Universidade Federal do Rio Grande do Norte
Centro de Tecnologia
Graduação em Engenharia de Computação

Adaptação da StyleGAN2 para Geração Bidirecional de Faces Individuais via Transfer Learning

Jonas Peixoto da silva

Orientador: Prof. Dr. Adrião Duarte Dória Neto

Coorientador: Prof. Dr. Keylly Eyglys Araújo dos Santos

Natal, RN, 9 de dezembro de 2025

Universidade Federal do Rio Grande do Norte
Centro de Tecnologia
Graduação em Engenharia de Computação

Adaptação da StyleGAN2 para Geração Bidirecional de Faces Individuais via Transfer Learning

Jonas Peixoto da Silva

Orientador: Prof. Dr. Adrião Duarte Dória Neto

Coorientador: Prof. Dr. Keylly Eyglys Araújo dos Santos

Trabalho de Conclusão de Curso de Graduação na modalidade Monografia, submetido como parte dos requisitos necessários para conclusão do curso de Engenharia de Computação pela Universidade Federal do Rio Grande do Norte (UFRN/CT).

Natal, RN, 9 de dezembro de 2025

Universidade Federal do Rio Grande do Norte - UFRN
Sistema de Bibliotecas - SISBI
Catalogação de Publicação na Fonte. UFRN - Biblioteca Central Zila Mamede

Silva, Jonas Peixoto da.
Adaptação da StyleGAN2 para geração bidirecional de faces individuais via transfer learning / Jonas Peixoto da Silva. - 2025.
37 f.: il.

Monografia (graduação) - Universidade Federal do Rio Grande do Norte, Centro de Tecnologia, Curso de Engenharia de Computação, Natal, RN, 2025.
Orientação: Prof. Dr. Adrião Duarte Dória Neto.
Coorientação: Prof. Dr. Keylly Eyglys Araújo dos Santos.

1. Redes generativas adversárias - Monografia. 2. StyleGan - Monografia. 3. Transfer learning - Monografia. 4. Síntese facial - Monografia. I. Dória Neto, Adrião Duarte. II. Santos, Keylly Eyglys Araújo dos. III. Título.

RN/UF/BCZM CDU 004

Elaborado por Jackeline dos Santos Pinheiro da Silva Maia Cavalcanti - CRB-15/317

Adaptação da StyleGAN2 para Geração Bidirecional de Faces Individuais via Transfer Learning

Jonas Peixoto da Silva

Monografia aprovada em 24 de novembro de 2025, pela banca examinadora composta pelos seguintes membros:

Prof. Dr. Adrião Duarte Doria Neto (orientador) ..... DCA/UFRN

Prof. Dr. Keylly Eyglys Araújo dos Santos (coorientador) ..... IFRN

Prof. Dr. Daniel Lopes Martins ..... IMD/UFRN

Prof. Dr. Thiago Henrique Freire de Oliveira ..... IFRN

AGRADECIMENTOS

A realização deste projeto foi uma caminhada repleta de apoio e carinho de pessoas muito especiais que estiveram ao meu lado durante toda a jornada.

Primeiramente, sou grato a Deus por me conceder forças, saúde e sabedoria para chegar até aqui.

Ao meu pai, João Maria da Silva, expresso gratidão por sua dedicação incansável ao trabalho, proporcionando tudo o que foi necessário para que eu pudesse me dedicar aos estudos e seguir firme nesta caminhada. Sou grato também pelas palavras de incentivo, pelo apoio constante e pelo amor que sempre esteve presente.

Expresso minha gratidão à minha mãe, Josivânia Peixoto Bezerra da Silva, que sempre cuidou de mim com muito amor, carinho e dedicação incondicional. Sua presença tornou meus dias mais leves e me deu forças para seguir em frente.

Minha gratidão também à minha namorada, Nathalia Ariele, pela paciência, carinho, amor e cuidado durante toda essa caminhada, sempre me apoiando e me incentivando a continuar.

Ao meu amigo Matheus Teodósio, deixo um agradecimento pelo incentivo constante, pelas palavras de motivação e por acreditar em mim desde o início desta trajetória.

Expresso ainda minha gratidão ao professor Adrião Duarte Dória Neto, meu orientador, e ao professor Keylly Eyglys Araújo, meu coorientador, pelos ensinamentos, pela dedicação nas reuniões de alinhamento sempre conduzidas com muita atenção e pelos conhecimentos técnicos, que foram essenciais para a construção deste projeto.

Por fim, estendo minha gratidão a todos os familiares, em especial à minha irmã Josy Karoline, e aos amigos que contribuíram de forma positiva para a concretização deste objetivo.

RESUMO

Este trabalho apresenta uma abordagem para otimizar a geração de rostos individuais utilizando *transfer learning* com base na rede *StyleGAN2*. O objetivo é reduzir o tempo de treinamento e evitar a necessidade de iniciar cada geração facial do zero, mantendo ao mesmo tempo alta qualidade nas imagens geradas. Foi adaptada uma arquitetura acoplada ao gerador pré-treinado, com o objetivo de promover a especialização, assim como congelamento seletivo de partes do gerador original e do discriminador para permitir a otimização do processamento. Em sequência foi aliada a um *encoder* independente responsável por mapear imagens reais para o espaço latente, possibilitando a reconstrução dos rostos a partir de seus vetores latentes. Os resultados obtidos demonstraram gerações e reconstruções consistentes, evidenciando o potencial da metodologia para aplicações em síntese facial personalizada.

Palavras-chave: redes generativas adversárias; stylegan; transfer learning; síntese facial; encoder.

ABSTRACT

This work presents an approach to optimize the generation of individual faces using transfer learning based on the *StyleGAN2* network. The objective is to reduce training time and avoid the need to start each facial generation from scratch while maintaining high quality in the generated images. An architecture was adapted and coupled to the pretrained generator, aiming to promote specialization, as well as the selective freezing of parts of the original generator and the discriminator to enable processing optimization. Subsequently, the approach was combined with an independent encoder responsible for mapping real images to the latent space, allowing the reconstruction of faces from their latent vectors. The obtained results demonstrated consistent generations and reconstructions, evidencing the potential of the methodology for personalized facial synthesis applications.

Keywords: generative adversarial networks; stylegan;transfer learning; facial synthesis; encoder.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização Teórica

As redes generativas adversárias (GANs) têm, desde sua proposta original (GOODFELLOW et al., 2014), ganhado cada vez mais destaque no campo de aprendizado profundo. Essas redes têm uma alta capacidade de geração de dados sintéticos que se assemelham aos dados reais. A arquitetura das mesmas, em sua essência, é simples: dois modelos competindo entre si, um com a capacidade de gerar imagens e outro para avaliar a veracidade dessa geração. Essa interação adversarial resulta em modelos capazes de produzir imagens com elevado grau de realismo, proporcionando aplicações em diversas áreas, incluindo a síntese de rostos.

Em paralelo, técnicas baseadas em transferência de aprendizado (*transfer learning*), têm se destacado por possibilitar a transferência de conhecimento de um modelo previamente treinado, para outro. Essa abordagem é relevante, pois permite que o novo modelo possa herdar informações já adquiridas anteriormente, sem a necessidade de ser treinado completamente do zero. Como resultado, a utilização de transfer learning tende a acelerar o treinamento de uma nova tarefa, reduzindo o tempo de treinamento e diminuindo o consumo dos recursos computacionais.

1.2 Motivação

Modelos de síntese de rostos têm se destacado em diversas pesquisas no campo científico para solucionar diversos problemas. Um exemplo relevante é o trabalho de (SANTOS, 2024), que explora a geração de sínteses faciais como estratégia para otimizar a transmissão de dados em interações online. Essa abordagem demonstra o potencial das gerações sintéticas como solução para reduzir o volume de dados em interações virtuais. Esse trabalho será melhor detalhado na subseção 1.3.2. É importante destacar que um dos objetivos do estudo do presente trabalho é apresentar uma solução alternativa à proposta de (SANTOS, 2024), propondo-se a superar alguns gargalos identificados no processo do treinamento, como o alto tempo de treinamento que, segundo o autor, pode levar cerca de **24 horas** para um treinamento eficiente, bem como a busca por melhorias na qualidade das imagens geradas, uma vez que a geração das imagens foi realizada na resolução $\mathbf{64 \times 64}$.

Outro ponto de destaque é que apesar de as GANs serem capazes de gerar imagens sintéticas com alto nível de qualidade, isso está diretamente relacionado à complexidade de sua arquitetura. Um maior nível de detalhamento de uma imagem como por exemplo,

uma alta resolução exige uma rede mais complexa e, consequentemente, um tempo de treinamento maior. No contexto de uma GAN especializada em rostos individuais, que é um dos objetivos principais deste trabalho, a rede precisaria aprender completamente do zero todas as vezes que fosse necessário se especializar em cada rosto, o que resultaria em um alto número de parâmetros treináveis e um tempo de processamento elevado, enfrentando esses desafios a cada novo treinamento.

A possibilidade de aproveitar modelos pré-treinados, como o *StyleGAN* e adaptá-los por meio de transferência de aprendizado, representa uma solução para esse desafio, permitindo a especialização individualizada sem treinar a rede totalmente do zero. Além disso, combinar essa técnica de geração com outras técnicas de reconstrução de imagens, como por exemplo, a utilização de *encoders* pode possibilitar a reconstrução individualizada em alta qualidade, possibilitando também aplicações e otimização de transmissão de dados sintéticos, como destacado por (SANTOS, 2024).

1.3 Trabalhos Relacionados

Este capítulo apresenta trabalhos relacionados ao desenvolvimento deste projeto, destacando avanços relevantes nas áreas de síntese de imagens com GANs e aplicações com representações faciais.

1.3.1 Evolução das GANs para Síntese de Imagens

A geração de imagens sintéticas avançou significativamente desde a introdução das GANs por Goodfellow et al. (2014). Um dos primeiros modelos amplamente utilizados foi o *DCGAN* (RADFORD; METZ; CHINTALA, 2016), que estabeleceu arquiteturas baseadas em convoluções, removendo camadas totalmente conectadas e incorporando o uso de *batch normalization*, o que resultou em um treinamento mais estável e favoreceu a aplicação das GANs em tarefas de aprendizado não supervisionado. Posteriormente, o modelo *Progressive Growing of GANs*, proposto por Karras et al. (2017), trouxe avanços significativos ao introduzir o treinamento progressivo, no qual a rede aumenta gradualmente sua resolução ao longo do processo de treinamento. Essa estratégia permitiu a geração de rostos de alta resolução, alcançando imagens de até 1024×1024 pixels. A evolução continuou com o surgimento do *StyleGAN* (KARRAS; LAINE; AILA, 2018) e, posteriormente, do *StyleGAN2* (KARRAS et al., 2019), que introduziram mecanismos de controle estilístico, ampliando a qualidade e o controle na síntese de imagens, incluindo a geração de faces em alta qualidade.

1.3.2 Representação Facial para Interações Online

O trabalho de Santos (2024) apresenta uma proposta relevante ao explorar o uso das redes generativas para a otimização da transmissão de dados. A técnica de rede utilizada pelo autor baseia-se nas Redes Generativas Adversárias Bidirecionais (BiGANs). Dessa forma, o autor utiliza a capacidade das mesmas para gerar representações faciais compactas das imagens, possibilitando aplicação, por exemplo, em ambientes virtuais, e reduzindo a quantidade de dados necessários e transmitidos durante uma reunião online.

Na Figura 1, é possível observar a metodologia aplicada pelo autor, representada em um fluxograma que descreve o funcionamento da rede. O processo inicia-se com a captura de frames, seguida pelas etapas de separação, aumento de dados e treinamento bidirecional. Em seguida, ocorre a separação das redes geradora e encoder.

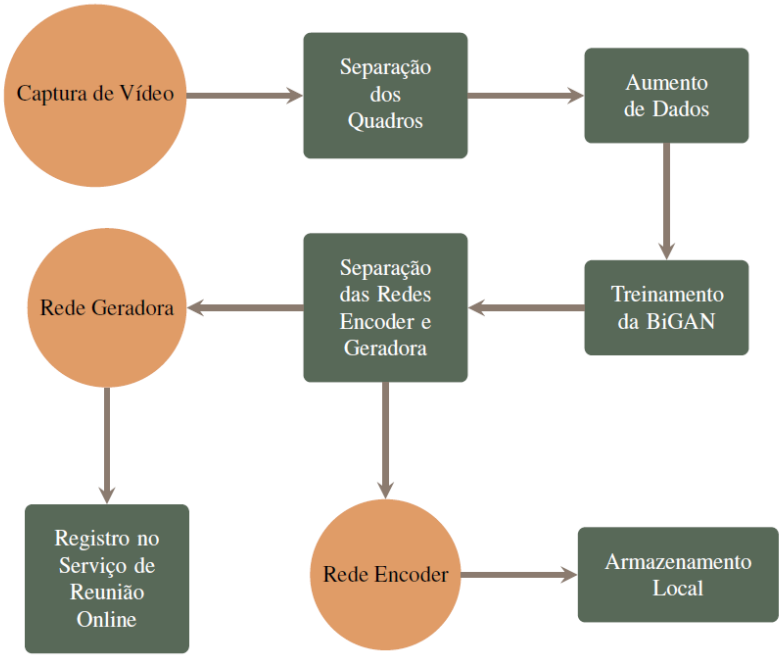

Figura 1 – Fluxo funcionamento rede BIGAN

Fonte: Adaptado de Santos (2024)

A rede encoder é responsável por codificar as características faciais dos participantes em um vetor latente, enquanto a rede geradora será utilizada posteriormente para reconstruir as imagens a partir desses vetores.

Na Figura 2, é possível visualizar a forma de aplicação dessa rede. Observa-se que cada usuário possui sua própria rede encoder armazenada localmente, enquanto os demais usuários mantêm as redes geradoras correspondentes. O fluxo de funcionamento ocorre a partir da captação de uma imagem pela rede encoder e da reconstrução dessa imagem pelas redes geradoras de cada usuário.

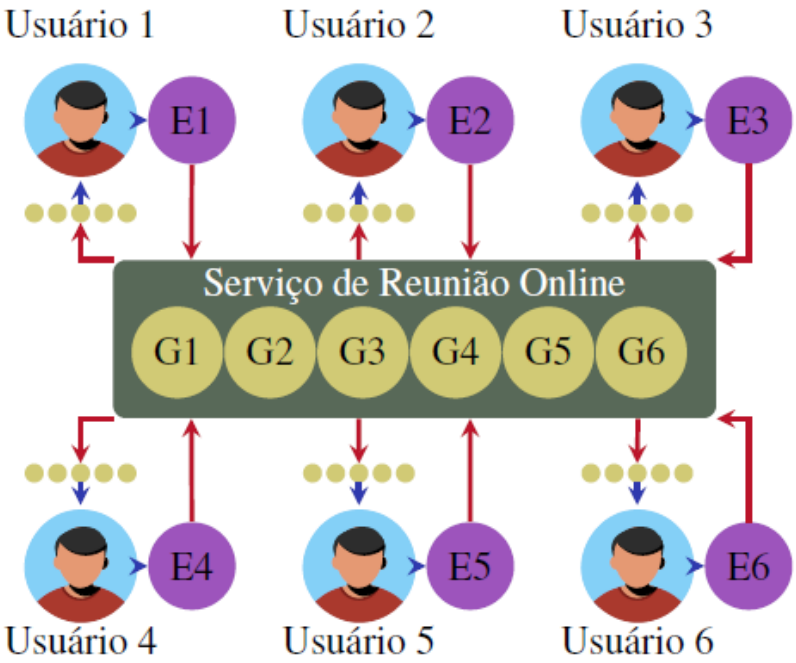

Figura 2 – Fluxo de Interação

Fonte: Adaptado de Santos (2024)

O método proposto apresentou uma redução no tamanho dos dados transmitidos, passando de cerca de 577.536 e 67.584 bytes utilizando o padrão de compressão MPEG-2 para apenas 13.008 bytes com a metodologia adotada. Essa redução demonstra uma boa compressão, resultando economia de largura de banda durante, por exemplo, uma transmissão de vídeo. Mesmo com essa compactação, o modelo manteve um índice de medida da similaridade estrutural (SSIM) de 0,6945. O SSIM varia de 0 a 1; valores mais próximos de 1 indicam que a imagem está muito semelhante à original, enquanto valores mais próximos de 0 mostram que houve perda significativa de qualidade, evidenciando que, apesar da redução no volume de dados, foi possível manter um nível adequado de semelhança entre as imagens.

1.4 Objetivos

Nesta seção, serão apresentados os objetivos deste trabalho, que guiam o desenvolvimento da metodologia proposta. Primeiramente, será apresentado o objetivo geral, seguido pelos objetivos específicos, que detalham as etapas do estudo.

1.4.1 Objetivo geral

Desenvolver uma metodologia baseada em transferência de aprendizado para adaptar modelos pré-treinados, em especial utilizando a *StyleGAN*, visando a geração otimizada

de rostos individuais. Assim como fazer utilização de encoders para fazer o mapeamento inverso das imagens para possibilitar a reconstrução do rosto alvo.

1.4.2 Objetivos Específicos

- Coletar dados individuais de rostos para treinamento.
- Desenvolver uma arquitetura adaptada da *StyleGAN*, que mantenha uma base genérica congelada e permita a criação de uma rede alternativa especializada para a geração de rostos individuais, evitando assim o treinamento completo do zero.
- Construir um encoder responsável por mapear qualquer imagem de entrada de rosto para o espaço latente correspondente no conjunto de treinamento do gerador.
- Desenvolver um modelo, capaz de gerar imagens faciais com boa qualidade visual, mantendo eficiência no processamento.
- Treinar e validar o modelo adaptado, avaliando a qualidade das imagens geradas.

1.5 Estrutura do Trabalho

Este trabalho apresenta uma introdução sobre o tema, mostrando os fatores que motivam a realização da pesquisa, além da motivação e dos objetivos. Em sequência, o Capítulo 2 aborda o referencial teórico, revisando os conceitos fundamentais de redes generativas adversárias, transferência de aprendizado e *StyleGAN*.

O Capítulo 3, por sua vez, explica a metodologia utilizada, detalhando a adaptação do gerador e discriminador pré-treinado para geração de rostos sintéticos assim como a construção do *encoder* para realizar o mapeamento inverso.

O Capítulo 4 trata dos resultados obtidos, apresentando as métricas e imagens geradas, assim como analisando qualitativamente a resolução e a fidelidade dos rostos produzidos.

O Capítulo 5 apresenta a conclusão do trabalho, sintetizando as principais contribuições, discutindo limitações encontradas e sugerindo direções para trabalhos futuros.

2 CAPÍTULO 2

2.1 *Machine Learning e Deep Learning*

A aprendizagem de máquina *(machine learning)* pode ser definida como um campo da inteligência artificial que se concentra no desenvolvimento de programas de computador capazes de aprender padrões a partir de dados. Diferentemente de algoritmos tradicionais, os modelos de machine learning ajustam seus parâmetros através de exemplos observáveis. Esse tipo de abordagem é útil quando não se conhecem algoritmos exatos para resolução de um problema. Embora algoritmos possam resolver diversas tarefas, existem problemas complexos, como prever comportamentos futuros, nos quais o método mais adequado é assumir que o futuro não será muito diferente do passado e utilizar exemplos passados para aprendizado (ALPAYDIN, 2014).

No campo do aprendizado de máquina, se têm um conceito fundamental: as redes neurais. Essas redes, por padrão, consistem em um conjunto de neurônios conectados entre si produzindo sequências de ativações entre eles. Os neurônios da camada de entrada recebem sinais de sensores, enquanto os intermediários recebem sinais ponderados dos neurônios anteriores. O objetivo do aprendizado é ajustar os pesos para que a rede possa aprender o comportamento desejado para a tarefa à qual foi destinada (SCHMIDHUBER, 2014).

Segundo Schmidhuber (2014) o aprendizado em redes neurais está relacionado ao problema fundamental da atribuição de crédito, isto é, identificar quais componentes modificáveis do sistema são responsáveis pelo sucesso ou fracasso do aprendizado, e como ajustá-los para melhorar o desempenho. O uso do deep learning destaca-se por torna a atribuição desses créditos de forma mais precisa, por meio das múltiplas camadas conectadas.

2.1.1 Conceitos Fundamentais de Deep Learning

O *Deep Learning* pode ser entendido como um ramo do aprendizado de máquina que faz uso de redes neurais artificiais com múltiplas camadas, modeladas para aprender padrões complexos. Esse aprendizado tem o papel de descobrir estruturas em grandes conjuntos de dados, utilizando algoritmos que indicam como devem ser alterados os pesos internos empregados nas diversas camadas da arquitetura (LECUN; BENGIO; HINTON, 2015)

Segundo LeCun, Bengio e Hinton (2015, p. 437), “uma arquitetura de aprendizado

profundo é uma pilha multicamadas de módulos simples, todos (ou a maioria) sujeitos a aprendizado". Assim, partindo das camadas de entrada, cada camada subsequente é responsável por processar informações em diferentes níveis de abstração até alcançar o objetivo final. Nesse contexto, os pesos das conexões entre os neurônios precisam ser ajustados de forma eficiente. Como descrevem LeCun, Bengio e Hinton (2015, p. 437):

> Para ajustar adequadamente o vetor de peso, o algoritmo de aprendizagem calcula um vetor gradiente que, para cada peso, indica em quanto o erro aumentaria ou diminuiria se o peso fosse aumentado em um pequeno valor. O vetor de peso é então ajustado na direção oposta ao vetor gradiente.

Durante o funcionamento do aprendizado de uma rede neural profunda, ocorre um processo chamado propagação de ativações, no qual os neurônios recebem sinais ponderados da camada anterior e transmitem suas saídas para as próximas camadas. Esse fluxo é regulado por funções de ativação, permitindo que a rede aprenda representações complexas a partir dos dados (SCHMIDHUBER, 2014). Nas primeiras camadas, por exemplo, redes aplicadas a imagens podem identificar bordas e formas simples, nas camadas intermediárias, podem detectar motivos e combinações, enquanto nas camadas finais, surgem representações mais abstratas que correspondem a partes ou mesmo a objetos inteiros. O aspecto fundamental é que essas representações são detectadas automaticamente a partir dos dados de entrada por meio do procedimento de aprendizado(LECUN; BENGIO; HINTON, 2015).

Modelos com várias camadas sucessivas, destacam-se desde meados de 1960. Na década de 80 o método de gradiente descendente para aprendizado supervisionado foi aplicado denominado de retropropagação, no entanto o treinamento se mostrou difícil aplicado às redes profundas (SCHMIDHUBER, 2014)

A retropropagação ou *backpropagation* é o processo de calcular o gradiente do erro e ajustar os pesos da rede para reduzi-lo. Como aponta Popescu et al. (2009), este método minimiza a diferença entre saída desejada e saída real usando gradiente descendente, propagando este erro para trás na rede, permitindo o aprendizado.

Diante dos conceitos de aprendizado de máquina, podemos destacar dois tipos: aprendizado supervisionado e aprendizado não supervisionado. No aprendizado supervisionado, cada entrada utilizada no treinamento é rotulada. Assim, a rede observa o conjunto de entrada e a categoria associada a cada amostra. Durante o treinamento, a máquina atualiza seus pesos e produz uma saída em forma de pontuações para cada categoria, sendo o objetivo final atribuir a maior pontuação à categoria correta. Um fator importante nesse processo é o cálculo da função objetivo, que mede o erro entre as pontuações previstas e os valores reais. A máquina ajusta os parâmetros para minimizar esse erro e para realizar esse ajuste de maneira adequada, utiliza o vetor de gradiente, que indica o quanto o erro

afeta os pesos, se deve aumentá-los ou diminuí-los, e a direção na qual o vetor de pesos deve ser ajustado(LECUN; BENGIO; HINTON, 2015).

Nesse tipo de aprendizado, é comum utilizar um conjunto de validação para medir o desempenho do sistema. O modelo é apresentado a novos dados, que não foram vistos durante o treinamento. Esse procedimento é utilizado para avaliar a capacidade de generalização do modelo.

A aprendizagem não supervisionada é definida como um processo de aprendizado que ocorre sem a utilização de rótulos. Nessa arquitetura, o modelo é capaz de identificar e extrair características a partir dos dados de entrada, sem depender de uma resposta esperada. Diferentemente do aprendizado supervisionado, que tem como base pares de entrada e saída. O aprendizado não supervisionado é normalmente utilizado para gerar uma codificação mais compacta e conveniente dos dados brutos, como imagens ou áudios. Schmidhuber (2014) destaca que códigos mais compactos, podem posteriormente, ser alimentados em maquínas de apredizado supervisionado.

2.1.2 Aplicações Deep Learning

O aprendizado profundo, por meio múltiplas camadas, permite aprender representações de dados em diversos níveis. Esses modelos em si foram capazes de aprimorar, por exemplo, o reconhecimento de objetos e de fala, assim como ser um poderoso fator diante da ciência para descobertas de medicamentos. O *Deep Learning* consegue resultados ainda mais impressionantes, como a previsão de atividades em dados de aceleradores de partículas, a reconstrução de circuitos cerebrais, além da compreensão de linguagem natural e da análise de sentimentos (LECUN; BENGIO; HINTON, 2015)

2.1.3 Redes CNN

As redes neurais convolucionais (CNN) têm extrema importância no campo do aprendizado profundo. Essas redes são capazes de extrair características dos dados por meio de operações de convolução, o que se mostra promissor, pois não é mais necessária a extração manual de características, como em métodos tradicionais(LI et al., 2020).

As CNN trouxeram avanços significativos dentro das mais diversas áreas. LeCun, Bengio e Hinton (2015) relata, por exemplo, avanços no processamento de imagem, vídeo, fala e áudio. Li et al. (2020) destacam que a CNN é uma das redes mais significativas dentro do aprendizado profundo desde que trouxe inovação nas áreas de visão computacional e processamento de linguagem natural despertando interesse desde de indústrias a instituições acadêmicas. Destaca também que a visão computacional baseada nas redes convolucionais tornou possível a realização de tarefas que eram consideradas extremamente complexas, como veículos autônomos.

Diante dos cenários de funcionamento das CNN, LeCun, Bengio e Hinton (2015, p. 439) relatam "Há quatro ideias principais por trás das *ConvNets* que aproveitam as propriedades dos sinais naturais: conexões locais, pesos compartilhados, agrupamento e o uso de muitas camadas". Nas CNN, os neurônios limitam suas conexões a regiões locais da camada anterior, reduzindo o número de parâmetros envolvidos no modelo. Além disso, o compartilhamento de pesos otimiza ainda mais a quantidade de parâmetros treináveis, enquanto a etapa de agrupamento promove a redução da dimensionalidade dos dados, preservando as informações mais relevantes para o processo de aprendizagem (LI et al., 2020).

A etapa de funcionamento e a arquitetura da CNN são essenciais para compreender o processo envolvido durante o aprendizado em redes convolucionais, uma vez que sua estrutura é formada por componentes fundamentais que permitem a extração e o processamento das informações da entrada. Nesse sentido, conforme apontado por Li et al. (2020, p. 2):

> Para construir um modelo CNN, quatro componentes são normalmente necessários. A convolução é uma etapa fundamental para a extração de características. As saídas da convolução podem ser chamadas de mapas de características. Ao definir um kernel de convolução com um determinado tamanho, perderemos informações na borda. Portanto, o preenchimento é introduzido para ampliar a entrada com valor zero, o que pode ajustar o tamanho indiretamente. Além disso, para controlar a densidade da convolução, o passo é empregado. Quanto maior o passo, menor a densidade. Após a convolução, os mapas de características consistem em um grande número de características que são propensas a causar problemas de sobreajuste. Como resultado, o agrupamento (também conhecido como subamostragem) é proposto para evitar redundância, incluindo agrupamento.

Diante da importância das redes convolucionais, elas se mostram fundamentais para a construção das (GANs), pois grande parte desses modelos utilizam em sua arquitetura camadas convolucionais. No contexto deste trabalho, compreender essas bases é essencial para a geração de rostos sintéticos.

2.2 Redes Generativas Adversariais (GANs)

As redes generativas adversárias, conhecidas como GANs (Generative Adversarial Networks), foram propostas por (GOODFELLOW et al., 2014). Esse tipo de rede é composta por dois modelos distintos: o gerador (G) e o discriminador (D), que são treinados simultaneamente em um processo competitivo. O gerador tem como objetivo criar amostras que se assemelham aos dados reais, enquanto o discriminador tenta distinguir entre as amostras reais e as geradas.

A dinâmica de treinamento das GANs é, conceitualmente, simples. O gerador recebe como entrada um vetor de ruído z e o transforma em uma amostra sintética. Essa amostra é

então enviada ao discriminador, que tenta determinar se ela pertence ao conjunto de dados reais ou se foi gerada pelo próprio gerador.A partir dessa avaliação, realiza-se o processo de retropropagação para atualização dos pesos, conforme ilustrado na Figura 3. No artigo original, (GOODFELLOW et al., 2014) comparam esse processo a uma analogia em que o gerador atua como um falsificador de moedas tentando produzir cópias convincentes, enquanto o discriminador representa a polícia, cuja função é identificar se a moeda é legítima ou falsificada. O objetivo do gerador, portanto, é enganar o discriminador, fazendo com que este classifique as amostras sintéticas como reais.

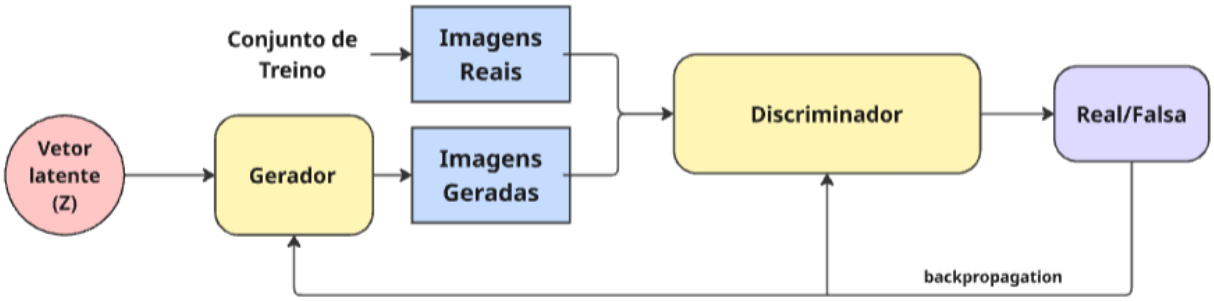

Figura 3 – Arquitetura Rede GAN

Fonte: Autoria própria

2.2.1 Dificuldades e Técnicas de Estabilização no Treinamento de GANs

O treinamento de redes GAN consiste em encontrar um ponto de equilíbrio entre o gerador e o discriminador, conhecido como equilíbrio de Nash. Ambos os modelos buscam minimizar suas próprias funções de custo. No entanto, alcançar esse equilíbrio é um problema muito difícil, pois envolve o uso de funções de custo não convexas e um espaço de parâmetros de alta dimensionalidade (SALIMANS et al., 2016)

Devido aos desafios observados e enfrentados no processo de treinamento das GANs, (SALIMANS et al., 2016) propõem, em seu artigo, uma série de estratégias para mitigar essas dificuldades. Entre essas abordagens, podemos destacar dois conceitos particularmente relevantes: o *Feature Matching* e o *Minibatch Discrimination*.

Feature Matching é uma técnica que tem como objetivo tratar as instabilidades observadas nas GANs. Agora, ao invés do gerador tentar maximizar diretamente a saída do discriminador, sua nova função é gerar dados que se assemelham estatisticamente aos dados reais. Para isso, o gerador utiliza as ativações das camadas intermediárias do discriminador como referência, buscando seguir a melhor distribuição indicada por essas camadas.

Outro aspecto relevante e que representa uma dificuldade no treinamento de GANs é o problema do colapso de modo. Nesse problema, o gerador começa a produzir sempre

as mesmas imagens ou imagens muito semelhantes. Isso acontece porque os gradientes do discriminador acabam apontando sempre para uma única direção ou direções muito próximas. Como o discriminador processa cada exemplo de forma independente, não há uma força que incentive o gerador a produzir imagens diferentes, logo as amostras tendem a convergir para aquele ponto que o discriminador acredita ser o mais realista. Uma forma de tentar resolver esse problema é usar o *mini batch discrimination*. Nessa abordagem, o discriminador deixa de olhar cada imagem isoladamente e passa a observar um conjunto de exemplos simultaneamente, identificando quando eles são muito parecidos. Com isso, o gerador é forçado a criar amostras mais variadas.

2.2.2 Aplicações das Redes GANs

As GANs são empregadas para gerar dados sintéticos de alta qualidade, incluindo imagens, vídeos e sons, além de possibilitar o aprimoramento da resolução de imagens existentes. Elas têm sido aplicadas também a criação de dados quando os originais são escassos ou sensíveis. No contexto deste trabalho, as GANs serão utilizadas especificamente para a síntese de imagens faciais com objetivo de gerar imagens sintéticas realistas do rosto alvo.

2.3 StyleGAN: conceitos e variações

2.3.1 Arquitetura da StyleGAN

Apesar dos grandes avanços nas redes GANs nos últimos anos, capacitando-as de realizar gerações de imagens cada vez mais realistas essas, limitam-se a tratar seus geradores como caixas pretas com pouco entendimento sobre a origem das variações estocásticas e sobre as propriedades do espaço latente. A StyleGAN surgiu com uma arquitetura inovadora de separar automaticamente os atributos de alto nível como, pose e identidade de rostos humanos. Como destacam os próprios autores Karras, Laine e Aila (2018, p. 1):

> Redesenhamos a arquitetura do gerador de forma a expor novas maneiras de controlar o processo de síntese de imagens. Nosso gerador começa a partir de uma entrada constante aprendida e ajusta o 'estilo' da imagem em cada camada de convolução com base no código latente, controlando assim diretamente a intensidade das características da imagem em diferentes escalas. Combinado com o ruído injetado diretamente na rede, essa mudança arquitetural leva à separação automática e não supervisionada de atributos de alto nível (por exemplo, pose, identidade) em relação à variação estocástica (por exemplo, sardas, cabelo) nas imagens geradas, e possibilita operações intuitivas de mistura e interpolação específicas por escala.

Na Figura 4 é apresentada a arquitetura do *StyleGAN*. A imagem compara geradores tradicionais com a proposta do *StyleGAN*. No lado esquerdo da imagem (a) pode-se

visualizar um gerador tradicional, onde o vetor z é inserido diretamente na primeira camada convolucional. Já no lado direito (b) é apresentada a proposta do *StyleGAN*. Nesse modelo, o vetor de entrada z é inicialmente processado por várias camadas totalmente conectadas, gerando o vetor intermediário w, que é então utilizado para controlar as camadas progressivas do gerador.

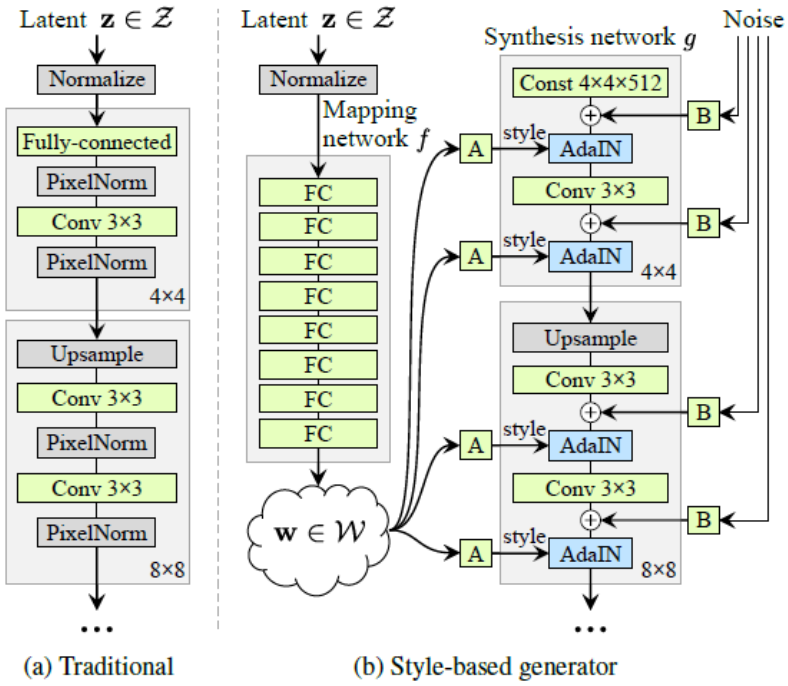

Figura 4 – Arquitetura Original Modelo StyleGAN

Fonte: Karras et al. (2019)

A segunda versão do modelo *StyleGAN*, o *StyleGAN2*, introduziu mudanças significativas em sua arquitetura. Nesta versão, foi removida a técnica de crescimento progressivo; assim, o *StyleGAN2* mantém as camadas fixas desde o início, porém com foco inicial nas primeiras camadas do modelo e posteriormente focando nos outros blocos sucessivos.

Além disso, houve uma alteração na normalização das camadas. Enquanto o modelo anterior utilizava a normalização adaptativa por instância (AdaIN), o *StyleGAN2* passou a empregar a técnica de demodulação de pesos(KARRAS et al., 2019). Essa arquitetura, a *StyleGAN2* é a utilizada no presente projeto, sendo fundamental para a síntese de imagens faciais desenvolvida neste estudo.

A *StyleGAN2* dentre os seus mais diversos conjuntos de treinamentos, foi treinada também em dados de rostos humanos. Seu poderoso poder arquitetural foi capaz de gerar faces sintéticas com alto realismo, assim como resoluções elevadas. Diante do escopo deste trabalho, onde o objetivo é utilizar uma rede pré-treinada,essa eficiência é imprescindível

para essa realização, devido tanto ao conjunto alvo serem semelhantes (rostos humanos) como também a disponibilização dos pesos resultantes do seu treinamento original.

Em capítulos seguintes, será apresentada a abordagem adotada para a adaptação da rede *StyleGAN2* na geração de imagens sintéticas, destacando a metodologia e resultados obtidos durante o processo.

2.4 Aprendizado por Transferência

O Aprendizado por transferência (*transfer learning*) é uma técnica de aprendizado que consiste em aproveitar o conhecimento já adquirido em uma tarefa para se especializar em outra. Quando as redes neurais profundas são treinadas em um grande conjunto de dados essas redes aprendem informações úteis que podem ser reutilizadas em novos objetivos. Isso é capaz de reduzir a necessidade de grande conjunto de dados assim como o tempo total de treinamento(PAN; YANG, 2010)

Segundo (YOSINSKI et al., 2014) redes profundas treinadas com imagens, por exemplo, as camadas iniciais já aprendem a identificar padrões genéricos das imagens não se especializando muito no conjunto de dados, mas em aspectos gerais. Já as camadas mais profundas se responsabilizam pelas características mais específicas da tarefa original. O autor ainda relata que mesmo para tarefas entre domínios diferentes a transferibilidade ainda pode ser mais eficiente do que começar a rede com pesos totalmente aleatórios.

Dentro do escopo das técnicas de transfer learning, (MO; CHO; SHIN, 2020) propuseram o método FreezeD, uma técnica aplicada ao discriminador de redes GAN. A ideia principal dessa abordagem é que as camadas convolucionais iniciais do discriminador capturam características genéricas dos conjuntos de treinamento o que se alinha com conceitos já destacados neste relatório. Logo, não seria mais necessário retreinar essas camadas para um novo conjunto de dados. No entanto, as camadas superiores são responsáveis por definir se a imagem é real ou falsa. Essas camadas mais altas devem continuar sendo treinadas, pois são elas que permitem que o discriminador se adapte a um novo domínio.

Os autores destacam que essa técnica melhora a estabilização do treinamento e reduz o risco de *overfitting* em conjuntos com um número pequeno de dados. Outro ponto é que, com essa técnica, foi possível obter um FID (*Fréchet Inception Distance*) menor em comparação com outras abordagens de treinamento, além de acelerar o processo de treinamento.

3 MODELO PROPOSTO

3.1 Introdução

No presente trabalho, o objetivo é desenvolver um modelo generativo capaz de sintetizar imagens de rostos, utilizando *transfer learning*, com intuito de alcançar otimizações do tempo de processamento, assim como geração de imagens de alta qualidade. Para isso foi utilizado a rede generativa *StyleGAN2*, adotando-a como ponto de partida pré-treinada para a síntese do nosso conjunto-alvo.

A escolha desse modelo justifica-se por dois fatores principais:

1. **Domínio de treinamento compatível:** o modelo foi originalmente treinado no domínio de rostos humanos, alinhando-se diretamente ao objetivo deste trabalho.
2. **Qualidade de síntese:** o *StyleGAN2* é capaz de gerar imagens de altíssima fidelidade.

Além disso, será desenvolvido um modelo *encoder*, responsável por realizar o mapeamento inverso do espaço latente, permitindo a reconstrução fiel das imagens a partir de seus vetores latentes.

Diante da proposta do modelo deste trabalho, é possível observar, na Figura 5, uma visão geral da arquitetura desenvolvida para a rede generativa. A ilustração apresenta a ideia central do projeto: a construção de um gerador treinável, adaptado a partir do gerador original da *StyleGAN2*, juntamente com a implementação de técnicas de congelamento das camadas dos modelos originais (gerador e discriminador), com o objetivo de otimizar o treinamento.

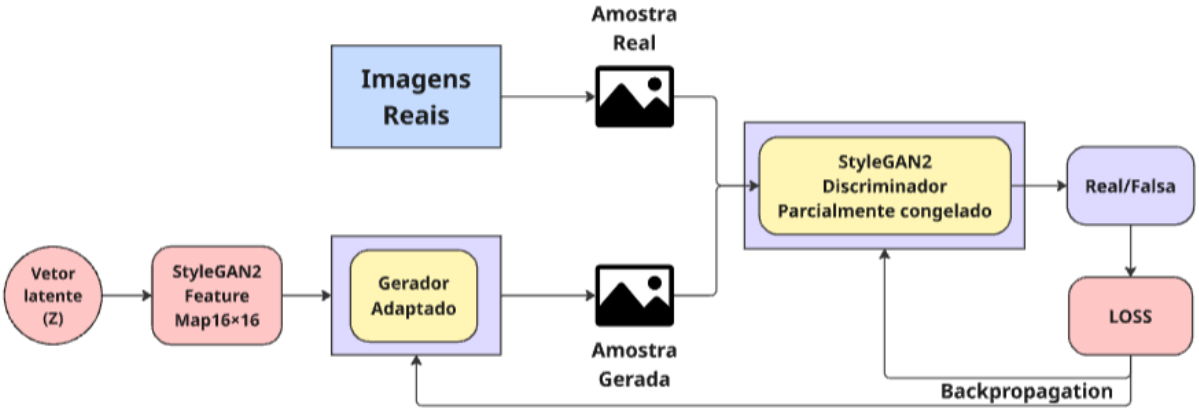

Figura 5 – Rede Generativa Proposta

Fonte: Autoria própria

Na Figura 6 pode-se visualizar a estrutura do modelo encoder proposto, cujo objetivo é realizar o mapeamento inverso das imagens, a fim de obter os vetores latentes necessários para a reconstrução das mesmas. O treinamento do encoder é realizado de forma independente do modelo adaptado desenvolvido na Figura 5. No entanto, o encoder utiliza o gerador já treinado, o qual é mantido completamente congelado, a fim de aprender a encontrar o melhor vetor latente que representa a imagem de entrada. Diante disso, nessa etapa apenas os pesos do encoder são treinados.

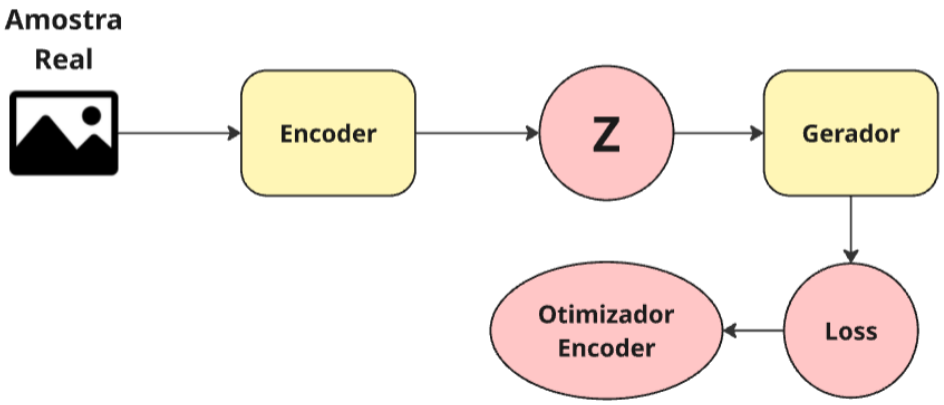

Figura 6 – Encoder Proposto

Fonte: Autoria própria

Nas próximas subseções deste capítulo, serão detalhadas todas as etapas do desenvolvimento do projeto, contemplando as configurações do modelo, as adaptações implementadas e a integração do *encoder* à arquitetura principal.

3.2 Coleta de Dados

A primeira etapa do nosso projeto consiste na coleta de dados que é fundamental para a qualidade dos resultados posteriores. O processamento dessa etapa ocorre primeiramente na captura de vídeo, essa captura é decomposta em vários frames. Em seguida, é aplicado um algoritmo de alinhamento para padronizar as imagens obtidas de forma que fiquem semelhante ao conjunto de dados utilizado, originalmente no treino da Stylegan.

Nessa etapa, foi definida a quantidade de frames necessária para ser utilizada durante o treinamento. Optou-se por escolher um conjunto de treinamento com 800 imagens de um único rosto, todas capturadas em um mesmo ambiente, que serviu como conjunto alvo para o treinamento posterior.

3.3 Adaptação do Gerador

Para a geração das sínteses faciais, foi proposta uma adaptação a partir da interceptação de camadas intermediárias da *StyleGAN2*. A arquitetura do *StyleGAN2*, é composta por camadas progressivas que vão de resoluções de **4×4 até 1024×1024**, porém existem versões da rede que têm como resolução final **até 256×256** que foi a versão utilizada neste projeto. Assim como em outras redes generativas tradicionais sua entrada é um **vetor latente** (z) de dimensão **512**.

Neste trabalho, foi encapsulado o gerador da rede *StyleGAN2* em uma classe *Python* e realizamos os seguintes procedimentos:

1. O gerador original foi completamente congelado, impedindo sua atualização durante o treinamento.

2. Foi selecionada uma camada para a interceptação da saída intermediária. A camada escolhida foi a **b16.conv1**, que corresponde à segunda camada convolucional do bloco **b16** da rede *StyleGAN2*, associada à resolução de **16×16**.

A captura dessas camadas foi realizada por meio de *hooks* do *PyTorch*. Os hooks são uma maneira de interceptarmos camadas intermediárias em tempo de execução quando as ativações de um camada são calculadas, logo conseguimos capturar o ponto de corte ideal desejado.

A escolha dessa interceptação se justifica pelo fato de que a resolução **16×16** representa um ponto intermediário no processo de geração. Nessa camadas a saída não é tão abstrata a ponto de não conseguir aproveitar as features, nem é tão detalhada a ponto de preservar excessivamente identidade do conjunto de treinamento original, assim pode-se aproveitar a captura de características estruturais globais do rosto. oferecendo um equilíbrio entre a preservação do conhecimento do conjunto de treino original e a adaptação ao novo alvo.

Após a captura da camada específica, foi utilizada a saída do gerador original pré -treinado, como entrada para uma nova sub rede a qual foi desenvolvida. Foi construído um novo gerador adaptado e acoplado ao gerador original *StyleGAN2*, O esboço resumido da arquitetura pode ser melhor visualizado na Figura 7. O objetivo é usar a captura da rede *StyleGAN2* como entrada e ir aumentando a resolução até **256×256**.

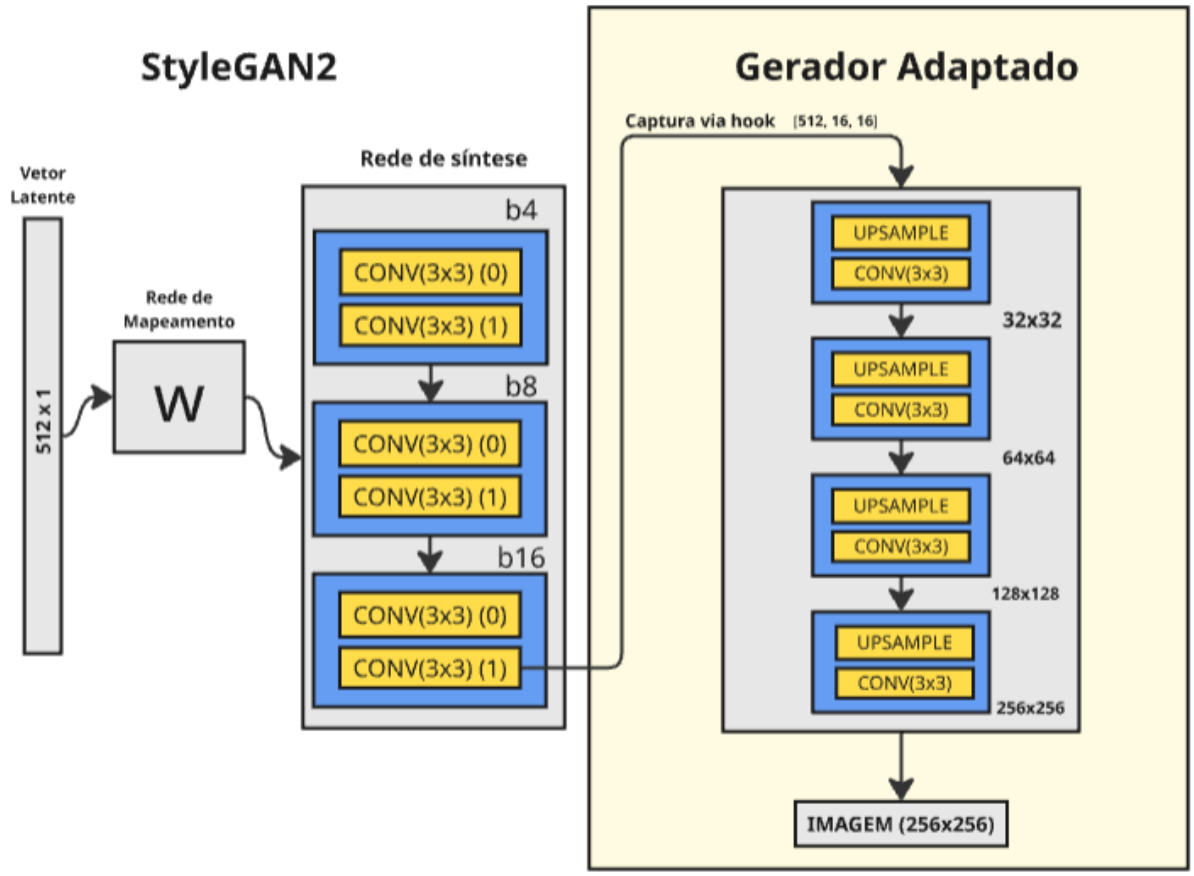

Figura 7 – Captura de camada utilizando *hooks* do *PyTorch*

Fonte: Autoria própria

Durante o treinamento, apenas os parâmetros dessa sub rede são ajustados, enquanto o gerador original permanece com suas partes congeladas. Para efeito de comparação, o gerador original da *StyleGAN2* completo teria cerca de **24 milhões** de parâmetros treináveis. A subrede desenvolvida tem aproximadamente **4 milhões** de parâmetros treináveis, o que garante um tempo de processamento menor.

A rede desenvolvida foi projetada para receber o mapa de características de 512 canais proveniente da *StyleGAN2*, correspondente à resolução 16×16. Ela é composta por etapas de *upsampling* usando o método *nearest neighbor*, que amplia a resolução duplicando os pixels vizinhos, seguidas por camadas convolucionais 2D com *kernel* 3×3 e ativação *ReLU*. Esse processo permite refinar as informações visuais e aumentar progressivamente a resolução, produzindo uma imagem final com três canais RGB e resolução de 256×256. Por fim, aplica-se a função de ativação *Tanh*, que normaliza os valores para o intervalo $[-1, 1]$, mantendo compatibilidade com a escala utilizada no conjunto de treinamento.

3.4 Adaptação do Discriminador

Na arquitetura do nosso modelo GAN, utilizamos o discriminador original da *StyleGAN2*, sem nenhuma modificação em sua estrutura. Devido ao treinamento original desse modelo, o discriminador, assim como o gerador, também foi treinado em um conjunto amplo de imagens, o que lhe oferece um vasto conhecimento na identificação de rostos humanos.

Ao carregarmos o modelo discriminador no projeto proposto, o procedimento realizado foi apenas o congelamento das camadas de maiores resoluções. No modelo original, como citado anteriormente, as resoluções chegam a **256×256**. O procedimento adotado consistiu em congelar justamente essas camadas de maiores resoluções, ou seja, as camadas de **256×256**, **128×128** e **64×64**. A escolha de congelar essas camadas parte do princípio de que tais resoluções já possuem a capacidade para identificar traços faciais, não havendo necessidade de reaprendizado. Já para as camadas subsequentes, responsáveis por resoluções mais baixas, optamos por mantê-las treináveis.

Durante o treinamento, apenas as camadas de resoluções **4×4**, **8×8**, **16×16** e **32×32** são treinadas. O fato das outras camadas estarem congeladas pode possibilitar ao modelo proposto um tempo de treinamento menor, já que há menos parâmetros para ajustar do que se todas as camadas estivessem livres, consequentemente também uma melhor estabilidade.

3.5 Treinamento

Para o treinamento do modelo, foi implementado um processo adversarial. Nesse processo, foram ajustados alguns hiperparâmetros: o *batch size* foi definido em 32, a taxa de aprendizado de 2×10^{-4} tanto para o gerador, como para o discriminador. O otimizador utilizado foi o Adam.

Para garantir que os pesos destinados permanecessem , nesta etapa do treinamento realizou-se a interação sobre as camadas do gerador e do discriminador, escolhendo manualmente as camadas que ficariam livres ou congeladas.

Nessa etapa do treinamento, foram sendo geradas grades de imagens sintéticas, permitindo a análise visual dos resultados obtidos pelo modelo. Junto a grade de visualização, para uma análise que não seja apenas parametrizada visualmente, foi colocado junto ao modelo a métrica de FID (Fréchet Inception Distance). Essa métrica mede a distância estatística entre as distribuições das imagens reais e das imagens geradas. Valores menores de FID indicam maior similaridade entre os dois conjuntos e, portanto, melhor qualidade de geração. No presente trabalho, o cálculo do FID foi realizado a cada 30 épocas de treinamento.

3.6 Mapeamento Inverso

Durante as etapas do modelo proposto, foi desenvolvido um encoder responsável por realizar o mapeamento inverso das imagens para o espaço latente. Essa etapa tem como principal objetivo possibilitar a reconstrução das imagens de entrada a partir das representações latentes aprendidas durante o treinamento do gerador. Assim, dado uma imagem de entrada, o encoder deve ser capaz de encontrar o vetor latente que melhor a represente e, a partir disso, permitir que o gerador reconstrua a imagem correspondente.

A arquitetura utilizada para a construção do *encoder* foi realizada utilizando camadas convolucionais, com o objetivo de transformar imagens no espaço latente de dimensão $Z = 512$. Nessa etapa, segue-se uma estrutura sequencial composta por camadas `Conv2d` com *kernel* 4×4, *stride* 2 e *padding* 1, reduzindo a resolução progressivamente de 256×256 até 4×4. As camadas convolucionais são seguidas por *Batch Normalization* e pela função de ativação *LeakyReLU*. Após a etapa convolucional, os mapas de ativação são processados e achatados por duas camadas totalmente conectadas com ativações *ReLU*, até a obtenção do vetor final.

O processo pode ser resumido como:

$$256 \times 256 \;\rightarrow\; 128 \;\rightarrow\; 64 \;\rightarrow\; 32 \;\rightarrow\; 16 \;\rightarrow\; 8 \;\rightarrow\; 4$$

e depois a camada de achatamento, resultando em um vetor de $Z = 512$.

Durante o treinamento do modelo foram utilizadas duas funções de perda: o erro quadrático médio (MSE) e o LPIPS (Learned Perceptual Image Patch Similarity). O MSE é responsável por medir a diferença ponto a ponto entre os pixels da imagem original e a reconstruída. Já o LPIPS, permite captar aspectos perceptuais da imagem.

A função de perda total foi definida como uma combinação ponderada das duas métricas, utilizando pesos de 0.3 para o MSE e 0.5 para o LPIPS. Esses pesos funcionam como fatores que ajustam a intensidade da contribuição de cada termo na perda total. Os valores foram definidos empiricamente com base nos testes realizados e nos resultados observados.

Na etapa de otimização foi utilizado o algoritmo Adam, com taxa de aprendizado de 0,0001. O treinamento desta etapa foi realizado por 30 épocas, levando a um tempo total aproximado de 15 minutos.

4 RESULTADOS

4.1 Introdução

Neste capítulo, são apresentados os principais resultados obtidos durante o desenvolvimento do projeto, bem como as imagens geradas e as métricas utilizadas para avaliar a qualidade do modelo de geração.

4.2 Geração

Nesta seção, serão apresentados os resultados de geração obtidos durante o treinamento, de forma visual, destacando a adaptação ao nosso domínio.

4.2.1 Geração Inicial

Uma das principais discussões deste trabalho refere-se ao uso de transfer learning no modelo de GANs, possibilitando um ponto de partida para a rede não começar sua geração totalmente do zero. Na Figura 8, é possível notar que com o conhecimento prévio da adquirido, a rede já parte inicialmente nas primeiras épocas do modelo, de uma estrutura facial previamente formada. Nota-se que, ao invés de a rede iniciar com uma imagem totalmente ruidosa, como ocorre normalmente nas primeiras épocas de uma rede GAN generativa, no modelo proposto a rede começa com uma representação já semelhante a um rosto humano.

Observando-se a ilustração, percebe-se algumas características faciais já pré-definidas, notando alguns detalhes sutis já nessa primeira época. São eles: contornos faciais, estrutura da cabeça, pose e brevemente posições de olhos, boca e nariz. Assim como também é possível perceber que na primeira época, o modelo tenta colocar algumas caracteristicas de adaptação ao novo domínio.

Figura 8 – Imagem gerada na época 0

Fonte: Autoria própria

4.2.2 Geração Intermediária

Ao longo do treinamento, pode-se observar uma melhor adaptação do modelo ao longo das épocas. Na Figura 9, que ilustra uma geração intermediária na época 8, apesar de se tratar ainda de uma época inicial, nota-se que o modelo reconhece e preserva bem o formato do rosto e a pose, mantendo-os de forma consistente com o conhecimento prévio adquirido. Outro ponto a ser observado é que, nessa época de treinamento, apesar da alta resolução gerada (256×256), já é possível notar uma boa textura de pele nos resultados obtidos.

Figura 9 – Imagem gerada na época 8

Fonte: Autoria própria

Nas Figuras 10 e 11, observa-se que o modelo já é capaz de gerar resultados semelhantes ao nosso conjunto de dados. Nessa etapa de geração, percebe-se que, devido às partes congeladas do gerador e do discriminador, as sínteses carregam também características do conjunto de treino original.

Destaca-se a época de treinamento representada na Figura 11, pois, a partir de meados desse estágio, já é possível identificar maior fidelidade ao conjunto de dados alvo do presente projeto, notando-se boas Posições de olhos, nariz, texturas de pele e até mesmo detalhes como barba tornam-se mais nítidos nesta etapa de treino.

Figura 10 – Imagem gerada na época 35

Fonte: Autoria própria

Figura 11 – Imagem gerada na época 72

Fonte: Autoria própria

4.2.3 Geração Final

No decorrer do treinamento, com o auxílio de métricas como o FID e a qualidade visual das imagens, foi escolhido um ponto de encerramento para o modelo. A Figura 12 destaca esse ponto, correspondente à época 180.

Na geração apresentada, fica evidente que o modelo já é capaz de expressar detalhes finos nas imagens. Observa-se uma boa posição de cada elemento facial, marcas de expressão bem definidas, assim como detalhes mais específicos, como sorrisos.

Figura 12 – Imagem gerada na época 180

Fonte: Autoria própria

4.3 Métricas

O FID é reconhecido como uma das principais métricas para a avaliação da qualidade das imagens geradas por modelos generativos. Na proposta deste trabalho, ao longo das diferentes épocas de geração, foi elaborada uma tabela de métricas, na qual se inclui, o cálculo do FID, permitindo acompanhar a evolução do desempenho do modelo ao longo do treinamento. Na figura 13 pode-se visualizar o resultado ao longo das épocas de treinamento.

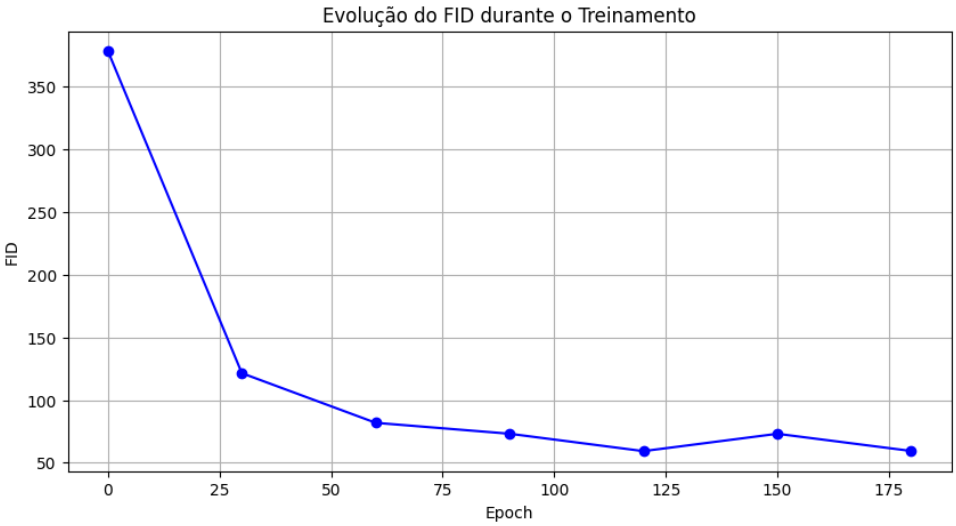

Figura 13 – FID de treinamento

Fonte: Autoria própria

Diante do gráfico apresentado, podem-se destacar alguns pontos essenciais para o desempenho do modelo. Observa-se que o treinamento inicia com um valor de FID relativamente elevado, acima de 350. No entanto, ao longo do progresso do treino, esse valor passa a diminuir. Por volta da 30ª época, já se nota um declínio acentuado, atingindo rapidamente valores próximos de 130. Nas épocas seguintes, a inclinação da queda diminui gradualmente, o que é esperado, uma vez que o modelo começa a se atentar aos detalhes, chegando a valores em torno de 80 por volta da 90ª época. No processo final, entre aproximadamente a 100ª e a 180ª época, o modelo entra em uma fase na qual o FID oscila em valores baixos, próximos de 50.

Em relação à estabilidade do modelo, notou-se, pelos testes aplicados, que este se manteve estável na maioria das avaliações. Foram testadas diversas taxas de aprendizado, tanto no gerador quanto no discriminador, na tentativa de se encontrar um equilíbrio e a convergência do modelo. Ao longo dos testes, observou-se que, na grande maioria dos casos, mesmo com diferentes taxas, o modelo apresentou boa convergência.

4.4 Inversão do Vetor Latente

Nessa seção será apresentado os resultados da etapa obtidos por meio do processo de inversão das imagens com objetivo de reconstruir as imagens originais por meio do encoder.

Na Figura 14 pode-se observar o resultado do modelo na reconstrução das imagens com base no mesmo conjunto de treino do modelo. Observa-se que o modelo foi capaz

de aprender e reconstruir bem imagens originalmente passadas, obedecendo às posições e expressões da imagem original e de seu par reconstruído.

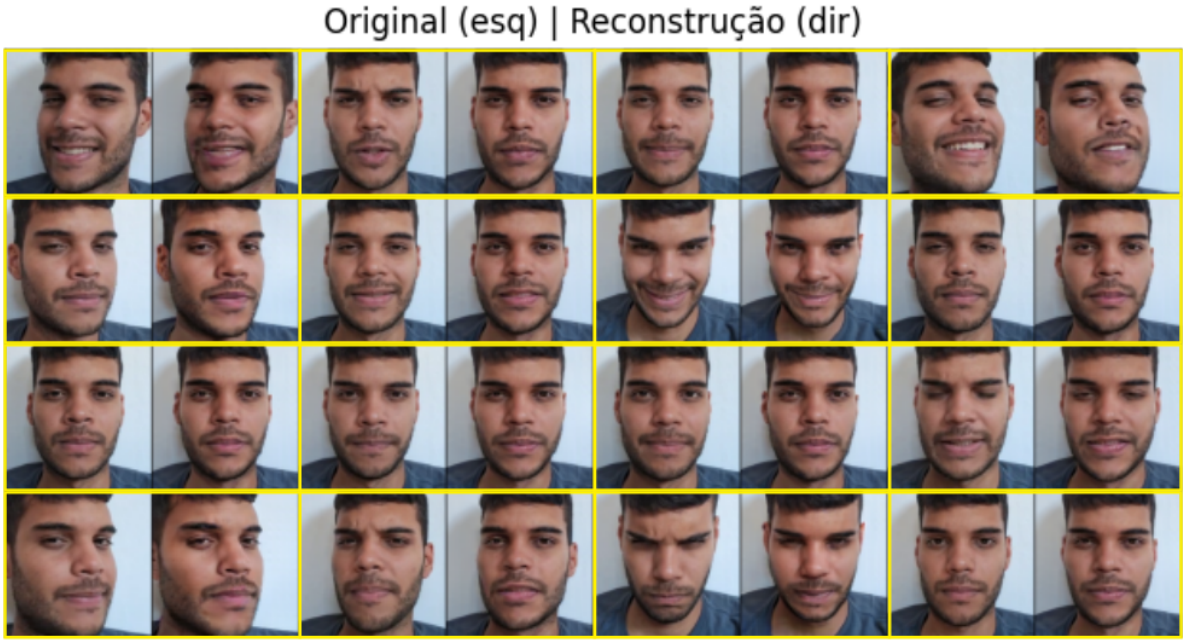

Figura 14 – Reconstrução Conjunto de Treinamento

Fonte: Autoria própria

Para avaliar o desempenho da reconstrução das imagens de forma quantitativa, foi utilizada a métrica de SSIM. No experimento, o modelo alcançou um SSIM médio de 0.6712, o que corresponde a um nível moderado de similaridade estrutural entre as imagens originais e as reconstruídas.

Diante dos objetivos, deste trabalho um deles é permitir a reconstrução no domínio do gerador, a partir de uma entrada qualquer, diferente do conjunto de treinamento, foram testadas novas entradas com aspectos diferentes do conjunto de treinamento, mas mantendo a mesma pessoa.

Na Figura 15 pode-se observar os resultados obtidos pelo modelo, mostrando que, dado qualquer entrada do indivíduo, o modelo é capaz de mapear o melhor vetor latente que a representa e, consequentemente, reconstruir a imagem. A imagem à esquerda representa a entrada fornecida ao modelo, enquanto a imagem à direita corresponde à imagem reconstruída mais semelhante presente no conjunto de treinamento.

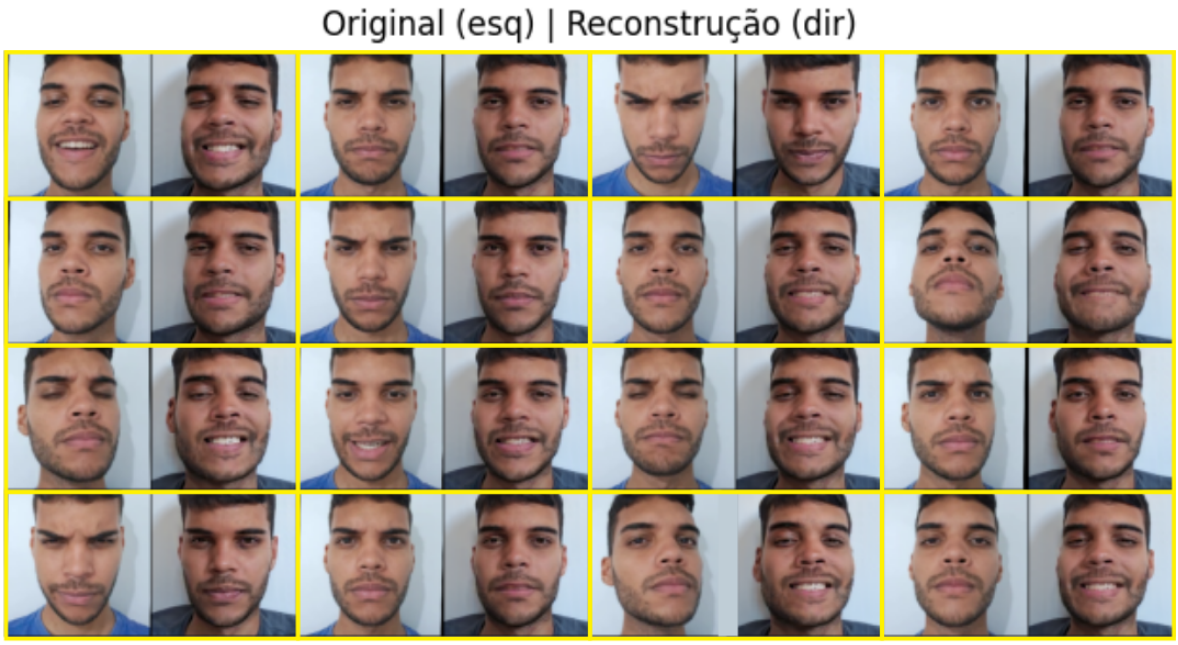

Figura 15 – Reconstrução Novas entradas

Fonte: Autoria própria

Dos resultados apresentados para novas imagens, foi observado que o modelo conseguiu mapear as reconstruções para o espaço latente aprendido durante o treinamento do gerador. Algumas das novas entradas apresentaram pares bem representados, enquanto outras não alcançaram o mesmo nível de correspondência. No geral, as poses foram preservadas nas reconstruções.

5 CONSIDERAÇÕES FINAIS

Este trabalho apresentou uma abordagem de aprimoramento no processo de geração de rostos utilizando redes generativas adversárias baseadas na arquitetura StyleGAN2, combinada com técnicas de transfer learning. A aplicação da metodologia abordada para síntese facial individualizada permitiu a geração de imagens sintéticas com características correspondentes aos dados de entrada definidos.

Durante o treinamento da rede generativa, o modelo demonstrou uma boa convergência, seguida de métricas de avaliação quantitativa que indicaram um bom desempenho, como a redução progressiva do FID ao longo das épocas, chegando a resultados próximos de 50. Para uma a avaliação visual, o modelo também apresentou boa qualidade nas imagens geradas.

Para a reconstrução das imagens diretamente a partir do vetor latente, o modelo permitiu reconstruções próximas às imagens originais. Observou-se que através da rede encoder, foi possível reconstruir imagens fielmente em comparação com as originais. As métricas de avaliação, incluindo o SSIM, atingiram aproximadamente 0,67, indicando consistência na reconstrução das imagens.

É importante destacar que com a implementação do presente projeto, foi possível otimizar o tempo de processamento e a geração de imagens em maior resolução, reduzindo alguns dos gargalos observados em trabalhos relacionados, como o de (SANTOS, 2024), no qual é relatado um tempo de treinamento de cerca de 24 horas e geração de imagens com resolução de 64×64. O modelo proposto, treinado no mesmo hardware utilizado por (SANTOS, 2024), apresentou um tempo médio de treinamento da GAN de cerca de 1 hora e 25 minutos para geração de imagens em resolução de 256x256 e mais 15 minutos para o encoder, resultando em um tempo total de aproximadamente 1 hora e 40 minutos. Outra métrica de comparação é o SSIM, que no trabalho citado foi de aproximadamente 0,69, enquanto no modelo proposto atingiu cerca de 0,67, o que mostra que, mesmo gerando imagens em resolução superior, o modelo manteve uma métrica similar de qualidade.

Durante o treinamento do modelo, algumas dificuldades foram observadas. Notou-se que para o gerador, camadas de resolução muito alta apresentaram maior dificuldade de convergência para o alvo específico, tornando o treinamento mais lento. Esse comportamento levou a este trabalho a optar por uma camada de resolução menor, garantindo maior estabilidade no processo de aprendizagem.

Dentre as contribuições deste trabalho, destaca-se o seu potencial para possibilitar a compressão de dados e viabilizar o processo de avatarização em ambientes virtuais,

possibilitando novas aplicações e aprimoramentos em modelos de síntese facial.

Como trabalhos futuros, aponta-se a possibilidade de continuidade desta linha de pesquisa com a exploração de novas estratégias que envolvam tanto variações arquiteturais de GANs quanto modelos de difusão, os quais também têm se destacado por sua capacidade de gerar imagens com elevado realismo.

Por fim, destaca-se que o presente trabalho atingiu os objetivos propostos, apresentando resultados satisfatórios e abrindo perspectivas para estudos futuros.

REFERÊNCIAS

ALPAYDIN, E. *Introduction to Machine Learning*. 3. ed. Cambridge: The MIT Press, 2014.

GOODFELLOW, I. J. et al. Generative adversarial nets. *Advances in neural information processing systems*, 2014. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf. Acesso em: 27 nov. 2025.

KARRAS, T. et al. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. Disponível em: https://arxiv.org/abs/1511.06434. Acesso em: 02 nov. 2025.

KARRAS, T.; LAINE, S.; AILA, T. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018. Disponível em: https://arxiv.org/abs/1812.04948. Acesso em: 27 nov. 2025.

KARRAS, T. et al. Analyzing and improving the image quality of stylegan. *CoRR*, abs/1912.04958, 2019. Disponível em: http://arxiv.org/abs/1912.04958. Acesso em: 27 nov. 2025.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, v. 521, p. 436–444, 05 2015. Disponível em: https://10.1038/nature14539. Acesso em: 27 nov. 2025.

LI, Z. et al. A survey of convolutional neural networks: Analysis, applications, and prospects. *CoRR*, abs/2004.02806, 2020. Disponível em: https://arxiv.org/abs/2004.02806. Acesso em: 27 nov. 2025.

MO, S.; CHO, M.; SHIN, J. Freeze discriminator: A simple baseline for fine-tuning gans. *CoRR*, abs/2002.10964, 2020. Disponível em: https://arxiv.org/abs/2002.10964. Acesso em: 27 nov. 2025.

PAN, S. J.; YANG, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, v. 22, n. 10, p. 1345–1359, 2010. Disponível em: https://10.1109/TKDE.2009.191. Acesso em: 27 nov. 2025.

POPESCU, M.-C. et al. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, v. 8, 07 2009.

RADFORD, A.; METZ, L.; CHINTALA, S. *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. 2016. Disponível em: https://arxiv.org/abs/1511.06434. Acesso em: 02 nov. 2025.

SALIMANS, T. et al. *Improved Techniques for Training GANs*. 2016. Disponível em: http://arxiv.org/abs/1606.03498. Acesso em: 27 nov. 2025.

SANTOS, K. E. A. dos. Representação facial para interações online utilizando redes generativas adversárias bidirecionais (bigans). *Universidade Federal do Rio Grande do Norte*, 2024. Disponível em: https://repositorio.ufrn.br/server/api/core/

bitstreams/4ac1a59c-3409-4c6b-ba87-85047a8804e1/content. Acesso em: 27 nov. 2025.

SCHMIDHUBER, J. Deep learning in neural networks: An overview. *CoRR*, abs/1404.7828, 2014. Disponível em: http://arxiv.org/abs/1404.7828. Acesso em: 27 nov. 2025.

YOSINSKI, J. et al. How transferable are features in deep neural networks? *CoRR*, abs/1411.1792, 2014. Disponível em: http://arxiv.org/abs/1411.1792. Acesso em: 27 nov. 2025.