Revisão sistemática sobre técnicas e algoritmos de IA para geração de animações

Leonardo Biagiotti Beloti[1], **Prof. Ms. Joaquim Pessoa Filho**[1]

[1]Ciência da Computação
Faculdade de Computação e Informática
Universidade Presbiteriana Mackenzie
São Paulo – SP – Brasil

10390339@mackenzista.com.br, joaquim@mackenzie.br

Resumo. *Esta revisão sistemática explora os avanços recentes nas técnicas e algoritmos de inteligência artificial aplicados à geração de animações 2D e 3D. O principal objetivo é oferecer uma análise aprofundada de como a IA tem contribuído para melhorar os fluxos de trabalho de animação ao aumentar a eficiência, o realismo e a criatividade. A metodologia envolveu uma busca abrangente em bases de dados acadêmicas utilizando palavras-chave específicas e operadores booleanos. Foram selecionados artigos seguindo critérios de inclusão rígidos e bem estruturados para garantir a qualidade da pesquisa. Os resultados apontam para o potencial transformador das GANs na síntese de conteúdos visuais realistas, a eficácia das CNNs no reconhecimento e manipulação de padrões para fluxos de trabalho em animação, a capacidade da transferência de estilo de combinar elementos artísticos com estruturas visuais e a aplicação inovadora da geração de expressões faciais a partir de áudio em avatares virtuais e ambientes digitais. Estudos futuros devem focar no desenvolvimento de modelos mais eficientes, adaptáveis e interpretáveis para superar as limitações atuais e ampliar as aplicações práticas da IA na indústria da animação.*

Palavras-chave: animação, algoritmos, inteligência artificial, redes adversárias generativas, redes neurais convolucionais, transferência de estilo de imagem, geração de expressões faciais

1. Introdução

A evolução da tecnologia nos últimos anos tem transformado significativamente a indústria da animação, especialmente com o advento de técnicas de Inteligência Artificial (IA) [Ren and Sheng 2022]. Animações 2D e 3D, antes completamente dependentes de processos manuais e exaustivos, agora podem ser desenvolvidas com maior precisão e eficiência graças a algoritmos de IA que automatizam partes do fluxo de trabalho.

A inteligência artificial aplicada à animação abrange uma série de abordagens, cada uma com seu próprio conjunto de algoritmos e técnicas. Entre as mais notáveis, destacam-se as redes neurais, a aprendizagem profunda (*deep learning*) e os modelos generativos, como as Redes Generativas Adversárias (GANs) [Ren and Sheng 2022]. Essas tecnologias não só têm o potencial de aumentar a eficiência no desenvolvimento de animações, mas também de reduzir significativamente os custos associados à criação de

conteúdos visuais de alta qualidade [Zhang et al. 2023]. Esse progresso tecnológico tem captado o interesse de estúdios de animação que veem na IA um aliado poderoso para inovar e melhorar a produção de filmes e séries animadas [Zhang et al. 2023].

Apesar desses avanços, a aplicação de IA na geração de animações ainda apresenta desafios consideráveis. Alguns dos principais obstáculos incluem a qualidade e a coerência das ilustrações gerados automaticamente, a necessidade de grandes volumes de dados para treinar algoritmos de aprendizagem profunda e a dificuldade em capturar a "intenção artística" por meio de técnicas puramente algorítmicas [Sang and Xu 2022]. Esses fatores exigem que a pesquisa continue a explorar formas de integrar melhor a IA ao processo criativo, de modo a alcançar um equilíbrio entre automação e controle criativo.

Neste cenário, torna-se importante uma revisão sistemática das técnicas e algoritmos de IA aplicados à geração de animações. Esta pesquisa visa fornecer uma análise das contribuições mais recentes nesse campo, avaliando os avanços mais promissores, os desafios que ainda precisam ser superados e as tendências emergentes da área.

O principal objetivo deste estudo é explorar os algoritmos de inteligência artificial usados atualmente na otimização dos processos de animação, focando nos algoritmos mais bem-sucedidos e nos casos de uso mais emblemáticos.

A estrutura do artigo é organizada da seguinte forma: na seção de Metodologia e Procedimentos, será descrito o processo de coleta e análise dos artigos que compõem esta revisão. Na seção de Resultados será apresentado um breve resumo sobre cada técnica discutida e suas principais contribuições para o campo da animação. A seção de Discussão será dedicada a uma análise crítica desses resultados, destacando as limitações e desafios encontrados na literatura. Por fim, a Conclusão oferecerá um resumo das principais descobertas e sugestões para futuras pesquisas.

2. Metodologia

A metodologia adotada para este estudo baseia-se em uma revisão sistemática de literatura [Page et al. 2021], sendo projetada para garantir uma abordagem rigorosa, por meio dos critérios estabelecidos, na identificação, seleção e análise de artigos relevantes sobre técnicas e algoritmos de IA para o auxílio da geração de animações.

Primeiro foi definida a pergunta científica da revisão sistemática "O objetivo dessa revisão é determinar se existem técnicas e algoritmos de IA que possam auxiliar o processo de trabalho dos animadores", a pergunta foi definida de forma ampla para obter uma visão mais abrangente da área e proporcionar uma maior flexibilidade no processo de confecção da revisão. Para garantir uma revisão abrangente e representativa das contribuições mais recentes, foram selecionadas quatro bases de dados acadêmicos: *Web of Science*, *Scopus*, *Science Direct* e *IEEE Xplore*. Essas bases foram escolhidas por serem reconhecidas como fontes confiáveis de publicações científicas de alta qualidade nas áreas de tecnologia e de engenharia.

A estratégia de busca envolveu a combinação de palavras-chave relacionadas à IA e à geração de animações, todas as palavras foram definidas antes do início da pesquisa. As palavras-chave utilizadas foram: "*animation*", "*animation production*", "*animation generation*", "*2D animation*", "*3D animation*", "*generate*", "*algorithm*", "*deep learning*", "*artificial intelligence*" e "*computer generated imagery*". Para refinar os resul-

tados, o operador *booleano* "*AND*" foi utilizado para combinar essas palavras-chave em diferentes variações. Isso permitiu a identificação de artigos que abordassem especificamente o uso de IA na geração de animações. A figura abaixo mostra as palavras-chave que foram utilizadas para realizar as pesquisas nas bases científicas e quantos artigos foram retornados de cada busca.

Tabela 1. Tabela de buscas

Base de Dados	Palavras-chave	Resultados
Web of Science	"animation generation" AND "deep learning"	11
Web of Science	"animation generation" AND "neural network"	11
Web of Science	"animation generation" AND "artificial intelligence"	11
Web of Science	"animation production" AND algorithm	23
Web of Science	"animation production" AND "deep learning"	10
Web of Science	"animation production" AND "neural network"	5
Web of Science	"animation production" AND "artificial intelligence"	15
Web of Science	"2D animation" AND generate	15
Web of Science	"animation" AND "computer-generated imagery"	13
Scopus	"animation generation" AND "deep learning"	21
Scopus	"animation generation" AND "neural network"	18
Scopus	"animation generation" AND "artificial intelligence"	9
Scopus	"animation production" AND algorithm	59
Scopus	"animation production" AND "deep learning"	26
Scopus	"animation production" AND "neural network"	15
Scopus	"animation production" AND "artificial intelligence"	25
Scopus	"2D animation" AND generate	20
Scopus	"animation" AND "computer-generated imagery"	30
Science Direct	"animation generation" AND "deep learning"	19
Science Direct	"animation generation" AND "neural network"	19
Science Direct	"animation generation" AND "artificial intelligence"	5
Science Direct	"animation production" AND algorithm	63
Science Direct	"animation production" AND "deep learning"	24
Science Direct	"animation production" AND "neural network"	25
Science Direct	"animation production" AND "artificial intelligence"	16
Science Direct	"2D animation" AND generate	79
Science Direct	"animation" AND "computer-generated imagery"	58
IEEE Xplore	"animation generation" AND "deep learning"	8
IEEE Xplore	"animation generation" AND "neural network"	9
IEEE Xplore	"animation generation" AND "artificial intelligence"	5
IEEE Xplore	"animation production" AND algorithm	16
IEEE Xplore	"animation production" AND "deep learning"	9
IEEE Xplore	"animation production" AND "neural network"	11
IEEE Xplore	"animation production" AND "artificial intelligence"	8
IEEE Xplore	"2D animation" AND generate	12
IEEE Xplore	"animation" AND "computer-generated imagery"	5

Fonte: Autor.

Como parte da estratégia de busca, foram considerados apenas artigos publicados nos últimos 10 anos (de 2014 a 2024). Essa delimitação temporal visa garantir que as contribuições analisadas reflitam as inovações mais recentes. Além disso, foram incluídos somente estudos publicados em inglês. Essa decisão foi tomada para garantir a uniformidade no idioma utilizado, facilitando a leitura, análise e comparação entre os trabalhos.

3. Procedimentos

Com base na metodologia definida, os seguintes procedimentos foram adotados para a seleção e análise dos artigos. Para garantir que a revisão sistemática fosse focada em contribuições significativas, foi estabelecido um conjunto de critérios de inclusão e exclusão. Os artigos selecionados para a análise final deveriam atender aos seguintes critérios de inclusão:

- Descrever precisamente técnicas baseadas em IA utilizadas na geração de animações 2D ou 3D.
- Explicar como as técnicas funcionam em detalhes, incluindo os algoritmos empregados.
- Apresentar exemplos práticos ou exemplos visuais que demonstrem a eficácia das técnicas discutidas.

Os critérios de exclusão incluíram:

- Artigos publicados em outros idiomas além do inglês.
- Artigos duplicados em bases de dados.
- Estudos que foram retratados.
- Estudos sem o texto completo disponível.

As buscas nas bases de dados acadêmicas foram realizadas seguindo as estratégias de busca previamente elaborados durante os dias 18 a 23 de abril de 2024.

Foram inicialmente identificados 728 artigos de potencial relevância para a pesquisa. Após filtrá-los com base nos critérios de inclusão e exclusão mencionados, foram selecionados 40 artigos para análise mais detalhada. Para tornar o estudo mais focado, foi buscado o tema de maior relevância dentre dos artigos selecionados, sendo identificado o tópico GANs (redes adversárias generativas), dessa forma, foram selecionados apenas artigos que abordassem esse tópico. Esse processo resultou em oito artigos, cujos principais temas foram examinados, permitindo a seleção de textos com abordagens semelhantes. Com isso, foram escolhidos cinco artigos que mais contribuíram para a pesquisa, abordando as principais inovações, desafios e tendências do uso de inteligência artificial no campo das animações.

Foram selecionados os artigos “Research on the Generation of Creative Animation Driven by Deep Learning Model” escrito por Sang e Xu (2022), “Speech driven facial animation generation based on GAN” escrito por Li et al. (2022), “Image Style Transfer Using Deep Learning Methods” escrito por Ren e Sheng (2022), “A Simulation Model Based on DCGAN to Generate 2D Animation Avatars” escrito por Chen et al. (2022) e “CBA-GAN: Cartoonization style transformation based on the convolutional attention module” escrito por Zhang et al. (2023).

Para organizar e gerenciar os artigos selecionados, foi criado um banco de dados na ferramenta *Excel* para catalogar as informações essenciais de cada artigo, incluindo título, autores, ano de publicação, base de dados de origem, número de citações, palavras-chave principais, resumo e uma breve descrição das técnicas e algoritmos discutidos. Essa organização facilitou a comparação entre os artigos e permitiu uma análise mais sistemática dos temas emergentes.

4. Resultados

Nesta seção, serão apresentados os principais resultados da revisão sistemática, organizados nos quatro tópicos de maior ênfase nos cinco artigos selecionados: Aplicações de Redes Adversárias Generativas na geração de animações, Uso de Redes Convolucionais em animações e modelagem visual, Transferência de estilo de imagem e Geração de expressões faciais. Para cada técnica, será apresentado um breve resumo do seu funcionamento e suas principais contribuições para a animação.

Os resultados apresentados a seguir foram obtidos com base nos critérios de inclusão e exclusão estabelecidos na seção de Metodologia.

4.1. Aplicações de Redes Adversárias Generativas na geração de animações

As redes adversárias generativas (GANs) representam uma das inovações mais marcantes no campo da inteligência artificial (IA) e aprendizado profundo, as GANs revolucionaram a geração de dados sintéticos, permitindo a criação de imagens, vídeos, textos e outras formas de dados realistas [Sang and Xu 2022]. O conceito principal por trás das GANs é o treinamento de dois modelos neurais em conjunto: uma rede geradora (ou gerador) e uma rede discriminadora (ou discriminador), onde ambos os modelos se aprimoram mutuamente [Sang and Xu 2022] [Ren and Sheng 2022].

4.1.1. Conceitos Fundamentais

O princípio básico das GANs se fundamenta em um processo competitivo entre dois agentes: o gerador e o discriminador [Sang and Xu 2022] [Ren and Sheng 2022]. A função do gerador é criar novos exemplos de dados que se assemelhem ao conjunto de dados de treinamento. Inicialmente, o gerador começa com tentativas aleatórias de gerar dados a partir de um vetor de ruído [Sang and Xu 2022]. A função do discriminador, por sua vez, é identificar se os exemplos apresentados a ele são reais (provenientes do conjunto de dados de treinamento) ou falsos (gerados pelo gerador).A figura abaixo mostra como é a estrutura de um modelo GAN.

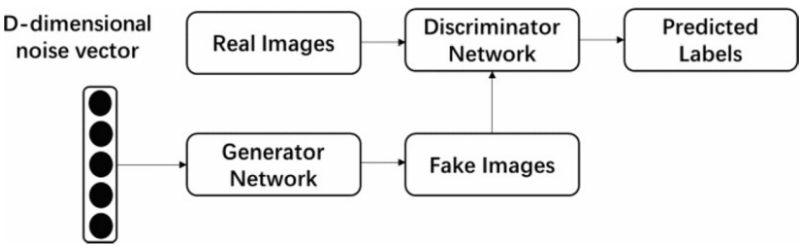

Figura 1. Estrutura de um modelo GAN. Fonte: (Ren and Sheng, 2022)

O treinamento das GANs ocorre de forma iterativa, com ambos os modelos se aprimorando ao longo do tempo. O gerador tenta "enganar" o discriminador, criando dados mais realistas, enquanto o discriminador tenta se tornar mais eficiente em detectar as falsificações. Essa competição leva a um equilíbrio em que o gerador se torna tão eficiente que os exemplos gerados são praticamente indistinguíveis dos dados reais [Sang and Xu 2022]. O objetivo final das GANs é minimizar a diferença entre os dados reais e os dados gerados, e, quando bem treinadas, as GANs são capazes de gerar amostras de dados altamente convincentes [Sang and Xu 2022].

4.1.2. Principais Componentes

O gerador tem como objetivo mapear um vetor de entrada, geralmente um vetor de ruído ou uma distribuição gaussiana, para uma amostra de dados no mesmo formato do conjunto de dados de treinamento. Em tarefas de geração de imagens, por exemplo, o gerador transforma esse vetor aleatório em uma imagem que se parece com as do conjunto de dados original. Durante o treinamento, ele tenta minimizar a capacidade do discriminador de identificar suas amostras como falsas [Ren and Sheng 2022].

O discriminador é uma rede neural que age como um classificador binário. Seu objetivo é diferenciar entre dados reais, extraídos do conjunto de dados de treinamento, e dados falsos, gerados pelo gerador [Ren and Sheng 2022]. O discriminador é treinado para maximizar sua precisão em distinguir entre essas duas fontes de dados, o que força o gerador a criar exemplos cada vez mais realistas [Chen et al. 2022].

Redes neurais convolucionais (CNN) possuem vantagens específicas no processamento de dados de imagem, sendo amplamente utilizadas como discriminadores em modelos GAN, já CNNs com uma estrutura de convolução transposta são habitualmente empregadas como geradoras [Sang and Xu 2022].

A função de perda das GANs é conhecida como perda adversarial, que pode ser interpretada como uma função de custo conjunta para os dois agentes. O gerador é treinado para minimizar a função de perda, enquanto o discriminador tenta maximizá-la, criando uma competição entre as duas redes.

4.1.3. Trabalhos relacionados

Semelhante ao modelo GAN tradicional, o modelo de transferência de estilo de ilustração proposto por Sang e Xu (2022) é composto por geradores e discriminadores. Para preservar melhor o conteúdo original das imagens e alcançar uma transferência eficaz de estilos artísticos nas ilustrações de animação, foi desenvolvida uma estrutura de gerador baseada no modelo básico de uma rede ResNet-18.

Para avaliar a capacidade de generalização do modelo aprimorado, Sang e Xu (2022) realizaram experimentos utilizando o conjunto de dados CelebFace. Nesse conjunto, foram coletadas 10.200 amostras e 202.677 imagens faciais com uma ampla variedade de estilos. Experimentos comparativos também foram realizados utilizando os modelos DCGAN e WGAN (Figura 2) originais.

Figura 2. Imagens geradas pelo modelo DCGAN e WGAN, respectivamente. Fonte: (Sang and Xu, 2022)

Figura 3. Imagens geradas pelo modelo GAN proposto. Fonte: (Sang and Xu, 2022)

Sang e Xu (2022) observaram que, em comparação com outros métodos, o conteúdo da imagem de animação gerado pelo DCGAN tradicional carece de autenticidade, e os detalhes faciais dos personagens de animação gerados são seriamente perdidos, o que dá às pessoas uma sensação de desarmonia, e toda a imagem parece o fenômeno do colapso de informações. A imagem gerada pelo modelo WGAN tem um bom desempenho no brilho de cor de toda a imagem, mas a qualidade da imagem é significativamente inferior aos outros dois métodos, e as características faciais do avatar da animação na imagem gerada não são claras.

Ren e Sheng (2022) discutem as principais descobertas de pesquisa e aplicações de aprendizado profundo em transferência de estilo de imagem das perspectivas de Redes Adversariais Generativas (GAN), incluindo CGAN, CycleGAN e Cartoon-GAN.

Uma extensão do GAN original, a Conditional Generative Adversarial Network (CGAN) foi proposta para superar as limitações dos modelos tradicionais, que não permitem controlar os resultados da tradução no domínio de destino e frequentemente produzem resultados sem diversidade. Atualmente, o CGAN tem apresentado resultados promissores em tarefas como a síntese de texto, extração de características de domínio e conversão de imagem. Diferentemente do GAN, o CGAN incorpora informações adicionais tanto no gerador quanto no discriminador, como rótulos de categorias ou dados provenientes de outras modalidades. Essas informações extras são inseridas como uma camada adicional de entrada para ambos os componentes. No gerador, o ruído de entrada é combinado com as informações suplementares em uma representação oculta compartilhada, enquanto no discriminador essas informações atuam como funções discriminantes, aprimorando a capacidade de diferenciação.

O CycleGAN é um modelo amplamente conhecido para tradução não supervisionada, uma técnica de aprendizado de máquina em que o objetivo é traduzir ou transformar dados de um domínio para outro sem ter pares diretos de exemplos correspondentes entre os dois domínios. Seu funcionamento é baseado em uma arquitetura de CGANs com a chamada ”perda de consistência cíclica”. Ele implementa dois GANs, cada um contendo um gerador e um discriminador, resultando em um total de quatro modelos na arquitetura. Diversos experimentos comprovaram sua eficácia na tradução não pareada de imagens, como pode ser observado na Figura 4.

Já o *Cartoon-GAN* é uma técnica projetada para transformar fotografias de cenas reais em imagens com estilo de desenho animado, como pode ser visto na Figura 5. No entanto, o estilo *cartoon*, que se caracteriza por um alto grau de abstração, simplificação, contornos nítidos, cores suaves e texturas simples, apresenta desafios para as abordagens tradicionais. Esses fatores complicam o processo de geração de imagens satisfatórias, especialmente ao lidar com funções de perda que dependem do descritor de textura.

Chen et al. (2022) propõe um algoritmo baseado em GAN para a geração de ilustrações em estilo de anime. Este algoritmo foi dividido em três etapas: *web crawling*, interceptação facial, e treinamento com visualização. Na primeira etapa, foram extraídas 8.000 ilustrações para serem usadas como conjunto de dados para o treinamento do modelo. A interceptação facial realiza a detecção de rostos nas imagens e as ajusta, recortando-as em avatares de formato quadrado e tamanho padronizado. Na etapa final, é aplicado um modelo baseado em CGAN.

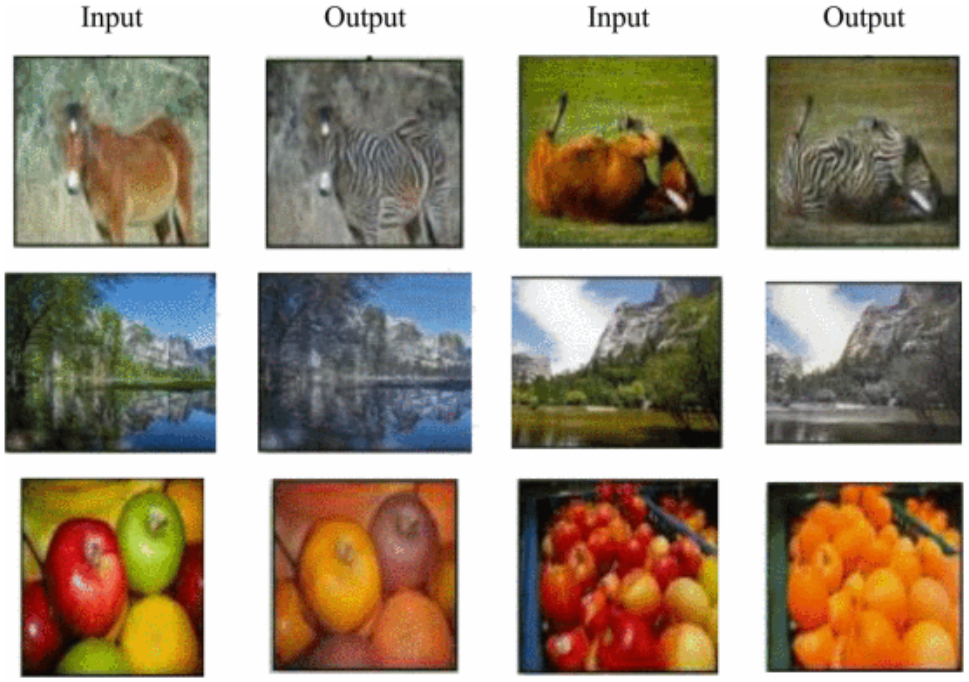

Figura 4. Exemplo de funcionamento do modelo CycleGAN. Fonte: (Ren and Sheng 2022)

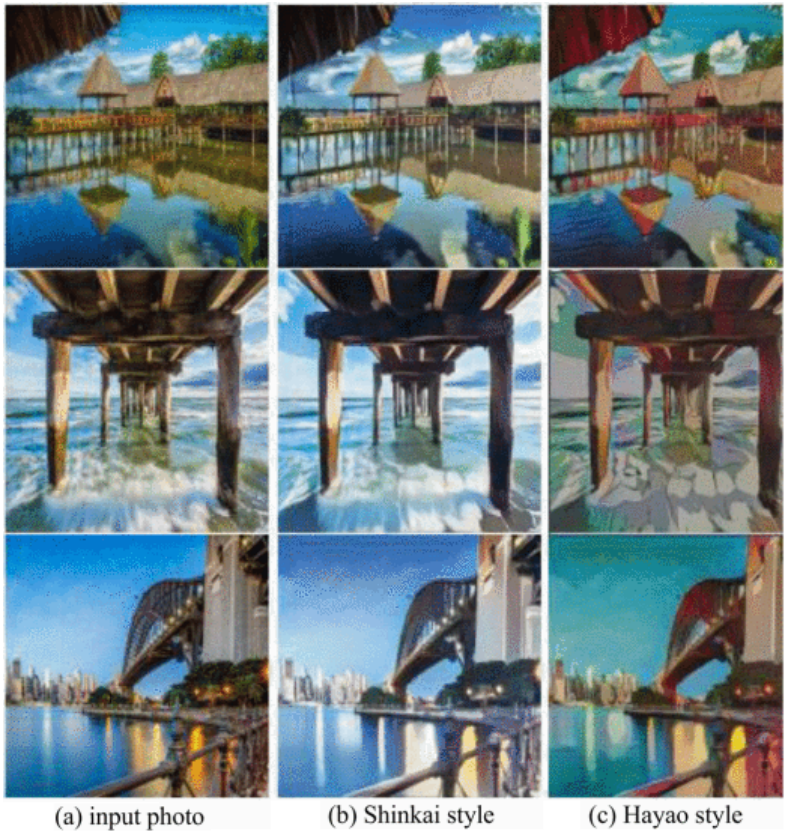

Figura 5. Exemplo de funcionamento do modelo Cartoon-GAN. Fonte: (Ren and Sheng 2022)

Chen et al. (2022) observaram que o algoritmo proposto pode gerar ilustrações com excelente qualidade visual. Após 200 épocas de treinamento, foi observada uma melhoria significativa nos resultados, com as imagens apresentando características faciais e expressões mais detalhadas. Por exemplo, os autores afirmam que os cabelos dos personagens exibem mais camadas e complexidade em comparação com os resultados anteriores.

Nos últimos anos, de acordo com Zhang et al. (2023), as redes adversárias generativas (GANs) têm demonstrado um desempenho notável na criação de imagens. Com base nessa tecnologia, muitos pesquisadores ao redor do mundo começaram a aplicar GANs para transformar fotos reais em desenhos animados, utilizando métodos como *CycleGAN*, *CartoonGAN*, *White-box cartoonize* e *AnimeGAN*. Segundo os autores, esses métodos geralmente incorporam perda de conteúdo semântico e perda de borda no modelo, com o objetivo de destacar os contornos e remover sombras. No entanto, o estilo *cartoon* resultante dessas transformações nem sempre é satisfatório — algumas imagens perdem conteúdo ou carecem de detalhes. Além disso, esses métodos exigem o treinamento de modelos específicos para cada estilo de desenho animado, o que limita sua flexibilidade.

Zhang et al. (2023) apresenta um modelo baseado em *CartoonGAN* denominado *CBA-GAN* para a transferência de estilo de imagem. Esse modelo utiliza aprendizado não supervisionado e é treinado com fotografias do mundo real e imagens de desenho animado. Para aprimorar a expressividade do modelo, acelerar o processo e melhorar a qualidade das ilustrações geradas, os elementos mais importantes da imagem recebem maior atenção, enquanto os detalhes menos relevantes são suprimidos. O *CBA-GAN* também se destaca por preservar com precisão as bordas, texturas e cores das imagens reais, com áreas sombreadas da saída final mantendo grande similaridade com as da imagem original. Além disso, Zhang et al. (2023) propõe um método de aprimoramento das bordas, garantindo que as imagens processadas com redução de ruído sejam convertidas em bordas no estilo de desenho animado.

Na fase de avaliação do modelo, Zhang et al. (2023) organizaram conjuntos de dados inspirados nos estilos de ilustração de três renomados diretores japoneses: Hayao Miyazaki, Makoto Shinkai e Mamoru Hosoda. Esses conjuntos foram empregados para treinar a transferência de estilo em paisagens. Para a transferência de estilo em rostos, foram utilizados os conjuntos de dados *Kyoto-face* e *PA-style*. Experimentos em estilo *cartoon* foram realizados utilizando imagens de rostos, alimentos, animais e paisagens, seus resultados podem ser observados na Figura 6.

De acordo com Zhang et al. (2023), os resultados experimentais obtidos demonstram a superioridade do *CBA-GAN* quando comparado a modelos similares, validando sua eficácia no processamento de cor, textura, borda e sombra na cartoonização de fotografias reais. Além disso, o modelo mostrou boa expansibilidade, sendo fácil adaptá-lo para cartoonização em vídeos. Apesar dos resultados promissores, Zhang et al. (2023) afirma o modelo apresenta algumas limitações, o treinamento utiliza um método de leitura aleatória dos dados, dificultando a aplicação de uma avaliação quantitativa consistente. Além disso, tanto o estilo quanto a paleta de cores dos conjuntos de dados influenciam diretamente os resultados, gerando instabilidade nos treinamentos.

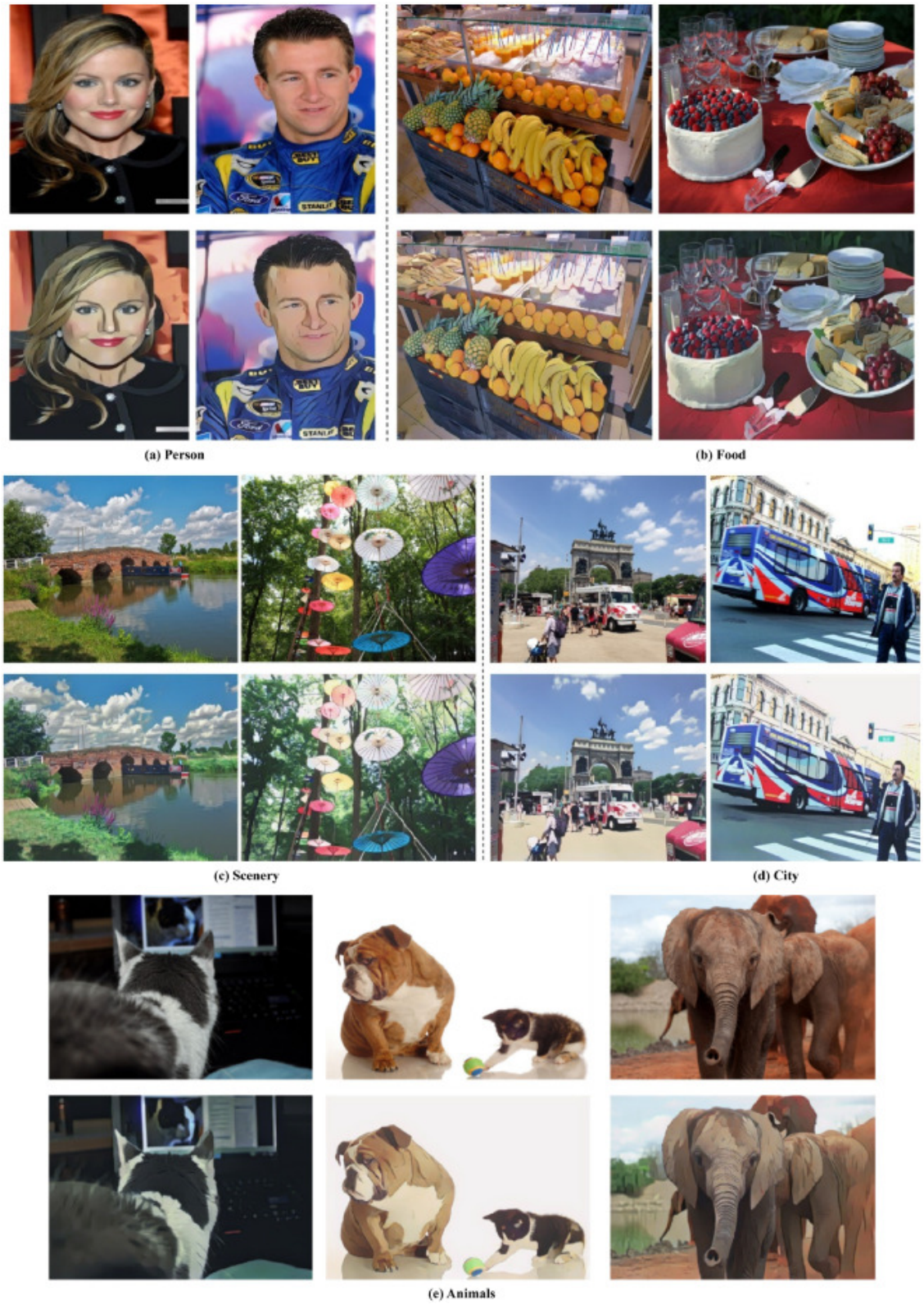

Figura 6. Imagens geradas pelo modelo CBA-GAN para cada uma das categorias. Fonte: (Zhang et al., 2023)

4.2. Uso de Redes Convolucionais em animações e modelagem visual

As Redes Neurais Convolucionais (CNNs) configuram-se como uma das arquiteturas mais robustas no campo do aprendizado profundo (*deep learning*), tendo suas raízes inspiradas na estrutura do córtex visual humano [Ren and Sheng 2022]. Redes convolucionais têm a capacidade de decompor e reorganizar tanto o conteúdo quanto o estilo de qualquer imagem, além de oferecer técnicas eficazes para a criação de imagens [Ren and Sheng 2022]. Estas redes são amplamente reconhecidas pela eficácia no processamento de dados em formato de grade [Ren and Sheng 2022], como imagens e vídeos, o que as torna especialmente úteis em aplicações de animação por computador e visão computacional.

4.2.1. Conceitos Fundamentais

Uma CNN é uma rede neural projetada para detectar padrões espaciais em dados visuais [Ren and Sheng 2022]. Sua principal distinção em relação a outras arquiteturas de redes está na utilização da operação de convolução [Ren and Sheng 2022], que permite à rede extrair características de forma hierárquica, desde características de baixo nível, como bordas e texturas, até características mais complexas, como formas e objetos inteiros.

As CNNs consistem em camadas de diferentes tipos, a Camada Convolucional é o componente central da CNN. Nesta camada, um filtro (ou kernel) é aplicado à imagem de entrada, deslocando-se por ela e gerando um mapa de características (*feature map*). Após as camadas convolucionais, as Camadas de *Pooling* reduzem a dimensionalidade dos mapas de características gerados, reduzindo a complexidade computacional sem perder as características mais importantes. Posteriormente, as Camadas de Ativação aplicam funções de ativação, como a *ReLU* (*Rectified Linear Unit*), para introduzir não-linearidade ao modelo, permitindo que ele capture padrões complexos nos dados. Por fim, as Camadas Totalmente Conectadas conectam todas as unidades da camada anterior com todas as unidades da próxima camada, essa conexão total é crucial para a combinação de características aprendidas, culminando na decisão final do modelo, como a classificação de uma imagem.

4.2.2. Principais Componentes

Os Filtros (*Kernels*) são pequenas matrizes aplicadas nas imagens de entrada para identificar padrões, como bordas ou texturas. Os Mapas de Características (*Feature Maps*) são resultados das operações de convolução que representam os padrões detectados na imagem. O *Stride* refere-se ao número de *pixels* pelos quais o filtro se desloca ao longo da imagem. *Strides* maiores resultam em saídas de resolução menor, enquanto *strides* menores oferecem maior detalhamento. O *Padding* é adição de bordas à imagem de entrada, permitindo que o filtro aplique a convolução em *pixels* nas extremidades, preservando o tamanho da imagem original.

4.2.3. Trabalhos relacionados

Ren e Sheng (2022) empregaram 16 camadas *pooling* e cinco camadas de agrupamento ao invés da camada totalmente conectada. Através da CNN, o modelo extraiu com sucesso os

recursos para formar o mapa de conteúdo do estilo da pintura a óleo, atingindo resultados ótimos em termos de transferência de estilo visual.

Além disso, os autores Sang e Xu (2022) demonstraram que, utilizando a ativação oculta de uma rede neural convolucional pré-treinada, é possível alcançar uma transferência de estilo visualmente atraente. O conteúdo e as informações de estilo foram interligados ao substituir as texturas em uma única camada. Embora haja apenas um nível de restrição, os autores afirmam que essa técnica gerou resultados interessantes. Além disso, o método mostrou-se simples de aplicar em vídeos, quadro a quadro, devido à sua consistência e ao ajuste intuitivo das configurações. Durante os testes, os pesquisadores constataram que a tecnologia podia ser usada tanto para criar novas imagens com estilo quanto para ajustar imagens estilizadas existentes, sem a necessidade de treinar novamente o modelo.

Segundo Ren e Sheng (2022), a maioria dos métodos de transferência de estilos visuais utiliza um modelo CNN previamente treinado em grandes conjuntos de imagens. Embora esses métodos consigam realizar a transferência de estilo com sucesso, a vasta quantidade de parâmetros necessários impõe limitações significativas à sua aplicação, o que acaba por desacelerar o avanço no campo da migração de estilo de imagem. A compressão e aceleração desses modelos têm sido áreas que também carecem de maior interesse. Implementar a compressão de modelos é um desafio futuro importante para o desenvolvimento da migração de estilo de imagem.

4.3. Transferência de estilo de imagem

A transferência de estilo de imagem é uma técnica avançada de manipulação visual baseada em inteligência artificial que visa combinar o conteúdo de uma imagem com o estilo visual de outra [Sang and Xu 2022] [Ren and Sheng 2022]. O conceito, impulsionado pelo uso de redes neurais profundas, tornou-se popular a partir do trabalho pioneiro de Gatys et al. (2015), que demonstrou a capacidade de redes convolucionais de separar e recombinar conteúdo e estilo de imagens de maneira eficaz.

Segundo Zhang et al. (2023), a produtividade da animação tradicional é relativamente baixa e os custos de produção são relativamente altos. Para economizar custos e acelerar a produção de animação, várias empresas de quadrinhos começaram a desenvolver algoritmos inteligentes para produzir ilustrações de alta qualidade.

4.3.1. Conceitos Fundamentais

A transferência de estilo de imagem envolve a combinação de duas imagens distintas: uma imagem de conteúdo e uma imagem de estilo [Ren and Sheng 2022]. A imagem de conteúdo mantém a estrutura e os elementos principais, enquanto a imagem de estilo fornece as características de textura e visuais, como cores e padrões. O objetivo é gerar uma nova imagem que preserve os elementos estruturais da primeira, mas reproduza a estética da segunda [Sang and Xu 2022] [Ren and Sheng 2022].

4.3.2. Principais Componentes

A técnica de transferência de estilo de imagem depende de vários componentes essenciais que trabalham em conjunto para alcançar resultados de alta qualidade. O principal componente é o modelo CNN [Sang and Xu 2022], que serve como o núcleo do processo de transferência de estilo. Ele é utilizado para extrair representações das características tanto da imagem de conteúdo quanto da imagem de estilo [Sang and Xu 2022]. O modelo mais comumente utilizado para esse propósito é o *VGG-19* [Gatys et al. 2015], uma arquitetura de rede neural pré-treinada em grandes conjuntos de dados, capaz de capturar detalhes de baixo nível, como bordas e texturas, além de estruturas de alto nível, como formas e objetos.

O processo de transferência de estilo é baseado na otimização de uma função de perda que equilibra duas componentes principais: a perda de conteúdo e a perda de estilo. A perda de conteúdo mede a diferença entre a imagem gerada e a imagem de conteúdo nas camadas mais profundas da rede, garantindo que a estrutura da imagem original seja preservada. Por outro lado, a perda de estilo mede a diferença entre as características de estilo da imagem gerada e da imagem de estilo, utilizando matrizes de correlação (conhecidas como gram matrices) para capturar as relações espaciais entre as características visuais extraídas nas camadas intermediárias da rede.

Esse processo é formulado como um problema de otimização, no qual a imagem gerada é ajustada iterativamente para minimizar a função de perda total. O algoritmo de gradiente descendente é frequentemente empregado para ajustar os pixels da imagem gerada, de modo a aproximá-la tanto do conteúdo quanto do estilo desejado.

4.3.3. Trabalhos relacionados

Sang e Xu (2022) afirmam que, na área de transferência de estilo de ilustração, a avaliação da qualidade dos resultados não deve se basear apenas no julgamento subjetivo da visão humana, mas também deve passar por uma análise quantitativa. Essa avaliação considera dois aspectos principais: a qualidade da imagem gerada, ou seja, se o conteúdo é realista e os detalhes são nítidos, e a diversidade das imagens geradas, ou seja, um bom modelo de geração deve produzir uma variedade de imagens, em vez de gerar repetidamente tipos semelhantes.

Segundo Ren e Sheng (2022), apesar dos avanços significativos na transferência de estilo de imagem utilizando aprendizado profundo, ainda existem desafios a serem superados. Um dos principais problemas é que cada modelo de treinamento exige ajustes manuais de parâmetros, o que torna o processo mais complexo e, frequentemente, resulta em uma qualidade de imagem abaixo do ideal. Assim, uma área crítica de pesquisa futura envolve o desenvolvimento de sistemas mais fáceis de controlar e que garantam maior qualidade nas imagens geradas.

Além disso, Ren e Sheng (2022) afirmam que a maioria dos métodos de transferência de estilo visual depende de modelos de redes neurais convolucionais (CNNs) pré-treinados em grandes conjuntos de imagens. Embora esses métodos consigam realizar a transferência de estilo com sucesso, o grande número de parâmetros nas CNNs limita

a sua aplicabilidade e desacelera o avanço da área. Portanto, a compressão de modelos é outro desafio importante a ser resolvido no futuro.

Outro aspecto crítico avaliado por Ren e Sheng (2022) são os resultados da transferência de estilo, uma vez que o conceito de estilo é abstrato e subjetivo. Embora existam índices de avaliação de qualidade de imagem, como Similaridade Estrutural (SSIM), Relação Sinal-Ruído de Pico (PSNR) e erro quadrático médio (MSE), os resultados frequentemente não correspondem à percepção humana. Esses índices falham em oferecer uma medida universal e precisa da qualidade, sendo necessário o desenvolvimento de uma técnica de avaliação padronizada que auxilie na melhoria dos algoritmos existentes.

4.4. Geração de expressões faciais

A geração de expressões faciais a partir de clipes de áudio de fala e imagens de rosto representa uma das fronteiras mais avançadas da síntese de animação facial, envolvendo a interseção de técnicas de visão computacional, processamento de áudio e aprendizado profundo. Esta tarefa tem como objetivo mapear automaticamente os movimentos faciais de um indivíduo com base no conteúdo de um clipe de áudio e em uma imagem estática, capturando tanto a sincronização labial quanto as expressões faciais dinâmicas [Li et al. 2022] associadas às nuances emocionais da fala.

4.4.1. Conceitos Fundamentais

A síntese de expressões faciais a partir de dados de áudio é um problema clássico de geração de imagens condicionada, onde o conteúdo de fala, representado pelo clipe de áudio, serve como uma entrada condicional para guiar a geração ou animação das expressões faciais [Li et al. 2022]. A entrada geralmente inclui uma imagem estática do rosto do indivíduo, que atua como referência para o modelo de geração, garantindo que as características faciais (como a forma dos olhos, boca e nariz) sejam preservadas, ao mesmo tempo em que a fala induz o movimento apropriado dos lábios e demais partes do rosto. Para realizar essa tarefa, o uso de redes neurais profundas tem se mostrado essencial.

4.4.2. Principais Componentes

A geração de expressões faciais realistas envolve a integração de diferentes componentes tecnológicos, que trabalham em conjunto para alcançar resultados sincronizados e naturais [Li et al. 2022]. O áudio da fala é transformado em uma representação numérica apropriada para ser utilizada como entrada em redes neurais, por meio da extração de características acústicas, como espectrogramas, *MFCCs* (*Mel Frequency Cepstral Coefficients*) ou *embeddings* de áudio, que capturam as propriedades fonéticas e prosódicas da fala. Simultaneamente, a imagem do rosto é processada por outra rede neural, que extrai uma representação vetorial latente. Essa representação serve como base para a geração da sequência de expressões faciais, garantindo que as características faciais individuais sejam preservadas ao longo da animação.

Um dos maiores desafios desse processo é garantir que os movimentos dos lábios estejam precisamente sincronizados com o conteúdo da fala. Isso requer uma modelagem

exata das relações temporais entre o áudio e as variações na forma da boca ao longo do tempo. Além da sincronização labial, é essencial que a geração de expressões faciais capture as emoções e nuances associadas à fala, como sorrisos, franzimentos de sobrancelha e movimentos oculares. Essas expressões globais são cruciais para transmitir as emoções e o contexto subjacente da comunicação verbal.

4.4.3. Trabalhos relacionados

Li et al. (2022) propuseram um modelo de geração de animação facial baseado em uma rede adversarial generativa (GAN), chamada *Facial Animation Adversarial Network* (FAAN), para produzir animações faciais sincronizadas com a fala a partir de gravações de fala humana e uma imagem facial. O modelo mapeia as características da imagem do rosto e da fala para um espaço comum durante o processo de codificação, gerando em seguida uma sequência de quadros de um rosto em movimento, de acordo com as características de coerência temporal presentes nos fragmentos de fala. A estrutura do modelo FAAN pode ser observada na Figura 7.

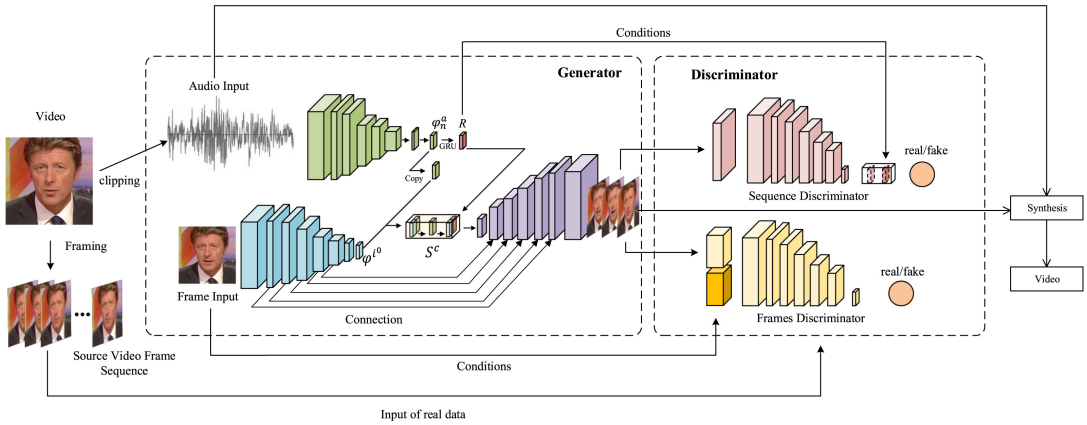

Figura 7. Estrutura do modelo FAAN. Fonte: (Li et al., 2022)

No modelo FAAN, o áudio da fala é segmentado em clipes, e os recursos de cada clipe são extraídos por um codificador. Esses recursos são então processados por uma Unidade Recorrente Fechada (GRU) para extrair uma sequência de recursos de áudio que preservam a coerência temporal. O FAAN utiliza tanto os clipes de áudio quanto uma única imagem facial como entradas. Como resultado, a sequência de quadros de vídeo gerada é sincronizada de forma precisa com o áudio original, e as características faciais do vídeo produzido são bem preservadas. Dessa forma, o modelo proposto por Li et al. (2022) se destaca na coordenação dos movimentos labiais na animação facial gerada.

A eficiência do modelo é avaliada utilizando o *Peak Signal-to-Noise Ratio* (PSNR), o *Structural Similarity Index Measure* (SSIM) e o *Frechet Inception Distance* (FID). Os resultados experimentais indicam que as pontuações de PSNR e SSIM do modelo FAAN superam as de outros modelos mais recentes que utilizam o conjunto de dados GRID. Por fim, a validade do modelo proposto é comprovada pela geração de um diagrama de fluxo óptico dos quadros de vídeo, o qual demonstra que o modelo é capaz

de conduzir de forma detalhada a animação dos movimentos da boca nos vídeos gerados.

5. Discussão

A presente seção analisa os principais avanços e desafios identificados na revisão sistemática, com foco nas implicações práticas, técnicas e teóricas das técnicas de IA para geração de animações. Além disso, são destacadas limitações observadas e possíveis caminhos para pesquisas futuras.

5.1. Aplicações de Redes Adversárias Generativas na geração de animações

Uma das aplicações mais notáveis das GANs é a criação de imagens artificiais que se assemelham a fotografias reais. Exemplos incluem rostos humanos, objetos, cenários urbanos e até mesmo obras de arte [Sang and Xu 2022]. GANs têm sido utilizadas em tarefas de super-resolução [Ren and Sheng 2022], onde imagens de baixa qualidade ou baixa resolução são aprimoradas para versões de alta qualidade. GANs também são amplamente utilizadas em transferência de estilo de imagem [Sang and Xu 2022] [Ren and Sheng 2022], como a aplicação do estilo de uma pintura famosa em uma fotografia, ou a transformação de uma imagem de inverno em uma versão de verão.

Apesar de seu enorme potencial, as GANs enfrentam diversos desafios. O primeiro e mais importante é o problema de treinamento instável [Sang and Xu 2022]. Como as GANs envolvem duas redes em competição, o treinamento pode ser delicado, levando a situações em que o gerador ou o discriminador se sobrepõem um ao outro, resultando em uma má qualidade dos dados gerados [Sang and Xu 2022]. Além disso, GANs ainda enfrentam dificuldades na generalização e na produção de resultados consistentes [Sang and Xu 2022]. Embora possam gerar exemplos altamente realistas em alguns casos, elas ainda podem falhar em tarefas onde a variação dos dados de entrada é alta ou onde a precisão dos detalhes é crucial.

A introdução das Wasserstein GANs (WGANs) [Sang and Xu 2022], que propõem uma nova função de perda baseada na distância de Wasserstein, melhora a estabilidade do treinamento e mitiga o problema do treinamento instável.

As CycleGANs [Ren and Sheng 2022] [Zhang et al. 2023] são outro avanço significativo, permitindo a transferência de estilo de imagens de um domínio para outro sem a necessidade de pares de imagens correspondentes. Essa abordagem tem sido amplamente utilizada em tradução de imagens, como na conversão de fotos de verão para fotos de inverno, ou na transformação de pinturas em fotos realistas.

O modelo proposto por Sang e Xu (2022) pode gerar imagens de animação com excelente qualidade visual. Em comparação com as imagens geradas pelos modelos como DCGAN e WGAN, o método proposto alcançou um equilíbrio superior entre o estilo da imagem e o conteúdo original, com os detalhes da imagem gerada sendo mais claros e obtendo uma saturação de cor mais alta.

5.2. Uso de Redes Convolucionais em animações e modelagem visual

Um dos principais desafios das CNNs é a necessidade de grandes volumes de dados rotulados para treinar os modelos de forma eficaz, o que pode ser uma barreira em domínios que não possuem amplos conjuntos de dados disponíveis. Outro desafio é o risco de

overfitting, que ocorre quando a rede se ajusta excessivamente aos dados de treinamento, prejudicando sua capacidade de generalizar para novos dados. Por fim, as CNNs, assim como muitas outras redes profundas, sofrem com a limitada explicabilidade de suas decisões, sendo frequentemente vistas como ”caixas-pretas”. Essa opacidade pode ser problemática em cenários mais complexos, como a criação de animações realistas, onde a compreensão dos mecanismos internos da rede poderia ser valiosa.

5.3. Transferência de estilo de imagem

A transferência de estilo de imagem oferece diversas vantagens em relação a técnicas tradicionais de manipulação de imagens, primeiramente, ela possibilita uma fusão flexível entre conteúdo e estilo, permitindo que elementos de diferentes fontes sejam combinados de maneira automática e coerente. Isso reduz a necessidade de intervenção manual e facilita a produção em larga escala, o que é especialmente valioso em indústrias como a de entretenimento e design gráfico. Outra vantagem importante é a capacidade de reutilizar modelos de CNN pré-treinados que já possuem um forte entendimento das características visuais de milhares de imagens. Isso significa que, mesmo com pouco ou nenhum treinamento adicional, esses modelos podem ser aplicados de forma eficaz em uma ampla variedade de tarefas de transferência de estilo.

A transferência de estilo de imagem tem uma ampla gama de aplicações, como por exemplo na arte digital [Ren and Sheng 2022], a técnica permite que artistas criem novas obras ao aplicar o estilo de pintores famosos, como Van Gogh ou Picasso, em suas próprias fotografias ou desenhos. Isso democratiza o acesso a estilos artísticos e facilita a experimentação criativa. Na produção de filmes e animações, a transferência de estilo pode ser usada para gerar efeitos visuais únicos [Ren and Sheng 2022], como transformar uma cena filmada em um estilo artístico específico, economizando tempo e recursos que seriam necessários para o design manual. Outra aplicação importante é na realidade aumentada e realidade virtual, onde a técnica pode ser utilizada para alterar o estilo visual de ambientes virtuais em tempo real, proporcionando uma experiência visual mais imersiva e personalizada. No campo da comunicação visual, essa técnica pode ser utilizada para adaptar materiais promocionais ao estilo visual de diferentes marcas ou campanhas.

Apesar de suas inúmeras vantagens, a transferência de estilo de imagem enfrenta desafios consideráveis, sendo um dos principais problemas o custo computacional. A técnica requer grande poder de processamento gráfico, o que pode ser inviável para sistemas com hardware limitado. Outro desafio significativo é a qualidade do resultado gerado [Zhang et al. 2023]. Embora a transferência de estilo tenha alcançado resultados notáveis, ainda existem dificuldades em preservar tanto o conteúdo quanto o estilo de forma perfeita. Em muitos casos, as imagens geradas podem apresentar artefatos visuais indesejados ou distorções.

Além disso, a técnica ainda apresenta limitações na generalização [Zhang et al. 2023]. Modelos de transferência de estilo geralmente funcionam bem com um conjunto restrito de estilos e conteúdos, mas podem ter dificuldades em lidar com estilos ou imagens de conteúdo muito divergentes das que foram originalmente utilizadas no treinamento da rede neural. Isso pode resultar em imagens de baixa qualidade ou incapacidade de capturar adequadamente as nuances de um estilo artístico específico.

5.4. Geração de expressões faciais

A geração de expressões faciais automatiza um processo complexo que, tradicionalmente, exigiria uma equipe dedicada de animadores e técnicas de captura de movimento, tornando-o mais acessível e eficiente. Além disso, permite a criação de expressões faciais altamente detalhadas e sincronizadas com a fala, melhorando a qualidade e o realismo da animação. Outro benefício significativo é a possibilidade de gerar animações faciais em tempo real, o que é crucial para interações em ambientes virtuais e assistentes digitais. Finalmente, essa técnica oferece uma flexibilidade considerável, pois pode ser aplicada a qualquer rosto humano ou personagem, desde que uma imagem de referência esteja disponível.

Segundo Li et al. (2022), apesar dos avanços na geração de expressões faciais, a maioria das abordagens ainda enfrenta desafios significativos. Um dos problemas mais comuns é o movimento restrito apenas à boca, já que muitos métodos geram esses movimentos com base no mapeamento de recursos de áudio ou na animação de pontos de referência a partir de vídeos existentes. Essas técnicas, contudo, não conseguem refletir mudanças mais amplas no rosto, limitando-se apenas à ação da boca.

As imagens geradas pelo modelo proposta por Li et al. (2022) conseguem refletir tanto as características de identidade do rosto quanto as mudanças de movimento durante a fala. Isso significa que a restauração das características não apenas preserva os traços originais da imagem facial de entrada, mas também aprimora o processo de geração, garantindo que a imagem facial resultante esteja em conformidade com a distribuição de características da imagem facial original.

6. Conclussão

A presente revisão sistemática revelou algoritmos e modelos de inteligência artificial que podem ajudar animadores a criar trabalhos melhores e com maior qualidade. As GANs destacaram-se como uma ferramenta poderosa para a criação de conteúdo visual sintético, possibilitando aplicações que vão desde a geração de imagens altamente realistas até a transferência de estilos artísticos específicos. As CNNs, por sua vez, mostraram-se essenciais na análise e manipulação de imagens, sendo amplamente aplicadas em tarefas que exigem detecção e reconhecimento de padrões visuais complexos, como na modelagem e na composição de cenas. A transferência de estilo, com sua capacidade de combinar a estrutura de uma imagem com as características estilísticas de outra, abriu novas possibilidades para a criação artística e a personalização visual em animações. Já a geração de expressões faciais sincronizadas com áudio, embora desafiante, demonstrou seu potencial para aplicações que exigem interação e resposta em tempo real, como assistentes digitais e avatares virtuais.

Apesar dos avanços substanciais, esta revisão também identificou desafios significativos. A estabilidade no treinamento de GANs e a necessidade de grandes conjuntos de dados rotulados para CNNs e outras arquiteturas profundas representam barreiras para a generalização e aplicação desses métodos. Além disso, o alto custo computacional e a limitada capacidade de controle sobre os resultados da transferência de estilo de imagem ainda precisam de desenvolvimento. A geração de expressões faciais a partir do áudio também enfrenta dificuldades para capturar nuances emocionais complexas e para

adaptar-se a características faciais variadas, exigindo refinamentos para alcançar uma performance mais robusta e generalizável.

Para o futuro, espera-se que sejam desenvolvidos modelos mais eficientes e adaptáveis, que possam lidar com a diversidade e complexidade das demandas da animação digital. O desenvolvimento de métodos que permitam um controle mais preciso sobre os parâmetros estéticos e técnicos da geração de conteúdo visual, bem como técnicas de treinamento que exijam menos dados rotulados, será essencial para consolidar de vez o uso da inteligência artificial na animação.

Referências Bibliográficas

Chen, K., Gao, C., and Cai, Y. (2022). A simulation model based on dcgan to generate 2d animation avatars. In *International Conference on Cloud Computing, Internet of Things, and Computer Applications (CICA 2022)*, volume 12303, pages 512–518. SPIE.

Gatys, L., Ecker, A. S., and Bethge, M. (2015). Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28.

Li, X., Zhang, J., and Liu, Y. (2022). Speech driven facial animation generation based on gan. *Displays*, 74:102260.

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *bmj*, 372.

Ren, S. and Sheng, Y. (2022). Image style transfer using deep learning methods. In *2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)*, pages 1190–1195. IEEE.

Sang, X. and Xu, L. (2022). Research on the generation of creative animation driven by deep learning model. *Scientific Programming*, 2022(1):5815693.

Zhang, F., Zhao, H., Li, Y., Wu, Y., and Sun, X. (2023). Cba-gan: Cartoonization style transformation based on the convolutional attention module. *Computers and Electrical Engineering*, 106:108575.