UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

BRUNO ABREU KEMMER

Emprego de modelos generativos para envelhecimento facial

São Paulo
2023

BRUNO ABREU KEMMER

Emprego de modelos generativos para envelhecimento facial

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 07 de junho de 2023. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientador: Prof. Dr. Clodoaldo Aparecido de Moraes Lima

São Paulo

2023

Ficha catalográfica elaborada pela Biblioteca da Escola de Artes, Ciências e Humanidades, com os dados inseridos pelo(a) autor(a)
Brenda Fontes Malheiros de Castro CRB 8-7012; Sandra Tokarevicz CRB 8-4936

Abreu Kemmer, Bruno
Emprego de modelos generativos para envelhecimento facial / Bruno Abreu Kemmer; orientador, Clodoaldo Aparecido de Moraes Lima. -- São Paulo, 2023.
114 p: il.

Dissertacao (Mestrado em Ciencias) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, 2023.
Versão corrigida

1. Envelhecimento facial. 2. Progressão facial. 3. Modelos generativos. 4. Redes generativas adversárias. 5. GANs. 6. Modelos de difusão. I. Lima, Clodoaldo Aparecido de Moraes, orient. II. Título.

Dissertação de autoria de Bruno Abreu Kemmer, sob o título **"Emprego de modelos generativos para envelhecimento facial"**, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em ____ de _____ de _____ pela comissão julgadora constituída pelos doutores:

Prof. Dr.
Instituição
Presidente

Prof. Dr.
Instituição

Prof. Dr.
Instituição

Prof. Dr.
Instituição

Prof. Dr.
Instituição

Dedico este trabalho a minha esposa Caroline Lipnharski e aos meus dois filhos, Arthur e Lucas que nasceram durante o período da elaboração do presente estudo e a minha mãe Rosa Marquart.

Agradecimentos

Gostaria de agradecer a minha esposa, Caroline Lipnharski, pelo suporte necessário durante o período da elaboração do presente estudo, pois foi um época turbulenta, já que houve uma pandemia e tivemos nossos dois filhos, e mesmo assim foi possível continuar a pesquisa. Também queria reconhecer o esforço de minha mãe, Rosa Marquart, incentivando o desenvolvimento de minha vida acadêmica e também ao meu pai, Ezequiel Kemmer. Gostaria de agradecer ao meu orientador Professor Dr. Clodoaldo Aparecido Lima em seus imprescindíveis comentários e reuniões frequentes para ajudar a direcionar o trabalho e assim incrementando a qualidade desta pesquisa. Conjuntamente, agradeço ao PPgSI pelas disciplinas ministradas no programa que proporcionaram a estrutura e conhecimento para efetuar essa dissertação. E gostaria de fazer um especial agradecimento ao meu colega Rodolfo Simões que me ajudou nos experimentos, em discussões técnicas, foi co-autor em um capítulo publicado na editora Springer, além de revisar os meus textos acadêmicos.

Resumo

Kemmer, Bruno Abreu. **Emprego de modelos generativos para envelhecimento facial**. 2023. 114 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2023.

Diversos fatores podem causar o envelhecimento facial: exposição solar, tabagismo, viver em um ambiente poluído, estresse e fatores genéticos. Esses elementos tornam a tarefa de envelhecimento facial por meio de algoritmos bastante complexa, e ao fazê-la, três atributos são desejados na face envelhecida, a saber, i) essa deve ter a idade esperada; ii) deve manter as características individuais; iii) a face sintética deve, preferencialmente, ter alta qualidade. Dois tipos de métodos são tradicionalmente utilizados nessa tarefa: métodos baseados em protótipos e métodos baseados em modelagem. Os primeiros calculam a diferença entre as médias de faixas de idades e os últimos utilizam modelos paramétricos para simular a passagem do tempo. Ambos falham em manter as características individuais ao levar a imagem de um domínio mais jovem para um envelhecido. Com os avanços em visão computacional, modelos generativos têm sido utilizados nessa tarefa, em especial, Redes Adversárias Generativas (GANs) e recentemente, modelos de difusão. Com essas abordagens é possível gerar imagens realistas de indivíduos envelhecidos. Essas redes podem codificar informações latentes possibilitando que o modelo gerador crie novas imagens condicionais a uma face dada como entrada. Isso pode levar a uma melhora no desempenho dos sistemas biométricos, auxiliar na busca de pessoas desaparecidas, na identificação de pessoas procuradas de forma automatizada e diversas outras aplicações no entretenimento. Na última década, houve um número crescente de publicações com foco na aplicação de modelos generativos para envelhecimento facial. A maioria dessas adaptaram uma arquitetura geral para atender aos desafios da tarefa adotada. Este trabalho descreve detalhadamente as arquiteturas mais importantes encontradas na literatura, as principais bases de dados e suas aplicações. Além disso, apresenta dois grupos de experimentos realizados: o primeiro compara os resultados de três modelos publicados entre 2017 e 2018 que utilizaram redes GANs, em três bases de dados: FG-NET, UTKFaces e CACD. O segundo grupo realiza um estudo comparativo entre dois modelos baseados em redes GANs, que foram publicados entre 2020 e 2021, com dois modelos de difusão condicionais que executam edição das imagens, os quais foram publicados entre 2022 e 2023 e empregaram base de imagens de alta resolução *FFHQ-Aging* (base de imagens FFHQ com a idade estimada das faces). Por fim, para medir a efetividade do envelhecimento facial nas imagens, modelos estimadores de idade e de verificação facial foram utilizados. Os resultados mostraram que GANs treinadas especialmente para essa tarefa têm obtido resultados superiores, porém, modelos de difusão condicionais genéricos, como os utilizados nesse último grupo de experimentos obtiveram resultados consideráveis, mesmo sem terem sido treinados para essa tarefa. Além do mais, muitos trabalhos recentes têm apresentado melhoria nos modelos de difusão, portanto, são esperados rápidos avanços em suas arquiteturas.

Palavras-chaves: Envelhecimento facial, Progressão facial, Modelos generativos, Redes generativas adversárias, GANs, Autoencoders, Modelos de difusão.

Abstract

Kemmer, Bruno Abreu. **Face aging using generative models**. 2023. 114 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2023.

Several factors can cause facial aging: sun exposure, smoking, living in a polluted environment, stress, and genetic factors. These elements make the task of facial aging through algorithms complex, and when doing it, three attributes are desired, the aged face has the expected age; the maintenance of individual characteristics, and realistic synthetic images. Traditionally, two kinds of modeling techniques were used in this task: prototype methods and modeling methods. The first calculates the mean difference between age groups, and the latter uses parametric models to simulate the change over time. However, both approaches fail to keep the individual characteristics when transforming a face from a younger domain to an aged one. With advances in computer vision, generative models have been applied to perform this task, especially generative adversarial networks (GANs) and, recently, diffusion models. With them, it becomes possible to generate realistic aged faces of individuals. These networks can encode latent information enabling a generator model to create new images conditional on an input image. This could lead to improvements in biometric systems. Doing so could help to search for missing persons and identify criminals in an automated way, in addition to multiple applications in entertainment. In the last decade, an increasing number of publications focused on applying state-of-the-art generative models on facial aging. Most of these works were done by customizing a general architecture to fit the aging problem's needs. This work details the most frequent architectures found in literature, the main available benchmark databases, and their applications. Additionally, two groups of experiments were tested: the first comparing the results of three models published between 2017 and 2018 that use GANs networks in three frequently used databases: FG-NET, UTKFaces, and CACD. The second group presents a comparative study between two models based on GANs networks which were published between 2020 and 2021 with two conditional diffusion models that perform image editing. These were published between 2022 and 2023 and use the high-resolution image base FFHQ- Aging (the image base FFHQ with the estimated age of the photos). Finally, to measure the effectiveness of facial aging in photos, age estimation, and facial verification models are used. The results showed that GANs specially trained for this task have obtained superior results, however, generic conditional diffusion models, such as those used in the last group of experiments, got considerable results, even without having been trained for this task. Furthermore, many recent works have shown improvement in diffusion models' components. Therefore, rapid advances in their architectures are expected.

Keywords: Face aging, Face progression, Generative models, Generative Adversarial Networks, GANs, Autoencoders, Diffusion models.

Lista de figuras

Lista de quadros

Lista de tabelas

Sumário

1 Introdução

Os métodos computacionais de envelhecimento facial têm como objetivo gerar uma face envelhecida mantendo as características individuais. Existem diversos fatores que podem causar o envelhecimento: excesso de exposição solar, tabagismo, ambiente poluído, estresse e também fatores genéticos. Além disso, procedimentos estéticos cirúrgicos e não cirúrgicos podem atenuar os efeitos do tempo, assim como o uso de substâncias cosméticas no momento da captura da imagem. Por esses motivos, podemos afirmar que o envelhecimento facial é um processo complexo e não determinístico.

Recentemente o tema dessa dissertação tem sido alvo de muitas publicações (YANG *et al.*, 2021; HELJAKKA; SOLIN; KANNALA, 2018; PALSSON *et al.*, 2018; GRIMMER; RAMACHANDRA; BUSCH, 2021; ALALUF; PATASHNIK; COHEN-OR, 2021; ZOSS *et al.*, 2022), pois pode auxiliar de forma automatizada na busca de pessoas desaparecidas, na identificação de criminosos, e também para fins de entretenimento. Seu uso também está presente em tarefas biométricas (identificação de indivíduos baseada em características físicas ou comportamentais), pois possibilita diminuir a distância entre as características individuais presentes no momento do treinamento e o estado atual das faces, principalmente se o treinamento tiver sido feito com imagens antigas (HUANG; HU, 2021; KHANNA *et al.*, 2020).

Os métodos tradicionais para efetuar o envelhecimento facial podem ser divididos em duas categorias: métodos baseados em protótipos e os baseados em modelagem (Yun Fu; Guodong Guo; HUANG, 2010). Nos métodos baseados em protótipos é estimada a média em grupos de idade previamente definidos, e as diferenças entre as médias dos grupos representa a variação do envelhecimento entre os mesmos. Calculadas as médias, é possível envelhecer uma imagem. Porém, ao fazer isso, é descartada a individualidade dos exemplos, consequentemente, esses métodos não são adequados se for requerido manter as características individuais das faces. Já os métodos baseados em modelagem, utilizam modelos paramétricos para simular as mudanças nos elementos que compõem a face (forma e textura). Porém, necessitam de várias imagens do mesmo indivíduo, e estas ao longo de um intervalo grande de idades, limitando drasticamente seu uso em algumas aplicações, como exemplo, a biometria.

Muitas publicações recentes têm utilizado modelos generativos para efetuar o envelhecimento facial, obtendo resultados bastante realistas. Inicialmente foram utilizados *autoencoders adversários* e posteriormente, os modelos generativos adversários (do inglês, *generative adversarial networks*, GANs) e recentemente, utilizando modelos de difusão.

1.1 Problema de pesquisa

Alguns trabalhos de revisão anteriores trataram o tema de envelhecimento facial (Yun Fu; Guodong Guo; HUANG, 2010; SHU *et al.*, 2016a; ALQAHTANI; KAVAKLI-THORNE; KUMAR, 2019). Em Yun Fu, Guodong Guo e Huang (2010), os autores apresentaram a evolução do envelhecimento facial e da estimação de idade. Entretanto, esses trabalhos apresentaram apenas uma visão geral dos métodos de envelhecimento facial, além de que, desde a sua publicação, há mais de dez anos, muitos trabalhos foram adicionados. Em Shu *et al.* (2016a), a publicação revisa o tema de envelhecimento facial, no entanto, os modelos generativos não são abordados pois foram publicados posteriormente. Em Alqahtani, Kavakli-Thorne e Kumar (2019), os autores revisaram diversas aplicações de GANs, incluindo em uma seção o envelhecimento facial, mas não de forma abrangente e profunda. Posteriormente, Grimmer, Ramachandra e Busch (2021) resumiu os principais avanços na utilização de modelos generativos para o envelhecimento facial.

Atualmente muitos trabalhos foram publicados utilizando modelos generativos para envelhecer faces, aplicando diversas arquiteturas gerais de modelos generativos adversários (do inglês, *generative adversarial networks*, GANs) e de *autoencoders* adaptadas para a tarefa. Porém, nas publicações são utilizadas bases de dados distintas, nas quais são aplicadas diferentes técnicas de pré-processamentos. Além disso, não são usados os mesmos métodos quantitativos e qualitativos de avaliação nas imagens resultantes. Por último, devido à recência dos trabalhos publicados em edições de imagens reais utilizando modelos de difusão, poucas publicações têm usado essa técnica na tarefa de envelhecimento facial.

1.2 Objetivo

Este trabalho teve como objetivo compilar as diferentes técnicas de envelhecimento facial que utilizaram modelos generativos, destacando os modelos utilizados como referência,

as principais bases de dados usadas, comparar trabalhos selecionados utilizando mesmas bases, e contrastar a utilização de modelos de difusão e modelos GANs na tarefa.

1.3 Método de pesquisa

Foi conduzida uma revisão bibliográfica da literatura sobre técnicas de envelhecimento facial que utilizem modelos generativos (Capítulo 3). Essas foram categorizadas com base nos modelos gerais generativos que utilizaram na tarefa. As principais bases de dados utilizadas estão descritas em suas quantidades de indivíduos, número de fotos, limites de idades, e se necessário, outras características.

Por fim, foram comparados dois grupos de algoritmos para mostrar a evolução das técnicas ao longo do tempo. O primeiro consiste em três modelos publicados entre 2017 e 2018 que utilizaram imagens em baixa resolução, as três bases de dados mais utilizadas foram pré-processadas usando a mesma metodologia e os trabalhos escolhidos foram treinados com essas bases executando o mínimo de alteração possível, apenas o necessário para possibilitar a comparação entre os modelos utilizados. Por exemplo, um trabalho pode ter utilizado intervalos de idades com alguns anos de diferença entre si, esses foram unificados de modo apropriado. Esse trabalho resultou na publicação de um capítulo no livro Aprendizado Adversário Generativo: Arquiteturas e Aplicações (do inglês, *Generative Adversarial Learning: Architectures and Applications*) (KEMMER; SIMÕES; LIMA, 2022).

O segundo grupo de experimentos emprega dois modelos de GANs publicados entre 2020 e 2021, que utilizaram imagens em alta resolução e foram comparados com dois modelos de difusão condicionais, os quais executam edição em imagens e foram publicados entre 2022 e 2023.

1.4 Avaliação

Foram feitas comparações qualitativas e quantitativas utilizando as mesmas fotos e as envelhecendo para idades desejadas semelhantes. Adicionalmente, foram utilizadas redes neurais previamente treinadas que executem estimação de idade e verificação facial.

O modelo de estimação de idade foi utilizado para medir o desempenho do envelhecimento facial, estimando a idade da face gerada e comparando com a idade desejada (no

caso de modelos que fazem uso de intervalos de idades, foi utilizada a média). Um modelo de verificação facial foi empregado para conferir se as características do indivíduo foram mantidas na imagem produzida.

1.5 Contribuições

Contribuições do trabalho:

- Um estudo comparativo detalhado sobre os elementos necessários para a realização de uma pesquisa em envelhecimento facial utilizando modelos generativos:
  - Investigação dos modelos generativos já utilizados em publicações anteriores.
  - Descrição das bases de dados que possam ser utilizadas.
  - Descrição detalhada das métricas de avaliação aplicadas ao envelhecimento facial.
- Uma análise detalhada dos modelos escolhidos, executando esses nas bases de dados selecionadas, aplicando a mesma forma de tratamento nas imagens de entrada, verificando os resultados de forma qualitativa e quantitativa e destacando seus pontos fortes e fracos.
- Um estudo comparativo verificando se os modelos de difusão condicionais escolhidos podem efetuar o envelhecimento facial comparando com dois modelos que utilizam GANs.

1.6 Limitações de escopo

O escopo da pesquisa contemplava descrever os métodos utilizados na tarefa de envelhecimento facial, detalhar as bases de dados utilizadas e comparação de modelos selecionados. Além disso, fazia parte do escopo da pesquisa utilizar modelos de estimação de idade e de verificação facial previamente treinados para medir os resultados obtidos.

1.7 Riscos e ameaças à validade

O treinamento dos modelos generativos para o envelhecimento é custoso e demorado, portanto, o cronograma poderia ter sido afetado caso não fosse possível utilizar máquinas de poder computacional razoável.

Outro risco teria sido de que os hiper-parâmetros publicados pelos autores que foram otimizados para as bases de dados que eles utilizaram, não gerassem imagens de boa qualidade em bases diferentes das treinadas pelos próprios. A probabilidade desse evento ocorrer não era alta, porém não poderia ser descartada.

Os modelos de estimação de idade e de verificação facial poderiam ter uma performance aquém do esperado nas imagens envelhecidas geradas. Para medir seu desempenho, eles foram executados nas imagens em que se tem a idade real e isso será adicionado na descrição dos experimentos.

Finalmente, é conhecido o fato de que as bases de dados disponíveis contém vieses para algumas etnias, gêneros e poucas imagens de crianças e pessoas de idade avançada, por tal motivo, era esperada razoável perda de qualidade nas imagens geradas nesses grupos.

1.8 Estrutura do documento

Incluindo esse capítulo de introdução, o estudo está estruturado da seguinte forma:

- **Capítulo 2** apresenta os principais modelos generativos utilizados nos trabalhos de envelhecimento facial e os modelos escolhidos para serem comparados neste estudo.
- **Capítulo 3** apresenta a revisão bibliográfica da literatura sobre técnicas de envelhecimento facial que utilizem modelos generativos.
- **Capítulo 4** descreve as bases de dados utilizadas na tarefa de envelhecimento.
- **Capítulo 5** expõe as principais métricas utilizadas no tema.
- **Capítulo 6** tem o objetivo mostrar trabalhos correlatos que possuem um grau de similaridade com o tema.
- **Capítulo 7** detalha e discute os resultados obtidos dos experimentos executados.
- **Capítulo 8** apresenta conclusões e tópicos de interesse para trabalhos futuros.

- **Apêndice A** descreve as questões de pesquisa e o protocolo da da revisão bibliográfica.
- **Apêndice B** detalha os resultados e condução da revisão bibliográfica.

2 Modelos generativos

Modelos generativos buscam estimar os parâmetros que definem a distribuição de probabilidade original dos dados. Utilizam como premissa o fato de que a amostra de treino são dados retirados dessa distribuição, e possuem por trás de si uma variável latente (oculta).

Três exemplos de modelos generativos são:

- *Autoencoders* variacionais: utilizam uma rede generativa (codificador) e uma rede de inferência (decodificador).
- Redes adversárias generativas (do inglês, *Generative adversarial networks*, GANs): são compostos de uma rede generativa (codificador) e uma rede discriminadora.
- Modelos de difusão: aprendem a adicionar e remover ruído de forma a gerar imagens realistas.

Como observado em Goodfellow, Bengio e Courville (2016) modelos generativos são mais difíceis de serem treinados quando comparados a modelos de classificação ou regressão, pois não somente necessitam mapear a saída y, dado a entrada x, mas também, posicionar a distribuição de uma variável oculta z de modo útil, e ainda mapear x dado a representação z.

2.1 Autoencoders

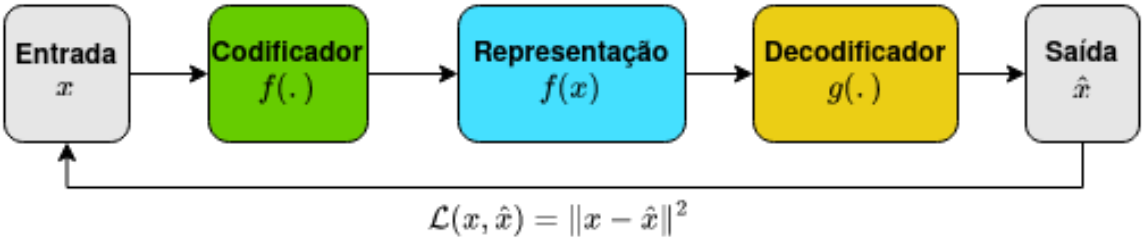
Autoencoders são modelos que tem como objetivo mapear a representação dos dados de entrada. O modelo é treinado para conseguir reproduzir em sua saída os dados de entrada.

Para que o mapeamento encontrado não seja o trivial (uma simples cópia da entrada em sua saída), são impostas restrições à rede, estas podem ser na camada de entrada (adição de ruído aos dados), ou restringir as camadas intermediárias, para que estas sejam de dimensão menor do que as camadas de entrada e saída.

Para isso, o modelo treina um codificador que mapeia os dados de entrada x para uma representação $f(x)$ e um decodificador, que a partir desta nova representação $f(x)$ consegue retornar os dados para uma aproximação da imagem original $\hat{x}$. O processo é ilustrado pela Figura 1. O *autoencoder* busca aproximar uma função que consiga descrever

o conjunto de treinamento e, como sua função de custo[1] o penaliza caso os dados na camada de saída não sejam parecidos com os da camada de entrada, o modelo irá buscar utilizar da melhor maneira que encontrar cada dimensão das camadas intermediárias a fim de melhor capturar as características mais importantes dos dados.

Figura 1 – Ilustração das etapas em um *autoencoder*: a entrada de alta dimensão x passa pelo codificador $f(.)$, por exemplo, uma rede neural, que a mapeia para um ponto no espaço de representação $f(x)$, geralmente em uma dimensão inferior. Em seguida, o decodificador é treinado para encontrar uma função que possibilite o mapeamento do espaço de representação para o espaço original $g(.)$, para obter uma aproximação da entrada $\hat{x}$. Neste exemplo, a função de custo é o erro quadrático médio.



Fonte: Kemmer, Simões e Lima (2022)

2.1.1 *Autoencoder* convolucional

Um tipo de *autoencoder* utilizado para fazer o mapeamento das principais características presentes em imagens, especialmente no caso de faces, é o *autoencoder* convolucional (MASCI *et al.*, 2011). A Figura 2 mostra um exemplo ilustrativo de arquitetura com seu codificador e decodificador[2] são redes neurais convolucionais. Nela, podemos ver os elementos que compõem o codificador (em verde), a representação mapeada é o vetor de 10 elementos h (em azul), e os componentes simétricos do decodificador (em amarelo).

Nesta arquitetura, a restrição está sendo imposta pelo fato de que a camada intermediária h (que está gerando a representação dos dados) tem uma dimensão consideravelmente menor comparada às camadas de entrada e de saída. Assim, a rede irá armazenar as características mais importantes, desprezando os ruídos e as informações

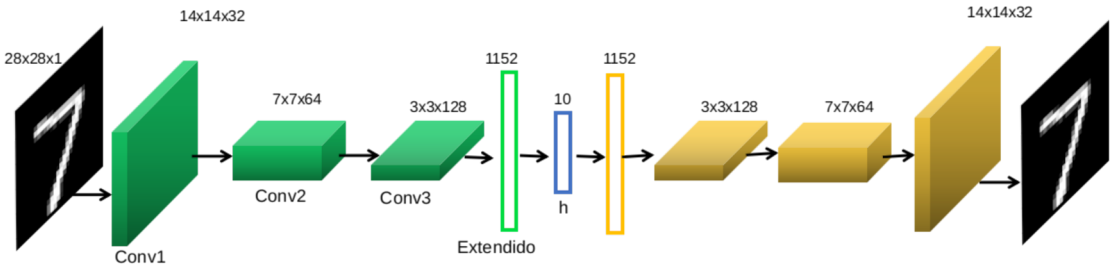
[1] A função de custo utilizada é apenas um exemplo.

[2] Alguns autores chamam as componentes convolucionais presentes no decodificador, de camadas deconvolucionais, ou também de convolução reversa.

que não são críticas para a recriação da imagem original, na sua melhor forma possível, na camada de saída.

De acordo com os autores (MASCI *et al.*, 2011), o principal fator proposto pelo *autoencoder* convolucional é a agressividade da restrição na dimensão da camada intermediária.

Figura 2 – Exemplo ilustrativo em que um *autoencoder* convolucional é utilizado para mapear as principais características de uma imagem. Neste caso, o codificador é composto por vários blocos de convolução, gerando a representação h, então o decodificador, consistindo em blocos de deconvoluções, restaura a imagem ao domínio original.



Fonte: Bruno Abreu Kemmer, 2023 (baseado em Guo *et al.* (2017))

2.1.2 *Autoencoder* variacional

Os *autoencoders* variacionais (do inglês, *variational autoencoder*, VAE) foram introduzidos por Kingma e Welling (2014) e são um dos tipos de *autoencoders* mais utilizados atualmente. São compostos de codificadores que buscam obter os parâmetros da distribuição dos dados originais, cujos dados de treinamento foram amostrados, ficando menos sobre ajustado ao conjunto de treinamento, e consequentemente, o torna mais robusto. Além disso, essa abordagem permite que, após encontrar os melhores parâmetros, seja possível manipular os valores de entrada, possibilitando gerar novos exemplos não existentes no conjunto de treinamento.

A Figura 3, mostra um *autoencoder* variacional. Nesse exemplo, as duas primeiras camadas buscam encontrar qual a melhor média μ e desvio padrão σ, para cada uma das variáveis. O objetivo do codificador é estimar o vetor de média e de desvios padrões condicionado aos dados de entrada. Desta distribuição será amostrado um vetor latente z, o qual, será utilizado pelo decodificador para reconstruir a imagem original x em sua aproximação $\hat{x}$.

Portanto, o codificador tem como objetivo estimar a distribuição dos exemplos condicionados a entrada x: $p_\phi(z|x)$. Já o decodificador, busca mapear x dado o exemplo amostrado z: $q_\theta(x|z)$. Em que p_ϕ é a função densidade de probabilidade mapeada pelo codificador e q_θ é a função densidade de probabilidade mapeada pelo decodificador.

A função de custo, descrita pela Equação 3, passa a ter duas componentes:

- Custo de reconstrução: penalização pela diferença entre a saída $\hat{x}$ e a entrada x. Assim como nos *autoencoders* tradicionais. Exemplo: $\mathcal{L}(x, \hat{x}) = \|x - \hat{x}\|_2$
- Termo de regularização: adiciona uma penalização entre quão distinta é a distribuição de probabilidade encontrada pelo modelo e uma distribuição a Priori[3]. $D(p_\phi(z|x) \parallel p(z))$.

  Esse termo evita que o modelo se sobre-ajuste aos dados, pois força que a distribuição se adéque a forma da distribuição a Priori, balanceando os erros ao longo das variáveis. Uma distribuição a Priori frequentemente utilizada é a distribuição gaussiana normal $p(z) = \mathcal{N}(\mu = 0, \sigma^2 = 1)$. Ao utilizar-la, a Divergência de Kullback-Leibler (KL)[4] é determinada pela Equação 2.

$$D(p_\phi(z|x) \parallel p(z)) = -\frac{1}{2}\sum_j (\sigma_j + \mu_j^2 - 1 - \log \sigma_j) \tag{2}$$

$$\mathcal{L}(\phi, \theta, x) = \mathcal{L}(x, \hat{x}) + D(p_\phi(z|x) \parallel p(z)) \tag{3}$$

Um ponto importante a ser observado é que, caso o vetor latente z fosse amostrado diretamente da distribuição definida pelos parâmetros μ e σ, poderíamos estimar a saída $\hat{x}$. Porém não seria possível retro-propagar o erro, já que esta camada não seria determinística. Ao aplicar a técnica de reparametrização para amostrar z, foi adicionado um componente aleatório apenas na variabilidade de z, logo, o valor da variável passa a ser definido e passível de ser retro-propagado. Isso é visto na Equação 4 e também pelas componentes cinzas no centro da Figura 3.

[3] Hipótese inicial da distribuição de probabilidade original dos dados.

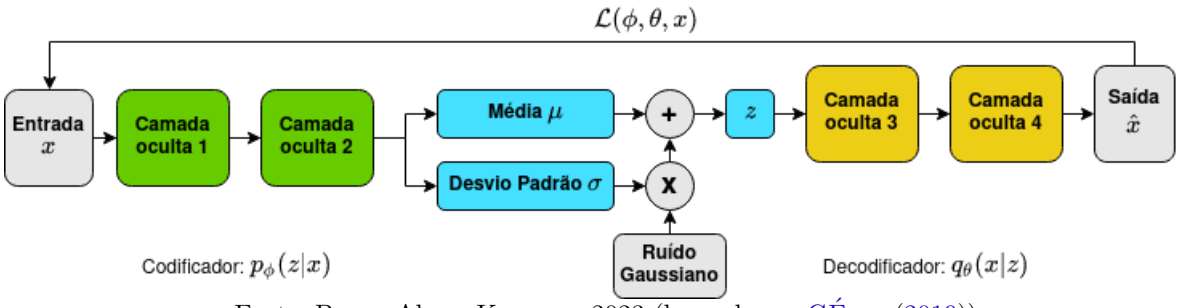
[4] A Divergência de Kullback-Leibler, $D_{KL}(P\|Q)$, é ganho de informação obtido se a distribuição de probabilidade Q for usada ao invés da distribuição P. Portanto, mede o erro de aproximação entre duas distribuições de probabilidade. É a métrica encontrada na maioria das implementações deste método.

$$D_{KL}(P\|Q) = \int q(x) \log \frac{q(x)}{p(x)} dx \tag{1}$$

$$z = \mu + \sigma \odot \epsilon, \ \ onde \ \epsilon \sim \mathcal{N}(0, 1) \tag{4}$$

em que $\odot$ é o multiplicador elemento a elemento.

Figura 3 – Ilustração das etapas em um *autoencoder* variacional: neste exemplo, o codificador e o decodificador consistem em redes neurais com duas camadas ocultas. O codificador tem como objetivo encontrar a melhor média μ e a melhor variância σ^2 que possam parametrizar a distribuição latente.



Fonte: Bruno Abreu Kemmer, 2023 (baseado em GÉron (2019))

Autoencoders variacionais também podem ser utilizados para determinar se existe viés na distribuição latente em que os modelos de aprendizado de máquina foram treinados, como podemos ver no trabalho de Amini *et al.* (2019). Assim é possível amostrar exemplos em todos os espectros das variáveis de entrada destes, evidenciando os vieses existentes.

Autoencoders variacionais podem ser usados no envelhecimento facial, como demonstrado por Pantraki, Kotropoulos e Lanitis (2019), em uma arquitetura baseada em VAE e GANs para fazer uma tradução imagem a imagem não supervisionada. Essa arquitetura utiliza uma pirâmide de codificadores locais e globais para se adaptar gradualmente ao efeito do envelhecimento. Como a representação latente z é compartilhada entre o codificador (VAE) e o gerador (GAN), isso garante a personalidade do indivíduo. Além disso, no trabalho de Chandaliya e Nain (2019) o gerador da GAN, que realiza o envelhecimento facial, vem do decodificador de um VAE.

2.1.3 *Autoencoders* adversários

Os *autoencoders* adversários (do inglês, *adversarial autoencoders*, AAE) (MAKHZANI *et al.*, 2016), são *autoencoders* probabilísticos com um método de regularização que obriga a representação latente do *autoencoder* ser parametrizada por uma distribuição a Priori arbitrária, e assim fazer com que a distribuição *a posteriori* siga essa codificação, possibilitando a geração de exemplos com sentido ao manipular a codificação de entrada.

Os autores definiram o problema como: z é o vetor de código latente, $q(z|x)$ e $p(x|z)$ são distribuições de codificação e decodificação, respectivamente. A função de codificação define uma distribuição posterior agregada de $q(z)$. O AAE é regularizado combinando $q(z)$, a uma distribuição a Priori arbitrária, $p(z)$. O método é semelhante ao VAE, no entanto, enquanto VAE usa uma penalidade de divergência KL para impor uma distribuição prévia no vetor de código latente do *autoencoder*, o AAE usa um procedimento de treinamento adversário, semelhante ao usado em GANs. Além disso, AAE captura melhor a variedade de dados em comparação com VAE (ZHANG; SONG; QI, 2017).

Os autores Zhang, Song e Qi (2017) condicionaram as distribuições utilizadas nos AAE, na arquitetura de *autoencoders* adversários condicionais (do inglês, *conditional adversarial autoencoders*, CAAE) ao adicionar vetores com as categorias (os rótulos) dos elementos, gerando imagens com qualidade superior.

2.2 Redes adversárias generativas, GANs

Redes adversárias generativas (do inglês, *Generative adversarial networks*, GANs) (GOODFELLOW *et al.*, 2014), são compostas de duas redes neurais: o gerador, o qual é uma rede com objetivo de gerar exemplos próximos aos dados reais; e o discriminador, cuja rede classificadora tem como alvo separar os dados gerados dos reais. Estas duas redes competem entre si durante a etapa de treinamento, em que o gerador tenta produzir exemplos tão próximos aos reais a ponto de enganar o discriminador, que tenta aprimorar a detecção das imagens geradas.

O método é treinado em duas fases: na primeira, é treinado o discriminador e sua entrada é composta de uma amostra dos dados reais junto com a mesma quantidade de exemplos gerados[5]. Esse é um problema de classificação binária em que os dados gerados tem classe 0 e os dados reais tem classe 1. A função de custo utilizada é a entropia cruzada. Nesta fase, somente os pesos da rede do discriminador são atualizados, já na segunda, o gerador é treinado. É gerado um conjunto de exemplos e todos esses recebem a mesma classe 1, e portanto para o discriminador, toda essa entrada é composta de dados reais. Assim como na etapa anterior, somente os pesos da rede generativa são ajustados. A Figura 4 ilustra os componentes presentes.

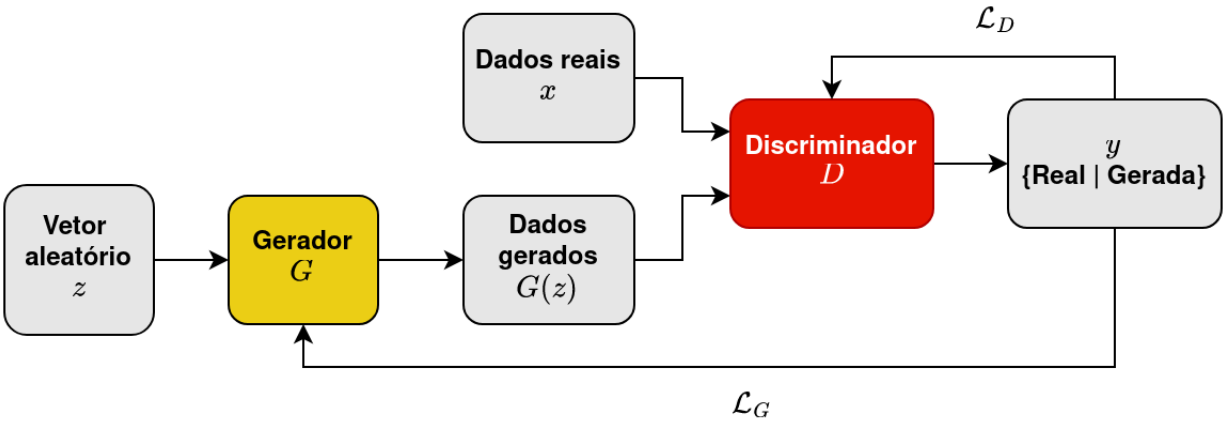
[5] Em geral, os exemplos são gerados com uma entrada de ruído Gaussiano na rede generativa.

O objetivo do gerador G é de que ele aprenda a real distribuição dos dados de treinamento, e no caso do discriminador D é de que ao final do treinamento, tenha uma performance aproximadamente igual ao acaso, ou seja, $\frac{1}{2}$ para ambas as classes. A otimização da GAN é executada com relação à função de custo conjunta de G e D, conforme a Equação 5. Em que $q_{data}(x)$ é a distribuição dos dados de treinamento e $p(z)$ é a distribuição da variável latente z.

$$L = \min_G \max_D \mathbb{E}_{x \sim q_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z))] \tag{5}$$

Essa arquitetura original também é chamada de redes adversárias generativas totalmente conectadas (do inglês, *Fully-connected generative adversarial networks*, FCGAN)- por ser composta de redes neurais com camadas ocultas totalmente conectadas.

Figura 4 – Ilustração das etapas em uma GAN: a rede do gerador G é alimentada com uma variável aleatória z. Seu objetivo é enganar a rede discriminadora, gerando imagens o mais próximo possível das reais. E a função da rede discriminadora D é determinar se a imagem y foi gerada ou é real.



Fonte: Kemmer, Simões e Lima (2022)

2.2.1 Problemas comuns

2.2.1.1 Colapso para a moda (do inglês, *mode collapse*)

Um problema comum apresentado na FCGAN (e em outras que compartilham de funções de custo similares) é que ao longo do treinamento da GAN, o gerador pode observar uma vulnerabilidade do discriminador em detectar se algumas classes são verdadeiras

ou falsas. Percebendo alguma classe que contenha uma vulnerabilidade, o gerador acaba focando nela, pois consegue enganar o discriminador mais facilmente, chegando ao ponto do gerador se especializar e apenas gerar essa classe, com boa qualidade. Porém, no momento em que o discriminador conseguir detectar o padrão das imagens geradas desta classe, o gerador estará em um mínimo local, e não conseguirá sair dele, parando o aprendizado do modelo.

Outro problema que pode ocorrer durante o treinamento, é que tradicionalmente GANs utilizam como função de custo uma variação derivada da entropia cruzada entre a distribuição real e a gerada, conforme Equação 5. Como a saída do discriminador é binária, pode limitar o aprendizado, já que o discriminador não transmite a informação para o gerador de quão falsa ele interpreta que a imagem é, somente se é falsa ou real. E já que sua tarefa é mais fácil, pode ocorrer o problema de dissipação do gradiente (do inglês, *vanishing gradients*).

Uma abordagem para resolver esse problema é utilizar outra função de custo, a perda de Wasserstein (ARJOVSKY; CHINTALA; BOTTOU, 2017). Nessa função (Equação 6), o crítico C substitui o papel do discriminador, já que não é empregado um classificador para determinar se a imagem é real ou gerada, mas sim uma crítica (um número real não limitado). O crítico também é uma rede neural e tem um requisito de que a função que ele representa tem que ser 1-Lipschitz (1-L) contínua, Equação 7. Isso significa que em todos os pontos, a norma do gradiente tem que ser no máximo 1, garantindo estabilidade, já que estará crescendo de forma linear. Quando isso ocorre, pode-se dizer que a função de perda de Wasserstein é válida.

$$\min_G \max_C |\mathbb{E}[C(x)] - \mathbb{E}[C(G(x))]| \tag{6}$$

$$\|\nabla f(x) \leq 1\|_2 \tag{7}$$

Dois métodos podem garantir a condição 1-L:

- **Corte de gradiente** (do inglês, *gradient clipping*): Após atualizar os gradientes e calcular o gradiente descendente, todos os valores acima dos limites são substituídos pelos valores limites. Um ponto negativo dessa abordagem é que limita o quanto o modelo pode aprender.

- **Regularização de gradiente** (do inglês, *gradient penalty*): Nesse caso, é adicionado um termo de regularização, sendo esse a interpolação entre a imagem gerada e a imagem real. A proporção de cada uma é definida pela variável aleatória ϵ. Essa técnica garante estabilidade ao treinamento por atender à restrição 1-L, possibilitando um aprendizado maior, comparado ao corte de gradiente.

2.2.1.2 Consistência local

Um problema observado na tentativa de gerar imagens maiores nas arquiteturas de GANs iniciais era que as imagens eram consistentes[6] globalmente, mas não localmente, existiam artefatos que um observador atento poderia identificar como não sendo de uma imagem real.

Para resolver isso Isola *et al.* (2017) utilizou uma *PatchGAN* em seu discriminador. Ao invés de receber a imagem e ter como saída apenas um valor para determinar se a imagem inteira é real ou gerada, o discriminador recebe segmentos da imagem e avalia cada um para determinar se o segmento é real. Toda a imagem acaba sendo avaliada utilizando uma janela deslizante e o resultado é uma matriz de probabilidades. Para o caso das imagens geradas, o objetivo deste discriminador é gerar uma matriz de zeros, representando que todos os segmentos são falsos. E para as imagens reais, uma matriz de valores unitários.

2.2.2 Tipos de GANs

Podemos classificar as arquiteturas de GANs em dois grupos relacionados ao modo de como as imagens são geradas: GANs condicionais e incondicionais. No primeiro grupo, as imagens geradas são condicionadas a uma classe previamente definida. No segundo, não existe restrição de qual classe será obtida, logo é um evento aleatório. Um benefício considerável das GANs incondicionais é que as bases de dados utilizadas não precisam ter seus exemplos rotulados.

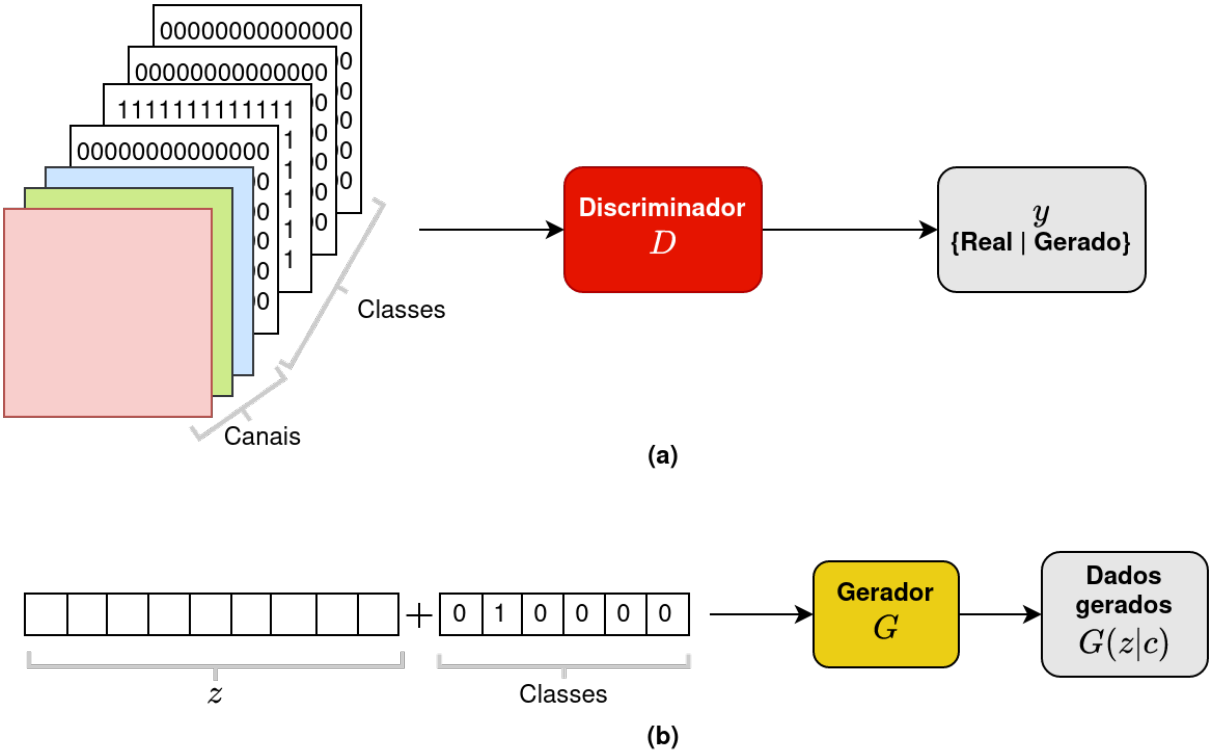
[6] Consistentes seriam imagens que um observador humano consideraria como reais.

2.2.2.1 GANs condicionais

Redes adversárias generativas condicionais (do inglês, *Conditional generative adversarial networks*, CGANs). Nessa variação, as redes são treinadas com dados rotulados, possibilitando que a rede generativa consiga criar dados de rótulos específicos. Assim, a rede generativa está modelando a distribuição dos dados condicionado a um rótulo.

A Figura 5 ilustra como o discriminador recebe tanto os dados reais quantos os gerados, condicionados a uma classe. Também mostra como o gerador recebe a variável latente z concatenada ao vetor condicional das classes, um rótulo codificado na forma de codificação binária (do inglês, *one-hot-encoding*). A rede generativa G aprenderá a gerar exemplos na forma $\hat{x} = G(z|c)$ e a informação da classe também será enviada ao discriminador.

Figura 5 – Ilustração da entrada do discriminador em uma arquitetura CGAN, consistindo nos canais da imagem (gerada ou real) concatenados com matrizes de codificação binária que contém a informação das classes (item a). No caso da entrada do gerador, o vetor latente z é concatenado a um vetor de codificação binária (item b).



Fonte: Kemmer, Simões e Lima (2022)

A aplicação de GANs condicionais no envelhecimento facial ocorre mapeando cada grupo de idade para uma classe. A imagem de entrada é codificada para seu vetor latente

z, concatenando com o vetor binário da classe com a idade que se deseja obter. Com o gerador já treinado, esse irá receber $z^* = [z, c]$, gerando uma versão sintética do indivíduo no intervalo de idade especificado pela classe, $G(z^*|c)$.

2.2.2.2 GANs incondicionais

Conforme descrito anteriormente, esse tipo de GAN irá gerar uma imagem em uma classe aleatória. Um modo de conseguir controlar os atributos da imagem obtida, consiste em manipular o vetor latente z, o qual será responsável por definir as características da imagem na saída do gerador. Podendo adicionar elementos (óculos, bigodes, etc) e modificar atributos (mudar a cor do cabelo, envelhecer, rejuvenescer, etc). Para que isso seja possível, é necessário encontrar dimensões no espaço latente de z que consigam controlar essas características. Essa abordagem não necessita de bases de dados rotuladas.

O espaço latente é variado até que se encontre uma imagem gerada com as características desejadas através de métodos de otimização (RICHARDSON *et al.*, 2021) ou classificadores pré-treinados (VIAZOVETSKYI; IVASHKIN; KASHIN, 2020), diferente das GANs condicionais em que o gerador é treinado com dados rotulados já com as classes pré-definidas.

Um dos problemas que pode ocorrer nesse processo é a correlação das dimensões da variável latente. Como não existe uma penalização para que as variáveis não estejam correlacionadas durante seu treinamento, quando duas características ocorrem simultaneamente na base de dados de treino, existe uma grande chance de essas estarem representadas em uma mesma dimensão de z. Por exemplo, ser homem e ter um bigode, e ao tentar manipular uma imagem gerada de uma mulher para conter um bigode, a deixará com feições masculinas.

Além disso, caso o vetor latente não tenha dimensões suficientes para armazenar separadamente os atributos a serem gerados, uma dimensão poderá controlar múltiplas características da imagem, dificultando ou impossibilitando um controle posterior.

Um método para obter vetores latentes controlados é utilizar o gradiente de um classificador pré-treinado que verifica se a imagem gerada tem ou não a característica desejada. Assim, possibilita mover o vetor latente z na direção que contém essa característica.

2.2.2.3 Tradução de imagem a imagem

A tradução de imagem a imagem (do inglês, *image-to-image translation*) consiste em obter uma imagem correspondente em outro domínio. Ela pode ocorrer de duas formas:

- Pareada: com pares de imagens, uma sendo de um domínio, e outra respectiva do outro domínio. Neste caso, as redes buscam manter o conteúdo e modificar o estilo das imagens. Um exemplo dessa arquitetura é a rede *pix2pix*.
- Não-pareada: são buscadas partes semelhantes de duas imagens de domínios distintos para transformar de um estilo no outro e vice-versa, mantendo o conteúdo similar. Demonstrado na *CycleGAN*.

2.2.3 Arquiteturas de referência

2.2.3.1 CGAN

Inicialmente proposta por Mirza e Osindero (2014), adicionou componentes condicionais a FCGAN, e a Equação 8 mostra como varia sua função de custo.

$$L_G = \min_G \max_D \mathbb{E}_{x|c\sim q_{data}(x|c)}[\log D(x)] + \mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z|c))] \tag{8}$$

Em Isola *et al.* (2017), foi utilizada uma CGAN que, dada uma imagem, gera uma variação com outros atributos. O título do trabalho é translação imagem a imagem (do inglês, *Image-to-image translation*).

Em Age-CGAN (ANTIPOV; BACCOUCHE; DUGELAY, 2018b), foi empregada uma CGAN para gerar faces envelhecidas, com a adição de duas redes neurais pré-treinadas, uma fazendo o reconhecimento facial e assim medindo quão distante a face gerada está da imagem original, também foi utilizada uma rede estimadora de idade para estimar a idade da imagem criada, penalizando caso ela esteja distante da desejada.

A arquitetura de uma Age-CGAN consiste em um codificador $E(x)$ que irá mapear a figura em alta dimensão para um vetor de baixa dimensão z. A rede de reconhecimento $FR(.)$ é utilizada para garantir a preservação da identidade. A Equação 9 mostra que a rede irá penalizar nos casos em que a imagem reconstruída $\hat{x}$ desvie muito da face original x, em termos da representação de suas características individuais. A rede generativa $G(z, y)$

irá gerar uma versão envelhecida da entrada condicional ao vetor binário de idade desejada y. Finalmente, a rede discriminadora D irá diferenciar entre as imagens reais e as geradas.

$$z^*_{IP} =_z \|FR(x) - FR(\hat{x})\|_2 \tag{9}$$

Adicionalmente, Yang *et al.* (2021) apresentou resultados promissores ao incorporar redes de verificação facial e estimadores de idades, na arquitetura da CGAN e uma estrutura piramidal de discriminadores.

2.2.3.2 Redes adversárias generativas convolucionais profundas, DCGANs.

Redes adversárias generativas convolucionais profundas (do inglês, *Deep convolutional generative adversarial networks*, DCGANs). Esta GAN incondicional, foi apresentada por Goodfellow *et al.* (2014) que utilizou camadas convolucionais para gerar pequenas imagens, usando as seguintes bases de dados: MNIST (LECUN; CORTES; BURGES, 2010), TFD (SUSSKIND; ANDERSON; HINTON, 2010) e CIFAR-10 (KRIZHEVSKY, 2009). Já ao final de 2015, Radford, Metz e Chintala (2016) adotou geradores com camadas de deconvolução e discriminadores com camadas de convolução, para gerar imagens grandes e realistas.

DCGANs são capazes de obter representações latentes dos dados, de forma que possibilitem operações aritméticas entre esses vetores. No artigo, os autores mostram ser possível calcular uma subtração entre a representação latente de um homem com óculos e de um homem sem óculos, e ao somar a esse resultado uma representação latente de uma mulher sem óculos, obtiveram uma representação latente de uma mulher com óculos.

Porém, ao tentar gerar imagens consideravelmente grandes utilizando DCGANs, o resultado foram imagens consistentes localmente, mas com falhas "globais", em detalhes perceptíveis a um olhar humano.

Um outro ponto positivo do trabalho foi mostrar técnicas empíricas que evitam problemas comuns no treinamento dos geradores e discriminadores em GANs. Substituindo camadas de *pooling* por convoluções espaçadas (do inglês, *strided convolutions*) e normalizações em lote (do inglês, *batch normalizations*).

Uma *DCGAN* condicionada ao sexo e à idade foi usada em Liu *et al.* (2019) para executar a progressão facial. Um codificador captura a imagem de alta dimensão em um

vetor latente com as características pessoais, conectado com um vetor condicional ao sexo e idade na entrada da *DCGAN*, para que os rostos gerados tenham individualidades condicionadas.

2.2.3.3 Tradução imagem a imagem com GANs condicionais, *Pix2Pix*.

Em tradução imagem a imagem com GANs condicionais (do inglês, *Image-to-image translation with conditional adversarial Networks*, *pix2pix*) (ISOLA *et al.*, 2017) temos uma tradução de imagem a imagem pareada. Foi utilizada uma GAN condicional para obter um par de imagens de dois domínios. Porém, ao invés de ter um vetor codificando as condições, as imagens dos dois domínios são enviadas para o discriminador analisar se são verdadeiras ou geradas.

Seu gerador utiliza uma rede-U (do inglês, *U-net*) (RONNEBERGER; FISCHER; BROX, 2015), essa consiste em um codificador e decodificador com conexões puladas (do inglês, *skip connections*) para evitar o sobre-ajuste. Originalmente essa rede foi utilizada com sucesso na tarefa de segmentação de imagens.

Logo, a imagem no domínio original será codificada em um vetor latente de menor dimensão, e o decodificador a recriará no outro domínio. Como são imagens pareadas, a imagem obtida deveria ser próxima de seu par real, e por isso é utilizada a distância L-1 entre elas na função de custo, o que também auxilia a obter atributos de baixa frequência corretos.

Um ponto interessante é que os autores não observaram grande vantagem em utilizar um vetor de ruído na entrada do gerador como em arquiteturas anteriores, mas adicionaram camadas de *dropout* para obter o efeito de estocasticidade. Esse trabalho foi o que definiu o termo *PatchGAN* que é utilizado em seu discriminador.

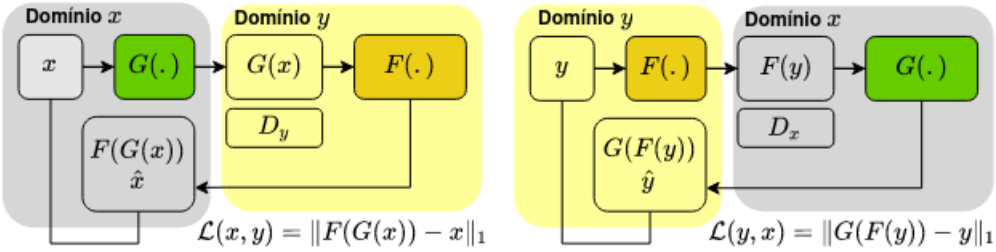
Em Wang *et al.* (2018), conhecido como *pix2pixHD*, melhorias na arquitetura permitiram a geração de imagens em alta resolução (2048x1024). Utilizou-se geradores e discriminadores em múltiplos tamanhos de imagem, sendo um gerador global e outro de aperfeiçoamento local. E três discriminadores que atuam em 1x, 2x e 4x das resoluções das imagens.

2.2.3.4 *CycleGAN*

Uma rede *CycleGAN* (ZHU *et al.*, 2017) faz a tradução de imagem a imagem não pareada. O objetivo é aprender uma função de mapeamento do domínio x para y e vice-versa. Isso é ilustrado pela Figura 6. Por não precisar de que a base de dados seja pareada com um exemplo de cada domínio auxilia na tarefa de envelhecimento facial, traduzindo de domínios mais jovens para mais velhos, sem exemplos correspondentes de cada domínio. Utiliza uma *PatchGAN* como discriminador com a função de perda, o erro de mínimos quadrados. Além disso, dois termos são adicionados à função de custo:

1. Perda por consistência de ciclo: distância entre os pixels das duas imagens já que apenas os estilos das mesmas deveriam ter mudado.
2. Perda de identidade: esse termo é opcional na *CycleGAN*. Também é a distância entre pixels entre uma imagem de um domínio ao passar no gerador que tem como objetivo gerar uma imagem desse mesmo domínio. Como ela já tem o estilo esperado, nada deveria mudar, idealmente essa distância deveria ser próxima de zero.

Figura 6 – A arquitetura da *CycleGAN*: o objetivo é aprender o mapeamento do domínio x para y, tornando as diferenças das imagens geradas $G(x)$ e das imagens reais y imperceptíveis ao discriminador D_y. A mesma abordagem é aplicada para encontrar o mapeamento do domínio y para x.



Fonte: Kemmer, Simões e Lima (2022)

Alguns trabalhos usaram o treinamento par-a-par de *CycleGANs* entre grupos de idade, Palsson *et al.* (2018) adicionou um modelo de predição de idade pré-treinado, o que melhora os resultados quando a lacuna entre a idade original e a desejada é pequena. Por fim, Yang e Lv (2020) dividiu os exemplos em faixas etárias e gênero, uma vez que os efeitos do envelhecimento diferem entre os dois, e usou *CycleGAN* para realizar a tradução do domínio.

2.2.3.5 PROGAN

Crescimento progressivo de redes adversárias generativas (do inglês, *Progressive growing of generative adversarial networks*, PROGAN) (KARRAS *et al.*, 2018), neste método tanto o discriminador quanto o gerador aumentam simultaneamente a complexidade da arquitetura da rede ao longo das iterações[7]. Eles começaram com uma imagem com poucos *pixels* e foram aumentando gradativamente: 4x4, 8x8, até 1024x1024. Isso permitiu a geração de imagens de alta resolução, uma vez que as redes começaram por aprender a estrutura das imagens e, progressivamente, abordam os detalhes menores em imagens de alta resolução.

Essa abordagem também reduz o tempo de treinamento, visto que a maioria das iterações ocorre quando a rede é mais simples e permite tamanhos de lote (do inglês, *batch size*) maiores devido às restrições de memória, pois, a princípio, terá imagens de menor resolução.

Em Shen *et al.* (2020), os autores propuseram um *framework* para interpretar a representação latente usada como entrada da *PROGAN* e da *StyleGAN* para edição semântica da face. O *framework* classificou os atributos no vetor latente usando um modelo supervisionado [8], permitindo a manipulação da variável latente para produzir imagens com as características desejadas.

2.2.3.6 *StyleGAN*

O trabalho apresentado em Karras, Laine e Aila (2019) mostrou uma arquitetura de GANs a qual consegue utilizar técnicas de transferência de estilo entre imagens. O objetivo principal é a geração de imagens de alta definição que apresentem realismo, além de serem diversas entre si. O método consegue obter variáveis latentes correspondentes à pose, textura, idade, entre outras.

As principais componentes utilizadas no trabalho foram:

- Crescimento progressivo das imagens, assim como utilizado em PROGAN.

[7] As iterações seguiram uma heurística definida pelos autores, descrita em seu trabalho.
[8] Os autores usaram uma SVM para realizar a tarefa de classificação e anotaram manualmente 4.000 exemplos.

- Uma rede neural (rede *perceptron* multicamadas, com 8 camadas) que efetua o mapeamento do vetor latente z para um outro vetor latente w, mas com as variáveis que controlam o estilo isoladas e podendo ser controladas.
- Normalização de estilo adaptativo a cada instância (do inglês, *adaptative instance normalization*, AdaIN) (HUANG; BELONGIE, 2019). É um método de normalização semelhante à normalização em lote, porém ocorre no nível da instância, para cada canal dessa. Feito isso, o vetor w é conectado a duas camadas totalmente conectadas que têm como saída dois termos: escala $\sigma(y)$ e o viés $\mu(y)$. Após a normalização da instância, Equação 10, é aplicado o estilo $\sigma(y)$, Equação 11, em que $\mu(y)$ é uma translação. Essa normalização ocorre em múltiplas camadas do gerador.
- Mistura de estilos (do inglês, *style mixing*) consiste em misturar diferentes vetores latentes w_1 e w_2 (obtidos de respectivos vetores latentes distintos iniciais, por exemplo, z_1 e z_2). Os estilos que cada w controla serão aplicados em diferentes camadas do gerador, podendo controlar quanto cada um impactará na imagem resultante. Como essa abordagem é utilizada durante o treinamento, aumenta a diversidade das imagens geradas.
- Variação estocástica (do inglês, *stochastic variation*), consiste em obter variações da imagem gerada ao concatenar ruído logo antes da normalização AdaIN. Ao afetar as camadas iniciais, modificam a estrutura das imagens sendo que as camadas finais mudam detalhes mais sutis.

$$x_{norm} = \frac{x - \mu(x)}{\sigma(x)} \tag{10}$$

$$AdaIN(x, y) = \sigma(y)\frac{x - \mu(x)}{\sigma(x)} + \mu(y) \tag{11}$$

Foi publicada uma nova versão deste método em Karras *et al.* (2020), com melhorias tanto na arquitetura das redes, quanto na metodologia adotada para seu treinamento.

Em Or-El *et al.* (2020a), uma *StyleGAN* foi usada para realizar o envelhecimento facial tanto nas texturas quanto na forma, um aspecto que a maioria dos trabalhos não consegue. Além disso, Viazovetskyi, Ivashkin e Kashin (2020), destilou aspectos da representação latente usada pela *StyleGAN* para gerar imagens, permitindo a troca de gênero, adicionando sorrisos e também executando o envelhecimento facial.

2.3 Modelos de difusão

Modelos de difusão têm como objetivo desconstruir a estrutura em uma distribuição de dados de forma lenta e sistemática em um processo iterativo de difusão no passo para frente (do inglês, *forward diffusion process*). Também é aprendido um processo de difusão reverso (do inglês, *reverse diffusion process*) que recria a estrutura original dos dados, gerando um modelo generativo dos dados muito flexível e computacionalmente tratável (SOHL-DICKSTEIN *et al.*, 2015).

Os autores Ho, Jain e Abbeel (2020) conseguiram bons resultados ao tentar prever o ruído $\mathcal{N}(\mu, \sigma^2)$ mas fixando σ^2. Também foi adicionado ruído de forma linear no passo para frente, usado um agendamento linear[9]. E a reconstrução foi feita utilizando uma arquitetura U-Net com blocos de atenção. Estes são os modelos de difusão probabilísticos de remoção de ruído (do inglês, *Denoising diffusion probabilistic models*, DDPM).

Em (NICHOL; DHARIWAL, 2021) os autores obtiveram melhores resultados com a utilização de aperfeiçoamentos como a adição no passo para frente de um ruído por um agendamento de co-senos, pois notaram que esse método desconstruía o sinal da imagem mais lentamente, melhorando o aprendizado. Também fizeram com que a rede neural que reconstrói os dados aprendesse os parâmetros em σ^2, aumentaram a profundidade das camadas escondidas e diminuíram a quantidade delas, ampliaram a quantidade de camadas de atenção e o número de cabeças de atenção, além de outros melhoramentos.

Em (SONG; MENG; ERMON, 2020), modelos de difusão implícitos removedores de ruído (do inglês, *Denoising Diffusion Implicit Models*, DDIM) foi mostrado que era possível otimizar a etapa de passo para frente de modelos DDPM ao utilizar uma abordagem não markoviana. Pois, ao invés de serem necessários 1.000 passos t de remoção de ruído, frequentemente utilizados nos modelos DDPM, era possível utilizar apenas um subconjunto (progressivo) dos passos t, acelerando consideravelmente o processo de amostragem, já que com 200 passos ou menos obtinham bons resultados. A Equação 12 mostra a adição de ruído no processo de amostragem obtida na imagem, x_t, no passo t, com ruído gaussiano ϵ, variância unitária α_t e a imagem original x_0. E a Equação 13 mostra o inverso, prevendo o ruído que será removido x_t em direção a x_0, em que $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

[9] No agendamento linear, o aumento da quantidade de ruído cresce de forma linear em relação aos passos t.

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t}\epsilon \tag{12}$$

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{\mathbf{x}_t - \sqrt{1 - \alpha_t} e_\theta^{(t)}(\mathbf{x}_t)}{\sqrt{\alpha_t}} \right) + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \epsilon_\theta^{(t)}(\mathbf{x}_t) + \sigma_t \epsilon_t \tag{13}$$

2.3.1 Modelos de difusão latente

Os modelos de difusão latente (do inglês, *Latent diffusion models*, LDM) ganharam notoriedade por poderem gerar imagens realistas guiadas por texto. Com a disponibilização do modelo de difusão latente pré-treinado *Stable diffusion* treinado com imagens de alta resolução presentes na base LAION (SCHUHMANN *et al.*, 2021) rapidamente ganhou destaque devido a ter seu código fonte aberto e alta qualidade na geração texto-para-imagem.

Foram propostos por Rombach *et al.* (2021), com um dos objetivos de diminuir o tempo de processamento e energia consumida no treinamento do modelo de difusão, esses fatores eram consideráveis nos modelos de difusão propostos anteriormente. A redução de dimensionalidade possibilita o uso de módulos de auto-atenção (do inglês, *self-attention*), já que esses aumentam a complexidade de forma quadrática com base nos dados de entrada.

O modelo proposto pelos autores foi divido em duas fases:

1. Na fase de compressão utilizaram um modelo *autoencoder* variacional para aprender o domínio perceptivo das imagens. Assim o modelo de difusão utilizado não tenta adicionar e remover ruído da imagem de entrada, em geral de alta dimensão, mas sim de um vetor latente (camada intermediária do *autoencoder* utilizado), reduzindo consideravelmente a complexidade computacional, pois diminui em oito vezes o tamanho de entrada. Este *autoencoder* variacional utiliza uma perda perceptiva e tem uma função objetivo adversária baseada em segmentos da imagem, assim, garantindo consistência em cada segmento.
2. Já na fase de aprendizagem generativa foi utilizada uma estrutura *U-Net* (RONNEBERGER; FISCHER; BROX, 2015) com mecanismos de atenção cruzada condicionados a diversas formas de entrada (texto e imagem) via um codificador específico do domínio.

2.3.1.1 Orientação livre de classificação (do inglês, *classifier free guidance*)

Uma técnica que melhora consideravelmente a qualidade dos exemplos produzidos por modelos de difusão é a orientação livre de classificação (do inglês, *classifier free guidance*) (HO; SALIMANS, 2022). Ela consiste em treinar o modelo de difusão condicional com uma componente de regularização parametrizável, a escala de orientação (do inglês, *guidance scale*) h_1 que controla quanto a componente ligada ao texto de entrada afeta a imagem final.

A entrada do modelo consiste em uma concatenação de duas componentes, uma condicional e uma incondicional. No caso de modelos de difusão condicionados a texto, são concatenadas duas representações do texto, a representação da sentença de entrada, e um vetor de mesmo tamanho com *tokens* que representam ausência de informação. Também são concatenados dois vetores latentes iguais que terão seu ruído removido. Logo, a saída terá duas componentes, uma condicionada ao texto de entrada $\hat{x}_{cond\_texto}$ e outra incondicional $\hat{x}_{incond}$, como pode ser visto na Equação 14. Valores de h_1 maiores forçam a imagem gerada a ser mais fiel ao texto de entrada, e valores menores dão mais liberdade na imagem final.

$$\hat{x} = \hat{x}_{incond} + h_1 * (\hat{x}_{cond\_texto} - \hat{x}_{incond}) \tag{14}$$

2.3.2 Edição de imagens utilizando modelos de difusão

Modelos de difusão demandam um alto poder computacional para serem treinados a ponto de poderem gerar imagens em alta resolução. Para mitigar essa complexidade, algumas metodologias podem ser utilizadas no caso de modelos de difusão condicionados a texto, possibilitando usar modelos pré-treinados para efetuar a edição de imagens. Duas dessas metodologias são:

- Métodos que utilizam a inversão DDIM para encontrar o vetor latente que melhor reconstrói a imagem original e executam a edição no passo para frente da remoção de ruído. Esses métodos não modificam os parâmetros dos modelos de difusão utilizados.
- Métodos que sobre-ajustam os pesos de modelos de difusão pré-treinados para se adequarem aos exemplos a serem editados.

A seguir, serão apresentados a técnica de inversão DDIM e alguns exemplos de modelos que utilizam essas duas abordagens.

2.3.2.1 Inversão DDIM

A técnica Inversão DDIM inicialmente apresentada por Song, Meng e Ermon (2020) é uma forma de edição de imagens reais, e consiste em encontrar a variável latente x_T que, ao percorrer o caminho de amostragem determinística, produzirá uma aproximação realista da imagem original x_0. Feito isso, passa a ser possível editar apenas uma parte da codificação que condiciona a geração de imagens, podendo substituir palavras mantendo características da imagem original. Outras formas de edição consistem em enviar uma máscara para editar apenas uma região de interesse.

2.3.2.2 EDICT

Inversão da difusão exata via transformações acopladas (do inglês, *Exact Diffusion Inversion via Coupled Transformations*, EDICT (WALLACE; GOKUL; NAIK, 2022)) permite edição de imagens reais usando um texto para guiar a edição, com resultados superiores aos obtidos, usando puramente a inversão DDIM, pois utilizam um par de vetores de ruído que são usados para um inverter o outro. Os autores mostraram seus resultados treinados com uma rede de difusão latente pré-treinada (ROMBACH *et al.*, 2021) nas bases de dados ImageNet 2012 e MS-COCO.

2.3.2.3 Inversão de texto vazio para edição de imagens reais usando modelos de difusão guiados

Em Inversão de texto vazio para edição de imagens reais usando modelos de difusão guiados, (do inglês, *Null-text Inversion for Editing Real Images using Guidade Diffusion Models*) (MOKADY *et al.*, 2022), tem como ponto de partida a trajetória de difusão $\{z_t^*\}_0^T$ de uma inversão DDIM. Os autores notaram que iniciar o processo de remoção de ruído na última camada latente z_T^* de uma inversão DDIM resulta em uma reconstrução insatisfatória quando a escala de orientação (do inglês, *guidance scale*) é alta, já que são adicionado erros de trajetória a cada passo quando é aplicada a técnica de orientação livre

de classificação. Para minimizar esses erros, essa trajetória foi otimizada apenas na parte incondicional, o que preservou a capacidade de edição por texto.

2.3.2.4 Tradução imagem a imagem com modelos de difusão condicionais (Pallete)

Em Pallete (SAHARIA *et al.*, 2021) os autores executaram a tradução imagem a imagem com modelos de difusão condicionais e avaliaram os resultados nas tarefas de re-colorização, completar partes faltantes de imagens e restauração de arquivos JPEG. Os autores compararam os resultados com GANs específicas treinadas para cada tarefa. Além disso, discutiram os efeitos da utilização de L_1 e L_2 nas funções objetivo dos modelos de difusão removedores de ruído, concluindo que L_1 produz imagens mais conservadoras (mais próximas da imagem original) e L_2 aumenta a diversidade do modelo.

Nas Seções 2.4.6 e 2.4.7, são detalhados outros dois modelos que executam a tradução imagem a imagem com modelos de difusão condicionais com abordagens diferentes, esses foram utilizados nos experimentos do Capítulo 7.

2.3.2.5 *DreamBooth*

O trabalho de Ruiz *et al.* (2022) teve como objetivo personalizar modelos de difusão texto-para-imagem possibilitando a utilização de elementos específicos (indivíduos, animais, itens, entre outros) na geração de novas imagens foto-realistas preservando as características originais que os distinguem. Para isso utilizou poucas imagens, entre três a cinco, e fez uma adaptação no modelo de difusão previamente treinado[10] para gerar um identificador único (*token*) e específico para aquele elemento. Isso possibilitou gerar novas imagens foto realistas em diferentes contextos de um elemento escolhido.

Dois pontos importantes da abordagem foram:

1. Visando evitar que a classe do elemento escolhido seja sobre-ajustada a ponto de só gerar novas imagens personalizadas desta, foi proposta uma função de perda de preservação de classe. Ela incentiva o modelo a produzir instâncias variadas da mesma classe.

[10] Foi utilizado o *Imagen* (SAHARIA *et al.*, 2022), mas o autor comenta ser possível aplicar em outros modelos de difusão

2. Foi ajustada a componente de super-resolução presente no modelo utilizado com pares de exemplos de baixa e alta-resolução das imagens de entrada. Garantindo que o modelo mantenha alta fidelidade de pequenos detalhes e assim garantindo a identidade dos mesmos.

Outro modelo que também sobre-ajusta os pesos da rede para possibilitar a personalização das imagens geradas é Inversão textual (do inglês, *Textual Invertion*) (GAL *et al.*, 2022).

2.3.3 Modelos auxiliares utilizados em conjunto com modelos de difusão

2.3.3.1 CLIP

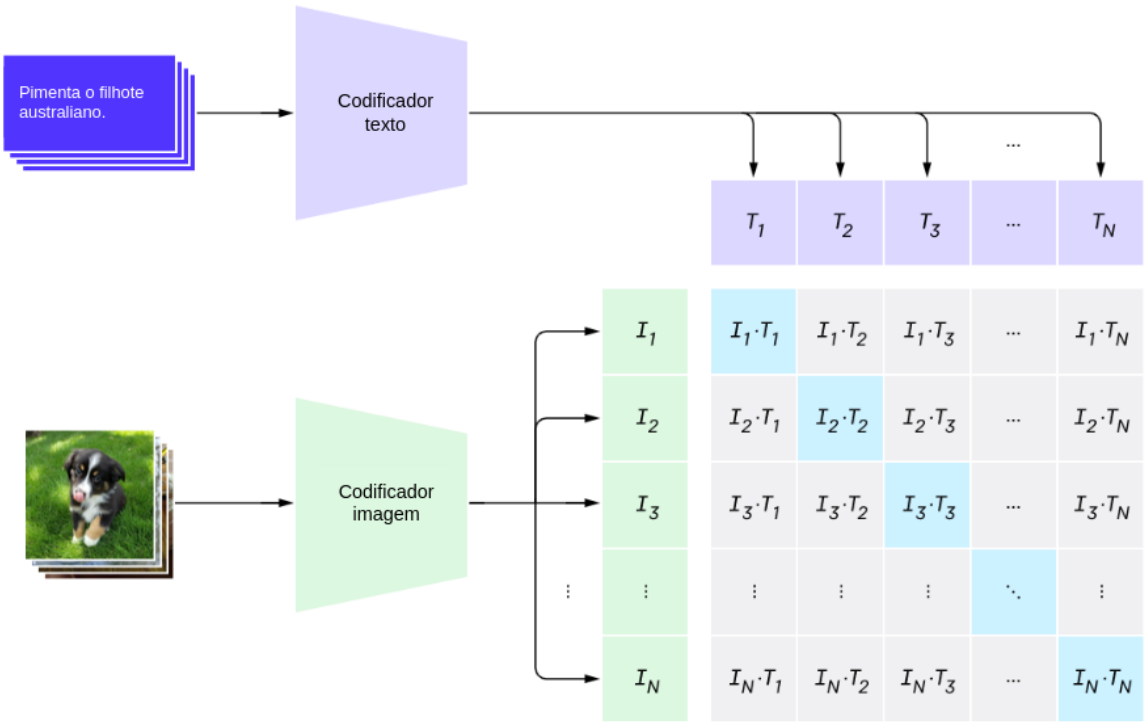
Pré-treinamento contrastante linguagem-imagem, (do inglês, *Contrastive Language-Image Pre-Training*, CLIP) (RADFORD *et al.*, 2021) é uma rede neural treinada em 400 milhões de pares de imagem na forma de imagem e suas legendas obtidas da internet. Essa rede foi proposta para ser usada sem a necessidade de ser novamente treinada para tarefas específicas, já que ela alinha em um mesmo espaço dimensional tanto a representação da imagem (após passar por um modelo codificador), I_n, quanto do texto (também após passar por um processo de geração de *tokens*, codificação e preenchimento para ter a mesma dimensão que a imagem), T_n, já que ambos têm a mesma dimensão e a matriz de resultado é o produto vetorial de ambos $I_n \cdot T_n$. Sua função objetivo tenta maximizar a similaridade dos pares correspondentes (diagonal azul na Figura 7) e minimizar a similaridade dos pares não relacionados (elementos em cinza).

2.4 Arquiteturas utilizadas nos experimentos

2.4.1 CAAE

Em *autoencoders* adversários condicionais (do inglês, *conditional adversarial autoencoders*, CAAE) (ZHANG; SONG; QI, 2017) os autores propuseram um método que aprende uma variação (*manifold*) de faces, que mapeará a progressão/regressão de idade de um determinado rosto por meio de uma estrutura que impõe discriminadores tanto no gerador, quanto no codificador.

Figura 7 – Processo de pré-treinamento contrastante. Sua função objetivo tenta maximizar a similaridade dos pares correspondentes (diagonal azul) e minimizar a similaridade dos pares não relacionados (elementos em cinza).

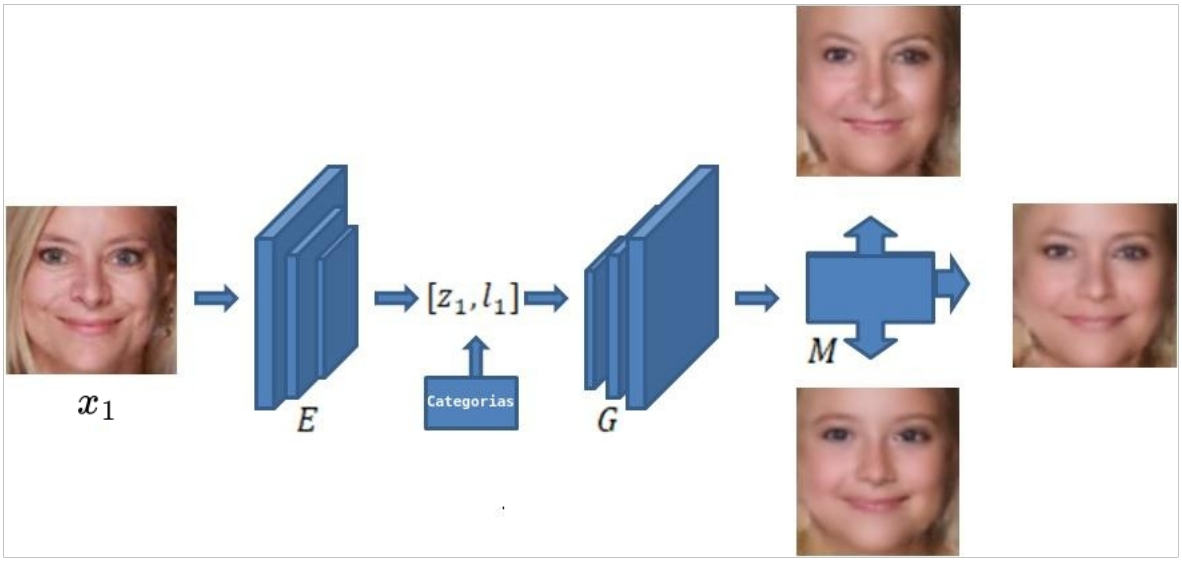


Fonte: Radford *et al.* (2021)

No trabalho, uma face x_1 foi mapeada para um vetor latente por meio de um codificador convolucional E, extrai as características pessoais z_1. O vetor foi concatenado com uma categoria de idade l_1, e um ponto foi gerado no espaço latente $[z_1, l_1]$. Observe que a identidade z e a idade l estão separadas no espaço latente, possibilitando a modificação da idade e preservando a personalidade. Através de outro gerador de mapeamento deconvolucional G, esses pontos foram mapeados para a variedade M - gerando uma série de imagens faciais, que apresentarão envelhecimento / rejuvenescimento da idade de x_1. O vetor latente preserva características personalizadas do rosto (ou seja, personalidade) e a condição imposta na idade controla a progressão ou regressão. O processo é ilustrado na Figura 8.

A principal diferença entre o AAE e o CAAE, é que o último, impõe discriminadores tanto nos codificadores quanto nos geradores. O discriminador do codificador garante uma transição suave no espaço latente, e o discriminador no gerador auxilia na geração de fotos realistas. Portanto, CAAE irá criar imagens de qualidade mais alta do que as geradas pelo AAE (ZHANG; SONG; QI, 2017).

Figura 8 – A face de entrada x_1 é codificada z_1 por um codificador E, representando a identidade. O vetor é concatenado com uma categoria de idade l_1, deste modo o vetor latente $[z_1, l_1]$ é construído. Por meio de outro gerador deconvolucional de mapeamento G, esses pontos são mapeados para a variedade (do inglês, *manifold*) M, representando a envelhecimento / rejuvenescimento de idade x_1. Na variedade, a imagem inferior ilustra o processo de regressão. A imagem à direita representa o mapeamento para a mesma idade, e a imagem superior refere-se ao processo de envelhecimento.



Fonte: Kemmer, Simões e Lima (2022)

2.4.2 IPCGAN

No modelo IPCGAN, os autores utilizaram uma CGAN para gerar a imagem envelhecida, e para garantir a identidade do rosto, uma perda perceptual $L_{identidade}$ foi adicionada a função objetivo da rede. Além disso, para forçar que as imagens geradas alcancem a idade desejada, foi utilizada uma rede neural estimadora de idade, caso o grupo estimado seja diferente do almejado L_{idade}, conforme mostrado na Equação 15. O termo L_G é a função de perda da CGAN definida na Equação 8.

$$G_{perda} = \lambda_1 L_G + \lambda_2 L_{identidade} + \lambda_3 L_{idade}, \tag{15}$$

Em que λ_1, λ_2 e λ_3 são hiperparâmetros que ponderam entre a manutenção da identidade e o efeito de envelhecimento.

2.4.3 RCRIIT

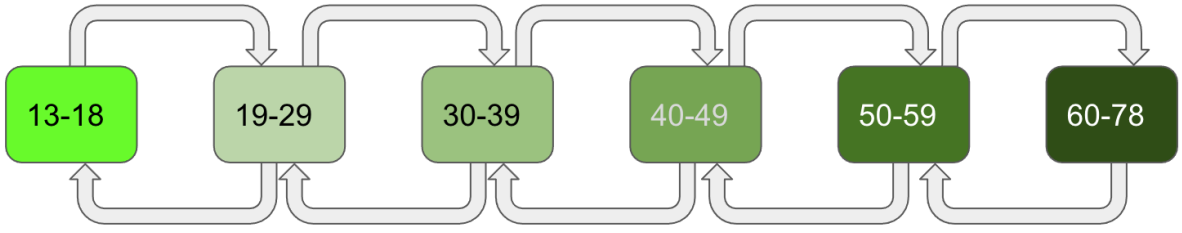
Heljakka, Solin e Kannala (2018) usaram uma translação encadeada entre faixas de idade, por exemplo, uma imagem entre 15 e 25 anos foi traduzida para uma imagem gerada com idade estimada entre 25 e 35 anos, até atingir a idade desejada. Para tanto, foram utilizados modelos par-a-par da *CycleGAN*, um para cada par de faixas etárias.

Em sua metodologia, os autores abordaram a modelagem e simulação de mudanças progressivas ao longo do tempo, como o envelhecimento do rosto humano. O artigo generalizou o mapeamento imagem a imagem para uma configuração sequencial. Ao tratar as fases de idade como uma sequência de domínios de imagem, os autores propuseram uma abordagem de adaptação sequencial do domínio adversário capaz de produzir transformações graduais para o envelhecimento humano. Como pode ser visto na Figura 9.

Portanto, no caso de envelhecimento, o primeiro par de GANs irá envelhecer para o próximo intervalo, e o resultado será utilizado como imagem de entrada do próximo par de GANs, e assim sucessivamente. No caso de rejuvenescimento, é feita a mesma abordagem mas na direção oposta.

Os autores usaram o conjunto de dados CACD e dividiram os dados nos grupos 2–18, 19–29, 30–39, 40–49, 50–59 e 60–78 (seis domínios, com cinco transformações diretas). Eles removeram amostras suficientes para que a idade média nos conjuntos ficasse em dez anos: 15, 25, 35, 45, 55 e 65.

Figura 9 – Ilustração das diversas *CycleGANs* treinadas na arquitetura, uma para cada par de grupos de idade. Portanto, no caso de envelhecimento, o primeiro par de GANs irá envelhecer para o próximo intervalo, e o resultado será utilizado como imagem de entrada do próximo par de GANs, e assim sucessivamente. No caso de rejuvenescimento, é feita a mesma abordagem mas na direção oposta.



Fonte: Kemmer, Simões e Lima (2022)

2.4.4 HRFAE

Em Edição de idades faciais em alta resolução (do inglês, *High Resolution Face Age Editing*, HRFAE) (YAO *et al.*, 2020), os autores utilizaram uma arquitetura codificador-decodificador para executar a edição de idade de fotos em alta resolução (1024x1024). O gerador G é composto por duas partes: um codificador E e um decodificador D. O modelo recebe a imagem de entrada x_0, essa passa pelo codificador gerando duas cópias de um vetor latente $E(x_0)$. O estimador de idade pré-treinado DEX (ROTHE; TIMOFTE; GOOL, 2015) é utilizado para determinar a idade da imagem de entrada α_0. Essa idade irá passar por um módulo de codificação binária com função de ativação *sigmoid*. Seu decodificador D tem duas tarefas: receber o vetor latente $E(z)$ e produzir uma imagem o mais similar possível à imagem de entrada $G(x_0, \alpha_0)$, e também fazer com que a imagem envelhecida $G(x_0, \alpha_1)$ seja realista e próxima da idade que se deseja atingir. A Figura 10 apresenta essa arquitetura descrita.

Para atingir esse objetivo, sua função de custo tem três componentes, conforme mostrado na Equação 16:

1. Perda adversária L_{GAN} utiliza a *PatchGAN* (ISOLA *et al.*, 2017) com a função objetivo da LSGAN (Mao *et al.*, 2017).
2. Perda de classificação de idade L_{class}, nela é utilizada o modelo pré-treinado DEX para estimar a idade obtida com uma função perda de entropia cruzada categórica.
3. Perda de reconstrução L_{recon}, monitora a capacidade do modelo conseguir reconstruir a imagem original na idade inicial. $L_{recon} = \|G(x_0, \alpha_0) - x_0)\|_1$

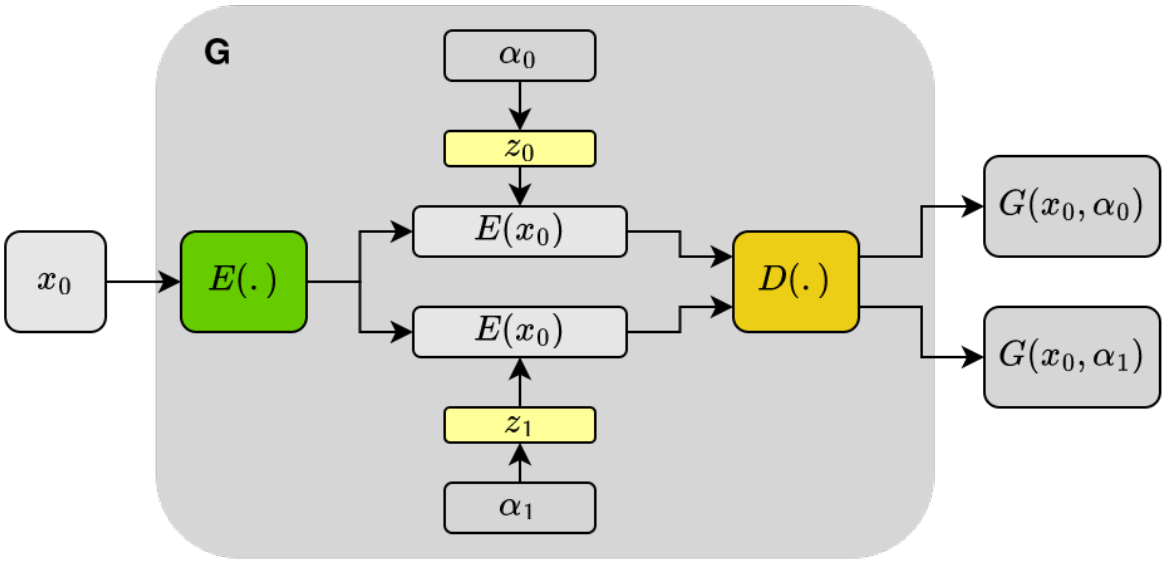
$$\lambda = \lambda_{recon} L_{recon} + \lambda_{class} L_{class} + L_{GAN}, \tag{16}$$

Em que λ_{recon} e λ_{class} são hiper-parâmetros que ponderam entre a manutenção da identidade (reconstrução da imagem original na idade inicial) e o efeito de envelhecimento (classificação na idade correta).

No treinamento executado pelos autores foi utilizada a base de imagens de alta definição FFHQ (Seção 4.9), como ela apresenta uma baixa quantidade de imagens de pessoas de idade mais avançada, utilizaram a rede StyleGAN para gerar 300.000 imagens sintéticas e assim obtiveram uma base de imagens balanceada no critério das idades presentes nela. Os autores somente utilizaram as imagens sintéticas nas faixas de idade em

que não havia imagens reais suficientes. No final, obtiveram 47.990 imagens com idades entre 20 e 69 anos.

Figura 10 – Arquitetura do modelo HRFAE em que x_0 é a imagem de entrada que ao passar pelo gerador G será envelhecida. O codificador E gera dois vetores latentes iguais que receberam a codificação z_0 referente à idade estimada da imagem original α_0 e também a codificação z_1 referente à idade esperada α_1. Esses vetores junto com as codificações passarão pelo decodificador D, e o resultado será a imagem reconstruída $G(x_0, \alpha_0)$ e a envelhecida $G(x_0, \alpha_1)$.



Fonte: Bruno Abreu Kemmer, 2023 (baseado em Yao *et al.* (2020))

2.4.5 SAM

Em Manipulação de idade baseada em estilo (do inglês *Style-Based Age Manipulation*, SAM) (ALALUF; PATASHNIK; COHEN-OR, 2021), os autores desenvolveram uma arquitetura que possibilita o envelhecimento ou rejuvenescimento facial tendo como entrada uma imagem real x e uma idade desejada α_t. Para conseguir isso, executaram uma translação imagem a imagem em que um codificador (pSp) previamente treinado (RICHARDSON *et al.*, 2021) extrai vetores latentes que representam x no espaço w^* de uma GAN incondicional (*StyleGAN*). Essa série de vetores de estilo possibilitou a reconstrução da imagem original ao passar pelo gerador da *StyleGAN*. Um segundo codificador, E_{idade}, foi treinado para capturar a diferença (resíduo) entre o vetor latente obtido pelo codificador (pSp) e a imagem envelhecida, as saídas dos codificadores são somadas e se tornam o vetor latente de entrada que a *StyleGAN* utiliza. Essa arquitetura pode ser vista na Figura 11.

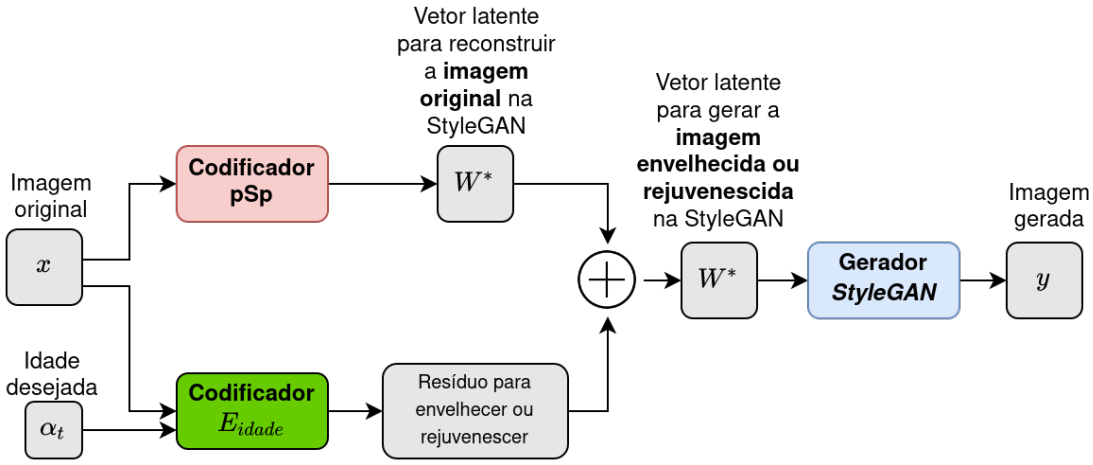
Para guiar o treinamento desse segundo codificador em obter a idade desejada e também manter as características individuais da imagem original, foram utilizadas:

- Uma rede pré-treinada DEX (ROTHE; TIMOFTE; GOOL, 2015) para regressão de idade, que utiliza a arquitetura VGG (SIMONYAN; ZISSERMAN, 2015).
- Uma rede pré-treinada para reconhecimento facial ArcFace (DENG *et al.*, 2019), auxiliando na manutenção da identidade.

Tais redes pré-treinadas não foram alteradas durante o treinamento do codificador E_{idade}, mantendo seus parâmetros fixos.

Adicionalmente, por se tratar de uma translação imagem a imagem, o modelo utiliza uma perda cíclica para tentar reconstruir a imagem original após uma passagem de consistência de ciclo, como mostra a Figura 12 e também detalhado na descrição da *CycleGAN* na Seção 2.2.3.4.

Figura 11 – Na arquitetura utilizada do modelo SAM, o codificador pSp gera o vetor latente que possibilita reconstruir a imagem original utilizando o gerador da *StyleGAN*. Um segundo codificador é treinado para determinar o resíduo que precisa ser adicionado a esse vetor latente para envelhecer ou rejuvenescer a imagem. A soma destes vetores passa a ser a entrada do gerador da *StyleGAN* que possibilitará obter a imagem na idade desejada.

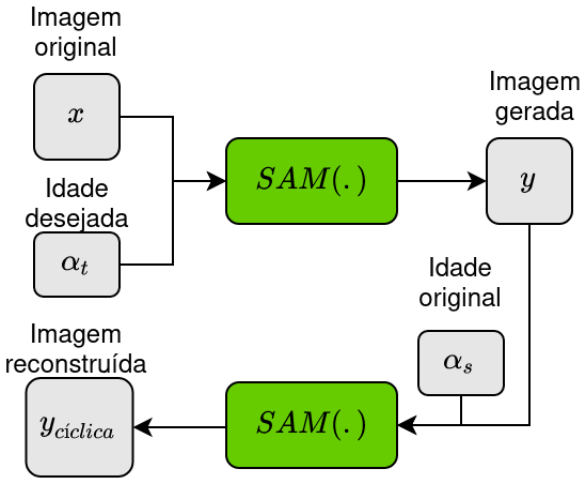


Fonte: Bruno Abreu Kemmer, 2023 (baseado em Alaluf, Patashnik e Cohen-Or (2021))

As funções de custo utilizadas foram:

- **Similaridade pixel a pixel:** $\mathcal{L}_2(x_{idade}) = \|x - SAM(x_{idade})\|_2$
- **Perda de similaridade perceptiva**: $\mathcal{L}_{LPIPS}(x_{idade}) = \|F(x) - F(SAM(x_{idade}))\|_2$ (Detalhada na Seção 5.5).

Figura 12 – Passagem de consistência de ciclo presente na arquitetura de SAM. Em que o modelo SAM recebe a imagem real x e a idade desejada α_t, gerando uma imagem envelhecida y, que será a entrada de outra passagem pelo modelo SAM junto com a idade original α_s, com o objetivo de reconstruir a imagem original.



Fonte: Bruno Abreu Kemmer, 2023 (baseado em Alaluf, Patashnik e Cohen-Or (2021))

- **Perda de regularização**: Essa regularização faz com que os vetores de estilo estejam perto da média dos vetores latentes. Os autores identificaram que sua utilização melhora a qualidade da imagem gerada por remover artefatos indesejados nas imagens produzidas.
- **Perda de identidade**: Diferença na similaridades de co-senos entre a imagem de saída e de entrada, ponderando pela quantidade de anos entre as imagens. Caso a diferença seja de muitos anos é esperado que exista uma perda de identidade, $\mathcal{L}_{ID}$.
- **Perda de envelhecimento**: Para verificar a qualidade do envelhecimento/rejuvenescimento na imagem gerada foi utilizada uma rede pré-treinada DEX, $\mathcal{L}_{idade} = \|\alpha_t - DEX(SAM(x_{idade}))\|_2$.

2.4.6 *Pix2pix-zero*

Em Translação imagem a imagem *Zero-shot* (do inglês, *Zero-shot Image-to-Image Translation*) (PARMAR *et al.*, 2023) os autores propuseram um método de edição de imagens reais que preservam as características das imagens originais. Seguindo os seguintes passos:

1. Executaram uma inversão DDIM para obter o vetor latente original que melhor representa essa imagem real no modelo utilizado[11].
2. Descobriram uma direção de edição. Utilizaram o modelo gerador de texto GPT-3 (BROWN *et al.*, 2020) para gerar sentenças com o termo de origem (exemplo: cachorro), e um termo de destino (exemplo: gato). Essas frases passam por um modelo CLIP (Seção 2.3.3.1) para obter as representações nesse domínio e a subtração da média dessas representações irá, teoricamente, ser a direção de edição das imagens. A Figura 14 exemplifica como isso ocorre.
3. Obtiveram uma legenda (um texto aderente a imagem) para que quando ocorrer a edição, ela aconteça na palavra que melhor represente o termo alterado. Como os módulos de atenção cruzada acabam gerando máscaras relacionando as palavras com a imagem, quando a edição usa essa informação ela mantém o que não está sendo alterado. No trabalho os autores utilizaram o modelo BLIP (LI *et al.*, 2022) para gerar o texto.
4. Executaram a edição através dos módulos de atenção cruzada. Para isso, Reconstruíram a imagem sem aplicar nenhuma edição, apenas usando o texto de entrada para obter os módulos de atenção cruzada para cada passo t. Feito isso, adicionaram a direção de edição e calcularam o gradiente da perda em relação a entrada x_t. Isso faz com que a edição se concentre na região representada por aquela palavra.

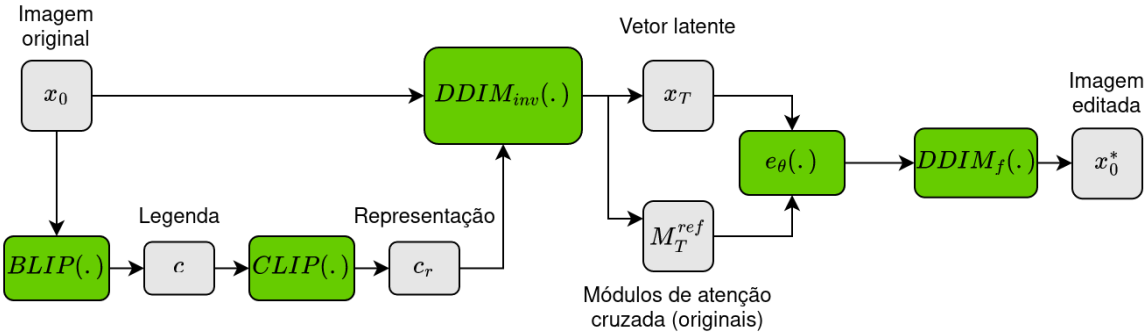
A Figura 13 ilustra os passos descritos acima.

2.4.7 *Instruct-pix2pix*

Em *Instruct-pix2pix* (BROOKS; HOLYNSKI; EFROS, 2022) foi apresentado um método para treinar um modelo que pode seguir instruções de edição humanas em imagens. O método recebe uma imagem de entrada e um texto com a instrução e executa a edição no passo para frente (do inglês, *forward pass*). Isso é possível pois os autores utilizaram o modelo GPT-3 (BROWN *et al.*, 2020) para gerar instruções de edição e legendas das imagens originais e editadas. Feito isso, os autores utilizaram a rede pré-treinada *Stable diffusion* (ROMBACH *et al.*, 2021) para gerar os pares de imagens referentes as legendas criadas, produzindo uma base de mais de 450.000 exemplos. Com esses exemplos, foi

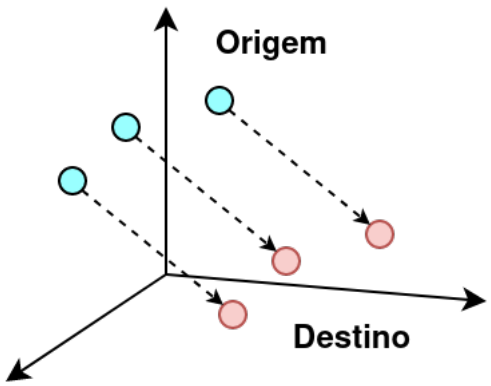
[11] Os autores utilizaram *Stable diffusion* 1.4.

Figura 13 – Etapas em uma edição *pix2pix-zero*. Nela a imagem original a ser editada x_0 é invertida x_T e automaticamente é gerada uma legenda c que será usada junto com os módulos de atenção cruzada para definir a região da imagem que será alterada. No momento de edição e_θ serão otimizados os módulos e após a remoção de ruído, usando a técnica de amostragem DDIM, será obtida a imagem editada x_0^*.



Fonte: Bruno Abreu Kemmer, 2023 (baseado em Parmar *et al.* (2023))

Figura 14 – Ilustração do processo de encontrar as direções de edição. Cada ponto representa o vetor no espaço latente do modelo CLIP (Com 768 dimensões por padrão) resultante após passar uma sentença com os termos desejados na direção que queremos encontrar. Sendo a direção a linha tracejada que leva os vetores latentes para outro contexto.



Fonte: Bruno Abreu Kemmer, 2023 (baseado em Parmar *et al.* (2023))

treinado um modelo de difusão para gerar imagens editadas dada uma imagem de entrada e uma instrução de edição.

Foi destacado pelos autores que em modelos de difusão texto-para-imagem (como *Stable diffusion*) podem gerar imagens drasticamente diferentes para textos levemente modificados, e para mitigar esse problema os autores utilizaram a técnica apresentada no trabalho *prompt-to-prompt* (HERTZ *et al.*, 2022) em que são calculados os pesos nos módulos de atenção cruzada relacionando as palavras com regiões da imagem, assim a edição fica restrita às regiões relacionadas às palavras que estão sendo editadas. Além disso, esse modelo tem um parâmetro ρ [12] que possibilita controlar a similaridade entre as duas imagens. Para fazer isso de forma automática foram amostrados 100 exemplos $\rho \sim \mathcal{U}(0,1;0,9)$ e filtrados com base em uma métrica de distância no espaço de representação de CLIP.

Outro ponto importante foi que os autores utilizaram dois parâmetros na orientação livre de classificação (do inglês, *classifier free guidance*): um que controla o quanto da imagem corresponde à imagem de entrada c_{imagem}, e que outro interfere em quanto à instrução será seguida $c_{instrucão}$. Como explicado na Seção 2.3.1.1.

[12] Relação de passos de difusão com pesos de atenção cruzada.

3 Revisão bibliográfica

Como detalhado na Seção 1.1, não foram encontradas revisões que detalhassem de forma sistemática como os modelos generativos são utilizados para o envelhecimento facial. Este trabalho visa preencher essa lacuna, ao mostrar as principais técnicas utilizadas, bases de dados e métodos de análise dos resultados, respondendo às seguintes perguntas:

- **Q1** - Quais técnicas são utilizadas para geração das faces envelhecidas?
- **Q2** - Quais bases de dados estão sendo utilizadas nesse procedimento?
- **Q3** - Como a performance do modelo está sendo medida?
- **Q4** - Como essa técnica está sendo integrada na tarefa biométrica?
- **Q5** - Está havendo uma melhora ao utilizar o modelo generativo na tarefa biométrica?

Para responder a essas questões foi realizada uma revisão bibliográfica aplicando a *string* de busca no motor de busca da plataforma Scopus, obtendo 122 estudos primários no dia 17 de janeiro de 2021 e após serem aplicados critérios de exclusão foram selecionados 54 trabalhos. O protocolo utilizado na revisão pode ser visto no Apêndice A e sua condução no Apêndice B.

***Q1** - Quais técnicas são utilizadas para geração das faces envelhecidas?*

Para responder a essa pergunta, foi mapeado[1] quais tipos de GANs ou *Autoencoders* existentes na literatura foram utilizadas (ou baseadas) na tarefa de envelhecimento facial. Na maioria dos casos, os autores citam qual modelo foi utilizado em sua seção de metodologia. No entanto, nos casos em que isso não ocorreu, e o modelo que foi utilizado estava claro, esse foi o utilizado. Também houve casos em que foram utilizadas múltiplas técnicas, nesses casos, ambas foram adicionadas. O mapeamento pode ser visto na Tabela 1, e a coluna ID(s) são as identificações dos trabalhos selecionados do Quadro 1 presente no Apêndice B.

Alguns trabalhos também utilizaram técnicas adicionais junto com os modelos generativos, a Tabela 2 apresenta os mais importantes.

[1] Os campos *Autoencoders* adversários incluem as técnicas condicionais (CAAE), e StyleGAN contém StyleGAN (KARRAS; LAINE; AILA, 2019) e StyleGAN2 (KARRAS *et al.*, 2020).

Tabela 1 – Técnicas utilizadas nos trabalhos selecionados.

Técnica	ID(s)	Total
CGAN	2, 7, 10, 11, 15, 17, 22, 25, 29, 30, 35, 38, 47, 51, 54	15
CycleGAN	5, 8, 9, 14, 15, 16, 23, 26, 27, 30, 32, 40, 44, 50	14
Autoencoders adversários	4, 11, 12, 13, 18, 28, 43, 48, 49, 53	10
DCGAN	24, 33, 42, 45, 54	5
Autoencoder Variacional	18, 29, 31, 37	4
StyleGAN	11, 19, 20, 21	4
StarGAN	14, 15, 37	3
RNN	6, 34, 52	3
PROGAN	20, 28	2
LSGAN	17, 38	2

Fonte: Bruno Abreu Kemmer, 2023.

Tabela 2 – Técnicas adicionais utilizadas nos trabalhos selecionados.

Técnica	ID(s)	Total
Transferência de aprendizagem	2, 4, 7, 12, 23, 29, 33, 35, 36, 39, 40, 41, 47, 49	14
PatchGAN	6, 8, 14, 15, 24, 30, 40, 44	8
WGAN	7, 8, 10, 15, 24, 36, 42, 43	8
Autoencoder adicional	12, 19, 42, 51	4
PCA	13, 34	2
Super-resolução	5, 36	2
Mapas de atenção	6, 10	2
Transformada *Wavelet*	25	1
Transformer	50	

Fonte: Bruno Abreu Kemmer, 2023.

De modo geral, as arquiteturas utilizadas neste problema específico refletem os avanços obtidos em GANs, como um todo. Pode ser notado que, a partir da data de publicação de Gulrajani *et al.* (2017), muitos autores passaram a adotar a função de custo baseada na distância de Wasserstein (WGAN), já que essa proporciona um treinamento mais estável, minimizando o "modo colapso"(do inglês, *mode colapse*), nesse modo todas as imagens geradas são a mesma. Outro exemplo é o uso de *PatchGANs* nos discriminadores para garantir a consistência local das imagens. Podemos ver o uso dessas duas técnicas na Tabela 2.

Posteriormente, foram adicionadas novas técnicas como o uso de mapas de atenção (do inglês, *attention map*) (SHI *et al.*, 2020) e melhorias via pós-processamento ao utilizar a técnica de super-resolução (SHARMA; SHARMA; JINDAL, 2020) que utiliza uma ESRGAN (WANG *et al.*, 2018a). Outra abordagem é dividir a tarefa do gerador em partes

especializadas junto com uma global (YANG *et al.*, 2018; LI *et al.*, 2019; YANG; LV, 2020).

Dos trabalhos que se destacaram (adicionais aos já apresentados no Capítulo 2), Liu *et al.* (2017) utilizou uma DCGAN e uma CGAN para conseguir capturar os padrões de transição ao longo do envelhecimento.

Wang *et al.* (2018b) emprega uma LSGAN (Mao *et al.*, 2017) para gerar imagens envelhecidas em um intervalo de idades alvo. Para isso, adicionaram uma rede que garante a preservação das características individuais e um classificador de idade. A primeira previamente treinada Alex-Net (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) e segunda, estima a qual grupo de idade a imagem gerada pertence, penalizando caso esteja distante da desejada. Por preservar a identidade original, esse método se chamou IPCGAN (do inglês, *identity-preserved conditional generative network*.)

Li *et al.* (2018) executou a abordagem de dividir para conquistar, separando o gerador de faces envelhecidas em geradores menores especializados em partes da face: testa, olhos, boca mais nariz e um da face total. Outra modificação proposta por Li *et al.* (2019), mapeou o envelhecimento no domínio da frequência, empregando transformadas de *wavelets*, com o objetivo de capturar os efeitos da passagem do tempo em rugas e ao redor dos olhos. Novamente, separando o gerador em uma rede global e três especializadas em testa, olhos e boca.

Shi *et al.* (2020) utilizou um mapa de atenção (VASWANI *et al.*, 2017), para determinar quais regiões foram mais afetadas pelo envelhecimento. Nessas, a GAN deve efetuar mudanças mais impactantes, mostrando bons resultados em grandes diferenças de idade. Abordagem similar adotada por Zhu *et al.* (2020).

Até o momento, o envelhecimento facial era executado via uma idade específica, ou um intervalo de idades alvo. Em Zhou *et al.* (2020) isso mudou, pois a idade desejada passa a ser determinada por outra imagem de referência. A arquitetura adotada teve três elementos: uma agente de idade (rede neural pré-treinada que captura a codificação referente à idade das imagens), um agente de identidade e GAN (baseado na arquitetura de CAAE (ZHANG; SONG; QI, 2017)).

Q2 *- Quais bases de dados estão sendo utilizadas nesse procedimento?*

Na Tabela 3, estão listadas as bases de dados utilizadas por pelo menos 2 trabalhos selecionados. A base mais utilizada é a CACD, pois contém muitas imagens (163.446 imagens de 2.000 celebridades) possibilitando um melhor treinamento dos modelos generativos. FG-NET tem sua presença em segundo lugar, já que foi a primeira base utilizada na tarefa, e é utilizada como *benchmark* em muitos estudos. MORPH também contém uma quantidade de imagens considerável (55.134 imagens faciais em ambiente controlado de 13.617 indivíduos), todavia é comercializada. Um detalhamento de cada uma delas pode ser encontrado no capítulo 4.

Tabela 3 – Bases de dados utilizadas nos trabalhos selecionados.

Nome da base de dados	ID(s)	Total
CACD	2, 3, 4, 5, 6, 7, 9, 12, 13, 17, 18, 19, 24, 25, 32, 33, 34, 35, 38, 39, 41, 42, 45, 46, 50, 51, 52, 53, 54	29
FG-NET	2, 3, 5, 6, 12, 17, 18, 22, 23, 24, 25, 28, 31, 38, 41, 45, 46, 48, 49, 51, 52, 53, 54	23
MORPH	2, 3, 6, 9, 10, 12, 13, 14, 15, 18, 22, 25, 33, 34, 38, 41, 45, 46, 51, 52, 53, 54	22
UTKFace	1, 4, 5, 11, 13, 16, 18, 22, 23, 28, 42, 43, 44, 49, 51	15
IMDB-Wiki	5, 8, 27, 33, 37, 40, 42, 45, 47, 48, 49, 50, 51, 54	14
Baseados em LFW	3, 9, 16, 34, 35, 36, 48, 52, 54	9
FFHQ	1, 19, 21	3
AgeDB	18, 27	2

Fonte: Bruno Abreu Kemmer, 2023.

Q3 *- Como a performance do modelo está sendo medida?*

Avaliação de resultados, quando se trata de métodos generativos, não tem uma solução consolidada. Isso pode ser visto pela quantidade de métodos utilizados na Tabela 4. A análise qualitativa das imagens geradas versus as imagens originais foi feita em quase todos os trabalhos. Alguns destes, foram utilizadas as imagens presentes em

trabalhos correlatos para comparar o desempenho entre o método proposto e os publicados anteriormente.

Entrevistar pessoas externas é uma abordagem largamente utilizada, validando se a imagem tem aspecto real ou perguntando qual seria a idade da imagem gerada. Estudo de ablação (do inglês, *ablation study*), consiste em remover ou trocar partes do método utilizado e ver como essas mudanças impactam nos resultados. Alguns trabalhos utilizaram essa abordagem para medir o impacto que as partes do modelo gerador causavam nas faces envelhecidas.

Adicionalmente houveram trabalhos que executam métodos quantitativos de verificação e reconhecimento facial, na maior parte dos casos foi utilizada a API da plataforma Face++ (INC, 2017), que provê tanto estimação da idade dos indivíduos, quanto a similaridade entre duas faces. Algumas publicações estimaram a idade das imagens geradas, utilizando modelos próprios e comparando com as idades reais presentes nos metadados das bases de dados.

Tabela 4 – Métodos utilizados para mensurar os resultados obtidos nos trabalhos selecionados.

Método	Quantidade de trabalhos
Análise qualitativa	42
Verificação/Reconhecimento facial	30
Estimação de idade	20
Entrevistas	12
Estudo de ablação	8
Agrupamento	6
Outra métrica quantitativa	9

Fonte: Bruno Abreu Kemmer, 2023.

Q4 *- Como essa técnica está sendo integrada na tarefa biométrica?*

Como pode ser observado pela Tabela 4, alguns trabalhos utilizaram métodos quantitativos como verificação e reconhecimento facial para analisar seus experimentos. Esses, são treinados com bases de dados externas (diferentes das utilizadas para o treino dos modelos generativos), e reportam os valores alcançados, em geral, mostrando a diferença ao utilizar o modelo gerador proposto e sem ele.

Alguns trabalhos têm como objetivo específico executar o reconhecimento ou verificação facial. Desses, Bayramli *et al.* (2019) adicionou ao modelo gerador uma rede

neural classificadora para executar o reconhecimento facial que utiliza uma arquitetura ArcFace, baseada em ResNet.

Com o objetivo de executar a verificação facial invariante à idade, o trabalho de Huang, Chen e Hu (2019), separou a tarefa em partes: uma rede para reconhecer o indivíduo, e uma rede discriminadora de idade. Ambas compartilhando um mesmo extrator de características, (uma rede previamente treinada baseada na arquitetura ResNet) e a função de custo é a perda *triplet* (do inglês, *triplet loss*).

Outra abordagem interessante foi adotada por Orrù, Marcialis e Roli (2020), empregando uma técnica de agrupamento *k-means*. São geradas diversas faces do mesmo indivíduo (algumas destas, envelhecidas) e são calculados os centroides. É esperado que faces de um mesmo indivíduo se comportem como um agrupamento único, e que os agrupamentos estejam distantes entre si. Assim, a verificação facial passa a ser buscar o limiar de distância entre a imagem a ser verificada e os centróides presentes na base de dados.

Q5 *- Está havendo uma melhora ao utilizar o modelo generativo na tarefa biométrica?*

Os trabalhos apresentaram ganhos ao utilizar métodos de envelhecimento nas tarefas biométricas, porém, os modelos de classificação utilizados já apresentavam uma acurácia próxima à humana (mais de 95% em alguns casos), logo, os ganhos ao adicionar um módulo novo não foram relevantes.

4 Bases de dados

As bases de dados usadas para envelhecimento facial requerem fotos de rostos humanos em várias idades. A maioria dos métodos requer múltiplas imagens dos mesmos indivíduos. Abaixo estão as bases de dados mais usadas para esta tarefa.

4.1 FG-NET

A base de dados FG-NET (LANITIS; TAYLOR; COOTES, 2002) está disponível publicamente e é amplamente utilizada em trabalhos relacionados ao envelhecimento facial. Porém, em comparação com outras bases de dados utilizadas nessa tarefa, possui poucas imagens, aspecto que pode limitar o desempenho dos modelos nela treinados; assim, seu uso é mais direcionado para avaliação de métodos de envelhecimento facial. Contém 1.002 imagens de 82 pessoas com idade entre 0 e 69 anos. A Tabela 5 mostra a distribuição das fotos entre as idades e a predominância de menores de 30 anos.

4.2 UTKFace

A base de dados UTKFace (ZHANG; SONG; QI, 2017) contém mais de 20.000 imagens, e o número de indivíduos não é especificado, pois não são identificados. As imagens não estão em um ambiente controlado e contêm metadados de gênero, etnia e idade dos rostos. Além disso, também fornece 68 pontos de referência nas faces. A Tabela 5 mostra a distribuição das fotos entre as idades.

Tabela 5 – Estratificação por idade das bases de dados FG-NET, UTKFace, CACD e *FFHQ Aging*.

Base de dados	**0-10**	**11-20**	**21-30**	**31-40**	**41-50**	**51-60**	**60+**
FG-NET	371	359	143	69	39	14	7
	37%	36%	14%	7%	4%	1%	1%
UTKFace	3.260	1.833	7.803	4.343	2.103	2.225	2.444
	14%	8%	32%	18%	9%	9%	10%
CACD	0	9.647	40.750	43.176	39.819	28.491	1.563
	0%	5,9%	24,9%	26,4%	24,4%	17,4%	1%
FFHQ Aging	9.873	6.257	19.511	15.143	9.678	7.726	1.812
	14,1%	8,9%	27,9%	21,6%	13,8%	11%	2,6%

Fonte: Bruno Abreu Kemmer, 2023.

4.3 CACD

Essa base de dados (WU; TURAGA; CHELLAPPA, 2012) foi obtida via buscas pela internet (feitas entre 2004-2013), e é composta por fotos de celebridades. Ela contém 163.446 imagens de 2.000 celebridades, com idades entre 16 e 62 anos. A idade presente nos metadados não é exata, já que foi obtida via subtração entre a data da foto e a data de nascimento (que está pública, como os indivíduos são famosos). Porém, a base pode ser usada, tanto para envelhecimento, quanto para verificação e reconhecimento facial invariante à idade (SHU *et al.*, 2016b).

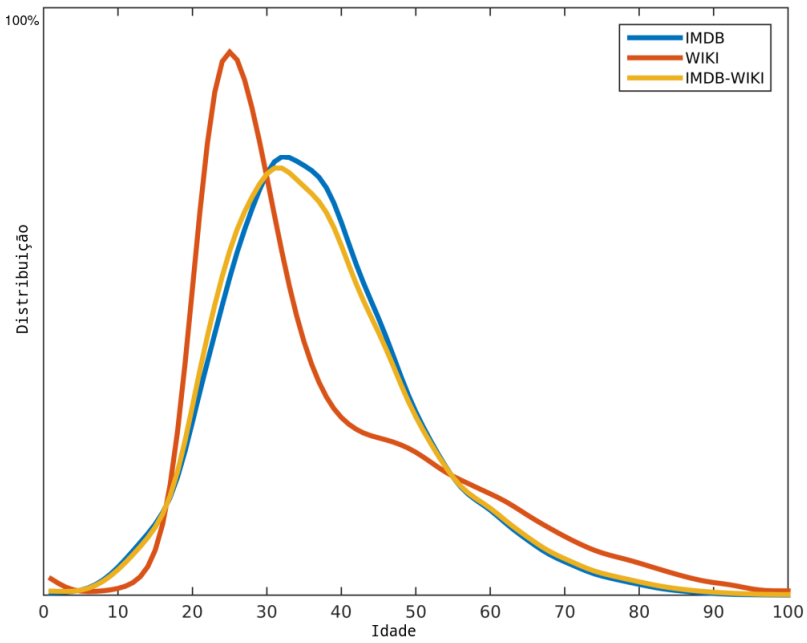
4.4 MORPH

A base de dados MORPH (RICANEK; TESAFAYE, 2006), também é amplamente utilizada nessa tarefa. Essa base está dividida em dois grupos: álbum 1 e álbum 2. O primeiro contém 1.690 imagens de 515 indivíduos, tanto homens quanto mulheres. O álbum 2 contém 94.000 imagens de 24.000 indivíduos, com idades entre 16 a 77 anos, em média 4 fotos por pessoa. Os metadados contém gênero e raça. A versão acadêmica é um subconjunto do álbum 2, contendo 55.000 imagens de 13.000 indivíduos, e a versão comercial tem 202.000 imagens. Ambas são comercializadas.

4.5 IMDB-Wiki

A base de dados IMDB-Wiki (ROTHE; TIMOFTE; GOOL, 2018) também foi extraída da internet, de dois sites: IMDb (portal com informações sobre filmes e artistas) e Wikipédia (enciclopédia digital de licença livre e colaborativa). Do primeiro foram obtidas 460.723 imagens, e do segundo 62.328 imagens. A base contém idade e gênero dos indivíduos, e a distribuição das idades contidas está descrita na Figura 15. Um subconjunto desta base é Celeb-A (LIU *et al.*, 2015), com 202.599 imagens.

Figura 15 – Distribuição das idades presentes na base de dados IMDB-Wiki.



Fonte: Rothe, Timofte e Gool (2018)

4.6 Cross-Age LFW (CALFW)

Cross-Age LFW (CALFW) (ZHENG; DENG; HU, 2017), é um subconjunto da base de dados LFW (HUANG *et al.*, 2007), muito utilizada em verificação e reconhecimento facial. Contém 3.000 pares de imagens com uma diferença de idades, do mesmo indivíduo.

4.7 AgeDB

AgeDB (Moschoglou *et al.*, 2017) é uma base de dados manualmente coletada com 16.488 imagens de 568 indivíduos com idades entre 0 e 101 anos.

4.8 CelebA-HQ

CelebA-HQ (KARRAS *et al.*, 2018) consiste em 30.000 imagens de resolução 1024x1024 extraídas da base de dados Celeb-A (LIU *et al.*, 2015). Como a base original tinha fatos que iam de 43x55 até 6732x8984, existiam alguma das 202.599 imagens que poderiam virar fotos de alta resolução necessárias para a PROGAN. Para conseguir utilizar as imagens, os autores usaram duas redes pré-treinadas, um *autoencoder* convolucional para

remover os artefatos JPEG das imagens, e uma rede de super-resolução para aumentar o tamanho da imagem. Além disso, dado que a base de dados original era não controlada, os autores efetuaram algumas etapas de pré-processamento (alinhamento da imagem usando 4 pontos, dois olhos e os cantos da boca, e completaram a imagem caso ela estivesse nos cantos), fizeram isso para as 202.599 imagens e selecionaram as 30.000 melhores.

4.9 Flickr-Faces-HQ (FFHQ)

Flickr-Faces-HQ (FFHQ) (KARRAS; LAINE; AILA, 2019) é uma base de dados com faces humanas em alta qualidade, com o objetivo de ser um *benchmark* para GANs. Consiste em 70.000 imagens em alta resolução de 1024x1024, no formato PNG. Essas imagens foram obtidas da plataforma Flickr, foram escolhidas as imagens que tinham permissão de compartilhamento, e foram cortadas e alinhadas usando a biblioteca dlib (KING; SONNENBURG,). Os autores comentam que essa base de imagens contém muito mais variações em termos de etnicidade e o fundo das fotos comparada a CelebA-HQ, e que também contém diversos acessórios, como óculos de grau e de sol, chapéus, etc.

Os autores Or-El *et al.* (2020a) publicaram junto com seus resultados, novos metadados das imagens presentes contendo pose, predição da idade e do gênero, e a segmentação das regiões das faces presentes em FFHQ em que chamaram de *FFHQ Aging*.

Importante notar que os intervalos de idade presentes nos metadados de *FFHQ Aging* foram predições obtidas pelos autores ao utilizar a plataforma Appen, que além de prover o intervalo de idades, entrega também uma métrica, chamada pela plataforma de intervalo de confiança da predição. Para reduzir possíveis erros ao utilizar predições incorretas, nos experimentos presentes no Capítulo 7, só foram utilizadas imagens com intervalo de confiança de 100%.

5 Métricas de avaliação

A avaliação da qualidade das imagens sintéticas geradas por modelos generativos é uma tarefa complexa. Caso seja feita de forma manual por humanos, essa atividade se torna trabalhosa e cara, ao utilizar-se de plataformas para terceirizar o serviço (plataformas de *crowdsourcing*[1]), além da chance de múltiplos avaliadores terem diferentes critérios de qualidade.

Algumas formas automáticas de avaliação são utilizadas na literatura, mas todas contém pontos negativos que devem ser considerados caso sejam utilizadas.

5.1 Distância de pixels

Consiste em subtrair os valores de cada pixel entre a imagem sintética (gerada) e a real. Porém, uma simples translação da imagem já altera em grande medida a métrica, não sendo recomendável sua utilização.

5.2 Inception Score (IS)

Publicada originalmente em Salimans *et al.* (2016), a métrica utiliza uma rede neural previamente treinada na base de dados ImageNet[2], usando a arquitetura InceptionNet. Sua saída é um vetor da probabilidade de cada uma das mil classes. O objetivo da métrica é medir a fidelidade (quanto uma classe se destaca comparada às outras) e a diversidade (quantas classes estão sendo geradas pela GAN, mais especificamente seu gerador). Na avaliação, todas as imagens geradas passam pela rede neural, e os vetores de saída são utilizados.

$$D_{KL}(p(y|x)||p(y)) = p(y|x)\log(\frac{p(y|x)}{p(y)}) \tag{17}$$

$$IS = \exp(\mathbb{E}_{x\sim p} D_{KL}(p(y|x)||p(y)) \tag{18}$$

As Equações 17 e 18 mostram seu cálculo, com y sendo a classe e x o exemplo gerado. Em $p(x|y)$ é calculada a probabilidade condicional (medindo a fidelidade), sendo um valor

[1] Exemplo: Amazon Mechanical Turk (MTurk).

[2] Base de imagens obtida da internet com mil classes de objetos e mais de um milhão de exemplos

maior caso uma classe tenha um valor muito superior às outras e p(y) a diversidade, medindo a distribuição das diferentes classes.

Um ponto negativo explicado em Barratt e Sharma (2018), é que essa métrica em nenhum momento compara as imagens geradas com as imagens reais. O que está sendo medido é se o classificador consegue identificar alguma das mil classes de objetos. Nada impede de uma imagem conter atributos da classe mas em locais que a torna irreal, e ainda assim, continua recebendo uma boa nota.

5.3 Similaridade de co-senos de representações faciais

Um abordagem utilizada em diversos trabalhos (ALALUF; PATASHNIK; COHEN-OR, 2021; OR-EL *et al.*, 2020b; WANG *et al.*, 2018b) para medir de forma quantitativa a manutenção da identidade nas imagens geradas, envelhecidas ou rejuvenescidas, é o uso de representações[3] de redes neurais de reconhecimento facial pré-treinadas. Já que essas foram otimizadas para encontrar atributos nas faces que tornem os indivíduos únicos. Logo, ao calcular a similaridade de co-senos[4] entre a representação da imagem original e a representação da imagem envelhecida ou rejuvenescida, pode-se capturar a preservação da identidade.

5.4 Distância de Fréchenet (FID)

A distância de Fréchenet (do inglês, *Fréchent Inception Distance*) foi proposta por Heusel *et al.* (2017) é o método de avaliação de GANs mais utilizado atualmente e também utiliza o classificador previamente treinado com arquitetura InceptionNet. Porém, no cálculo da métrica é removida a última camada da rede (totalmente conectada), resultando em sua saída um vetor de dimensão 2.048, sendo uma representação da imagem dos atributos detectados por essa rede. Estatísticas dessas representações, tanto das imagens reais quanto das sintéticas (geradas) são usadas durante o cálculo da FID. A distância de Fréchenet entre duas distribuições normais multivariadas é dada pela Equação 19. Sendo μ_X e Σ_X a média e covariância das imagens geradas, μ_Y e Σ_Y a média e covariância das imagens reais.

[3] Em geral, é utilizada o resultado da última camada totalmente conectada como representação facial.
[4] Lembrando que a similaridade de co-senos tem o valor 1 representando uma similaridade máxima, 0 para vetores ortogonais e -1 uma relação inversa.

$$d(X,Y) = \|\mu_X - \mu_Y\|_2 + \mathrm{Tr}\left(\Sigma_X + \Sigma_Y - 2\sqrt{\Sigma_X \Sigma_Y}\right) \tag{19}$$

Baixos valores de FID significam que as distribuições estão próximas, o que se deseja que aconteça ao comparar a distribuição das imagens geradas e as reais. Pontos negativos: ao aumentar o tamanho da amostra utilizada para o cálculo, o valor da FID diminui, o que não é desejável para uma métrica de avaliação. Por necessitar de cálculos complexos, demora para ser executada. Por fim, utiliza apenas duas estatísticas, média e covariância, das distribuições.

5.5 Perceptual path lenght (PPL)

Essa métrica foi introduzida por Karras, Laine e Aila (2019) e serve para avaliar como gerador da GAN consegue interpolar imagens no espaço latente z.

Similaridade perceptiva é uma métrica que também utiliza uma rede previamente treinada para obter a representação das características detectadas, mas ao contrário da métrica FID, utiliza a rede VGG16 ao invés da InceptionNet. Seus autores citam que essa métrica é superior a outras tradicionais como relação sinal-ruído de pico (do inglês, *peak signal-to-noise ratio*, PSNE) e índice de similaridade estrutural (do inglês, *structural similarity index*, SSIM), por uma larga margem na percepção humana de similaridade (ZHANG *et al.*, 2018).

6 Trabalhos correlatos

Na revisão feita por Grimmer, Ramachandra e Busch (2021), são apresentados os principais modelos que executam o envelhecimento facial usando redes profundas, a evolução da quantidade de publicações, uma taxonomia das técnicas existentes e são apresentadas as principais bases de dados utilizadas. No trabalho, os autores avaliaram três modelos de envelhecimento para imagens em alta definição publicados recentemente: HRFAE (YAO *et al.*, 2020), LIFE (OR-EL *et al.*, 2020b) e SAM (ALALUF; PATASHNIK; COHEN-OR, 2021). Na avaliação, utilizam fotos externas de 4 pessoas jovens, entre 24 e 32 anos, e apresentam como os modelos se comportam ao tentar envelhecer as fotos para três grupos de idades: 65, 50-69 e 60 anos.

Para fazer uma análise quantitativa, os autores estimaram a idade da foto original utilizando um modelo pré-treinado ArcFace (DENG *et al.*, 2019) e também de cada foto gerada. Além disso, a métrica FID é utilizada para quantificar a percepção de realidade das imagens, e a distância de co-senos é utilizada entre o resultado da imagem original e a envelhecida na última camada da rede pré-treinada.

Como esse trabalho é uma revisão sistemática de envelhecimento facial profundo, foi dado mais destaque para a descrição de metodologias utilizadas na tarefa, categorização dos trabalhos selecionados na taxonomia apresentadas e as bases utilizadas por cada um. Logo, o espaço reservado para a comparação dos três modelos que os autores selecionaram ficou limitada, visto que foi utilizada apenas 4 imagens na análise.

Os experimentos realizados no Capítulo 7 se assemelham as comparações feitas nos trabalhos apresentados abaixo pois os autores verificam a performance da abordagem apresentada com às existentes.

6.1 GANs

6.1.1 HRFAE

O modelo HRFAE (YAO *et al.*, 2020) (detalhado na Seção 2.4.4), utilizou uma arquitetura codificador-decodificador para o envelhecimento facial. O treinamento do modelo utilizou a base FFHQ e imagens sintéticas[1] para ter uma base de imagens em alta

[1] Obtidas com a rede StyleGAN.

resolução balanceada. Compararam os resultados com IPCGAN (WANG *et al.*, 2018b), S2GAN (YANG *et al.*, 2018) e FaderNet (LAMPLE *et al.*, 2017) na base de imagens CACD. Um detalhamento do modelo pode ser encontrado na Seção 2.4.4.

6.1.2 SAM

Em SAM (ALALUF; PATASHNIK; COHEN-OR, 2021) (descrito na Seção 2.4.5), os autores compararam seu modelo com LIFE (OR-EL *et al.*, 2020b) e HRFAE (YAO *et al.*, 2020), já que na data de publicação eram considerados trabalhos no estado da arte. A base de dados utilizadas para a avaliação dos resultados é a CelebA-HQ, pois contém a idade das celebridades. Assim como em (GRIMMER; RAMACHANDRA; BUSCH, 2021) os autores utilizaram ArcFace (DENG *et al.*, 2019) para calcular a similaridade de co-senos de cada par de imagens. No estudo foi também apresentada uma análise qualitativa entre as imagens geradas de cada trabalho. Para a análise quantitativa, os autores geraram 80 imagens de LIFE, e escolheram a imagem em que o modelo ArcFace teve a predição mais próxima da desejada. Os modelos SAM e HRFAE podem gerar faces para uma idade pré-definida, porém, para fins de comparação o mesmo protocolo foi utilizado. Por fim, como os autores utilizaram ArcFace no treinamento de seu modelo, para fazer a predição das imagens das fotos, eles utilizaram Microsoft Azure Face API. Também foi feita uma pesquisa em que humanos avaliavam fotos do mesmo indivíduo, e respondiam qual das imagens eles preferiam, usando a idade desejada e a qualidade da imagem gerada como métricas. Outros detalhes desse modelo podem ser vistos na Seção 2.4.5.

6.2 Modelos de difusão

Os trabalhos recentes que apresentam edição de imagens utilizando modelos de difusão trazem análises dos resultados mais genérica do que os trabalhos que utilizam GANs especificamente para o envelhecimento facial. Tanto que (PARMAR *et al.*, 2023) e (BROOKS; HOLYNSKI; EFROS, 2022) utilizados nos experimentos realizados no Capítulo 7 não trazem métricas quantitativas nem qualitativas específicas para essa tarefa.

Um trabalho que aplica modelos de difusão na tarefa de envelhecimento facial é (WOLLEB *et al.*, 2022). No trabalho, os autores utilizaram um modelo de difusão DDPM,

com um esquema de amostragem DDIM (Seção 2.3.2.1) e orientação de gradiente (Seção 2.2.2.3) para efetuar a tarefa de tradução imagem a imagem não pareada condicional (Seção 2.2.2.3). Nesse trabalho, como no cálculo da orientação de gradiente é utilizada uma função externa, essa pode ser aplicada somente no processo de remoção de ruído (amostragem) não sendo necessário treinar o modelo de difusão novamente.Nos experimentos, os autores treinaram um modelo de difusão DPPM e um modelo de regressão com mecanismos de atenção. Utilizaram uma base de dados de faces com idades obtida da plataforma Kaggle[2] e uma base de dados de ressonâncias magnéticas.

Os autores mostraram análises qualitativas de 2 indivíduos de 40 anos, em que a edição modifica para as seguintes idades: 10, 20, 60 e 80 anos.

[2] Essa base de dados não foi descrita no Capítulo 4, pois não foi disponibilizada as informações para descrever as imagens presentes, como pode ser visto em ⟨https://www.kaggle.com/datasets/mariafrenti/age-prediction⟩.

7 Resultados experimentais

Foram realizados dois estudos comparativos entre modelos generativos que efetuam a tarefa de envelhecimento facial.

No primeiro, foi aferida a performance de um *autoencoder* adversário condicional, CAAE, e duas redes adversárias generativas (GANs), IPCGAN e RCRIIT, envelhecendo esses modelos para idades esperadas e comparando os resultados obtidos com imagens em baixa resolução, pois estas eram as principais bases utilizadas na época de suas publicações.

No segundo estudo comparativo, que utilizou imagens de alta definição, foi medido o desempenho de quatro modelos: dois modelos de arquitetura que utilizavam GANs, HRFAE e SAM, e dois modelos de difusão condicionais que possibilitavam a edição de imagens, *Pix2pix-zero* e *Instruct-pix2pix*.

7.1 Experimentos em baixa resolução

7.1.1 Introdução

Nestes experimentos foram comparados três métodos de envelhecimento facial: CAAE (ZHANG; SONG; QI, 2017), IPCGAN (WANG *et al.*, 2018b), e RCRIIT (HELJAKKA; SOLIN; KANNALA, 2018). Esses trabalhos foram recorrentemente citados por publicações posteriores e tiveram seus códigos disponibilizados publicamente. Esses códigos foram utilizados pelo autor deste trabalho para treinar novamente os modelos e obter os resultados de experimentos aqui apresentados. Para medir o reconhecimento facial foram utilizadas duas bases de dados de imagens com idades anotadas: FG-NET e UTKFaces. Ambas são frequentemente usadas como referência em artigos de envelhecimento facial e estão disponíveis de forma gratuita.

Os resultados descritos na seção abaixo foram publicados em Kemmer, Simões e Lima (2022).

7.1.2 Configuração dos experimentos

Para a detecção dos rostos presentes nas bases de dados FG-NET e UTKFaces foi utilizado um algoritmo de detecção de faces *Multitask Cascaded Convolutional Networks*,

MTCNN[1] (Zhang *et al.*, 2016). Esse modelo tem como saída um conjunto de faces detectadas (na forma de retângulos), a probabilidade de conter uma face, e alguns pontos faciais de referência. Durante o pré-processamento foi utilizado o retângulo com maior probabilidade para recortar a imagem, mantendo uma margem externa para não perder partes da face. Os pontos de referência retornados pelo modelo foram usados para alinhar a face (os dois pontos centrais dos olhos) e para centralizá-la (centro do nariz). Finalmente, a imagem foi reduzida para 128×128, para que todas tivessem o mesmo formato e devido aos modelos de envelhecimento facial utilizados nessa comparação terem sua arquitetura feita para essas dimensões de entrada.

Para avaliar os três modelos, os exemplos foram divididos em seis faixas de idade, possibilitando a comparação entre cada transformação do envelhecimento ou rejuvenescimento. Um estimador de idade[2] (CAO; MIRJALILI; RASCHKA, 2020) foi usado para verificar se a transformação foi precisa.

Foi utilizado o *autencoder* adversário condicional CAAE (definido na Seção 2.4.1) com a utilização de uma implementação pública do mesmo[3]. Nos experimentos, os quatro blocos (codificador, gerador e dois discriminadores) são atualizados de forma alternada com um tamanho de mini-lote de 80 por meio de um gradiente descendente estocástico, ADAM (KINGMA; BA, 2014) ($\alpha = 0,0002, \beta_1 = 0,5$) em 100 épocas, seguindo os hiper-parâmetros utilizados pelos autores. Durante o teste, apenas E e G estão ativos.

Foram treinados e comparados dois processos de envelhecimento facial com essa configuração experimental, mas usando conjuntos de dados diferentes, UTKFace e FG-NET. Além disso, dividimos os dados em dez grupos: 0 - 5, 6–10, 11–15, 16-20, 21-30, 31-40, 41–50, 51–60, 61-70 e 70+. Isso permite comparação direta com (ZHANG; SONG; QI, 2017).

Também foi empregue o modelo IPCGAN (definido na Seção 2.4.2) com a implementação disponibilizada pelo autor[4] e com os hiper-parâmetros descritos ($\lambda_1 = 75$, $\lambda_2 =$ 5E-4 e $\lambda_3 = 30$)[5]. Uma diferença entre esse experimento e o feito pelos autores é que o

[1] Implementação em ⟨https://github.com/ipazc/mtcnn⟩

[2] A rede neural estimadora de idade foi pré-treinada na base de dados CACD e o código está disponível em: ⟨https://github.com/Raschka-research-group/coral-cnn/tree/master/single-image-prediction--w-pretrained-models⟩.

[3] Disponível em: ⟨https://github.com/mattans/AgeProgression/tree/v1.0.0⟩.

[4] Disponível em: https://github.com/dawei6875797/Face-Aging-with-Identity-Preserved-Conditional-Generative-Adversarial-Networks

[5] Utilizando os mesmos hiper-parâmetros foram geradas imagens com qualidade semelhante às publicadas pelos autores.

primeiro utiliza a base de dados UTKFaces para treinamento e os autores utilizaram a base de dados CACD com cinco classes. E para permitir a comparação deste modelo com os outros presentes no experimento o conjunto de dados foi dividido em seis categorias (11-20, 21-30, 31-40, 41-50, 51-60 e 60+).

Finalmente, com o modelo RCRIIT (descrito na Seção 2.4.3) foi usada a metodologia de divisão dos dados detalhada nos experimentos dos autores, o conjunto de dados UTKFace foi dividido em seis grupos. Foi reduzido o intervalo na primeira faixa etária entre 13 e 18 anos, sendo escolhida aleatoriamente 1.000 imagens para cada faixa etária. Assim como na divisão dos autores, foi mantida a idade média nos conjuntos com dez anos de diferença. Portanto, também são necessários cinco módulos de transformação. O primeiro é formado pelas duas primeiras faixas etárias (13-18 e 19-29), o segundo módulo é formado pela segunda e terceira faixas etárias (19-29 e 30-39) e assim por diante.

Os experimentos considerados usam o *pipeline* [6] disponibilizado pelos autores. A cadeia completa obteve cinco módulos sucessivos treinados de forma independente. Cada módulo foi treinado com um tamanho de mini-lote de 1 com ADAM ($\alpha = 0,0002, \beta_1 = 0,5, \beta_2 = 0,999$) em 100 épocas. Também seguindo os hiper-parâmetros publicados.

7.1.3 Resultados e discussão

As Figuras 16 e 17 apresentam as imagens geradas a partir do processo de envelhecimento e rejuvenescimento utilizando o modelo CAAE através do jogo de envelhecimento[7]. É possível detectar que as imagens geradas a partir do modelo treinado com UTKFace são mais nítidas. Isso deve ser causado pela maior variedade de dados de treinamento em cada grupo do que o FG-NET, principalmente nos últimos grupos (conforme descrito na Tabela 5).

[6] Disponível online em: ⟨https://github.com/AaltoVision/img-transformer-chain⟩

[7] Método de visualização do envelhecimento facial nomeado pelos autores com seu código disponível em: ⟨https://github.com/mattans/AgeProgression/blob/v1.0.0/aging_game.ipynb⟩

Figura 16 – Jogo de envelhecimento com conjunto de dados FG-NET utilizando o modelo CAAE: a primeira coluna mostra as faces de entrada, as colunas restantes são imagens geradas de progressão e regressão de idade de acordo com a coluna de grupo de idade. As caixas brancas indicam o mapeamento para a mesma faixa etária da imagem original.

Figura 17 – Jogo de envelhecimento com conjunto de dados UTKFace: a primeira coluna mostra as faces de entrada, as colunas restantes são imagens geradas de progressão e regressão de idade de acordo com a coluna de grupo de idade. As caixas brancas indicam o mapeamento para a mesma faixa etária da imagem original.

A Figura 18 mostra as imagens geradas pelo modelo IPCGAN a partir do processo de envelhecimento e rejuvenescimento e a Figura 19 expõe uma ilustração dos resultados de envelhecimento ao utilizar o modelo RCRIIT.

Figura 18 – A primeira coluna mostra as faces de entrada (retângulos vermelhos); as colunas restantes são geradas a partir da progressão e regressão de idade de acordo com a coluna de grupo de idade.

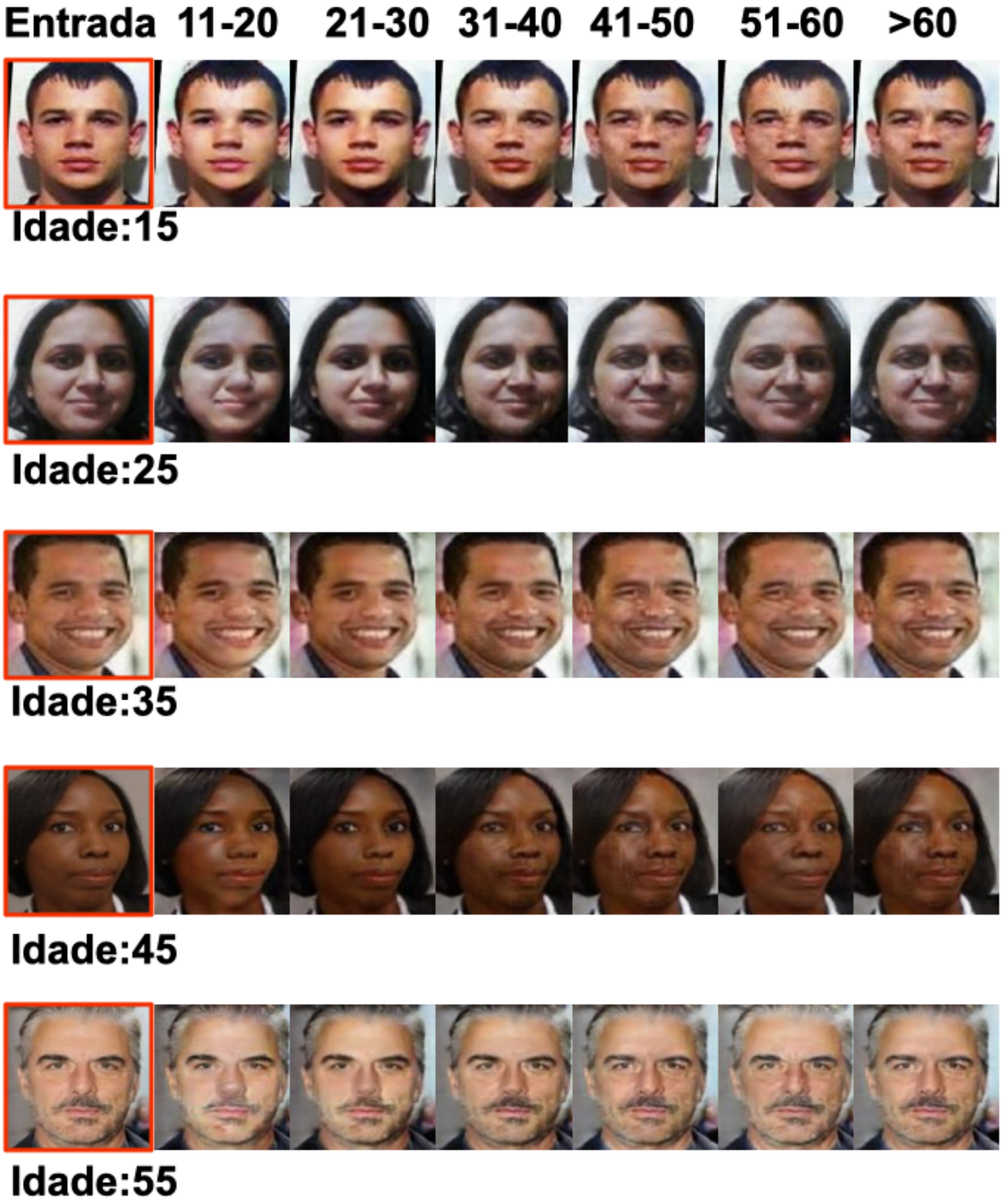
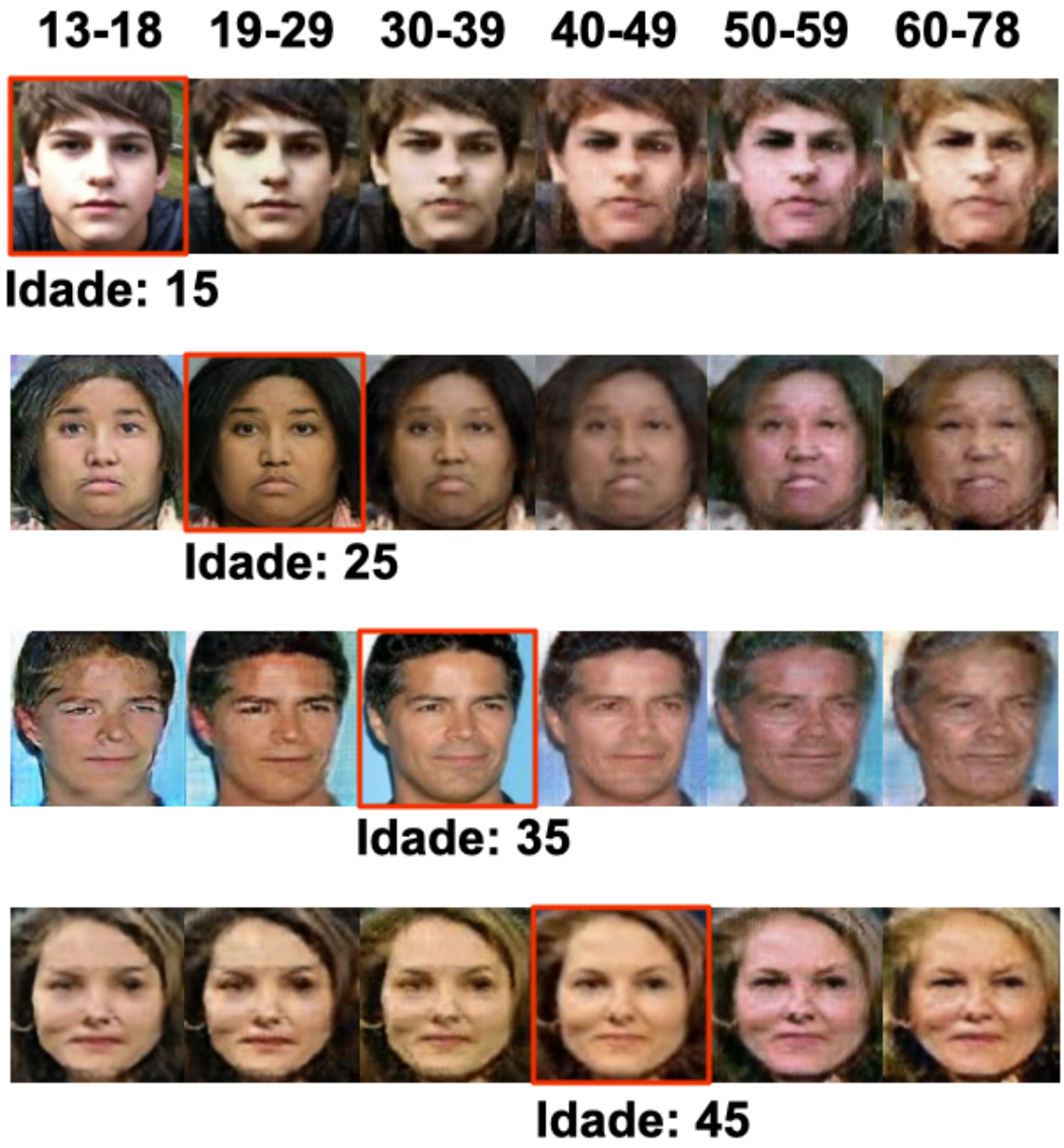


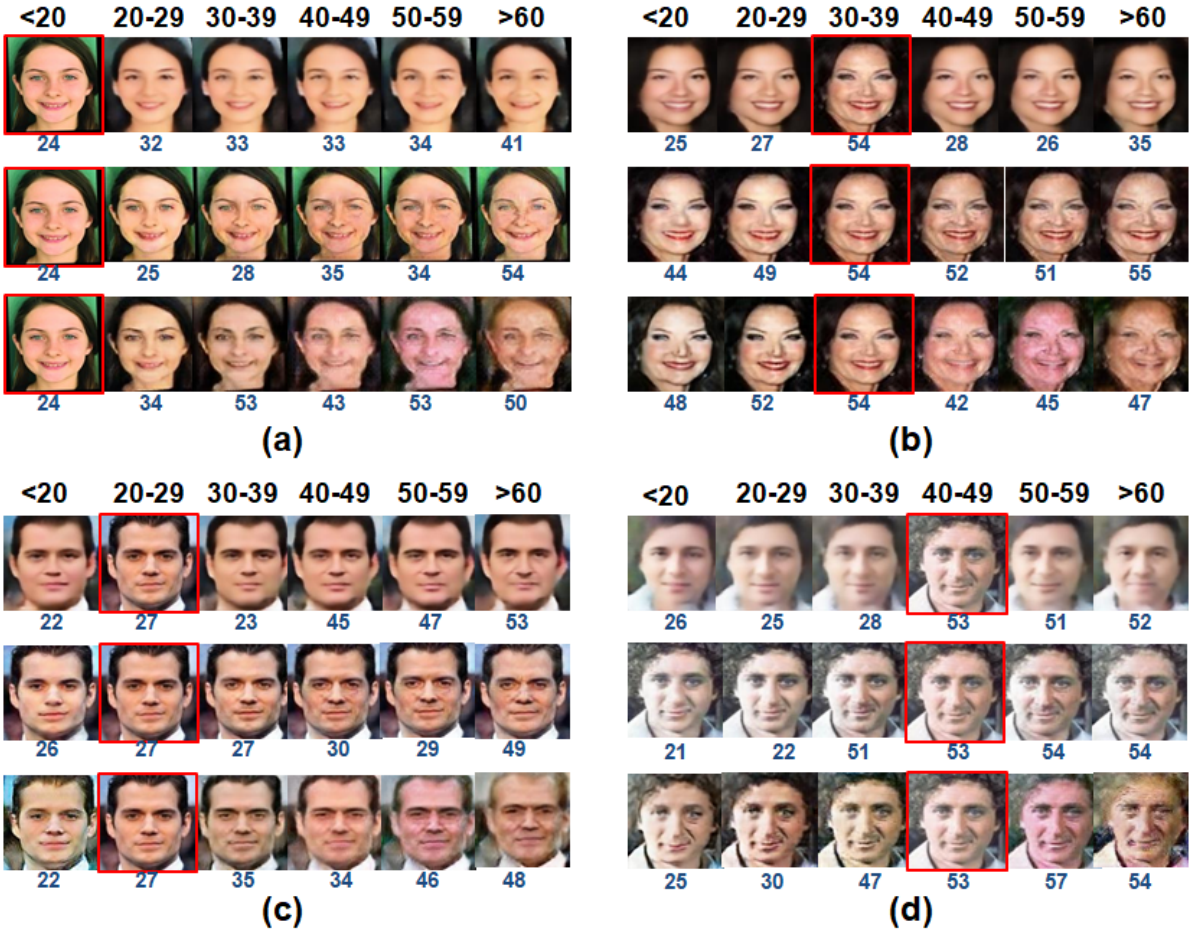
Figura 19 – As faixas etárias consideradas são especificadas em cada coluna, e as caixas vermelhas indicam a imagem de entrada. A linha 1 mostra uma imagem de entrada com 15 anos relacionada à faixa etária de 13 a 18 anos, onde foi realizado o processo de envelhecimento. A linha 2 mostra uma imagem de entrada com 25 anos relacionada à faixa etária 19-29, na qual uma etapa de rejuvenescimento foi realizada para gerar a imagem da primeira coluna e quatro etapas foram realizadas para o processo de envelhecimento (terceira coluna em diante). As linhas 3 e 4 mostram imagens de entrada das faixas etárias 30-39 e 40-49, nas quais as etapas de rejuvenescimento e envelhecimento foram realizadas para gerar as imagens nas colunas anteriores e posteriores.



Uma comparação dos resultados dos três modelos pode ser vista na Figura 20. Possibilitando notar as transformações para cada grupo de idade, e também a estimativa

de idade da imagem gerada feita pelo estimador de idade. O estimador de idade não é exato, porém funciona como uma boa métrica do envelhecimento facial.

Figura 20 – Os grupos de transição estão especificados por coluna, e as caixas vermelhas indicam a imagem de teste original. Figuras (a), (b), (c), e (d) representam exemplos de imagens de teste de 15, 25, 35, e 45 anos. As linhas 1, 2, e 3 mostram as transições de envelhecimento e rejuvenescimento para cada modelo CAAE, IPCGAN, e RCRIIT, respectivamente. O número abaixo de cada imagem de rosto, em azul, indica a predição do estimador de idade.



Qualitativamente, é possível verificar que as três abordagens realizam algumas transformações de envelhecimento e rejuvenescimento, com destaque para as abordagens GANs. Quantitativamente, ao utilizar o estimador de idade, podemos ver que existe um envelhecimento ou rejuvenescimento, porém abaixo da idade esperada, não envelhecendo ou rejuvenescendo o suficiente.

Observando novamente a Figura 20 também é possível realizar uma comparação qualitativa para verificar se a personalidade foi preservada ao longo das transformações de envelhecimento e rejuvenescimento, pode-se perceber que a identidade não foi preservada

pelo método CAAE (linha 1). O modelo obteve uma performance melhor no envelhecimento facial comparado à preservação da identidade.

7.2 Experimentos em alta resolução

7.2.1 Introdução

Nesse segundo grupo de experimentos foram avaliados quatro modelos que executam a tarefa de envelhecimento facial. Dois modelos são baseados em GANs: HRFAE (Seção 2.4.4) e SAM (Seção 2.4.5), e outros dois baseados em modelos de difusão, *Pix2pix-zero* (Seção 2.4.6) e *Instruct-pix2pix* (Seção 2.4.7). Os dois primeiros foram explicitamente treinados para gerar uma versão envelhecida da face de uma pessoa, e os outros dois tem uma geração de tiro zero (do inglês, *zero-shot generation*), em outras palavras, eles são modelos genéricos que executam diferentes tarefas, incluindo o envelhecimento facial. Como modelos de difusão tem ganhado atenção por causa da alta qualidade e diversidade das imagens geradas, comparar os resultados deles com modelos treinados e desenvolvidos especificamente para a tarefa com métricas relevantes é essencial. Portanto, nesse experimento foram utilizadas as faces presentes na base de dados *FFHQ Aging* (Seção 4.9), com os intervalos de idades preditas presentes nos metadados dela.

7.2.2 Configuração dos experimentos

As métricas empregadas nesta comparação entre os modelos foram o erro médio, erro médio absoluto, similaridade de co-senos (Seção 5.3) e FID (Seção 5.4).

Para o cálculo do erro médio, foi calculada a média da idade predita por grupo e seu erro médio, com esse sendo a média da subtração entre a idade predita (Equação 20) e a presente nos metadados da base, e o erro médio absoluto (Equação 21). Em que y_i é a idade estimada pelo modelo DEX na imagem inicial e $\hat{y}_i$ é a média da idade predita por grupo.

$$ME = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i - y_i \tag{20}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{21}$$

Também foi calculada a similaridade de co-senos entre as representações (descrita na Seção 5.3) obtidas ao utilizar a rede pré-treinada FaceNet[8] (SCHROFF; KALENICHENKO; PHILBIN, 2015). FaceNet tem uma performance similar a ArcFace (DENG *et al.*, 2019), porém como SAM utiliza a ArcFace em sua função de custo, a comparação fica mais justa utilizando outra rede pré-treinada.

Além disso, foi calculada a métrica FID por ser prevalente nas publicações de redes generativas (detalhes sobre a métrica podem ser consultados na Seção 5.4). Um ponto relevante a ser considerado é que ao aumentarmos a quantidade de exemplos tanto das imagens originais quanto das imagens geradas, o valor da métrica diminui, e quanto menor, mais próximas estão as distribuições. Portanto, foram comparados os 100 exemplos da categoria de entrada original (as 50 imagens escolhidas para mulheres e 50 para homens) usando a métrica FID, com as respectivas imagens geradas envelhecidas, também com a mesma quantidade.

A base de imagens utilizada foi a *FFHQ Aging* (Seção 4.9) em que as fotos têm uma idade estimada em grupos já definidos. Foram selecionadas faixas de idades próximas às utilizadas pelos autores em seus trabalhos. A Tabela 6 detalha como foram estruturados os grupos na comparação, em cada grupo foram sorteadas 50 imagens de indivíduos únicos tanto para mulheres quanto para homens. Foi utilizado um intervalo de 20 anos entre as faixas para envelhecimento, com exceção do último que tem uma diferença de 60 anos, feito para verificar um caso extremo.

Após testar o mesmo estimador de idade Coral (CAO; MIRJALILI; RASCHKA, 2020) usado na comparação do primeiro grupo de experimentos (Seção 7.1), notou-se que existia um viés na média das idades das bases de dados utilizadas nos treinamentos (tanto do modelo treinado com a base CACD quanto com a base MORPH2 tiveram um viés para a faixa de 30-40 anos, pois a maioria de imagens dessas bases tem essa idade). Logo, foi utilizado o modelo pré-treinado DEX (ROTHE; TIMOFTE; GOOL, 2015) que obteve estimativas próximas às faixas de idade presentes na base *FFHQ-Aging*, como pode ser visto na Tabela 7. Esse viés menor pode ser explicado devido ao modelo DEX ter sido

[8] O modelo FaceNet tem acurácia de 99,63% na tarefa de reconhecimento facial na base de dados LFW, logo é esperado o vetor de 512 posições resultante da saída de sua última camada consiga capturar a individualidade das faces.

Tabela 6 – Experimentos realizados utilizando o modelo *Pix2pix-zero*. Em que a coluna De-Para é a concatenação entre a idade da imagem Original e a Idade Destino. Origem é a faixa de idade estimada original presente na base *FFHQ Aging*, Destino* a faixa de idade estimada do grupo de faces utilizado para encontrar a direção de edição (no modelo *pix2pix-zero*) e Destino para os modelos em que era possível condicionar a uma idade alvo específica.

De-Para	Origem	Destino*	Destino
10 - 30	10 - 14	30 - 39	30
20 - 40	20 - 29	40 - 49	40
30 - 50	30 - 39	50 - 69	50
40 - 70	40 - 49	70 - 120	70
10 - 70	10 - 14	70 - 120	70

Fonte: Bruno Abreu Kemmer, 2023.

treinado com a base IMDB-Wiki, que contém mais de 460.000 exemplos e sua distribuição de idade pode ser vista na Figura 15.

Tabela 7 – Tabela mostra a idade média dos resultados dos estimadores de idade escolhidos nas 50 imagens selecionadas para cada faixa na coluna Origem da base *FFHQ-Aging*. Coral (CACD) utiliza um modelo pré-treinado com a base CACD, Coral (MORPH2) utiliza um modelo pré-treinado com a base MORPH2 e DEX que foi treinado com a base IMDB-WIKI.

Origem	Coral (CACD)	Coral (MORPH2)	DEX (IMDB-Wiki)
10 - 14	35,0	36,3	17,4
20 - 29	35,4	36,3	26,1
30 - 39	40,9	37,1	35,3
40 - 49	45,4	38,6	46,2

Fonte: Bruno Abreu Kemmer, 2023.

Foram utilizados os modelos previamente treinados, e disponibilizados pelos autores[9].

Nos experimentos realizados que utilizaram o modelo *Pix2pix-zero*, foi seguidos os passos descritos na Seção 2.4.6. Foi feita a inversão DDIM, para encontrar as respectivas direções de edição foram gerados dois grupos de setenças, um com o gênero (homem e mulher) e idade, e outro somente com a idade. Essa separação foi feita pois homens e mulheres tem características faciais e envelhecem de formas distintas.

[9] HRFAE: modelo pré-treinado presente em ⟨https://github.com/InterDigitalInc/HRFAE⟩ e com os hiper-parâmetros descritos na Seção 4.2 do trabalho de Yao *et al.* (2020). SAM: modelo pré-treinado presente em ⟨https://github.com/yuval-alaluf/SAM⟩, e com seus hiper-parâmetros descritos no Apêndice A da publicação de Alaluf, Patashnik e Cohen-Or (2021). *Pix2pix-zero*: modelo pré-treinado presente em ⟨https://github.com/pix2pixzero/pix2pix-zero⟩ com detalhamento e e hiper-parâmetros no Apêndice D do artigo de Parmar *et al.* (2023). *Instruct-pix2pix*: modelo pré-treinado presente em ⟨https://github.com/timothybrooks/instruct-pix2pix⟩ e com sua implementação e hiper-parâmetros detalhados na Seção 3 do Apêndice A do trabalho de Brooks, Holynski e Efros (2022).

Além disso, os modelos *Pix2pix-zero* e *Instruct-pix2pix* possibilitam armazenar os modelos em 16 bits[10] porém, para garantir que qualquer perda de qualidade nas imagens não estaria relacionada com erros de aproximação, os experimentos realizados utilizaram 32 bits.

Nos experimentos realizados, foi utilizado o valor padrão de 1,5 para c_{imagem} e variando $c_{instrução}$ nos seguintes valores: 0; 1,5; 3 e 10. Seguindo o que foi demonstrado pelos autores.

A instrução utilizada na inferência do modelo *Instruct-pix2pix* foi, no caso para envelhecer uma mulher para sesenta anos: "Faça uma mulher de sessenta anos", na língua inglesa *"make a woman sixty years old"*. Quando eram homens foi trocada a palavra em inglês *woman* por *men*.

Importante ressaltar que foram analisados os resultados separando as imagens de entrada entre homens e mulheres, porém, como não foram notadas diferenças significativas nos resultados, portanto, os dados foram unificados.

7.2.3 Resultados e discussão

No caso do modelo *Pix2pix-zero* em uma das variações experimentadas foi utilizado o gênero e a idade das imagens para obter a direção de edição, e na outra, apenas a idade. A média da similaridade de co-senos na inversão DDIM usando essas abordagens não mudou, tendo o valor médio de 0,85. Observando pela Tabela 8, pode-se notar uma leve melhora quando não foram utilizados os gêneros para encontrar as direções de edição do modelo, logo, esses resultados serão utilizados nas comparações.

Nos modelo *Instruct-pix2pix* variando o valor de quanto a instrução de envelhecimento será seguida $c_{instrução}$ pode-se ver pela Tabela 9 que se o parâmetro tem o valor 0, a imagem não é envelhecida pois a instrução não é seguida, o que é esperado. Aumentando o valor de $c_{instrução}$ para 10, o erro cai consideravelmente, porém, as imagens deixam de estar próximas na representação da FaceNet, perdendo a identidade. Esses resultados mostram que o melhor valor para $c_{instrução}$ é 3, o mesmo publicado pelos autores[11], logo, nos resultados apresentados a seguir será mostrado esse.

[10] Podendo rodar em máquinas de menor quantidade de RAM em suas GPUs.
[11] No artigo, os autores utilizam S_I para c_{imagem} e S_T para $c_{instrução}$.

Tabela 8 – Resultados de dois experimentos que utilizaram o modelo *Pix2pix-zero* em um deles foi utilizado o gênero e a idade das imagens para obter a direção de edição, e no outro, apenas a idade.

		Experimentos realizados				
		10-30	**20-40**	**30-50**	**40-70**	**10-70**
Com gênero e idade	**Idade predita**	$25,8 \pm 6,9$	$27,0 \pm 5,5$	$32,0 \pm 6,0$	$43,0 \pm 9,0$	$29,3 \pm 10,5$
	Erro Médio	$4,2 \pm 6,9$	$13,0 \pm 5,5$	$18,0 \pm 6,0$	$27,0 \pm 9,0$	$40,7 \pm 10,5$
	Erro Abs. Médio	$6,3 \pm 5,0$	$13,3 \pm 4,7$	$18,0 \pm 6,0$	$27,0 \pm 9,0$	$40,7 \pm 10,5$
	Sim. Co-senos	$0,47 \pm 0,32$	$0,81 \pm 0,24$	$0,81 \pm 0,24$	$0,8 \pm 10,24$	$0,48 \pm 0,32$
	FID	108,51	79,52	81,88	91,28	103,58
Somente idade	**Idade predita**	$23,2 \pm 7,5$	$27,9 \pm 5,7$	$35,6 \pm 7,4$	$45,4 \pm 10,2$	$31,3 \pm 15,4$
	Erro Médio	$\mathbf{6,8 \pm 7,5}$	$\mathbf{12,1 \pm 5,7}$	$\mathbf{14,4 \pm 7,4}$	$\mathbf{24,6 \pm 10,2}$	$\mathbf{38,7 \pm 15,4}$
	Erro Abs. Médio	$8,1 \pm 6,0$	$12,3 \pm 5,1$	$14,8 \pm 6,7$	$24,6 \pm 10,2$	$38,9 \pm 15,1$
	Sim. Co-senos	$\mathbf{0,62 \pm 0,29}$	$\mathbf{0,82 \pm 0,23}$	$\mathbf{0,79 \pm 0,25}$	$\mathbf{0,81 \pm 0,20}$	$\mathbf{0,52 \pm 0,31}$
	FID	102,28	84,04	82,75	96,31	104,35

Erro Médio = Idade esperada - Idade predita | Sim. Co-senos = Similaridade de co-senos | c = $c_{instrução}$ | Idade predita = média da predição das idades | FID = Distância de Fréchenet, como essa métrica é calculada utilizando estatísticas da amostra, não temos o desvio padrão dela.

Ao observar os resultados apresentados pela Tabela 10, é possível notar que o modelo SAM consegue obter as menores médias de erros, com a exceção do modelo HRFAE na configuração "10-30". Para confirmar a hipótese de que os erros médios apresentados pelo modelos SAM eram menores do que de todos os outros modelos, foi executado um teste de Wilcoxon-Mann-Whitney (WILCOXON, 1945) comparando o modelo SAM com os demais, em cada uma das configurações que foram apresentadas, compondo ao todo 15 testes.

Sendo, H_0 a hipótese de que a diferença entre o erro médio de ambos os modelos comparados (para cada configuração) serem estatisticamente irrelevantes, e H_1 a hipótese de que a média do erro ao utilizar o modelo SAM, μ_1, seja menor do que a média do erro do modelo que está sendo comparado, μ_2.

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 < \mu_2$$

Todas as hipóteses nulas foram rejeitadas[12], com a exceção no caso do modelo HRFAE na configuração "30-50"em que o teste obteve o p-valor de 0,8. Portanto, podemos concluir que mesmo nessa configuração a performance de ambos os modelos foram

[12] Todos os testes obtiveram p-valor abaixo de 0,01%, com a exceção do caso relatado (modelo HRFAE na configuração "30-50".

Tabela 9 – Resultados de *Instruct-pix2pix* variando o valor de quanto a instrução de envelhecimento será seguida $c_{instrução}$.

		Experimentos realizados				
		10-30	**20-40**	**30-50**	**40-70**	**10-70**
c=0	**Idade predita**	$18,8 \pm 5,1$	$26,8 \pm 4,2$	$35,5 \pm 5,8$	$51,9 \pm 6,6$	$30,2 \pm 13,6$
	Erro Médio	$11,2 \pm 5,1$	$13,2 \pm 4,2$	$14,5 \pm 5,8$	$18,1 \pm 6,6$	$39,8 \pm 13,6$
	Erro Abs. Médio	$11,6 \pm 3,9$	$13,4 \pm 3,7$	$14,8 \pm 5,0$	$18,1 \pm 6,6$	$39,8 \pm 13,5$
	Sim. Co-senos	$0,88 \pm 0,14$	$0,90 \pm 0,13$	$0,87 \pm 0,18$	$0,82 \pm 0,14$	$0,78 \pm 0,18$
	FID	33,4	28,63	39,30	43,10	45,73
c=1,5	**Idade predita**	$20,9 \pm 6,0$	$28,0 \pm 5,2$	$39,1 \pm 7,1$	$57,1 \pm 5,7$	$39,7 \pm 14,6$
	Erro Médio	$\mathbf{9,1 \pm 6,0}$	$\mathbf{12,0 \pm 5,2}$	$\mathbf{10,9 \pm 7,1}$	$\mathbf{12,9 \pm 5,7}$	$\mathbf{30,3 \pm 14,6}$
	Erro Abs. Médio	$10,0 \pm 4,4$	$12,2 \pm 4,5$	$11,5 \pm 6,2$	$13,1 \pm 5,2$	$30,5 \pm 14,2$
	Sim. Co-senos	$0,88 \pm 0,11$	$0,95 \pm 0,08$	$0,90 \pm 0,19$	$0,84 \pm 0,12$	$0,75 \pm 0,16$
	FID	22.64	16.78	25.85	38.15	62.17
c=3	**Idade predita**	$19,8 \pm 6,6$	$28,7 \pm 4,7$	$39,4 \pm 5,6$	$51,1 \pm 5,7$	$21,2 \pm 7,1$
	Erro Médio	$\mathbf{10,2 \pm 6,6}$	$\mathbf{11,3 \pm 4,7}$	$\mathbf{10,6 \pm 5,6}$	$\mathbf{18,9 \pm 5,7}$	$\mathbf{48,8 \pm 7,1}$
	Erro Abs. Médio	$11,2 \pm 4,6$	$11,3 \pm 4,7$	$10,6 \pm 5,5$	$18,9 \pm 5,7$	$48,8 \pm 7,1$
	Sim. Co-senos	$\mathbf{0,93 \pm 0,03}$	$\mathbf{0,94 \pm 0,02}$	$\mathbf{0,93 \pm 0,03}$	$\mathbf{0,92 \pm 0,07}$	$\mathbf{0,92 \pm 0,03}$
	FID	45,43	38,89	40,74	42,52	52,47
c=10	**Idade predita**	$32,0 \pm 11,1$	$37,4 \pm 11,3$	$48,6 \pm 10,2$	$54,1 \pm 8,8$	$32,2 \pm 10,9$
	Erro Médio	$-2,0 \pm 11,1$	$2,6 \pm 11,3$	$1,4 \pm 10,2$	$15,9 \pm 8,8$	$37,8 \pm 10,9$
	Erro Abs. Médio	$8,1 \pm 7,7$	$9,7 \pm 6,3$	$8,1 \pm 6,2$	$16,1 \pm 8,4$	$37,8 \pm 10,9$
	Sim. Co-senos	$0,67 \pm 0,13$	$0,66 \pm 0,16$	$0,67 \pm 0,11$	$0,66 \pm 0,16$	$0,66 \pm 0,12$
	FID	184,86	170,90	176,90	182,96	182,65

Erro Médio = Idade esperada - Idade predita | Sim. Co-senos = Similaridade de co-senos | c = $c_{instrução}$ | Idade predita = média da predição das idades | FID = Distância de Fréchenet, como essa métrica é calculada utilizando estatísticas da amostra, não temos o desvio padrão dela.

equivalentes, e como o modelo SAM obteve erros estatisticamente menores em todas as outras configurações, podemos afirmar que, utilizando essa métrica, ele é o modelo que melhor consegue executar a tarefa de envelhecimento. Um possível motivo, é por ter um estimador de idade em sua função de custo.

Um ponto de atenção é que a similaridade entre a imagem envelhecida e a original é menor, comparada aos resultados de *Instruct-pix2pix*, que beiram uma similaridade total (valor 1). A alta similaridade entre as imagens obtidas pela HRFAE pode ser explicada pelo motivo de as mudanças executadas pela rede no envelhecimento serem baixas (exceto

na configuração "30-50"), logo, as imagens não estão sendo propriamente envelhecidas, como pode ser visto nos valores de erro e baixos valores de FID.

Outro ponto interessante de se notar na Tabela 10 é que o erro médio é estritamente positivo em todos os modelos, isso mostra que os modelos não conseguem alcançar a idade esperada, ficando sempre com uma idade menor do que a desejada, demonstrando um viés sistemático. Futuros trabalhos têm o potencial de utilizar esse ponto em suas arquiteturas.

Dois fatores podem afetar a manutenção das características dos indivíduos: um deles é o fato do modelo *Pix2pix-zero* usar direções de edição obtidas por meio de uma direção média única, isso pode fazer com que o modelo envelheça todas as imagens de modo incondicional, o que pode ser visto pelos baixos valores de similaridade e altos valores de FID. Outro fator é que esse modelo necessita reconstruir a imagem original por meio da técnica de inversão DDIM, nas imagens em que existam perdas de qualidade nessa etapa, pode comprometer a imagem final. A Figura 21 mostra exemplos de quando isso ocorre.

Tabela 10 – Comparação entre os resultados obtidos.

		Experimentos realizados				
		10-30	**20-40**	**30-50**	**40-70**	**10-70**
HRFAE	**Idade predita**	$18,9 \pm 2,7$	$24,9 \pm 2,6$	$47,8 \pm 4,9$	$43,9 \pm 4,6$	$25,4 \pm 5,9$
	Erro Médio	$11,1 \pm 2,7$	$15,1 \pm 2,6$	$\mathbf{2,2 \pm 4,9}$	$26,1 \pm 4.6$	$44,6 \pm 5,9$
	Erro Abs. Médio	$11,1 \pm 2,7$	$15,1 \pm 2,6$	$3,9 \pm 3,7$	$26,1 \pm 4,6$	$44,6 \pm 5,9$
	Sim. Co-senos	$0,85 \pm 0,06$	$0,91 \pm 0,04$	$0,88 \pm 0,05$	$\mathbf{0,94 \pm 0,03}$	$0,90 \pm 0,04$
	FID	28,32	22,16	21,11	20,26	22,02
SAM	**Idade predita**	$26,3 \pm 3,4$	$36,3 \pm 3,8$	$47,6 \pm 4,3$	$61,4 \pm 3,4$	$59,1 \pm 3,9$
	Erro Médio	$\mathbf{3,7 \pm 3,4}$	$\mathbf{3,7 \pm 3,8}$	$2,4 \pm 4,3$	$\mathbf{8,6 \pm 3,4}$	$\mathbf{10,9 \pm 3,9}$
	Erro Abs. Médio	$4,4 \pm 2,5$	$4,4 \pm 2,9$	$3,7 \pm 3,1$	$8,7 \pm 3,1$	$10,9 \pm 3,7$
	Sim. Co-senos	$0,77 \pm 0,14$	$0,75 \pm 0,09$	$0,71 \pm 0,16$	$0,68 \pm 0,11$	$0,49 \pm 0,12$
	FID	**114,21**	**109,87**	**121,06**	**107,24**	**135,25**
Pix2pix-zero	**Idade predita**	$23,2 \pm 7,5$	$27,9 \pm 5,7$	$35,6 \pm 7,4$	$45,4 \pm 10,2$	$31,3 \pm 15,4$
	Erro Médio	$6,8 \pm 7,5$	$12,1 \pm 5,7$	$14,4 \pm 7,4$	$24,6 \pm 10,2$	$38,7 \pm 15,4$
	Erro Abs. Médio	$8,1 \pm 6,0$	$12,3 \pm 5,1$	$14,8 \pm 6,7$	$24,6 \pm 10,2$	$38,9 \pm 15,1$
	Sim. Co-senos	$\mathbf{0,62 \pm 0,29}$	$\mathbf{0,82 \pm 0,23}$	$\mathbf{0,79 \pm 0,25}$	$\mathbf{0,81 \pm 0,20}$	$\mathbf{0,52 \pm 0,31}$
	FID	102,28	84,04	82,75	96,31	104,35
Instruct-pix2pix	**Idade predita**	$19,8 \pm 6,6$	$28,7 \pm 4,7$	$39,4 \pm 5,6$	$51,1 \pm 5,7$	$21,2 \pm 7,1$
	Erro Médio	$10,2 \pm 6,6$	$11,3 \pm 4,7$	$10,6 \pm 5,6$	$18,9 \pm 5,7$	$48,8 \pm 7,1$
	Erro Abs. Médio	$11,2 \pm 4,6$	$11,3 \pm 4,7$	$10,6 \pm 5,5$	$18,9 \pm 5,7$	$48,8 \pm 7,1$
	Sim. Co-senos	$\mathbf{0,93 \pm 0,03}$	$\mathbf{0,94 \pm 0,02}$	$\mathbf{0,93 \pm 0,03}$	$\mathbf{0,92 \pm 0,07}$	$\mathbf{0,92 \pm 0,03}$
	FID	45,43	38,89	40,74	42,52	52,47

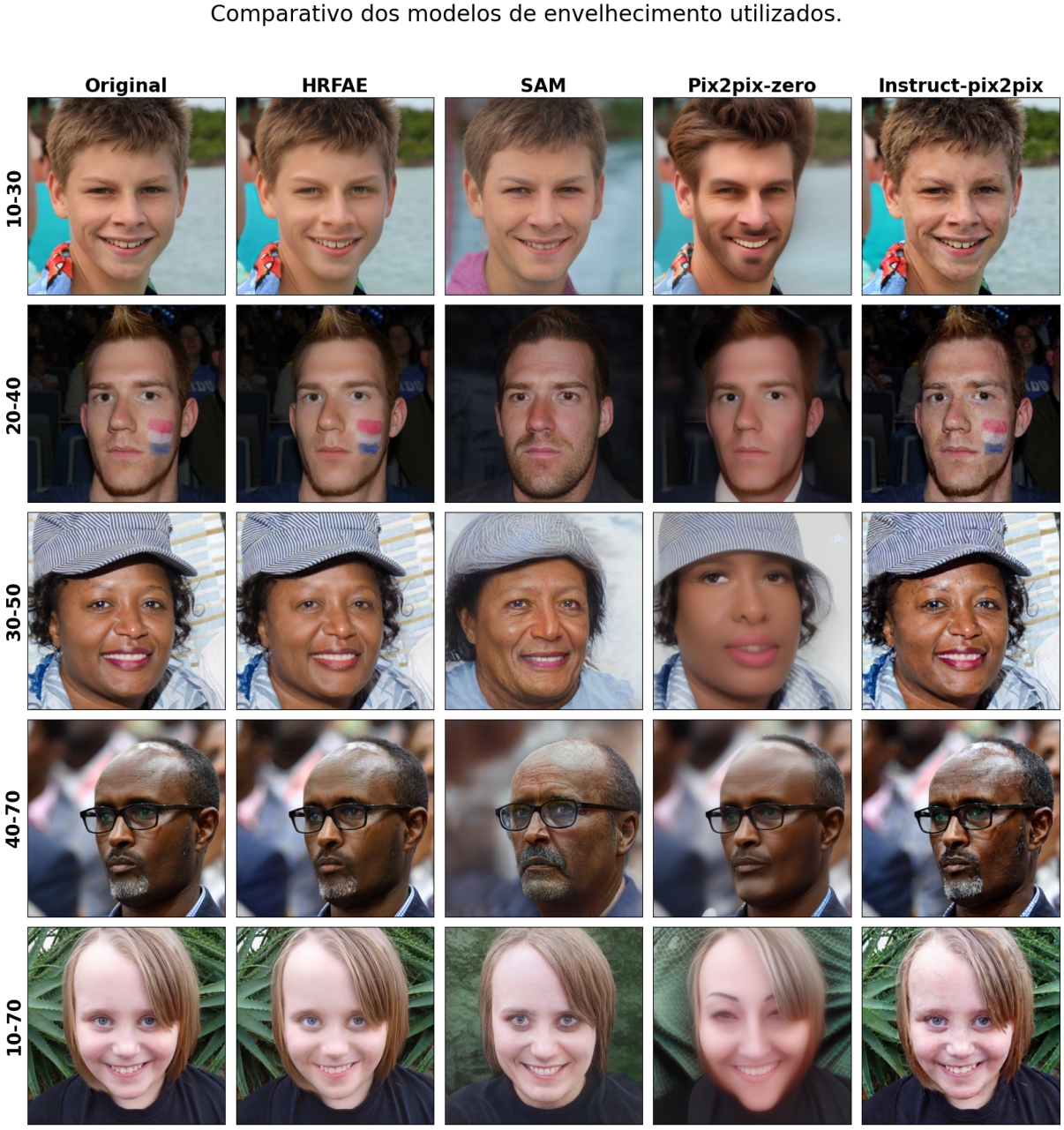
Erro Médio = Idade esperada - Idade predita | Sim. Co-senos = Similaridade de co-senos | c = $c_{instrução}$ | Idade predita = média da predição das idades | FID = Distância de Fréchenet, como essa métrica é calculada utilizando estatísticas da amostra, não temos o desvio padrão dela.

Observando as Figuras 22 e 23, o modelo SAM é o único que consegue levar a imagem original com aproximadamente 10 anos, para idades mais velhas. Como é uma fase de crescimento, existem muitas mudanças na face e os outros modelos não conseguiram capturar essas mudanças estruturais da face. O modelo *Pix2pix-zero* teve resultados esfumaçados em algumas imagens, o que é esperado por usar uma direção de edição média para levar de um domínio para o outro. *Instruct-pix2pix* manteve a identidade dos indivíduos, e obteve envelhecimento considerável, porém, em todos os casos degradou a textura da pele, algo que é comum ao passar dos anos (principalmente nos casos em que se tem muita incidência solar ou tabagismo), mas não é garantido que o envelhecimento será assim. O modelo SAM parece ter mais variabilidade no modo como envelhece as faces.

Figura 21 – Exemplos de imagens em que a etapa de reconstrução da imagens original usando a técnica de inversão DDIM não obteve bons resultados, podendo comprometer a qualidade das imagens envelhecidas geradas.

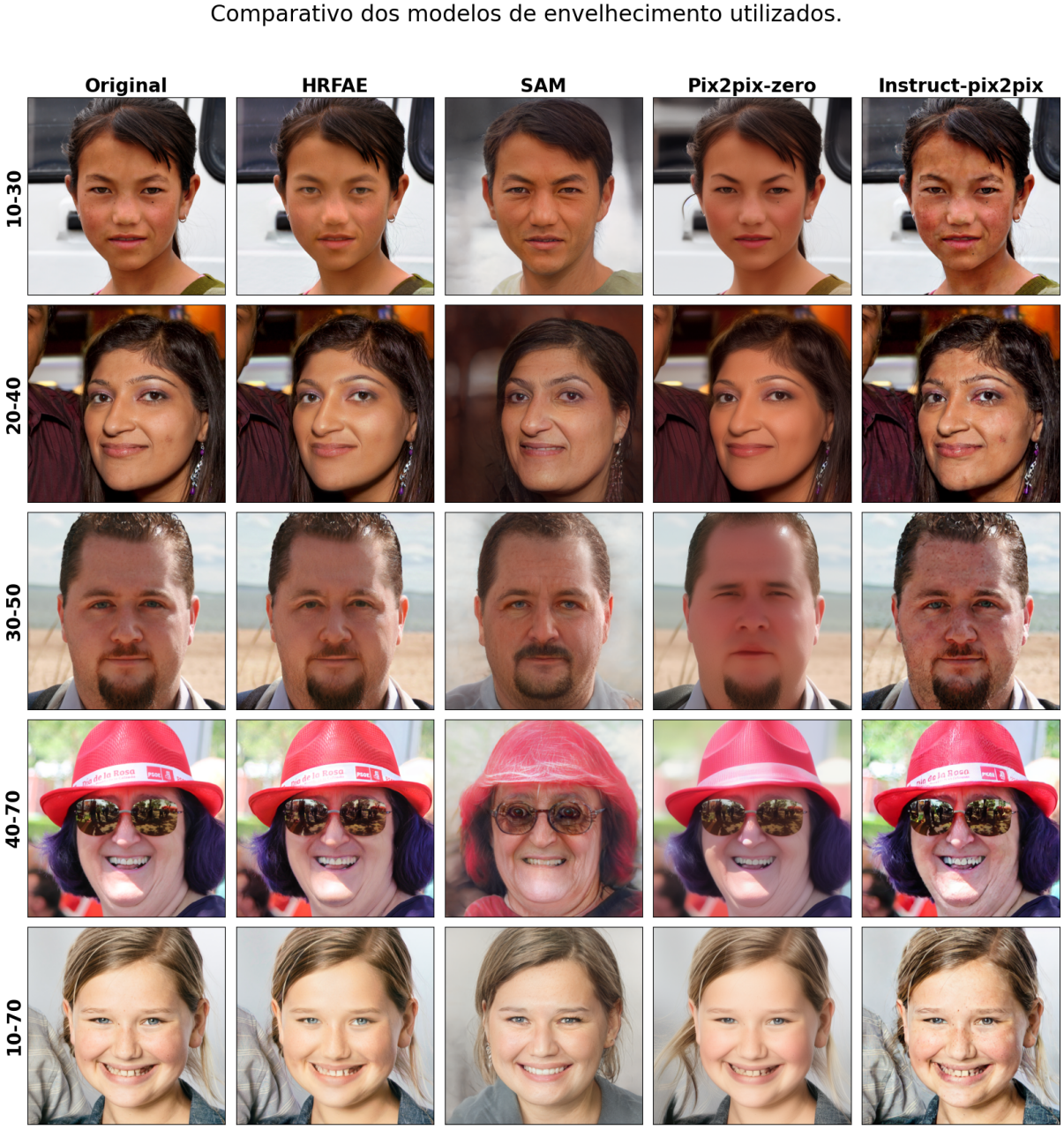
Fonte: Bruno Abreu Kemmer, 2023

Figura 22 – Comparativo dos resultados nas tarefas de envelhecimento facial em cada modelo. A primeira coluna Original, é a imagem de entrada com idade estimada inicial e as 4 imagens a sua direita são os resultados dos modelos tentando obter a imagem envelhecida com a idade esperada, por exemplo, 10-30 é a tarefa de envelhecer uma face com aproximadamente 10 anos para 30 anos.



Fonte: Bruno Abreu Kemmer, 2023

Figura 23 – Comparativo dos resultados nas tarefas de envelhecimento facial em cada modelo. A primeira coluna Original, é a imagem de entrada com idade estimada inicial e as 4 imagens a sua direita são os resultados dos modelos tentando obter a imagem envelhecida com a idade esperada.



Fonte: Bruno Abreu Kemmer, 2023

8 Conclusão e Trabalhos Futuros

8.1 Conclusão

Nessa dissertação foram apresentados os principais modelos generativos utilizados na tarefa de envelhecimento facial, as bases de dados relevantes disponíveis para treinamento, métricas de avaliação dos resultados e dois estudos comparativos.

O primeiro estudo comparativo foi efetuado com imagens em baixa resolução com um *autoencoder* adversário condicional, CAAE, e duas redes adversárias generativas (GANs), IPCGAN e RCRIIT, envelhecendo esses modelos para idades esperadas e comparando os resultados obtidos. Pode-se notar os ganhos que IPCGAN teve ao adicionar um estimador de idade na função de custo de sua GAN, elemento que foi replicado em vários estudos posteriores. Além disso, RCRIIT, obteve resultados satisfatórios na época de sua publicação, mas com degradação de performance quando a distancia entre a idade esperada e a atual aumentava.

O segundo estudo comparativo, que utilizou imagens de alta definição, analisou a performance, também na tarefa de envelhecimento facial quatro modelos: dois modelos de arquitetura que utilizavam GANs, HRFAE e SAM, e dois modelos de difusão condicionais que possibilitavam a edição de imagens, *Pix2pix-zero* e *Instruct-pix2pix*. É notável a evolução que os modelos apresentaram, comparado aos modelos do estudo anterior. Desses, dois modelos se sobressaíram nessa análise, SAM, um modelo baseado em uma GAN pré-treinada (StyleGAN2) e especializado na tarefa de envelhecimento e rejuvenescimento facial, tendo componentes de sua função de custo específicas para garantir a identidade e a mudança de idade. Esse foi o modelo que conseguiu atingir os menores erros entre a idade esperada e a medida por estimadores de idade pré-treinados, conseguindo bons resultados, por exemplo levando uma face com idade estimada de 10 anos para 60 anos. Porém, com uma pequena perda de identidade, medida pela similaridade de co-senos utilizada, comparando com o método *Instruct-pix2pix*, em que essa métrica obteve quase similaridade total. O segundo modelo que se destacou foi *Instruct-pix2pix*, um modelo de difusão condicional genérico, esse apenas recebe uma instrução em forma de texto, podendo editar as imagens de diversas formas. No estudo, foi somente enviada uma instrução para envelhecer a face para uma idade desejada, e mesmo assim conseguiu imagens envelhecidas realistas. Todavia, o modelo se valeu da abordagem de aplicar um

filtro na face no envelhecimento, degradando a textura da pele, algo que é comum ao passar dos anos, mas não é garantido que o envelhecimento ocorrerá assim. O modelo SAM demonstrou ter mais variabilidade no modo como envelhece e rejuvenesce as faces.

HRFAE não efetuou mudanças consideráveis nas faces, o que manteve a identidade, porém, não resolvendo a tarefa de envelhecimento ou rejuvenescimento. *Pix2pix-zero* necessitava de uma etapa de reconstrução DDIM a qual em alguns casos adicionava erros de reconstrução, os quais eram evidenciados ao aplicar a direção de edição que efetuava o envelhecimento. Tendo alguns casos que em que não era possível reconstruir a imagem original. Além disso, o modelo teve resultados esfumaçados em algumas imagens, o que é esperado por usar uma direção de edição média para levar de um domínio para o outro.

Atualmente, GANs especialistas ainda demonstram uma performance superior comparadas a modelos de difusão genéricos, como os utilizados nos experimentos. Isso demonstra uma grande capacidade nos modelo de difusão já que *Instruct-pix2pix* obteve resultados consideráveis mesmo não tendo sido treinado especialmente para essa tarefa. Além do mais, modelos de difusão ganharam notoriedade recentemente, com muitos trabalhos apresentando novas técnicas nos últimos anos, portanto, são esperados rápidos avanços em suas arquiteturas.

8.2 Trabalhos Futuros

As seguintes abordagens poderiam ser tentadas em trabalhos futuros, algumas delas demandam um custo computacional elevado:

- Treinar um modelo CLIP específico para a tarefa de estimação de idade. Nele, em vez de codificar o texto de entrada e maximizar a semelhança com a legenda recebida (como explicado na Seção 2.3.3.1), maximizar a semelhança entre características conhecidas que afetam o envelhecimento (por exemplo, sexo e etnia) e uma faixa de idades de 0 a 100 anos, em que cada ano seria um *token*. O modelo teria como objetivo maximizar o produto vetorial da face com essas variáveis binárias e seria esperado que encontrasse uma representação relevante nesse domínio. Algumas dificuldades para alcançar esse fim seriam juntar múltiplas bases de imagens com idades anotadas, isolar faces para não ter mais de uma por imagem, e o treinamento do modelo por CLIP utilizou 400 milhões de pares de imagens e legendas para obter a representação

que tem atualmente. Uma forma de mitigar essa questão, seria partir do modelo CLIP e sobre-ajustar os pesos da rede para a tarefa descrita.

- Gerar mais pares de imagens seguindo o protocolo descrito em *Instruct-pix2pix* (Seção 2.4.7) com instruções para envelhecer e rejuvenescer faces e retreinar o modelo como os autores fizeram, iniciando do modelo *Stable diffusion* já treinado ou partindo do modelo *Instruct-pix2pix* disponibilizado pelos autores.
- Treinar um modelo de difusão condicionado à idade e às características que afetam o envelhecimento, descritas acima (sexo e etnia), adicionando as funções de custo da rede a diferença entre a idade esperada e a alcançada (usando uma rede estimadora de idade previamente treinada) e um modelo de verificação facial pré-treinado. Novamente é interessante pontuar que modelos de difusão demandam um alto custo computacional para obterem imagens de boa qualidade.

Referências

ALALUF, Y.; PATASHNIK, O.; COHEN-OR, D. Only a matter of style: Age transformation using a style-based regression model. *ACM Trans. Graph.*, Association for Computing Machinery, v. 40, n. 4, 2021. Citado 8 vezes nas páginas 17, 50, 51, 52, 67, 69, 70 e 82.

ALQAHTANI, H.; KAVAKLI-THORNE, M.; KUMAR, G. Applications of generative adversarial networks (gans): An updated review. *Archives of Computational Methods in Engineering*, 2019. Citado na página 18.

AMINI, A.; SOLEIMANY, A. P.; SCHWARTING, W.; BHATIA, S. N.; RUS, D. Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. Honolulu HI USA: ACM, 2019. p. 289–295. ISBN 978-1-4503-6324-2. Citado na página 27.

ANTIPOV, G.; BACCOUCHE, M.; DUGELAY, J.-L. Boosting cross-age face verification via generative age normalization. In: . [S.l.: s.n.], 2018. v. 2018-January, p. 191–199. Citado na página 114.

ANTIPOV, G.; BACCOUCHE, M.; DUGELAY, J.-L. Face aging with conditional generative adversarial networks. In: . [S.l.: s.n.], 2018. v. 2017-September, p. 2089–2093. Citado 2 vezes nas páginas 34 e 114.

ANWAAR, M.; LOO, C.; SEERA, M. Face image synthesis with weight and age progression using conditional adversarial autoencoder. *Neural Computing and Applications*, v. 32, n. 8, p. 3567–3579, 2020. Citado na página 113.

ARJOVSKY, M.; CHINTALA, S.; BOTTOU, L. Wasserstein generative adversarial networks. In: PRECUP, D.; TEH, Y. W. (Ed.). *Proceedings of the 34th International Conference on Machine Learning*. International Convention Centre, Sydney, Australia: PMLR, 2017. (Proceedings of Machine Learning Research, v. 70), p. 214–223. Citado na página 30.

BARRATT, S.; SHARMA, R. *A Note on the Inception Score*. 2018. Citado 2 vezes nas páginas 67 e 107.

BAYRAMLI, B.; ALI, U.; QI, T.; LU, H. Fh-gan: Face hallucination and recognition using generative adversarial network. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 11953 LNCS, p. 3–15, 2019. Citado 2 vezes nas páginas 60 e 114.

BROOKS, T.; HOLYNSKI, A.; EFROS, A. A. *InstructPix2Pix: Learning to Follow Image Editing Instructions*. [S.l.]: arXiv, 2022. Citado 3 vezes nas páginas 53, 70 e 82.

BROWN, T. B.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A.; AGARWAL, S.; HERBERT-VOSS, A.; KRUEGER, G.; HENIGHAN, T.; CHILD, R.; RAMESH, A.; ZIEGLER, D. M.; WU, J.; WINTER, C.; HESSE, C.; CHEN, M.; SIGLER, E.; LITWIN, M.; GRAY, S.; CHESS, B.; CLARK, J.; BERNER, C.; MCCANDLISH, S.; RADFORD, A.; SUTSKEVER, I.; AMODEI, D. Language models are few-shot learners. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red

Hook, NY, USA: Curran Associates Inc., 2020. (NIPS'20). ISBN 9781713829546. Citado na página 53.

CAO, W.; MIRJALILI, V.; RASCHKA, S. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, v. 140, p. 325 – 331, 2020. ISSN 0167-8655. Citado 2 vezes nas páginas 73 e 81.

CHANDALIYA, P.; NAIN, N. Conditional perceptual adversarial variational autoencoder for age progression and regression on child face. In: . [S.l.: s.n.], 2019. Citado 2 vezes nas páginas 27 e 113.

CHEN, L.; HU, X.; ZHANG, Z. Face aging with boundary equilibrium conditional autoencoder. *IEEE Access*, v. 6, p. 54834–54843, 2018. Citado na página 114.

CHEN, S.; ZHANG, D.; YANG, L.; CHEN, P. Age-invariant face recognition based on sample enhancement of generative adversarial networks. In: . [S.l.: s.n.], 2019. p. 388–392. Citado na página 113.

DENG, J.; GUO, J.; XUE, N.; ZAFEIRIOU, S. Arcface: Additive angular margin loss for deep face recognition. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2019. p. 4685–4694. Citado 4 vezes nas páginas 51, 69, 70 e 81.

DUONG, C.; QUACH, K.; LUU, K.; LE, T.; SAVVIDES, M. Temporal non-volume preserving approach to facial age-progression and age-invariant face recognition. In: . [S.l.: s.n.], 2017. v. 2017-October, p. 3755–3763. Citado na página 114.

FANG, H.; DENG, W.; ZHONG, Y.; HU, J. Triple-gan: Progressive face aging with triple translation loss. In: . [S.l.: s.n.], 2020. v. 2020-June, p. 3500–3509. Citado na página 113.

GAL, R.; ALALUF, Y.; ATZMON, Y.; PATASHNIK, O.; BERMANO, A. H.; CHECHIK, G.; COHEN-OR, D. *An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion*. [S.l.]: arXiv, 2022. Citado na página 45.

GENOVESE, A.; PIURI, V.; SCOTTI, F. Towards explainable face aging with generative adversarial networks. In: . [S.l.: s.n.], 2019. v. 2019-September, p. 3806–3810. Citado na página 113.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. ⟨http://www.deeplearningbook.org⟩. Citado na página 23.

GOODFELLOW, I.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; BENGIO, Y. Generative Adversarial Nets. p. 9, 2014. Citado 2 vezes nas páginas 28 e 35.

GOU, D.; ZHANG, S.; NING, X.; WANG, W. A face aging network based on conditional cycle loss and the principle of homology continuity. In: . [S.l.: s.n.], 2019. p. 264–268. Citado na página 114.

GRIMMER, M.; RAMACHANDRA, R.; BUSCH, C. Deep face age progression: A survey. *IEEE Access*, v. 9, p. 83376–83393, 2021. Citado 4 vezes nas páginas 17, 18, 69 e 70.

GULRAJANI, I.; AHMED, F.; ARJOVSKY, M.; DUMOULIN, V.; COURVILLE, A. C. Improved training of wasserstein gans. *CoRR*, abs/1704.00028, 2017. Citado na página 57.

GUO, X.; LIU, X.; ZHU, E.; YIN, J. Deep Clustering with Convolutional Autoencoders. In: LIU, D.; XIE, S.; LI, Y.; ZHAO, D.; EL-ALFY, E.-S. M. (Ed.). *Neural Information Processing*. Cham: Springer International Publishing, 2017. v. 10635. ISBN 978-3-319-70095-3 978-3-319-70096-0. Citado na página 25.

GÉRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. 2. ed. [S.l.]: O'Reilly Media, 2019. 1150 p. Citado na página 27.

HAMZAH, N.; ZAMAN, F. H. K. Face aging on realistic photo in cross-dataset implementation. In: . [S.l.: s.n.], 2020. v. 917, n. 1. Citado na página 113.

HELJAKKA, A.; SOLIN, A.; KANNALA, J. Recursive chaining of reversible image-to-image translators for face aging. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 11182 LNCS, p. 309–320, 2018. Citado 4 vezes nas páginas 17, 48, 72 e 114.

HERTZ, A.; MOKADY, R.; TENENBAUM, J.; ABERMAN, K.; PRITCH, Y.; COHEN-OR, D. *Prompt-to-Prompt Image Editing with Cross Attention Control*. [S.l.]: arXiv, 2022. Citado na página 55.

HEUSEL, M.; RAMSAUER, H.; UNTERTHINER, T.; NESSLER, B.; HOCHREITER, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 6629–6640. ISBN 9781510860964. Citado na página 67.

HO, J.; JAIN, A.; ABBEEL, P. Denoising diffusion probabilistic models. In: LAROCHELLE, H.; RANZATO, M.; HADSELL, R.; BALCAN, M.; LIN, H. (Ed.). *Advances in Neural Information Processing Systems*. [S.l.]: Curran Associates, Inc., 2020. v. 33, p. 6840–6851. Citado na página 40.

HO, J.; SALIMANS, T. *Classifier-Free Diffusion Guidance*. [S.l.]: arXiv, 2022. Citado na página 42.

HUANG, G. B.; RAMESH, M.; BERG, T.; LEARNED-MILLER, E. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. [S.l.], 2007. Citado na página 64.

HUANG, X.; BELONGIE, S. Arbitrary style transfer in real-time with adaptive instance normalization. In: . [S.l.: s.n.], 2019. Cited By 1. Citado na página 39.

HUANG, Y.; CHEN, W.; HU, H. Age-puzzle facenet for cross-age face recognition. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 11366 LNCS, p. 603–619, 2019. Citado 2 vezes nas páginas 61 e 114.

HUANG, Y.; HU, H. A parallel architecture of age adversarial convolutional neural network for cross-age face recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 31, n. 1, p. 148–159, 2021. Citado 2 vezes nas páginas 17 e 113.

INC, M. *Face++ research toolkit*. 2017. ⟨www.faceplusplus.com⟩. Accessed: 2021-01-31. Citado na página 60.

ISOLA, P.; ZHU, J.-Y.; ZHOU, T.; EFROS, A. A. Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2017. Citado 4 vezes nas páginas 31, 34, 36 e 49.

JIA, L.; SONG, Y.; ZHANG, Y. Face aging with improved invertible conditional gans. In: . [S.l.: s.n.], 2018. v. 2018-August, p. 1396–1401. Citado na página 114.

KARRAS, T.; AILA, T.; LAINE, S.; LEHTINEN, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv:1710.10196 [cs, stat]*, fev. 2018. ArXiv: 1710.10196. Citado 2 vezes nas páginas 38 e 64.

KARRAS, T.; LAINE, S.; AILA, T. A Style-Based Generator Architecture for Generative Adversarial Networks. *arXiv:1812.04948 [cs, stat]*, mar. 2019. ArXiv: 1812.04948. Citado 4 vezes nas páginas 38, 56, 65 e 68.

KARRAS, T.; LAINE, S.; AITTALA, M.; HELLSTEN, J.; LEHTINEN, J.; AILA, T. Analyzing and Improving the Image Quality of StyleGAN. *arXiv:1912.04958 [cs, eess, stat]*, mar. 2020. ArXiv: 1912.04958. Citado 2 vezes nas páginas 39 e 56.

KEMMER, B.; SIMÕES, R.; LIMA, C. Face aging using generative adversarial networks. In: _____. *Generative Adversarial Learning: Architectures and Applications*. Cham: Springer International Publishing, 2022. p. 145–168. ISBN 978-3-030-91390-8. Citado 8 vezes nas páginas 19, 24, 29, 32, 37, 47, 48 e 72.

KHANNA, A.; THAKUR, A.; TEWARI, A.; BHAT, A. Cross-age face verification using face aging. In: . [S.l.: s.n.], 2020. p. 94–99. Citado 2 vezes nas páginas 17 e 113.

KING, D. E.; SONNENBURG, S. *Dlib-ml: A Machine Learning Toolkit*. Citado na página 65.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. Citado na página 73.

KINGMA, D. P.; WELLING, M. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*, maio 2014. Citado na página 25.

KITCHENHAM, B. Procedures for Performing Systematic Reviews. p. 33, jul. 2004. Citado na página 106.

KRIZHEVSKY, A. *Learning multiple layers of features from tiny images*. [S.l.], 2009. Citado na página 35.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: PEREIRA, F.; BURGES, C. J. C.; BOTTOU, L.; WEINBERGER, K. Q. (Ed.). *Advances in Neural Information Processing Systems*. [S.l.]: Curran Associates, Inc., 2012. v. 25, p. 1097–1105. Citado na página 58.

LAMPLE, G.; ZEGHIDOUR, N.; USUNIER, N.; BORDES, A.; DENOYER, L.; RANZATO, M. Fader networks: Manipulating images by sliding attributes. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 5969–5978. ISBN 9781510860964. Citado na página 70.

LANITIS, A.; TAYLOR, C. J.; COOTES, T. F. Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 24, n. 4, p. 442–455, abr. 2002. ISSN 1939-3539. Citado na página 62.

LECUN, Y.; CORTES, C.; BURGES, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, v. 2, 2010. Citado na página 35.

LI, J.; LI, D.; XIONG, C.; HOI, S. *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. [S.l.]: arXiv, 2022. Citado na página 53.

LI, P.; HU, Y.; HE, R.; SUN, Z. Global and local consistent wavelet-domain age synthesis. *IEEE Transactions on Information Forensics and Security*, v. 14, n. 11, p. 2943–2957, 2019. Citado 2 vezes nas páginas 58 e 113.

LI, P.; HU, Y.; LI, Q.; HE, R.; SUN, Z. Global and local consistent age generative adversarial networks. In: . [S.l.: s.n.], 2018. v. 2018-August, p. 1073–1078. Citado 2 vezes nas páginas 58 e 114.

LI, P.; HUANG, H.; HU, Y.; WU, X.; HE, R.; SUN, Z. Hierarchical face aging through disentangled latent characteristics. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 12348 LNCS, p. 86–101, 2020. Citado na página 113.

LIU, L.; YU, H.; WANG, S.; WAN, L.; HAN, S. Learning shape and texture progression for young child face aging. *Signal Processing: Image Communication*, v. 93, 2021. Citado na página 113.

LIU, S.; SUN, Y.; ZHU, D.; BAO, R.; WANG, W.; SHU, X.; YAN, S. Face aging with contextual generative adversarial nets. In: . [S.l.: s.n.], 2017. p. 82–90. Citado 2 vezes nas páginas 58 e 114.

LIU, X.; XIE, C.; KUANG, H.; MA, X. Face aging simulation with deep convolutional generative adversarial networks. In: . [S.l.: s.n.], 2018. v. 2018-January, p. 220–224. Citado na página 114.

LIU, X.; ZOU, Y.; XIE, C.; KUANG, H.; MA, X. Bidirectional face aging synthesis based on improved deep convolutional generative adversarial networks. *Information (Switzerland)*, v. 10, n. 2, 2019. Citado 2 vezes nas páginas 35 e 114.

LIU, Z.; LUO, P.; WANG, X.; TANG, X. Deep learning face attributes in the wild. In: *Proceedings of International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2015. Citado 2 vezes nas páginas 63 e 64.

MADHUKAR, P.; CHETAN, R.; PRASAD, S.; SHAYAN, M.; KRUPA, B. Age progression using generative adversarial networks. In: . [S.l.: s.n.], 2020. v. 2020-November, p. 1249–1254. Citado na página 113.

MAKHZANI, A.; SHLENS, J.; JAITLY, N.; GOODFELLOW, I. Adversarial autoencoders. In: *International Conference on Learning Representations*. [S.l.: s.n.], 2016. Citado na página 27.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y. K.; Wang, Z.; Smolley, S. P. Least squares generative adversarial networks. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2017. p. 2813–2821. ISSN 2380-7504. Citado 2 vezes nas páginas 49 e 58.

MASCI, J.; MEIER, U.; CIREşAN, D.; SCHMIDHUBER, J. Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. In: HONKELA, T.; DUCH, W.; GIROLAMI, M.; KASKI, S. (Ed.). *Artificial Neural Networks and Machine Learning – ICANN 2011*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. v. 6791, p. 52–59. ISBN 978-3-642-21734-0 978-3-642-21735-7. Series Title: Lecture Notes in Computer Science. Citado 2 vezes nas páginas 24 e 25.

MIRZA, M.; OSINDERO, S. Conditional Generative Adversarial Nets. *arXiv:1411.1784 [cs, stat]*, nov. 2014. ArXiv: 1411.1784. Citado na página 34.

MOKADY, R.; HERTZ, A.; ABERMAN, K.; PRITCH, Y.; COHEN-OR, D. *Null-text Inversion for Editing Real Images using Guided Diffusion Models*. [S.l.]: arXiv, 2022. Citado na página 43.

Moschoglou, S.; Papaioannou, A.; Sagonas, C.; Deng, J.; Kotsia, I.; Zafeiriou, S. Agedb: The first manually collected, in-the-wild age database. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. [S.l.: s.n.], 2017. p. 1997–2005. Citado na página 64.

NICHOL, A.; DHARIWAL, P. *Improved Denoising Diffusion Probabilistic Models*. [S.l.]: arXiv, 2021. Citado na página 40.

OR-EL, R.; SENGUPTA, S.; FRIED, O.; SHECHTMAN, E.; KEMELMACHER-SHLIZERMAN, I. Lifespan age transformation synthesis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 12351 LNCS, p. 739–755, 2020. Citado 3 vezes nas páginas 39, 65 e 113.

OR-EL, R.; SENGUPTA, S.; FRIED, O.; SHECHTMAN, E.; KEMELMACHER-SHLIZERMAN, I. Lifespan age transformation synthesis. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. [S.l.: s.n.], 2020. Citado 3 vezes nas páginas 67, 69 e 70.

ORRù, G.; MARCIALIS, G. L.; ROLI, F. A novel classification-selection approach for the self updating of template-based face recognition systems. *Pattern Recognition*, v. 100, p. 107121, abr. 2020. ISSN 00313203. Citado na página 61.

PALSSON, S.; AGUSTSSON, E.; TIMOFTE, R.; GOOL, L. V. Generative adversarial style transfer networks for face aging. In: . [S.l.: s.n.], 2018. v. 2018-June, p. 2165–2173. Citado 3 vezes nas páginas 17, 37 e 114.

PANTRAKI, E.; KOTROPOULOS, C.; LANITIS, A. Leveraging image-to-image translation generative adversarial networks for face aging. In: . [S.l.: s.n.], 2019. v. 2019-May, p. 8370–8374. Citado 2 vezes nas páginas 27 e 114.

PARMAR, G.; SINGH, K. K.; ZHANG, R.; LI, Y.; LU, J.; ZHU, J.-Y. *Zero-shot Image-to-Image Translation*. [S.l.]: arXiv, 2023. Citado 4 vezes nas páginas 52, 54, 70 e 82.

PHAM, Q.; YANG, J.; SHIN, J. Semi-supervised facegan for face-age progression and regression with synthesized paired images. *Electronics (Switzerland)*, v. 9, n. 4, 2020. Citado na página 113.

RADFORD, A.; KIM, J. W.; HALLACY, C.; RAMESH, A.; GOH, G.; AGARWAL, S.; SASTRY, G.; ASKELL, A.; MISHKIN, P.; CLARK, J.; KRUEGER, G.; SUTSKEVER, I. *Learning Transferable Visual Models From Natural Language Supervision*. [S.l.]: arXiv, 2021. Citado 2 vezes nas páginas 45 e 46.

RADFORD, A.; METZ, L.; CHINTALA, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In: . [S.l.: s.n.], 2016. Citado na página 35.

RICANEK, K.; TESAFAYE, T. MORPH: a longitudinal image database of normal adult age-progression. In: . [S.l.: s.n.], 2006. p. 341–345. Citado na página 63.

RICHARDSON, E.; ALALUF, Y.; PATASHNIK, O.; NITZAN, Y.; AZAR, Y.; SHAPIRO, S.; COHEN-OR, D. Encoding in style: A stylegan encoder for image-to-image translation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2021. p. 2287–2296. Citado 2 vezes nas páginas 33 e 50.

ROMBACH, R.; BLATTMANN, A.; LORENZ, D.; ESSER, P.; OMMER, B. *High-Resolution Image Synthesis with Latent Diffusion Models*. [S.l.]: arXiv, 2021. Citado 3 vezes nas páginas 41, 43 e 53.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: NAVAB, N.; HORNEGGER, J.; WELLS, W. M.; FRANGI, A. F. (Ed.). *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing, 2015. p. 234–241. ISBN 978-3-319-24574-4. Citado 2 vezes nas páginas 36 e 41.

ROTHE, R.; TIMOFTE, R.; GOOL, L. V. Dex: Deep expectation of apparent age from a single image. In: *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. [S.l.: s.n.], 2015. p. 252–257. Citado 3 vezes nas páginas 49, 51 e 81.

ROTHE, R.; TIMOFTE, R.; GOOL, L. V. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, Springer, v. 126, n. 2-4, p. 144–157, 2018. Citado 2 vezes nas páginas 63 e 64.

RUIZ, N.; LI, Y.; JAMPANI, V.; PRITCH, Y.; RUBINSTEIN, M.; ABERMAN, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022. Citado na página 44.

SAHARIA, C.; CHAN, W.; CHANG, H.; LEE, C. A.; HO, J.; SALIMANS, T.; FLEET, D. J.; NOROUZI, M. *Palette: Image-to-Image Diffusion Models*. [S.l.]: arXiv, 2021. Citado na página 44.

SAHARIA, C.; CHAN, W.; SAXENA, S.; LI, L.; WHANG, J.; DENTON, E.; GHASEMIPOUR, S. K. S.; AYAN, B. K.; MAHDAVI, S. S.; LOPES, R. G.; SALIMANS, T.; HO, J.; FLEET, D. J.; NOROUZI, M. *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*. [S.l.]: arXiv, 2022. Citado na página 44.

SAJID, M.; SHAFIQUE, T. Hybrid generative-discriminative approach to age-invariant face recognition. *Journal of Electronic Imaging*, v. 27, n. 2, 2018. Citado na página 114.

SALIMANS, T.; GOODFELLOW, I.; ZAREMBA, W.; CHEUNG, V.; RADFORD, A.; CHEN, X. Improved techniques for training gans. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2016. (NIPS'16), p. 2234–2242. ISBN 9781510838819. Citado na página 66.

SCHROFF, F.; KALENICHENKO, D.; PHILBIN, J. Facenet: A unified embedding for face recognition and clustering. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2015. p. 815–823. Citado na página 81.

SCHUHMANN, C.; VENCU, R.; BEAUMONT, R.; KACZMARCZYK, R.; MULLIS, C.; KATTA, A.; COOMBES, T.; JITSEV, J.; KOMATSUZAKI, A. *LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs*. [S.l.]: arXiv, 2021. Citado na página 41.

SHARMA, N.; SHARMA, R.; JINDAL, N. An improved technique for face age progression and enhanced super-resolution with generative adversarial networks. *Wireless Personal Communications*, v. 114, n. 3, p. 2215–2233, 2020. Citado 2 vezes nas páginas 57 e 113.

SHEN, Y.; GU, J.; TANG, X.; ZHOU, B. Interpreting the latent space of gans for semantic face editing. In: . [S.l.: s.n.], 2020. p. 9240–9249. Citado 2 vezes nas páginas 38 e 113.

SHENG, M.; MA, Z.; JIA, H.; MAO, Q.; DONG, M. Face aging with conditional generative adversarial network guided by ranking-cnn. In: . [S.l.: s.n.], 2020. p. 314–319. Citado na página 113.

SHI, C.; ZHANG, J.; YAO, Y.; SUN, Y.; RAO, H.; SHU, X. Can-gan: Conditioned-attention normalized gan for face age synthesis. *Pattern Recognition Letters*, v. 138, p. 520–526, 2020. Citado 3 vezes nas páginas 57, 58 e 113.

SHU, X.; XIE, G.-S.; LI, Z.; TANG, J. Age progression: Current technologies and applications. *Neurocomputing*, v. 208, p. 249–261, out. 2016. ISSN 09252312. Citado na página 18.

SHU, X.; XIE, G.-S.; LI, Z.; TANG, J. Age progression: Current technologies and applications. *Neurocomputing*, v. 208, p. 249–261, 2016. ISSN 0925-2312. Citado na página 63.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. In: BENGIO, Y.; LECUN, Y. (Ed.). *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. [S.l.: s.n.], 2015. Citado na página 51.

SOHL-DICKSTEIN, J.; WEISS, E.; MAHESWARANATHAN, N.; GANGULI, S. Deep unsupervised learning using nonequilibrium thermodynamics. In: BACH, F.; BLEI, D. (Ed.). *Proceedings of the 32nd International Conference on Machine Learning*. Lille, France: PMLR, 2015. (Proceedings of Machine Learning Research, v. 37), p. 2256–2265. Citado na página 40.

SONG, J.; MENG, C.; ERMON, S. Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020. Citado 2 vezes nas páginas 40 e 43.

SONG, J.; ZHANG, J.; GAO, L.; LIU, X.; SHEN, H. Dual conditional gans for face aging and rejuvenation. In: . [S.l.: s.n.], 2018. v. 2018-July, p. 899–905. Citado na página 114.

SUN, Y.; TANG, J.; SHU, X.; SUN, Z.; TISTARELLI, M. Facial age synthesis with label distribution-guided generative adversarial network. *IEEE Transactions on Information Forensics and Security*, v. 15, p. 2679–2691, 2020. Citado na página 113.

SUN, Y.; TANG, J.; SUN, Z.; TISTARELLI, M. Facial age and expression synthesis using ordinal ranking adversarial networks. *IEEE Transactions on Information Forensics and Security*, v. 15, p. 2960–2972, 2020. Citado na página 113.

SUSSKIND, J. M.; ANDERSON, A. K.; HINTON, G. E. The toronto face database. *Department of Computer Science, University of Toronto, Toronto, ON, Canada, Tech. Rep*, v. 3, 2010. Citado na página 35.

THENGANE, V.; GAWANDE, M.; DUDHANE, A.; GONDE, A. Cycle face aging generative adversarial networks. In: . [S.l.: s.n.], 2018. p. 125–129. Citado na página 114.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, ; POLOSUKHIN, I. Attention is All you Need. p. 11, 2017. Citado na página 58.

VIAZOVETSKYI, Y.; IVASHKIN, V.; KASHIN, E. Stylegan2 distillation for feed-forward image manipulation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 12367 LNCS, p. 170–186, 2020. Citado 3 vezes nas páginas 33, 39 e 113.

WALLACE, B.; GOKUL, A.; NAIK, N. *EDICT: Exact Diffusion Inversion via Coupled Transformations*. [S.l.]: arXiv, 2022. Citado na página 43.

WANG, C.-C.; LIU, H.-H.; PEI, S.-C.; LIU, K.-H.; LIU, T.-J. Face aging on realistic photos by generative adversarial networks. In: . [S.l.: s.n.], 2019. v. 2019-May. Citado na página 114.

WANG, S.; DING, Z.; FU, Y. Cross-generation kinship verification with sparse discriminative metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 41, n. 11, p. 2783–2790, 2019. Citado na página 113.

WANG, T.-C.; LIU, M.-Y.; ZHU, J.-Y.; TAO, A.; KAUTZ, J.; CATANZARO, B. High-resolution image synthesis and semantic manipulation with conditional gans. In: . [S.l.: s.n.], 2018. p. 8798–8807. Cited By 1140. Citado na página 36.

WANG, W.; YAN, Y.; CUI, Z.; FENG, J.; YAN, S.; SEBE, N. Recurrent face aging with hierarchical autoregressive memory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 41, n. 3, p. 654–668, 2019. Citado na página 114.

WANG, X.; YU, K.; WU, S.; GU, J.; LIU, Y.; DONG, C.; LOY, C. C.; QIAO, Y.; TANG, X. *ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks*. 2018. Citado na página 57.

WANG, Z.; TANG, X.; LUO, W.; GAO, S. Face aging with identity-preserved conditional generative adversarial networks. In: . [S.l.: s.n.], 2018. p. 7939–7947. Citado 5 vezes nas páginas 58, 67, 70, 72 e 114.

WILCOXON, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, [International Biometric Society, Wiley], v. 1, n. 6, p. 80–83, 1945. ISSN 00994987. Citado na página 84.

WOLLEB, J.; SANDKüHLER, R.; BIEDER, F.; CATTIN, P. C. *The Swiss Army Knife for Image-to-Image Translation: Multi-Task Diffusion Models*. [S.l.]: arXiv, 2022. Citado na página 70.

WU, T.; TURAGA, P.; CHELLAPPA, R. Age Estimation and Face Verification Across Aging Using Landmarks. *IEEE Transactions on Information Forensics and Security*, v. 7, n. 6, p. 1780–1788, dez. 2012. ISSN 1556-6021. Citado na página 63.

YANG, C.; LV, Z. Gender based face aging with cycle-consistent adversarial networks. *Image and Vision Computing*, v. 100, 2020. Citado 3 vezes nas páginas 37, 58 e 113.

YANG, H.; HUANG, D.; WANG, Y.; JAIN, A. Learning face age progression: A pyramid architecture of gans. In: . [S.l.: s.n.], 2018. p. 31–39. Citado 3 vezes nas páginas 58, 70 e 114.

YANG, H.; HUANG, D.; WANG, Y.; JAIN, A. Learning continuous face age progression: A pyramid of gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 43, n. 2, p. 499–515, 2021. Citado 3 vezes nas páginas 17, 35 e 113.

YAO, X.; PUY, G.; NEWSON, A.; GOUSSEAU, Y.; HELLIER, P. High resolution face age editing. *CoRR*, abs/2005.04410, 2020. Citado 5 vezes nas páginas 49, 50, 69, 70 e 82.

Yun Fu; Guodong Guo; HUANG, T. S. Age Synthesis and Estimation via Faces: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 32, n. 11, p. 1955–1976, nov. 2010. ISSN 0162-8828. Citado 2 vezes nas páginas 17 e 18.

ZENG, J.; MA, X.; ZHOU, K. Caae++: Improved caae for age progression/regression. *IEEE Access*, v. 6, p. 66715–66722, 2018. Citado na página 114.

ZENG, J.; MA, X.; ZHOU, K. Photo-realistic face age progression/regression using a single generative adversarial network. *Neurocomputing*, v. 366, p. 295–304, 2019. Citado na página 113.

ZENO, B.; KALINOVSKIY, I.; MATVEEV, Y. Identity preserving face synthesis using generative adversarial networks. In: . [S.l.: s.n.], 2019. Citado na página 114.

Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, v. 23, n. 10, p. 1499–1503, 2016. Citado na página 73.

ZHANG, R.; ISOLA, P.; EFROS, A. A.; SHECHTMAN, E.; WANG, O. *The Unreasonable Effectiveness of Deep Features as a Perceptual Metric*. 2018. Citado na página 68.

ZHANG, X.; WEI, P.; ZHENG, N. Face age transformation with progressive residual adversarial autoencoder. In: . [S.l.: s.n.], 2019. v. 2019-July. Citado na página 113.

ZHANG, Z.; SONG, Y.; QI, H. Age progression/regression by conditional adversarial autoencoder. In: . [S.l.: s.n.], 2017. v. 2017-January, p. 4352–4360. Citado 8 vezes nas páginas 28, 45, 46, 58, 62, 72, 73 e 114.

ZHAO, S.; LI, J.; WANG, J. Disentangled representation learning and residual gan for age-invariant face verification. *Pattern Recognition*, v. 100, 2020. Citado na página 113.

ZHENG, T.; DENG, W.; HU, J. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *CoRR*, abs/1708.08197, 2017. Citado na página 64.

ZHOU, Y.; HU, B.; HE, J.; GUAN, Y.; SHAO, L. Dual reference age synthesis. *Neurocomputing*, v. 411, p. 164–177, 2020. Citado 2 vezes nas páginas 58 e 113.

ZHU, H.; HUANG, Z.; SHAN, H.; ZHANG, J. Look globally, age locally: Face aging with an attention mechanism. In: . [S.l.: s.n.], 2020. v. 2020-May, p. 1963–1967. Citado 2 vezes nas páginas 58 e 113.

ZHU, J.-Y.; PARK, T.; ISOLA, P.; EFROS, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Computer Vision (ICCV), 2017 IEEE International Conference on*. [S.l.: s.n.], 2017. Citado na página 37.

ZOSS, G.; CHANDRAN, P.; SIFAKIS, E.; GROSS, M.; GOTARDO, P.; BRADLEY, D. Production-Ready Face Re-Aging for Visual Effects. *ACM Transactions on Graphics*, v. 41, n. 6, 2022. ISSN 0730-0301, 1557-7368. Citado na página 17.

Apêndices

Apêndice A – Protocolo da revisão sistemática

Tema

Foi executada uma revisão de trabalhos publicados no tema envelhecimento facial, que utilizam os seguintes métodos de modelos generativos: GANs e *autoencoders*.

Os objetivos desta revisão foram:

- Entender quais técnicas foram aplicadas, e em quais bases de dados públicas foram utilizadas na maioria destas publicações.
- Mapear o processo quando essas técnicas foram aplicadas em tarefas biométricas: reconhecimento[1] ou autenticação[2] facial.

Questões de pesquisa

As questões de pesquisa respondidas nesta revisão sistemática (RS) com base em Kitchenham (2004) são:

- **Q1** - Quais técnicas são utilizadas para geração das faces envelhecidas?
- **Q2** - Quais bases de dados estão sendo utilizadas nesse procedimento?
- **Q3** - Como a performance do modelo está sendo medida?
- **Q4** - Como essa técnica está sendo integrada na tarefa biométrica?
- **Q5** - Está havendo uma melhora ao utilizar o modelo generativo na tarefa biométrica?

A questão **Q1** tem como objetivo verificar dentre os dois métodos de modelos generativos (GANs e *autoencoders*), quais foram utilizados para o envelhecimento facial. No caso dos *autoencoders*, foram listados todos os tipos encontrados e quando utilizaram GANs, foram classificadas conforme as arquiteturas expostas no Capítulo 2.

Na questão **Q2**, houve mapeamento das bases utilizadas, e usando as referências para as mesmas, foi possível determinar:

- Quantidade de imagens de faces presentes.
- Quantidade de indivíduos distintos.

[1] A tarefa de reconhecimento facial consiste em identificar (reconhecer) uma ou mais pessoas em uma imagem utilizando uma base de dados de faces.

[2] A tarefa de autenticação facial consiste em: dado duas imagens distintas, determinar se ambas pertencem ao mesmo indivíduo.

- Se a base é de domínio público.
- Se as imagens presentes estão livres de ruído.
- Se foram obtidas em ambientes não-controlados.

A questão **Q3**, é necessária, pois medir a performance de modelos generativos não é uma atividade que tem uma solução consolidada. Existem alguns métodos quantitativos [3], mas com limitações publicadas (BARRATT; SHARMA, 2018). Alguns autores utilizam pesquisas qualitativas, entrevistando pessoas para tentar identificar quando uma imagem foi gerada e quando não.

Já no caso da questão **Q4**, essa foi respondida pelo subconjunto dos artigos que aplicam modelos generativos em tarefas biométricas. A motivação para a mesma foi verificar como o envelhecimento está integrado na tarefa biométrica. Se está sendo utilizado de forma integrada ou em duas etapas, primeiro envelhecendo a imagem do indivíduo presente na base de dados para depois utilizá-la, ou se outra abordagem foi utilizada. E por fim, na questão **Q5**, buscou-se identificar se houve uma melhora nas métricas das tarefas biométricas.

String de busca

Os trabalhos foram extraídos utilizando o motor de busca da plataforma Scopus[4]. A *string* de busca canônica está na Tabela 11.

Tabela 11 – String de busca

TITLE-ABS-KEY(((generative AND adversarial AND network*) OR gan OR (autoencoder OR auto-encoder)) AND ((face AND aging) OR (age AND (invariant OR progression OR synthesis)) OR (age-invariant OR face-aging OR age-progression OR age-systhesis)))

[3] *Inception Score* (IS) e *Fréchet Inception Distance* (FID)
[4] https://www.scopus.com

Critérios de inclusão (CI)

- **CI-1** Trabalhos que utilizam modelos generativos: GANs ou *autoencoders* (generativos).
- **CI-2** Trabalhos que utilizam estes modelos para envelhecimento de imagens do mesmo indivíduo.
- **CI-3** Trabalhos que utilizam bases de dados de faces humanas.

Critérios de exclusão (CE)

- **CE-1** Trabalhos em línguas diferentes do inglês.
- **CE-2** Trabalhos não revisados por pares.
- **CE-3** Trabalhos que não estejam em estágio final de publicação.
- **CE-4** Trabalhos fora das áreas de atuação da computação, engenharia ou ciências da decisão.
- **CE-5** Trabalhos que não foram revisados por pares.
- **CE-6** Mesmos trabalhos mas em formato reduzido ou versões anteriores.

Critérios de qualidade (CQ) dos estudos primários

- **CQ-1** Os objetivos do trabalho foram definidos de modo adequado?
- **CQ-2** A metodologia utilizada foi apresentada de maneira clara?
- **CQ-3** Os modelos generativos foram descritos de modo apropriado?
- **CQ-4** Os hiper-parâmetros utilizados foram detalhados para ser possível reproduzir o experimento, se necessário?
- **CQ-5** O código utilizado foi disponibilizado?
- **CQ-6** A abordagem utilizada foi comparada com outras similares?
- **CQ-7** As bases de dados utilizadas foram descritas ou apresentadas suas referências?
- **CQ-8** As bases de dados eram adequadas (variabilidade, distribuição das idades)?
- **CQ-9** O modelo mostrou resultados utilizando alguma das métricas quantitativas utilizadas pela comunidade científica para modelos generativos[5]?
- **CQ-10** Os resultados foram apresentados de maneira clara?

[5] IS ou FID

- **CQ-11** Foi apresentado o método de integração do modelo generativo com a tarefa biométrica?
- **CQ-12** Foi exposto a diferença entre utilizar ou não o modelo generativo na tarefa biométrica?

Os critérios de qualidade **CQ-11** e **CQ-12** somente foram analisados quando os trabalhos eram aplicados em tarefas biométricas.

Estratégia para extração de dados e síntese de resultados

A *string* de busca (Tabela 11) foi utilizada no motor de busca e foram extraídos os trabalhos que atendiam aos critérios de inclusão e exclusão. Todos os resumos foram lidos para verificar se realmente se enquadravam nos critérios propostos.

Nos trabalhos que atendiam a tais critérios foram catalogados os itens anteriormente detalhados nas questões de pesquisa. Além disso, foi mapeado para cada critério de qualidade, se ele foi atendido de forma suficiente (1 ponto), apenas parcialmente (0,5 ponto) ou não foi atendido (0 ponto).

Apêndice B – Condução da revisão sistemática

Ao aplicar a *string* de busca no motor de busca da plataforma Scopus, ainda sem aplicar os critérios de exclusão (CEs) foram retornados 122 estudos primários no dia 17 de janeiro de 2021. Após a aplicação dos CEs[1] diretamente nos filtros da plataforma, foram removidos 37 registros, restando 85.

Desses, todos os resumos foram revisados para verificar se estavam de acordo com os critérios de inclusão e exclusão, sendo que 31 não estavam e foram removidos, restando 54 trabalhos selecionados, como mostra a Figura 24. Sendo 36% publicações de periódicos e 64% artigos de conferências.

Os critérios de inclusão foram definidos conjuntamente com o professor orientador, buscando responder às questões de pesquisa, descritas anteriormente.

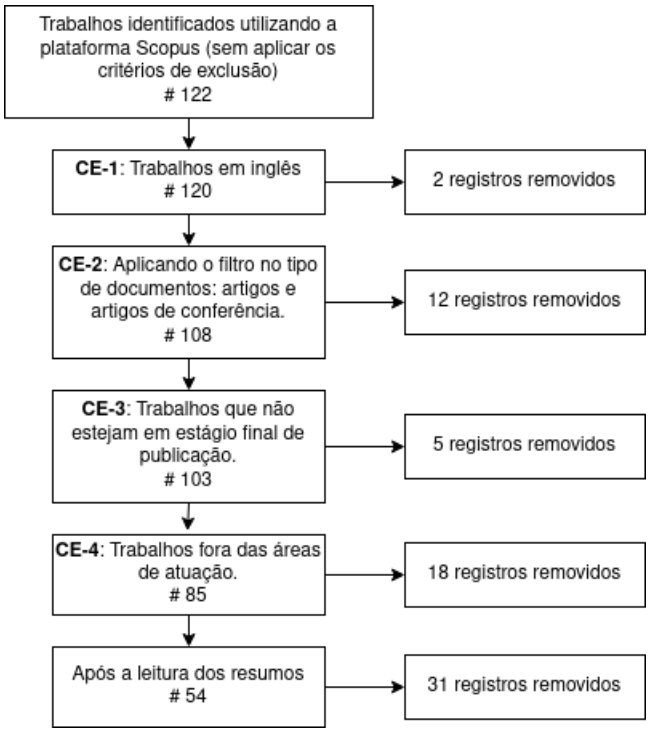
Um exemplo de artigos que foram descartados são os que tinham o acrônimo GAN, mas tratavam do significado da substância química conhecida como nitreto de gálio (GaN na tabela periódica), alguns artigos encontrados eram sobre esse tema. Durante a leitura dos resumos estes casos foram descartados manualmente. Os 54 trabalhos selecionados para a leitura integral estão listados no Quadro 1. A coluna ID (identificação) deste quadro será utilizada para mapear quais trabalhos utilizaram as diferentes técnicas, bases de dados e formas de análises dos resultados em tabelas e quadros subsequentes.

Podemos ver pela Figura 25 que houve um aumento expressivo na quantidade de publicações, e como foram selecionados somente os trabalhos que chegaram ao estágio final de publicação, e as buscas ocorreram durante o primeiro mês de 2021. Dessa forma, inferimos que a quantidade de publicações desse ano irá aumentar.

A Tabela 12, mostra a avaliação dos critérios de qualidade definidos no apêndice A. As referências da coluna IDs estão no Quadro 1. Pode ser notado que poucos trabalhos disponibilizam o código utilizado (critério de qualidade **Q5**).

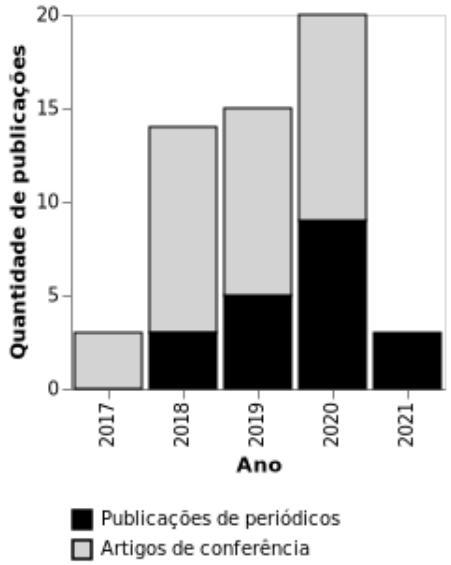
[1] O **CE-5** está contemplado nos anteriores.

Figura 24 – Condução da revisão na plataforma Scopus.



Fonte: Bruno Abreu Kemmer, 2023.

Figura 25 – Distribuição dos artigos selecionados ao longo dos anos.



Fonte: Bruno Abreu Kemmer, 2023.

Tabela 12 – Avaliação dos trabalhos selecionados pelos critérios de qualidade.

ID(s)	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
1, 4, 6, 14, 15, 18, 33, 34, 51	1	1	1	1	0	1	1	1	1	1	-	-
2, 24	1	1	1	1	0	0.5	1	1	1	1	1	1
3	1	1	1	0.5	0	1	1	1	1	1	1	1
5	1	0.5	1	1	0	1	1	1	1	0.5	-	-
7	1	1	1	1	0	1	1	0.5	0.5	1	1	1
8	1	1	1	0	0	0.5	1	0.5	0.5	1	-	-
9, 12, 26, 35, 36	1	1	1	1	0	1	1	1	1	1	1	1
10	1	1	1	1	1	1	0.5	1	1	1	-	-
11, 28	1	1	1	1	0	1	1	1	0.5	1	-	-
13	1	1	1	0	0	0	1	1	0	0.5	-	-
16	1	1	1	0	0	0	0.5	1	0	0	-	-
17	1	0.5	0	1	0	1	1	1	1	1	1	1
19	1	1	1	1	1	1	1	1	0.5	0.5	-	-
20	1	1	1	1	1	0	-	-	1	1	-	-
21	1	1	1	1	1	1	0.5	0.5	1	1	-	-
22	1	1	1	1	0	0.5	1	1	0.5	0.5	-	-
23	1	1	1	1	0	0.5	1	1	0.5	1	-	-
25	1	1	1	1	1	1	1	1	1	0.5	-	-
27	0.5	0.5	0.5	0	0	0	1	1	1	0.5	-	-
29	1	1	1	1	0	0.5	1	1	1	1	-	-
30	1	1	1	1	0	0.5	0	0	0	1	-	-
31	1	0.5	1	0	0	1	1	1	1	1	-	-
32	1	1	1	1	0	0	0.5	1	1	1	-	-
37	1	1	1	0.5	0	0	1	1	0.5	0.5	-	-
38	1	0.5	1	1	0	0.5	1	1	1	1	1	1
39	1	1	1	1	1	1	1	0.5	1	1	1	1
40	1	1	1	0.5	0	0.5	1	0.5	0.5	1	-	-
41	1	1	1	1	0	0.5	1	1	0.5	0.5	0.5	0.5
42	1	1	1	0	0	0.5	1	1	0	1	-	-
43	1	1	1	1	0	0.5	1	0.5	0.5	1	-	-
44	1	1	1	1	0	0	0.5	0.5	0	0.5	-	-
45	1	1	1	1	0	0	1	1	0.5	1	-	-
46	1	0.5	0.5	0	0	0.5	0.5	1	1	1	1	1
47	1	1	1	1	0	0	0.5	1	0.5	1	1	1
48	1	1	-	1	0	1	1	1	1	1	1	1
49	1	1	1	1	0	0.5	1	1	0	1	-	-
50	1	1	0.5	1	1	0	0.5	1	0	0.5	-	-
52	1	1	1	-	0	0.5	1	1	0	1	1	1
53	1	1	1	1	1	1	1	1	0.5	1	-	-
54	1	1	1	1	0	1	1	1	0.5	1	1	1
Média	**1,0**	**0,9**	**0,9**	**0,8**	**0,1**	**0,7**	**0,9**	**0,9**	**0,7**	**0,9**	**1,0**	**1,0**

Fonte: Bruno Abreu Kemmer, 2023.

Quadro 1 – Trabalhos selecionados.

ID	Referência	Título original
1	(LIU *et al.*, 2021)	Learning shape and texture progression for young child face aging
2	(YANG *et al.*, 2021)	Learning Continuous Face Age Progression: A Pyramid of GANs
3	(HUANG; HU, 2021)	A Parallel Architecture of Age Adversarial Convolutional Neural Network for Cross-Age Face Recognition
4	(ZHOU *et al.*, 2020)	Dual reference age synthesis
5	(SHARMA; SHARMA; JINDAL, 2020)	An Improved Technique for Face Age Progression and Enhanced Super-Resolution with Generative Adversarial Networks
6	(SHI *et al.*, 2020)	CAN-GAN: Conditioned-attention normalized GAN for face age synthesis
7	(SHENG *et al.*, 2020)	Face Aging with Conditional Generative Adversarial Network Guided by Ranking-CNN
8	(YANG; LV, 2020)	Gender based face aging with cycle-consistent adversarial networks
9	(FANG *et al.*, 2020)	Triple-GAN: Progressive face aging with triple translation loss
10	(ZHU *et al.*, 2020)	Look globally, age locally: Face aging with an attention mechanism
11	(PHAM; YANG; SHIN, 2020)	Semi-supervised facegan for face-age progression and regression with synthesized paired images
12	(ZHAO; LI; WANG, 2020)	Disentangled representation learning and residual GAN for age-invariant face verification
13	(ANWAAR; LOO; SEERA, 2020)	Face image synthesis with weight and age progression using conditional adversarial autoencoder
14	(SUN *et al.*, 2020b)	Facial Age and Expression Synthesis Using Ordinal Ranking Adversarial Networks
15	(SUN *et al.*, 2020a)	Facial Age Synthesis with Label Distribution-Guided Generative Adversarial Network
16	(HAMZAH; ZAMAN, 2020)	Face Aging on Realistic Photo in Cross-Dataset Implementation
17	(KHANNA *et al.*, 2020)	Cross-age face verification using face aging
18	(LI *et al.*, 2020)	Hierarchical Face Aging Through Disentangled Latent Characteristics
19	(OR-EL *et al.*, 2020a)	Lifespan Age Transformation Synthesis
20	(SHEN *et al.*, 2020)	Interpreting the latent space of GANs for semantic face editing
21	(VIAZOVETSKYI; IVASHKIN; KASHIN, 2020)	StyleGAN2 Distillation for Feed-Forward Image Manipulation
22	(MADHUKAR *et al.*, 2020)	Age progression using generative adversarial networks
23	(ZENG; MA; ZHOU, 2019)	Photo-realistic face age progression/regression using a single generative adversarial network
24	(CHEN *et al.*, 2019)	Age-invariant face recognition based on sample enhancement of generative adversarial networks
25	(LI *et al.*, 2019)	Global and Local Consistent Wavelet-Domain Age Synthesis
26	(WANG; DING; FU, 2019)	Cross-Generation Kinship Verification with Sparse Discriminative Metric
27	(GENOVESE; PIURI; SCOTTI, 2019)	Towards Explainable Face Aging with Generative Adversarial Networks
28	(ZHANG; WEI; ZHENG, 2019)	Face Age Transformation with Progressive Residual Adversarial Autoencoder
29	(CHANDALIYA; NAIN, 2019)	Conditional Perceptual Adversarial Variational Autoencoder for Age Progression and Regression on Child Face

Quadro 1 – Trabalhos selecionados. (*Continuação*)

ID	Referência	Título original
30	(GOU *et al.*, 2019)	A face aging network based on conditional cycle loss and the principle of homology continuity
31	(PANTRAKI; KOTROPOULOS; LANITIS, 2019)	Leveraging Image-to-image Translation Generative Adversarial Networks for Face Aging
32	(WANG *et al.*, 2019)	Face aging on realistic photos by generative adversarial networks
33	(LIU *et al.*, 2019)	Bidirectional face aging synthesis based on improved deep convolutional generative adversarial networks
34	(WANG *et al.*, 2019)	Recurrent Face Aging with Hierarchical AutoRegressive Memory
35	(HUANG; CHEN; HU, 2019)	Age-Puzzle FaceNet for Cross-Age Face Recognition
36	(BAYRAMLI *et al.*, 2019)	FH-GAN: Face hallucination and recognition using generative adversarial network
37	(ZENO; KALINOVSKIY; MATVEEV, 2019)	Identity preserving face synthesis using generative adversarial networks
38	(YANG *et al.*, 2018)	Learning Face Age Progression: A Pyramid Architecture of GANs
39	(WANG *et al.*, 2018b)	Face Aging with Identity-Preserved Conditional Generative Adversarial Networks
40	(PALSSON *et al.*, 2018)	Generative adversarial style transfer networks for face aging
41	(LI *et al.*, 2018)	Global and Local Consistent Age Generative Adversarial Networks
42	(JIA; SONG; ZHANG, 2018)	Face Aging with Improved Invertible Conditional GANs
43	(CHEN; HU; ZHANG, 2018)	Face Aging with Boundary Equilibrium Conditional Autoencoder
44	(THENGANE *et al.*, 2018)	Cycle Face Aging Generative Adversarial Networks
45	(LIU *et al.*, 2018)	Face Aging Simulation with Deep Convolutional Generative Adversarial Networks
46	(SAJID; SHAFIQUE, 2018)	Hybrid generative-discriminative approach to age-invariant face recognition
47	(ANTIPOV; BACCOUCHE; DUGELAY, 2018b)	Face aging with conditional generative adversarial networks
48	(ANTIPOV; BACCOUCHE; DUGELAY, 2018a)	Boosting cross-age face verification via generative age normalization
49	(ZENG; MA; ZHOU, 2018)	CAAE++: Improved CAAE for age progression/regression
50	(HELJAKKA; SOLIN; KANNALA, 2018)	Recursive Chaining of Reversible Image-to-Image Translators for Face Aging
51	(SONG *et al.*, 2018)	Dual conditional GANs for face aging and rejuvenation
52	(DUONG *et al.*, 2017)	Temporal Non-volume Preserving Approach to Facial Age-Progression and Age-Invariant Face Recognition
53	(ZHANG; SONG; QI, 2017)	Age progression/regression by conditional adversarial autoencoder
54	(LIU *et al.*, 2017)	Face aging with contextual generative adversarial nets

Fonte: Bruno Abreu Kemmer, 2023.