proposto por AV et al. (2024) uma nova técnica de geração de *deepfakes* chamada Latent Flow Diffusion, capaz de gerar vídeos a partir da imagem de um alvo e um vídeo fonte existente. A abordagem proposta utiliza uma sequência de fluxo óptico no espaço latente, com base no vídeo da fonte, para distorcer a imagem do alvo, resultando em uma aparência espacial realista e dinâmicas temporais consistentes. Os experimentos demonstraram que o modelo LFD supera consistentemente os modelos anteriores de difusão de vídeo para geração de *deepfakes*.

Recentemente, um grande avanço em geração de mídia sintética foi alcançado pelo setor de pesquisa em inteligência artificial da Google (GOOGLE, 2025), com o anúncio dos modelos de acesso fechado Imagen 4 [1], Veo 3 [2] e Lyria 2 [3]. Esses modelos avançados, baseados em difusão latente, são capazes de gerar imagens, vídeos com áudio e música, respectivamente, de qualidade extremamente alta a partir de simples *prompts* como entrada. Embora não sejam voltados necessariamente para a geração de *deepfakes*, ainda são motores poderosos para esta tarefa. Sendo assim, é necessário que os métodos de detecção continuem evoluindo para seja possível distinguir entre realidade e mídias criadas por estes modelos, que representam o estado da arte atual neste contexto.

3.2 Detecção de Deepfakes

Existem diversas abordagens possíveis para o desafio de detecção de *deepfakes*. Nesta seção são apresentadas algumas das metodologias existentes na literatura.

Rossler et al. (2019) apresentam uma abordagem para detecção de imagens sintéticas baseada na arquitetura XceptionNet (CHOLLET, 2017) e utilizando o conjunto de dados proposto no mesmo trabalho, FaceForensics++. O modelo foi treinado e avaliado no conjunto de dados com diferentes níveis de qualidade, apresentando acurácia de 99.26% no FF++ sem compressão, 95.73% com compressão e imagens de alta qualidade (FF++ C23) e 81% com maior compressão e imagens de menor qualidade (FF++ C40).

Nguyen, Yamagishi e Echizen (2019) utilizam uma arquitetura que conta com uma rede VGG-19 utilizada para extração de atributos que, por sua vez, serão utilizados para alimentar uma CapsNet que irá diferenciar entre imagens reais ou falsas. O método proposto alcança as

[1] https://deepmind.google/models/imagen/
[2] https://deepmind.google/models/veo/
[3] https://deepmind.google/models/lyria/

acurácias máximas de 95.93% na detecção a nível de quadros e de 99.23% na detecção a nível de vídeo no conjunto de dados proposto por Afchar et al. (2018).

Guarnera, Giudice e Battiato (2020) afirmam que a detecção através de CNNs não é tão robusta, apresentando bom desempenho apenas no contexto em que foi treinada. Desta forma, é proposta uma nova abordagem baseada em um algoritmo de Expectation-Maximization treinado para extrair os traços convolucionais (*convolutional traces*) deixados durante o processo de geração de imagens por meio de GANs. Esta técnica de detecção baseada nos CT se manteve robusta quando testada com diversos tipos de ataques, alcançando acurácia média de 98% ao detectar mais de 10 arquiteturas diferentes de GANs. Além disso, atingiu 93% de acurácia contra *fakes* gerados por meio do FACEAPP. A utilização dos CT dá indícios não só de que a imagem foi gerada por uma GAN, portanto é sintética, mas pode ser útil até mesmo para identificar qual foi a arquitetura específica de GAN utilizada. Além disso, em comparação com os métodos baseados em redes neurais, esta abordagem é computacionalmente mais rápida.

Na tarefa de detecção de imagens geradas e manipuladas, Baek, Yoo e Bae (2020) apresentam uma nova abordagem através de um *ensemble* de redes generativas adversárias focadas principalmente na habilidade do discriminador de detectar imagens sintéticas. Quando colocada a prova contra os métodos DeepFake (DEEPFAKES,), Face2Face (THIES et al., 2016), Faceswap (FACESWAP,) e NeuralTextures (THIES; ZOLLHÖFER; NIESSNER, 2019), esta abordagem foi capaz de alcançar desempenho competitivo e próximo de métodos do estado da arte, como Xception (ROSSLER et al., 2019) e MesoNet (AFCHAR et al., 2018) e superando outros métodos. Sua aparente principal vantagem é uma redução significativa do viés para imagens reais e falsas em comparação com outros métodos.

O trabalho FakeLocator, desenvolvido por Huang et al. (2022), vai além da tarefa de detectar *deepfakes*, é capaz ainda de localizar onde existem manipulações na imagem. Isto é alcançado através de um mapeamento em escala de cinza capaz de evidenciar inconsistências de textura na imagem geradas pela etapa de *upsampling* de GANs. A eficácia deste método inovador foi testada nos conjuntos de dados FaceForensics$++$ (ROSSLER et al., 2019) e DFFD (DANG et al., 2020), bem como diversos outros métodos de manipulação como: STGAN (LIU et al., 2019), StarGAN (CHOI et al., 2018), AttGAN (HE et al., 2019), StyleGAN2 (KARRAS et al., 2020), entre outros.

Para a tarefa de detecção de vídeos *deepfake*, Al-Adwan et al. (2024) utilizam uma rede neural híbrida de CNN e RNN com um algoritmo de otimização de enxame de partículas (*particle swarm optimization*). O método proposto foi testado nos conjuntos de dados Celeb-DF e DFDC, alcançando uma acurácia média de 97.26% e 94.2%, respectivamente.

Outra abordagem que também se beneficia de meta-heuristicas é o método desenvolvido por Cunha et al. (2024). É proposto um método híbrido utilizando EfficientNet-Gated Recurrent Unit (GRU), aprendizado de transferência baseado em EfficientNet-B0 para a tarefa de classificação de vídeos falsos e uma variação do algoritmo PSO para auxiliar na otimização

dos hiperparâmetros utilizados pelas redes. Nos experimentos apresentados pela pesquisa, o método proposto otimizado com o algoritmo PSO foi capaz de performar melhor que outros métodos para diversos conjuntos de dados de *deepfake*.

São desenvolvidas abordagens de detecção de *deepfakes* recentes e poderosas com base em transformadores. Khan e Dai (2021) apresentam um detector de vídeos *deepfake* baseado em um transformador combinado com aprendizado incremental. Além disso, o método utiliza técnicas de reconstrução facial 3D para gerar mapas de textura UV a partir de imagens de rosto, possibilitando capturar melhor características relevantes do rosto analisado. A eficácia desta abordagem foi testada em 5 conjuntos de dados públicos, ultrapassando resultados do estado da arte, alcançando as melhores acurácias de 99.79%, 99.28% e 91.69% nos conjuntos FaceForensics$++$, DeepFake Detection e DFDC respectivamente.

É apresentado por Zhang et al. (2022) uma rede robusta e eficiente baseada em transformadores chamada TransDFD. O método é capaz de capturar traços sutis de falsificação e implementa um módulo de atenção espacial escalável, capaz de destacar características mais relevantes enquanto suprime outras menos importantes. Além disso, possui um módulo *transformer* que extrai características locais e globais refinadas. É uma abordagem proposta para ser computacionalmente eficiente. Sua eficiência e resultados superam abordagens do estado da arte em robustez e eficiência computacional em diversos conjuntos de dados públicos, apresentando AUC de 98.40% no conjunto FaceForensics$++$.

Um novo método baseado em CNN e ViT é proposto por Lin et al. (2023). A abordagem foca em melhorar o desempenho em conjuntos de dados de baixa qualidade e em avaliações entre diferentes conjuntos de dados. O modelo incorpora um módulo de convoluções multiescala capaz de extrair características intrínsecas que podem revelar manipulações nos rostos analisados. As características extraídas são utilizadas pelo transformador, possibilitando que aprenda melhor as relações contextuais presentes no rosto, sendo utilizado para realizar a classificação das imagens. Experimentos mostram que o método supera modelos modernos em detecção tanto em dados de alta quanto de baixa qualidade, além de apresentar boa capacidade de generalização entre conjuntos de dados, alcançando os melhores resultados no conjunto Celeb-DF, com acurácia de 99.80%.

Zhang et al. (2022) exploram o uso de uma arquitetura com transformador para análise de inconsistências espaço-temporais em regiões faciais ao longo dos quadros. A proposta reorganiza os vídeos em *patches*, que são processados por um ViT. Para melhorar a robustez e generalização, é introduzida uma operação de *dropout* espaço-temporal que explora pistas espaço-temporais e atua como uma forma de aumento de dados. Experimentos demonstram que a abordagem supera 25 métodos do estado da arte, mostrando impressionante robustez, generalização e capacidade de representação, com destaque para as pontuações alcançadas em AUC de 99.8%, 99.1% e 97.2% nos conjuntos FaceForensics$++$, DFDC e Celeb-DF.

Coccomini et al. (2022) utilizam uma combinação de EfficientNet-B0 como extrator de

características multiescala e a combinação de dois transformadores visuais para detecção de *deepfakes* em vídeos. É empregado um procedimento de inferência simples com votação para processar múltiplos rostos em uma mesma cena de vídeo. O modelo alcançou AUC de 95.1% e *F1-score* de $88,0\%$.

Khan e Dang-Nguyen (2022) apresentam uma rede híbrida de transformador que utiliza uma estratégia de fusão de características para detecção de vídeos *deepfake*. As redes XceptionNet and EfficientNet-B4 são utilizados como extratores de características junto com o transformador. Além disso, são propostas duas novas técnicas de aumento de dados, *face cut-out* e *random cut-out*, que melhoram o desempenho do modelo e reduzem o *overfitting*. Os melhores resultados obtidos pela arquitetura foram utilizando a técnica *random cut-out*, alcançando o melhor valor de acurácia de 98.24% no conjunto DFDC e apresentando também que o modelo proposto pode aprender eficazmente com quantidades limitadas de dados.

É apresentado por Heo, Yeo e Kim (2023) um modelo transformador visual robusto, combinando características extraídas por CNN com *patch-embedding*, permitindo a interação entre todas as posições para identificar regiões de artefatos. Os melhores resultados foram obtidos no conjunto Celeb-DFv2, alcançando 99.3% de AUC e 97.8% de *F1-score*.

3.3 Mamba

Nesta seção serão apresentados alguns dos artigos mais relevantes que possuem abordagens baseadas em Mamba e SSMs.

Mamba, um novo modelo SSM foi proposto por Gu e Dao (2023), propondo superar as limitações de eficiência computacional dos transformadores para processamento de sequências longas, possibilitando complexidade computacional linear, em contraste com complexidade quadrática do mecanismo de atenção encontrado em transformadores. Para este fim, Mamba utiliza parâmetros dependentes dos dados de entrada, possibilitando raciocínio baseado no conteúdo por SSMs e um algoritmo que possibilita processamento paralelo otimizado. Apresenta ainda um mecanismo de seleção melhorado, como uma alternativa do mecanismo de atenção. Mamba demonstra desempenho comparável ao estado da arte em modalidades como processamento de linguagem, áudio e genômica, fazendo isso com escalabilidade linear e $5\times$ mais *throughput* que transformadores. É apresentado ainda que, na tarefa de modelagem de linguagem, o Mamba-3B supera transformadores de tamanho equivalente, alcançando desempenho comparável a modelos duas vezes maiores.

Diferentes abordagens foram propostas com o objetivo de adaptar o uso de Mamba para tarefas visuais, tendo em vista que em seu estado original, é limitado ao processamento de sinais de 1 dimensão. Entre essas abordagens está Vision Mamba (ZHU et al., 2024), uma nova arquitetura para visão computacional baseada em modelos de estado, utilizando blocos Mamba bidirecionais. O ViM supera a necessidade de *self-attention* em tarefas visuais, combi-

nando *embeddings* de posição e compressão eficiente de representações visuais com os SSMs bidirecionais. Em *benchmarks* como ImageNet, COCO e ADE20k, ViM supera transformadores como o DeiT (TOUVRON et al., 2021) em desempenho, eficiência computacional e uso de memória, sendo 2,8 vezes mais rápido e economizando 86.8% de memória em imagens de alta resolução. Esses resultados destacam seu potencial como próxima geração de *backbones* para modelos de visão.

Hatamizadeh e Kautz (2024) desenvolveram MambaVision, um modelo híbrido que combina a arquitetura Mamba com blocos de transformadores de visão para tarefas de visão computacional. O trabalho apresenta uma reformulação do Mamba, adicionando blocos de *self-attention* nas camadas finais para capturar melhor dependências espaciais de longo alcance. O resultado disso é um processamento mais eficiente de características visuais. MambaVision atinge resultados do estado da arte em classificação de imagens no conjunto ImageNet-1K em acurácia e *throughput*, superando arquiteturas de tamanho similar em tarefas como detecção de objetos, segmentação de instâncias e segmentação semântica nos conjuntos MS COCO e ADE20K.

Liu et al. (2024) apresentam o *backbone* de visão VMamba, baseado em Mamba e projetado para eficiência computacional com complexidade linear. No núcleo do VMamba há uma pilha de blocos de Espaço de Estado Visuais (Visual State-Space) com módulos de Selective Scan 2D. Os módulos SS2D servem para integrar os dados de diferentes dimensões provenientes do scan seletivo, que são de 1 dimensão e os dados de imagem de 2 dimensões. A arquitetura demonstra desempenho promissor em diversas tarefas de percepção visual, com destaque para sua escalabilidade superior em comparação com outras abordagens.

O trabalho de Pei, Huang e Xu (2024) propõe EfficientVMamba, um modelo leve e eficiente que combina a capacidade de extração de informações globais e locais utilizando blocos VSS. EfficientVMamba adota uma varredura seletiva com amostragem eficiente (*atrous-based selective scan*), possibilitando a representação eficiente tanto de características globais quanto locais. Além disso, o modelo integra blocos SSM com convoluções para aprimorar ainda mais o desempenho. Experimentos mostram resultados competitivos em tarefas visuais e com menor custo computacional, com o EfficientVMamba-S com 1.3G FLOPs superando o Vim-Ti com 1.5G FLOPs em 5,6% de acurácia no ImageNet.

Abordagens baseadas em Mamba estão sendo desenvolvidas também para o contexto de geração de imagens. Alguns trabalhos em específico focam no uso de Mamba e técnicas de difusão para essa tarefa, como DiM (TENG et al., 2024), que combina a eficiência de Mamba com as capacidades dos modelos de difusão para síntese eficiente de imagens em alta resolução. Uma estratégia de treinamento *weak-to-strong* é adotada, pré-treinando o modelo em imagens de baixa resolução antes de ajustá-lo em resoluções maiores. Além disso, são utilizadas técnicas de *upsampling*, possibilitando a geração de imagens de maior resolução sem a necessidade de treinamento adicional.

No mesmo contexto, ZigMa (HU et al., 2024) aborda as limitações de escalabilidade e complexidade quadrática dos modelos de difusão, especialmente os baseados em transformadores, propondo o uso de Mamba para geração de imagens. O estudo aborda um ponto de melhoria dos métodos visuais baseados em Mamba, a continuidade espacial no esquema de varredura. A partir disso é proposto ZigMa, uma solução que apresenta melhorias em velocidade e utilização de memória em comparação com métodos baseados em transformadores. A escalabilidade da arquitetura foi testada nos conjuntos FacesHQ 1024×1024, UCF101, MultiModal-CelebA-HQ e MS COCO 256×256.

Existem ainda trabalhos que aplicam Mamba em diferentes tarefas para processamento e classificação de imagens. Vim4Path (NASIRI-SARVI et al., 2024), utiliza ViM e um *framework* DINO para aprendizado de representações em patologia computacional, alcançando resultados superiores em relação ao ViT utilizando o conjunto de dados Camelyon16, com aumento de 8,21 na métrica ROC AUC. Adicionalmente, foi conduzida uma análise de explicabilidade que mostrou alinhamento do ViM com o fluxo de trabalho de patologistas, reforçando seu potencial para diagnósticos práticos.

Ainda no contexto de estudos médicos, Tao, Wang e Li (2024) propõem uma rede de fusão de características *dual-branch* para a classificação de imagens gastrointestinais, combinando ResNet50, Feature Pyramid Network e Vision Mamba. O modelo utiliza um módulo de fusão com atenção espacial, atenção por canal e conectividade residual para integrar características globais e locais extraídas pelos dois ramos. Os experimentos realizados com os conjuntos de dados Kvasir-Dataset e Gastrovision mostram que o modelo proposto supera métodos existentes em performance.

Em aplicações industriais, Hue et al. (2024) apresentam ConvMamba, uma nova arquitetura de rede neural projetada para a classificação de defeitos em rolamentos de maquinário. O modelo inclui melhorias no processamento de sinais para aumentar a robustez das previsões. Foram apresentadas comparações com o modelo Hybrid-CNN-MLP, nas quais ConvMamba demonstrou um aumento de até 33,56% na acurácia em condições de carga variáveis, alcançando um máximo de 99,20%.

Os trabalhos RawBMamba (CHEN et al., 2024b) e BiCrossMamba-ST (KHEIR; POLZEHL; MÖLLER, 2025) apresentam abordagens baseadas em Mamba para detecção de áudios sintéticos e *deepfakes* de fala. RawBMamba é capaz de capturar características de curto alcance de áudios forjados, através de camadas sinc (SincNet) e convolucionais, e de longo alcançe através de uma adaptação bidirecional de Mamba. Enquanto BiCrossMamba-ST utiliza uma análise de espectro de áudio por meio de blocos Mamba bidirecionais e blocos de atenção cruzada, capturando indicativos sutis de fala sintética nos áudios. As duas abordagens apresentam performance competitiva em conjuntos de dados de áudios forjados.

Os autores Peng et al. (2025) desenvolveram WMamba, que utiliza a arquitetura Mamba como base para realizar análises de imagens por meio de *wavelets*, possibilitando a

captura e identificção de características de imagens forjadas que são imperceptíveis através apenas da análise espacial. Este método inovador alcançou AUC de 96.29% e 82.97% na detecção de faces forjadas nos conjuntos de dados CelebDFv2 e DFDC, respectivamente.

Para detecção de vídeos sintéticos, o trabalho de Chen et al. (2024a) introduz um módulo chamado Detail Mamba (DeMamba), focado em melhorar a detecção de conteúdo gerado por inteligência artificial através da análise de inconsistências espaciais e temporais em vídeos. Os resultados de seus experimentos apresentam grande robustez e capacidade de generalização. Além do módulo de detecção, o trabalho apresenta um novo conjunto de dados em larga escala gerado por inteligência artificial, GenVideo, contendo tanto vídeos reais quanto sintetizados por uma grande variedade de técnicas de geração.

4 Metodologia

Este capítulo apresenta a metodologia de aquisição e processamento dos conjuntos de dados, configuração dos modelos e condução dos experimentos. A Figura 2 apresenta o diagrama da metodologia.

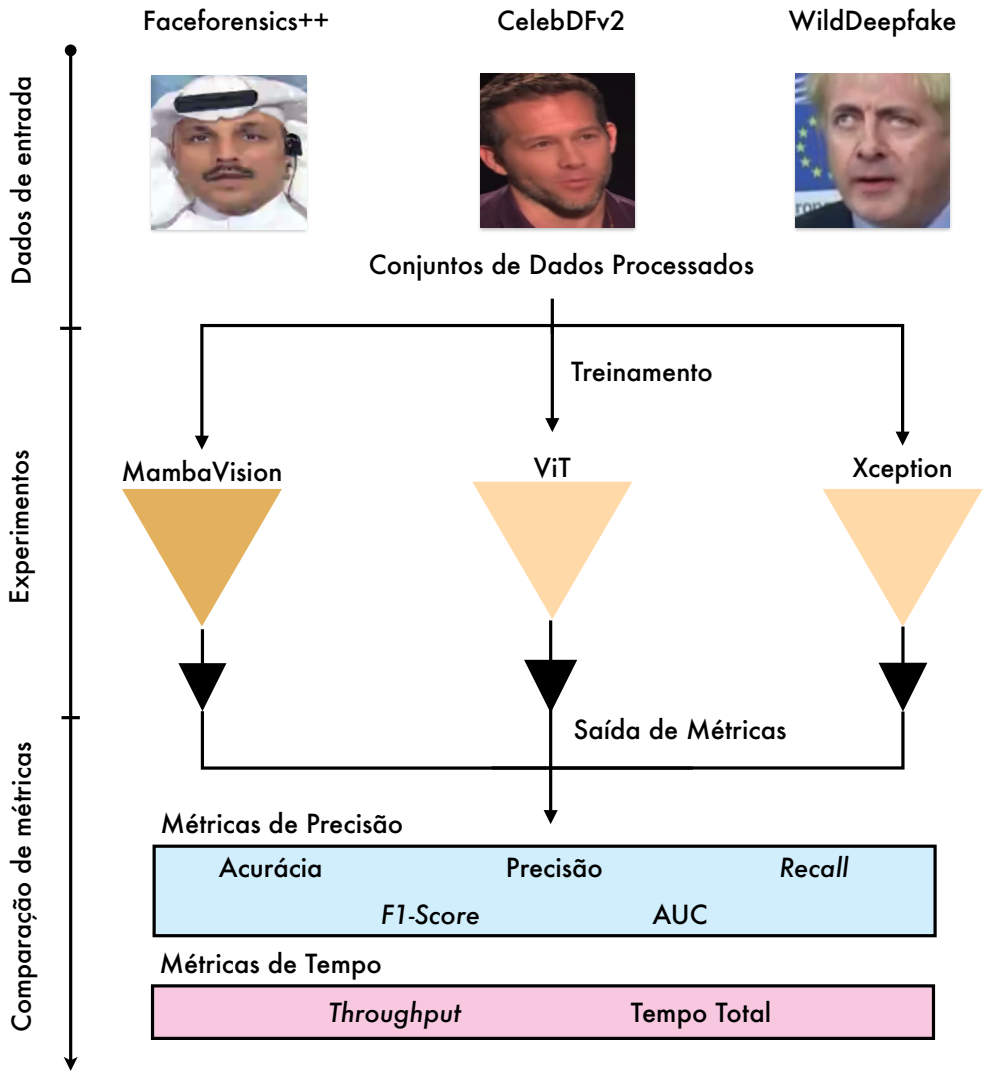

Figura 2 – Metodologia adotada para os experimentos. Os conjuntos de dados FaceForensics++, CelebDFv2 e WildDeepfake, após pré-processamento, são usados como dados para treinamento dos três métodos propostos para os experimentos: MambaVision, ViT e Xception. As métricas resultantes, relacionadas a precisão e a tempo, são utilizadas para comparação a fim de obter as conclusões do trabalho. Diagrama do autor, 2025.

4.1 Conjuntos de Dados

São utilizados três importantes conjuntos de dados no contexto de detecção de *deepfakes*, já apresentados por outros trabalhos da área (PASSOS et al., 2024), compondo uma base robusta para comparação de resultados.

4.1.1 FaceForensics++

FaceForensics++ (ROSSLER et al., 2019) é um conjunto de dados público que possui 1000 vídeos reais, em sua maioria extraídos do Youtube, bem como suas versões manipuladas utilizando os métodos DeepFakes, Face2Face, FaceShifter, FaceSwap e NeuralTextures, resultando em 5000 sequências manipuladas. Além disso, o conjunto de dados inclui 363 sequências originais e 3068 vídeos adicionais manipulados com o método DeepFakes, que foram integrados a partir do conjunto de dados DeepFakeDetection, somados resultam no total de 1363 sequências reais e 8068 sequências falsas. Vale dizer que 55% do vídeos são na resolução VGA, 32,5% em HD e 12,5% em Full-HD, sendo que 60% dos vídeos apresentam indivíduos do sexo masculino e 40% feminino.

4.1.2 Celeb-DFv2

Celeb-DFv2 (LI et al., 2020) é um conjunto de dados de *deepfakes* de grande escala que possui 890 vídeos reais de celebridades coletados do Youtube e 5639 vídeos sintéticos correspondentes de alta qualidade de 59 celebridades diferentes. Cada vídeo tem aproximadamente 13 segundos de duração, a uma taxa média de 30 quadros por segundo. O conjunto é distribuído da seguinte forma: Por etnia, 88.1% Caucasianos, 5.1% Asiáticos e 6.8% são Afro-Americanos; 6.4% abaixo de 30 anos, 28.0% entre 30 e 40 anos, 26.6% entre 40 e 50 anos, 30.5% entre 50 e 60 anos, 8.5% 60 anos ou mais; 56.8% do sexo masculino e 43.2% do femininino.

Figura 3 – Amostras do conjunto Celeb-DFv2. Imagens em verde, à esquerda, são originais, enquanto as imagens em vermelho são manipuladas.

4.1.3 WildDeepfake

WildDeepfake (ZI et al., 2020) é um conjunto de dados que foi proposto para representar melhor os *deepfakes* reais encontrados na Internet. Conta com 3805 sequências de vídeos verdadeiros e 3509 falsos, totalizando 7314 instâncias extraídas de diversas fontes na Internet.

4.2 Pré-processamento

O pré-processamento dos conjuntos de dados é realizado com o objetivo de adequar os dados para o treinamento, reduzir o *overfitting* e melhorar a generalização dos métodos. Os conjuntos de dados processados são usados como dados de entrada para os experimentos com os modelos. O processo de aquisição e processamento é apresentado detalhadamente.

4.2.1 FaceForensics++

O conjunto de dados FaceForensics++ com 40 de compressão (C40) foi utilizado para os experimentos. O *dataset* completo foi obtido utilizando o código oficial fornecido [4] e inclui instâncias do conjunto de dados DeepFakeDetection. Os vídeos baixados foram decompostos em quadros com um fator de amostragem de 25, ou seja, apenas 1 a cada 25 quadros do vídeo é mantido nos diretórios de imagens resultantes, com o objetivo de reduzir tanto o tamanho total de armazenamento quanto o tempo de treinamento dos modelos. O próximo passo é a extração das faces, realizada nos quadros resultantes, utilizando o modelo de detecção facial pré-treinado de rede neural profunda do OpenCV [5], resultando em imagens com resolução de 224×224 *pixels*, agora recortadas para conter apenas os segmentos das faces. O dataset é então dividido em subconjuntos de treinamento, teste e validação nas proporções de 70%, 15% e 15% respectivamente, cada um contendo exemplos reais e falsos nas mesmas proporções do *dataset* completo, visando realizar uma classificação binária ao invés de identificar cada tipo de manipulação separadamente.

Além disso, uma última etapa de aumento de dados (*data augmentation*) é aplicada com o objetivo de reduzir *overfitting* ao longo dos experimentos. Para esse fim, a biblioteca Torchvision do PyTorch é utilizada para aplicar aleatoriamente recortes (*cropping*), rotação horizontal (*horizontal flipping*), alteração de cor (*color jittering*) e eliminação aleatória de regiões (*random erasing*) nas imagens da pasta de treinamento, cada técnica conta com a probabilidade de 10% de ser aplicada por imagem, criando uma cópia adicional daquela instância transformada pelas aumentações. Esse processo resulta em um aumento de aproximadamente 40% no tamanho do conjunto de dados. Todo o processamento e organização dos conjuntos é realizado previamente e não de forma separada durante a execução de cada modelo, garantindo

[4] https://github.com/ondyari/FaceForensics/blob/master/dataset/README.md
[5] https://github.com/sr6033/face-detection-with-OpenCV-and-DNN

que todos os modelos utilizados nos experimentos sejam treinados e avaliados sobre exatamente as mesmas instâncias e divisões de dados.

Vale ressaltar que o desbalanceamento de classes presente nos conjuntos de dados processados é resultados das distribuições originais dos *datasets* e não é tratado para os experimentos, essa decisão visa avaliar como os modelos se comportam com dados desbalanceados, mais próximos do que se encontra em situações reais. O conjunto FaceForensics++ processado contém um total de 218.558 imagens, e conta com a distribuição de classes aproximada de 84% de instâncias falsas e 16% reais.

4.2.2 Celeb-DFv2

O processamento do conjunto de dados Celeb-DFv2 segue exatamente a mesma lógica descrita anteriormente para o FaceForensics++, com a única diferença de que os vídeos foram obtidos diretamente do repositório Celeb-DFv2 no Kaggle [6]. O conjunto de dados Celeb-DFv2 processado totaliza 119.952 imagens, com a distribuição de classes sendo de aproximadamente 86% instâncias falsas e 14% reais.

4.2.3 WildDeepfake

O conjunto de dados WildDeepfake foi obtido através do repositório oficial no HuggingFace[7] e, em contraste com os outros dois conjuntos de dados descritos, os vídeos já estão divididos em quadros (*frames*) e armazenados em diferentes pastas compactadas no formato *.tar* quando o conjunto é baixado. Desta forma, o primeiro passo foi executar um código de reorganização para extrair e dividir as imagens em pastas de treinamento, teste e validação, mantendo a mesma proporção dos outros dois conjuntos de dados: 70%, 15% e 15%, respectivamente. O conjunto original é composto por mais de 1 milhão de imagens, portanto se faz necessário realizar a amostragem do conjunto com um fator de 10, mantendo apenas 1 a cada 10 quadros, reduzindo efetivamente o conjunto de dados para 10% do tamanho original.

As etapas de extração de faces e aumento de dados seguem a mesma lógica aplicada nos outros dois conjuntos de dados. O conjunto de dados WildDeepfake processado contém um total de 151.549 imagens, com a distribuição de classes de aproximadamente 63% de instâncias falsas e 37% reais, apresentando o menor desbalanceamento de classes entre os três conjuntos de dados utilizados nos experimentos.

4.3 Execução dos Experimentos

O método baseado em Mamba para processamento visual MambaVision (HATAMIZADEH; KAUTZ, 2024), apresentado na Revisão de Literatura, foi adaptado e treinado na tarefa

[6] https://www.kaggle.com/datasets/reubensuju/celeb-df-v2
[7] https://huggingface.co/datasets/xingjunm/WildDeepfake

de detecção de *deepfakes*. Ainda dentro das referências apresentadas na seção de Revisão de Literatura, foram escolhidas para compor a base de comparação com métricas de tempo as abordagens que possuem código disponível publicamente, para que possam ser utilizados para calcular as métricas *throughput* e tempo de execução. A CNN selecionada foi a abordagem de detecção Xception, apresentada por Rossler et al. (2019). O método baseado em ViT selecionado foi o de Khan e Dai (2021).

Os treinamentos, testes e captura das métricas de avaliação dos modelos MambaVision, CNN e ViT foram realizados sob as mesmas configurações de *hardware*. As abordagens baseadas em transformador e CNN seguiram os detalhes de implementação e hiperparâmetros encontrados em suas respectivas publicações, sofrendo apenas as adaptações necessárias para seguir a metodologia de experimentos proposta. O ambiente foi equipado com uma GPU RTX 3060 com 12 GB de VRAM, e os modelos foram implementados principalmente em PyTorch, utilizando o *framework* CUDA da NVIDIA. Todos os três modelos foram inicialmente pré-treinados no conjunto de dados ImageNet-1K. Em seguida, retomamos o treinamento específico para *deepfakes* a partir dos *checkpoints* pré-treinados, que serviram como base para o ajuste fino subsequente. Os dados utilizados para treinamento específico de *deepfakes* de todos os modelos foram os conjuntos de dados obtidos após pré-processamento realizado em 4.2.

São apresentadas as modificações e configurações específicas para o treinamento de cada modelo.

4.3.1 MambaVision

O modelo MambaVision-T (variante *Tiny*) foi inicializado com pesos pré-treinados no ImageNet-1K e treinado por 50 épocas utilizando o otimizador AdamW, com uma taxa de aprendizado de 1^{-3} e função de perda calculada com Binary Cross-Entropy. A cabeça de classificação foi modificada em relação ao código original do MambaVision para produzir saídas para 2 classes: real e falso.

4.3.2 Xception

O modelo Xception segue os detalhes de implementação de Rossler et al. (2019). Após o carregamento dos pesos pré-treinados no ImageNet-1K, o modelo foi treinado por 50 épocas utilizando o otimizador Adam, com taxa de aprendizado de 2^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$ e $\varepsilon = 1^{-8}$, com função de perda Cross-Entropy.

4.3.3 ViT

O modelo ViT para classificação binária de *deepfakes* baseia-se em Khan e Dai (2021). A inicialização foi feita com pesos pré-treinados do ImageNet-1K e treinado por 15 épocas utilizando o otimizador Stochastic Gradient Descent, com taxa de aprendizado de 3^{-3}, momentum

de 0.9 e função de perda Cross-Entropy. O número reduzido de épocas de treinamento deve-se a limitações computacionais e de tempo de execução.

Os resultados e métricas obtidos são utilizados para comparação.

4.4 Métricas de Avaliação

Para avaliação do desempenho dos modelos e posterior comparação, são utilizadas as métricas apresentadas nesta seção.

- Acurácia: Proporção de todas as classificações corretas.

$$Acurácia = \frac{VN + VP}{VN + VP + FN + FP}. \tag{5}$$

- Precisão: Proporção de classificações positivas corretas.

$$Precisão = \frac{VP}{VP + FP}. \tag{6}$$

.

- *Recall*: Proporção de todos os positivos que foram classificados corretamente como positivos.

$$Recall = \frac{VP}{VP + FN}. \tag{7}$$

- *F1-score*: Média harmônica entre precisão e *recall*

$$F1 = 2 * \frac{Recall * Precisão}{Recall + Precisão} = \frac{2VP}{2VP + FN + FP}. \tag{8}$$

- *Area Under the ROC Curve* (AUC): Valor entre 0 e 1 calculado com base na área sob a curva ROC, que por sua vez é uma representação visual obtida a partir da taxa de verdadeiros positivos (*Recall*) e a taxa de falsos positivos. Para análise da AUC, quanto mais próximo de 1, melhor.

- Tempo de execução: Tempo de treino e tempo de inferência medidos em segundos (s) ou milissegundos (ms).

- *Throughput*: Quantidade de inferências por período de tempo.

$$Throughput = \frac{NumeroInferências}{Tempo(s)}. \tag{9}$$

5 Resultados

Seguindo a metodologia descrita para os experimentos, são apresentadas as métricas obtidas através do teste dos modelos. Os resultados estão organizados em duas tabelas. A Tabela 1 apresenta as métricas de precisão dos modelos, enquanto a Tabela 2 apresenta as métricas de relacionadas a tempo de cada modelo.

Tabela 1 – Comparativo de métricas de precisão obtidas com os experimentos

Conjunto	Modelo	Acc	Prec	Rec	F1	AUC
FaceForensics++	Xception	**0.9475**	**0.8771**	**0.7709**	**0.8206**	0.8755
	ViT	0.8417	0.4618	0.1043	0.1702	0.6939
	MambaVision	0.9279	0.8003	0.7155	0.7555	**0.9533**
WildDeepfakes	Xception	**0.8108**	**0.6991**	0.8244	**0.7566**	0.8139
	ViT	0.6503	0.5087	0.5674	0.5364	0.6841
	MambaVision	0.7940	0.6619	**0.8634**	0.7493	**0.8910**
CelebDFv2	Xception	0.9977	0.9922	**0.9917**	0.9919	0.9952
	ViT	0.9232	0.9273	0.4929	0.6437	0.9276
	MambaVision	**0.9979**	**0.9936**	0.9912	**0.9924**	**0.9999**

A Tabela 1 fornece uma comparação abrangente entre Xception, ViT e MambaVision em termos de acurácia, precisão, *recall*, *F1-score* e AUC entre três referências de *deepfake*. No conjunto FaceForensics++, que contém aproximadamente 84% de amostras falsas, Xception atinge a maior precisão (87.71%) e acurácia (94.75%), enquanto o MambaVision obtém AUC superior (95.33%), indicando uma discriminação geral mais forte apesar de sua acurácia alcançada ser ligeiramente inferior (92.79%).

No conjunto de dados WildDeepfakes, a precisão dos modelos Xception (69.91%) e MambaVision (66.19%) convergem, refletindo um controle comparável de falsos positivos. No entanto, o maior *recall* (86.34% vs. 82.44%) e AUC (89.10% vs. 81.39%) pertencem ao MambaVision, destacando sua capacidade aprimorada de detectar verdadeiros positivos. No conjunto CelebDFv2, ambos os modelos atingem uma precisão próxima ao teto ($> 99\%$) e acurácia ($> 99{,}7\%$), com o MambaVision demonstrando uma leve superioridade sobre o Xception em quase todas as métricas, exceto no *recall*. Notavelmente, os resultados obtidos pelo MambaVision no conjunto de dados CelebDFv2 são comparáveis em valores aos resultados de outras abordagens baseadas em transformadores, como (LIN et al., 2023), (ZHANG, 2022) e (HEO; YEO; KIM, 2023), que alcançaram resultados exemplares, com AUC de 99.80%, 97.21% e 99.30%, respectivamente.

O desempenho do modelo ViT foi substancialmente inferior ao de Xception e MambaVision, exibindo acurácia e AUC consistentemente mais baixas em todas as três referências. Por exemplo, no FaceForensics++, o ViT alcançou apenas 84.17% de acurácia, em comparação com 94.75% do Xception e 92.79% do MambaVision. Embora uma avaliação mais exaustiva deste modelo baseado em transformadores seja apresentada por Khan e Dai (2021), as adapta-

ções realizadas foram necessárias para garantir a consistência com o protocolo experimental adotado neste estudo, possivelmente impactando nas métricas de precisão. Adicionalmente, a diferença significativa das métricas de eficiência calculadas para o modelo ViT em comparação da abordagem baseada em Mamba, ressaltam sua vantagem prática em relação à alternativa baseada em transformador.

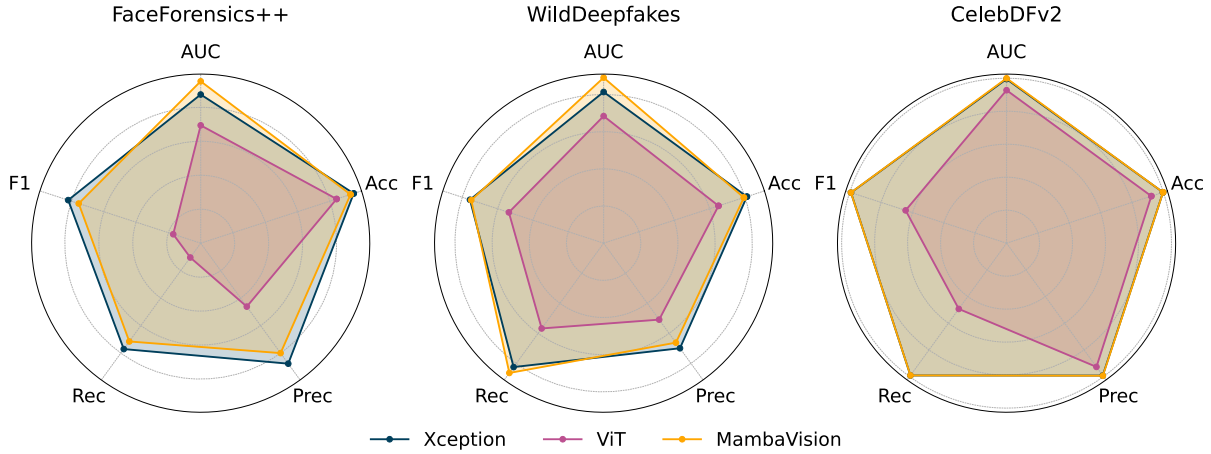

Figura 4 – Gráficos de radar comparando métricas de performance — acurácia (Acc), *recall* (Rec), precisão (Prec), *F1-score* (F1) e AUC—dos modelos nos três conjuntos de dados usados para os experimentos. Imagem do autor, 2025.

Na Figura 4, os gráficos de radar consolidam visualmente as cinco principais métricas de avaliação (Acc, Rec, Prec, F1 e AUC) para cada modelo em todos os conjuntos de dados. Pode-se observar que Xception exibe um pico mais pronunciado no eixo de precisão em comparação com MambaVision — indicando sua maior tendência a minimizar falsos positivos — mas apresenta uma ligeira queda no AUC em relação ao MambaVision. Por outro lado, o ViT sofre reduções acentuadas nos eixos de *recall* e F1, revelando sua menor sensibilidade a amostras falsas e ao desbalanceamento geral. O MambaVision adota uma postura intermediária, exibindo um pentágono mais uniforme com valores ligeiramente superiores de AUC e F1. O gráfico também revela que Xception e MambaVision atingem precisões praticamente idênticas, ambos superando amplamente ViT.

Tabela 2 – Comparativo de métricas de tempo obtidas com os experimentos

Conjunto	**Modelo**	**Throughput (img/s)**	**Tempo total (s)**
FaceForensics++	Xception	424.54	62.24
	ViT	7.93	3330.33
	MambaVision	**546.11**	**48.28**
WildDeepfakes	Xception	424.33	39.91
	ViT	7.53	2249.71
	MambaVision	**542.28**	**31.16**
CelebDFv2	Xception	423.47	34.28
	ViT	7.01	2070.84
	MambaVision	**535.31**	**27.02**

A Tabela 2 resume o desempenho temporal dos modelos Xception, ViT e MambaVision

em termos de *throughput* de inferência em imagens por segundo e tempo total de processamento em segundos. Com base nos resultados de desempenho de tempo obtidos nos testes realizados com os três conjuntos de dados (FaceForensics++, WildDeepfakes e CelebDFv2), o modelo MambaVision demonstrou desempenho significativamente superior em comparação aos modelos Xception e ViT, tanto em termos de *throughput* quanto de tempo total de processamento. No conjunto de dados FaceForensics++, o MambaVision atingiu um *throughput* aproximadamente 22.26% superior ao Xception e 98.55% superior ao ViT. Em relação ao tempo total de inferência, foi cerca de 29% mais rápido do que o Xception e impressionantes 68 vezes mais rápido do que o ViT. Resultados semelhantes foram observados no conjunto de dados WildDeepfakes, onde o MambaVision apresentou um throughput 21.75% superior ao Xception e 98.61% superior ao ViT, além de ser 28% mais eficiente em tempo total de processamento do que o Xception e mais de 71 vezes mais rápido do que o ViT. O conjunto de dados CelebDFv2 seguiu o mesmo padrão, com o *throughput* do MambaVision 20.89% superior ao Xception e 98.69% superior ao ViT, e o tempo total de processamento 27% mais rápido do que o Xception e mais de 75 vezes mais rápido do que o ViT. Esses resultados destacam o MambaVision como o modelo mais eficiente computacionalmente entre os três avaliados, tornando-o vantajoso para aplicações que exigem altas taxas de processamento e baixos tempos de resposta. Além de alcançar métricas de precisão sólidas, sua superioridade temporal o torna uma escolha promissora para cenários onde desempenho e escalabilidade são cruciais.

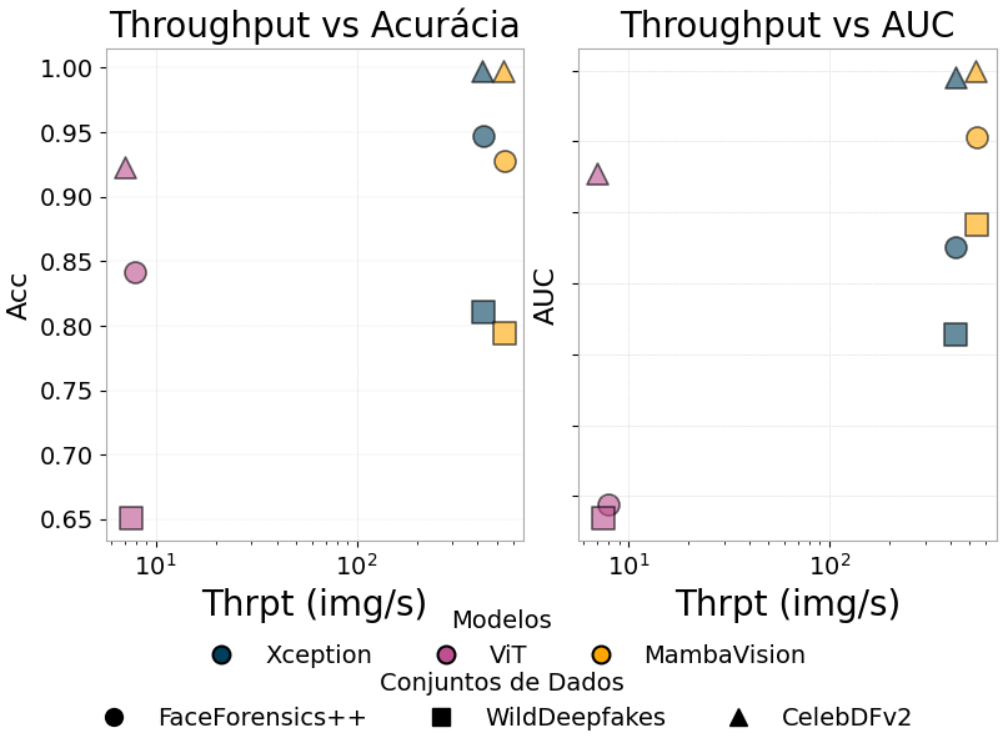

Figura 5 – Comparativo de throughput de inferência (Thrpt, img/s) contra métricas de performance — Acurácia (Acc) à esquerda e AUC à direita — dos modelos nos três conjuntos de dados avaliados. Imagem do autor, 2025.

Na Figura 5, o *throughput* de inferência — plotado em escala logarítmica — é comparado com as duas métricas principais para avaliação da qualidade das detecções, acurácia (à esquerda) e AUC (à direita), para cada conjunto de dados. O modelo ViT aparece consistentemente na região inferior esquerda, indicando taxas de processamento muito baixas e um poder de discriminação visivelmente inferior. O Xception ocupa uma faixa intermediária, equilibrando *throughput* estável com alta acurácia e AUC. O MambaVision, por outro lado, desloca-se para a direita — alcançando o maior *throughput* — enquanto seus pontos permanecem em um nível vertical semelhante ao do Xception. Em todas as três referências esse padrão evidencia que o MambaVision proporciona ganhos substanciais de velocidade sem perdas significativas de qualidade de detecção em acurácia ou AUC, enquanto o ViT não consegue atingir esse mesmo equilíbrio sob situações idênticas de equipamento.

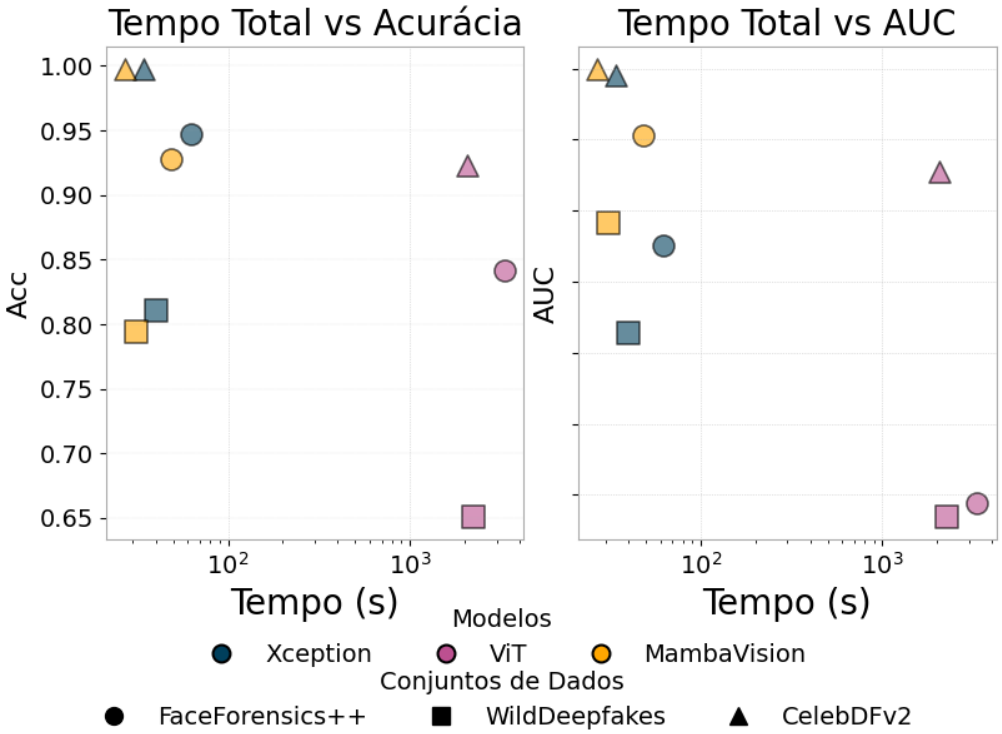

Figura 6 – Comparativo de tempo total de processamento (Tempo, s) contra métricas de performance — Acurácia (Acc) à esquerda e AUC à direita — dos modelos nos três conjuntos de dados avaliados. Imagem do autor, 2025.

Na Figura 6, o tempo total de processamento — novamente apresentado em escala logarítmica — é comparado à discriminação do modelo em termos de acurácia (à esquerda) e AUC (à direita) para cada conjunto de dados. O ViT ocupa claramente a extremidade direita, com tempos de execução na ordem de milhares de segundos, enquanto entrega os menores valores de acurácia e AUC entre as três arquiteturas avaliadas. O Xception aparece em uma faixa central, combinando tempos de inferência mais baixos com acurácia e AUC consistentemente altos. O MambaVision desloca-se para a esquerda, reduzindo o tempo total, mas sua acurácia e AUC permanecem em níveis equiparáveis aos do Xception.

6 Conclusão

Com a contínua evolução dos métodos de geração de *deepfakes*, soluções robustas para detecção tornam-se cada vez mais essenciais. Este artigo propôs realizar a detecção de *deepfakes* utilizando a recente arquitetura Mamba, tendo em vista que modelos baseados em Mamba oferecem um equilíbrio promissor entre acurácia e eficiência computacional, graças à sua complexidade linear. Experimentos conduzidos nos conjuntos de dados FaceForensics++, CelebDFv2 e WildDeepfake mostraram resultados encorajadores que reforçam essa afirmação. Entre todos os modelos avaliados, MambaVision destacou-se ao oferecer a melhor combinação de acurácia e desempenho temporal, tornando-o vantajoso em cenários onde desempenho e escalabilidade são cruciais e para aplicações que exigem altas taxas de processamento e baixos tempos de resposta.

O estudo demonstra que o uso de modelos de espaço de estado estruturados pode servir como uma alternativa mais leve enquanto mantém boa discriminação em comparação com redes totalmente baseadas em atenção para a detecção de mídias forjadas.

Trabalhos futuros envolvem expandir os experimentos para novos conjuntos de dados e analisar os *trade-offs* entre *throughput*, acurácia e consumo de memória frente às abordagens baseadas em transformadores, bem como outras arquiteturas relevantes. Estudos de ablação podem ajudar a expandir a compreensão sobre o impacto de alterações na profundidade do modelo e como as configurações da dimensão oculta influenciam o desempenho da detecção, oferecendo subsídios para equilibrar eficiência e eficácia na implementação de detectores baseados em Mamba em ambientes reais. Finalmente, comparações com variantes adicionais da arquitetura Mamba, como ViM (ZHU et al., 2024), VMamba (LIU et al., 2024) e EfficientVMamba (PEI; HUANG; XU, 2024) podem ser exploradas para aprofundar a avaliação da arquitetura na tarefa de detecção de mídias sintéticas, evidenciando ainda mais seus possíveis benefícios e limitações.

Esses esforços visam não apenas fortalecer a compreensão sobre os limites e potencialidades da arquitetura Mamba, mas também consolidá-la como uma solução eficiente e viável para enfrentar os desafios emergentes na detecção de mídias sintéticas — uma tarefa essencial para preservar a integridade e a confiança no mundo atual.

Referências

AFCHAR, D. et al. Mesonet: a compact facial video forgery detection network. In: IEEE. *2018 IEEE international workshop on information forensics and security (WIFS)*. [S.l.], 2018. p. 1–7.

AL-ADWAN, A. et al. Detection of deepfake media using a hybrid cnn–rnn model and particle swarm optimization (pso) algorithm. *Computers*, MDPI, v. 13, n. 4, p. 99, 2024.

ALEXEY, D. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.

AV, A. et al. Latent flow diffusion for deepfake video generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2024. p. 3781–3790.

BAEK, J.-Y.; YOO, Y.-S.; BAE, S.-H. Generative adversarial ensemble learning for face forensics. *Ieee Access*, IEEE, v. 8, p. 45421–45431, 2020.

BECKMANN, A.; HILSMANN, A.; EISERT, P. Fooling state-of-the-art deepfake detection with high-quality deepfakes. In: *Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security*. [S.l.: s.n.], 2023. p. 175–180.

BROWN, T. B. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

CHEN, H. et al. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *arXiv preprint arXiv:2405.19707*, 2024.

CHEN, Y. et al. Rawbmamba: End-to-end bidirectional state space model for audio deepfake detection. *arXiv preprint arXiv:2406.06086*, 2024.

CHOI, Y. et al. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2018. p. 8789–8797.

CHOLLET, F. Xception: Deep learning with depthwise separable convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 1251–1258.

COCCOMINI, D. A. et al. Combining efficientnet and vision transformers for video deepfake detection. In: SPRINGER. *International conference on image analysis and processing*. [S.l.], 2022. p. 219–229.

CUNHA, L. et al. Video deepfake detection using particle swarm optimization improved deep neural networks. *Neural Computing and Applications*, Springer, p. 1–37, 2024.

DANG, H. et al. On the detection of digital face manipulation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*. [S.l.: s.n.], 2020. p. 5781–5790.

DEEPFAKES. Disponível em: <https://github.com/deepfakes/faceswap>. Acesso em: 01 junho 2024.

DOLHANSKY, B. et al. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

FACESWAP. Disponível em: <https://github.com/MarekKowalski/FaceSwap/>. Acesso em: 01 junho 2024.

GARCIA, G. L. et al. Introducing bode: A fine-tuned large language model for portuguese prompt-based task. *arXiv preprint arXiv:2401.02909*, 2024.

GOODFELLOW, I. et al. Generative adversarial nets. *Advances in neural information processing systems*, v. 27, 2014.

GOOGLE. *Announcing Veo 3, Imagen 4, and Lyria 2 on Vertex AI*. 2025. [Online; acesso em 28. Jul. 2025]. Disponível em: <https://cloud.google.com/blog/products/ai-machine-learning/announcing-veo-3-imagen-4-and-lyria-2-on-vertex-ai>.

GU, A.; DAO, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

GU, A.; GOEL, K.; RÉ, C. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

GUARNERA, L.; GIUDICE, O.; BATTIATO, S. Fighting deepfake by exposing the convolutional traces on images. *IEEE Access*, IEEE, v. 8, p. 165085–165098, 2020.

HATAMIZADEH, A.; KAUTZ, J. Mambavision: A hybrid mamba-transformer vision backbone. *arXiv preprint arXiv:2407.08083*, 2024.

HE, Z. et al. Attgan: Facial attribute editing by only changing what you want. *IEEE transactions on image processing*, IEEE, v. 28, n. 11, p. 5464–5478, 2019.

HEO, Y.-J.; YEO, W.-H.; KIM, B.-G. Deepfake detection algorithm based on improved vision transformer. *Applied Intelligence*, Springer, v. 53, n. 7, p. 7512–7527, 2023.

HO, J.; JAIN, A.; ABBEEL, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, v. 33, p. 6840–6851, 2020.

HU, V. T. et al. Zigma: A dit-style zigzag mamba diffusion model. In: SPRINGER. *European Conference on Computer Vision*. [S.l.], 2024. p. 148–166.

HUANG, Y. et al. Fakelocator: Robust localization of gan-based face manipulations. *IEEE Transactions on Information Forensics and Security*, IEEE, v. 17, p. 2657–2672, 2022.

HUE, N. T. et al. Convmamba: A data-efficient neural network for bearing fault diagnosis. In: IEEE. *2024 7th International Conference on Green Technology and Sustainable Development (GTSD)*. [S.l.], 2024. p. 165–172.

JIANG, L. et al. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2020. p. 2889–2898.

KARRAS, T.; LAINE, S.; AILA, T. A style-based generator architecture for generative adversarial networks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2019. p. 4401–4410.

KARRAS, T. et al. Analyzing and improving the image quality of stylegan. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2020. p. 8110–8119.

KAUR, A. et al. Deepfake video detection: challenges and opportunities. *Artificial Intelligence Review*, Springer, v. 57, n. 6, p. 1–47, 2024.

KENTON, J. D. M.-W. C.; TOUTANOVA, L. K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: MINNEAPOLIS, MINNESOTA. *Proceedings of naacL-HLT*. [S.l.], 2019. v. 1, p. 2.

KHAN, S. A.; DAI, H. Video transformer for deepfake detection with incremental learning. In: *Proceedings of the 29th ACM international conference on multimedia*. [S.l.: s.n.], 2021. p. 1821–1828.

KHAN, S. A.; DANG-NGUYEN, D.-T. Hybrid transformer network for deepfake detection. In: *Proceedings of the 19th international conference on content-based multimedia indexing*. [S.l.: s.n.], 2022. p. 8–14.

KHEIR, Y. E.; POLZEHL, T.; MÖLLER, S. Bicrossmamba-st: Speech deepfake detection with bidirectional mamba spectro-temporal cross-attention. *arXiv preprint arXiv:2505.13930*, 2025.

LI, Y. et al. Celeb-df: A large-scale challenging dataset for deepfake forensics. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2020. p. 3207–3216.

LIN, H. et al. Deepfake detection with multi-scale convolution and vision transformer. *Digital Signal Processing*, Elsevier, v. 134, p. 103895, 2023.

LIU, M. et al. Stgan: A unified selective transfer network for arbitrary image attribute editing. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2019. p. 3673–3682.

LIU, Y. et al. Vmamba: Visual state space model. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. [S.l.: s.n.], 2024.

MASOOD, M. et al. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence*, Springer, v. 53, n. 4, p. 3974–4026, 2023.

NASIRI-SARVI, A. et al. Vim4path: Self-supervised vision mamba for histopathology images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2024. p. 6894–6903.

NEVES, J. C. et al. Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE Journal of Selected Topics in Signal Processing*, IEEE, v. 14, n. 5, p. 1038–1048, 2020.

NGUYEN, H. H.; YAMAGISHI, J.; ECHIZEN, I. Capsule-forensics: Using capsule networks to detect forged images and videos. In: IEEE. *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. [S.l.], 2019. p. 2307–2311.

NIRKIN, Y.; KELLER, Y.; HASSNER, T. Fsgan: Subject agnostic face swapping and reenactment. In: *Proceedings of the IEEE/CVF international conference on computer vision*. [S.l.: s.n.], 2019. p. 7184–7193.

PASSOS, L. A. et al. A review of deep learning-based approaches for deepfake content detection. *Expert Systems*, Wiley Online Library, v. 41, n. 8, p. e13570, 2024.

PEI, X.; HUANG, T.; XU, C. Efficientvmamba: Atrous selective scan for light weight visual mamba. *arXiv preprint arXiv:2403.09977*, 2024.

PENG, S. et al. Wmamba: Wavelet-based mamba for face forgery detection. *arXiv preprint arXiv:2501.09617*, 2025.

RAO, S.; VERMA, A. K.; BHATIA, T. A review on social spam detection: Challenges, open issues, and future directions. *Expert Systems with Applications*, Elsevier, v. 186, p. 115742, 2021.

ROMBACH, R. et al. High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2022. p. 10684–10695.

ROSSLER, A. et al. Faceforensics++: Learning to detect manipulated facial images. In: *Proceedings of the IEEE/CVF international conference on computer vision*. [S.l.: s.n.], 2019. p. 1–11.

TAO, X.; WANG, J.; LI, J. Application and improvement of dual-branch feature fusion network in multi-class classification of gastrointestinal images. In: IEEE. *2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS)*. [S.l.], 2024. p. 793–798.

TENG, Y. et al. Dim: Diffusion mamba for efficient high-resolution image synthesis. *arXiv preprint arXiv:2405.14224*, 2024.

THIES, J.; ZOLLHÖFER, M.; NIESSNER, M. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, ACM New York, NY, USA, v. 38, n. 4, p. 1–12, 2019.

THIES, J. et al. Face2face: Real-time face capture and reenactment of rgb videos. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 2387–2395.

TOUVRON, H. et al. Training data-efficient image transformers & distillation through attention. In: PMLR. *International conference on machine learning*. [S.l.], 2021. p. 10347–10357.

VASWANI, A. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

WESTERLUND, M. The emergence of deepfake technology: A review. *Technology innovation management review*, v. 9, n. 11, 2019.

ZAKHAROV, E. et al. Few-shot adversarial learning of realistic neural talking head models. In: *Proceedings of the IEEE/CVF international conference on computer vision*. [S.l.: s.n.], 2019. p. 9459–9468.

ZHANG, D. et al. Deepfake video detection with spatiotemporal dropout transformer. In: *Proceedings of the 30th ACM international conference on multimedia*. [S.l.: s.n.], 2022. p. 5833–5841.

ZHANG, T. Deepfake generation and detection, a survey. *Multimedia Tools and Applications*, Springer, v. 81, n. 5, p. 6259–6276, 2022.

ZHANG, Y. et al. A robust lightweight deepfake detection network using transformers. In: SPRINGER. *Pacific rim international conference on artificial intelligence*. [S.l.], 2022. p. 275–288.

ZHU, L. et al. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

ZI, B. et al. Wilddeepfake: A challenging real-world dataset for deepfake detection. In: *Proceedings of the 28th ACM international conference on multimedia*. [S.l.: s.n.], 2020. p. 2382–2390.