UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

UM ESTUDO SOBRE SÍNTESE DE ROSTO HUMANO GERADO POR GAN AO FILTRAR O CONJUNTO DE TREINAMENTO POR ATRIBUTOS FACIAIS CONSIDERANDO A PERCEPÇÃO HUMANA COMO CRITÉRIO DE AVALIAÇÃO DA QUALIDADE DA IMAGEM SINTETIZADA

Daniel da Silva Costa

Orientadora
Ana Cristina Bicharra Garcia

RIO DE JANEIRO, RJ - BRASIL
DEZEMBRO de 2020

UM ESTUDO SOBRE SÍNTESE DE ROSTO HUMANO GERADO POR GAN AO FILTRAR O CONJUNTO DE TREINAMENTO POR ATRIBUTOS FACIAIS CONSIDERANDO A PERCEPÇÃO HUMANA COMO CRITÉRIO DE AVALIAÇÃO DA QUALIDADE DA IMAGEM SINTETIZADA

Daniel da Silva Costa

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:

Ana Cristina Bicharra Garcia, D.Sc. - UNIRIO

Pedro Nuno de Souza Moura, D.Sc. - UNIRIO

Jean-Pierre Briot, D.Sc. – UNIRIO

Adriana Santarosa Vivacqua, D.Sc. – UFRJ

Igor Natal, D.Sc. – IFSP

Rio de Janeiro, RJ – BRASIL
DEZEMBRO de 2020

Catalogação informatizada pelo(a) autor(a)

C837 Costa, Daniel da Silva
Um estudo sobre síntese de rosto humano gerado por GAN ao filtrar o conjunto de treinamento por atributos faciais considerando a percepção humana como critério de avaliação da qualidade da imagem sintetizada / Daniel da Silva Costa. -- Rio de Janeiro, 2020.
86

Orientadora: Ana Cristina Bicharra Garcia.
Dissertação (Mestrado) - Universidade Federal do Estado do Rio de Janeiro, Programa de Pós-Graduação em Informática, 2020.

1. Redes Adversárias Generativas. 2. Aprendizado Profundo. 3. Percepção Humana. I. Garcia, Ana Cristina Bicharra, orient. II. Título.

Dedico essa dissertação à minha esposa.

Semper ascendens.

Agradecimentos

A Deus, por sempre me permitir crescer moral e intelectualmente.

À minha esposa, Michele Velloso Martins Costa, pelo companheirismo e por não me deixar desistir nos momentos mais difíceis e delicados da caminhada terrena.

À minha orientadora, Dra. Ana Cristina Bicharra Garcia, pelo exemplo de pesquisadora e pela paciência em me orientar.

Aos amigos e familiares, por sempre me apoiarem.

À toda a equipe PPGI UNIRIO: professores, colegas e funcionários, pelo excelente trabalho que realizam neste programa. Aprendi muito com todos.

A todos que, de forma direta ou indireta, me ajudaram em mais esta etapa da minha vida. Todos fizeram a diferença.

Costa, Daniel da Silva. Um Estudo sobre Síntese de Rosto Humano Gerado por GAN ao Filtrar o Conjunto de Treinamento por Atributos Faciais Considerando a Percepção Humana como Critério de Avaliação da Qualidade da Imagem Sintetizada. UNIRIO, 2020. 90 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

Trabalhos recentes têm apresentado resultados impressionantes na geração de novos exemplos de imagens usando *Generative Adversarial Networks* (GANs). No entanto, a geração de uma imagem que seja a síntese facial de um determinado conjunto de imagens requer uma grande quantidade de imagens de treinamento. Nesse sentido, esta pesquisa investigou o *trade-off* entre o tamanho do conjunto de dados de treinamento e a variabilidade das imagens neste conjunto de treinamento e o impacto na qualidade das imagens sintéticas geradas usando um modelo de *Deep Convolutional Generative Adversarial Networks* (DCGAN). Reportamos a configuração do tamanho do *batch size* e do espaço latente. Experimentamos diminuir a variação da imagem escolhendo imagens de rostos do conjunto de dados CelebA com atributos específicos: maçãs do rosto salientes, sobrancelhas arqueadas, rosto oval e lábios grandes. Para cada configuração DCGAN, o melhor modelo foi selecionado usando a métrica *Fréchet Inception Distance* (FID). Além da pontuação de FID, avaliamos as imagens sintéticas geradas de acordo com a percepção humana usando a plataforma de *crowdsourcing* Appen. Nossos resultados mostram que é possível obter melhores imagens sintéticas com um conjunto de dados de treinamento menor, reduzindo a variabilidade dos atributos faciais neste conjunto de dados ao produzir imagens compatíveis com a percepção humana.

Palavras-chave: Redes Adversárias Generativas, Aprendizado Profundo, Percepção Humana.

ABSTRACT

Recent works have shown impressive results in generating new examples of images using Generative Adversarial Networks (GAN). However, generating an image that is the facial synthesis of a given dataset requires a large amount of training images. In this sense, we investigate the trade-off between the size of training dataset and variance of the images in this training set and the impact on the quality of synthetic images generated using a Deep Convolutional Generative Adversarial Network (DCGAN) model. We report on the configuration of the batch size and latent space. We experiment diminishing the image variance by choosing facial images from CelebA dataset with specific attributes: high cheekbones, arched eyebrows, oval face and big lips. For each DCGAN configuration, the best model was selected using the Fréchet Inception Distance (FID) metric. Beyond the FID score, we evaluated the generated synthetic images according to the human perception using Appen crowdsourcing platform. Our results show that it is possible to obtain better synthetic images with a smaller training dataset by reducing the variance of facial features on this dataset when generating images compatible with human perception.

Keywords: Generative Adversarial Networks, Deep Learning, Human Perception.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Nomenclaturas

API	*Application Programming Interface* (Interface de Programação de Aplicações)
CNN	*Convolutional Neural Networks* (Rede Neuras Convolucionais)
DCGAN	*Deep Convolutional Generative Adversarial Networks* (Redes Adversárias Generativas Convolucionais Profunda)
DL	*Deep Learning* (Aprendizado Profundo)
FID	*Fréchet Inception Distance*
GAN	*Generative Adversarial Networks* (Redes Adversárias Generativas)
GPU	*Graphics Processing Unit (Unidade de Processamento Gráfico)*
HOG	*Histogram of Oriented Gradients* (Histograma de Gradientes Orientados)
LeakyReLU	*Leaky Rectified Linear Units*
ML	*Machine Learning* (Aprendizado de Máquina)
ReLU	*Rectified Linear Units*
RGB	*Red Green Blue* (Vermelho Verde Azul)

1. Introdução

1.1 Motivação

A geração automática de novas imagens que sintetizem imagens reais é um importante desafio e útil em diversas áreas.

No campo jurídico, sistemas inteligentes podem permitir a geração de imagens de rostos de suspeitos de crimes. Por exemplo, em (ZALTRON et al., 2020), foi proposto e desenvolvido um sistema que permitia às testemunhas oculares criarem composições de rosto facilmente, sem a necessidade de um profissional especializado, gerando as imagens dos suspeitos diretamente no sistema. Foi aplicada a computação generativa e a evolucionária, o que possibilitou ao sistema aprender e representar os rostos a partir de exemplos selecionados pelos usuários.

Hoje, é possível gerar novos exemplos de imagens histopatológicas de câncer de mama (KARRAS et al., 2020) o que pode, por exemplo, permitir às pessoas e aos sistemas de informação conhecerem variações de células doentes e, consequentemente, facilitar o treinamento das pessoas e dos modelos computacionais auxiliando nas avaliações dos exames de saúde.

No caso do entretenimento, VOLZ et al. (2018) utilizaram uma abordagem envolvendo *Generative Adversarial Networks* (GAN) (GOODFELLOW et al., 2014) e *Procedural Content Generation* (PCG), para a criação automática de fases para o jogo digital Mario. E em (JIN et al., 2017) foi utilizado GAN para gerar automaticamente personagens de anime.

Outras áreas podem se beneficiar da utilização de modelos generativos, como pode ser visto em (GOODFELLOW, 2016), por exemplo: (i) aprendizagem por reforço baseada em modelos generativos; (ii) sintetização de imagem de alta resolução a partir de uma imagem de baixa resolução; e (iii) conversão de imagens de um estilo em outro.

1.2 **Contexto e Justificativa**

Nas últimas décadas, as abordagens de *Deep Learning* (DL) vem se tornando protagonistas entre as técnicas de Inteligência Artificial, e as razões para esse sucesso talvez sejam explicadas por: (i) disponibilidade de dados massivos; (ii) disponibilidade de poder computacional eficiente e acessível; e (iii) avanços técnicos (BRIOT et al., 2017).

Um trabalho relevante, que também ajuda a justificar o recente interesse das técnicas de DL, foi o resultado alcançado por (KRIZHEVSKY et al., 2012) na competição *ImageNet*, que usou *Deep Convolutional Neural Networks* (DCNN) e superou outras técnicas.

Uma característica importante do DL é a sua generalidade: os modelos de DL são agnósticos ao aprender, de forma não supervisionada, diversas características dos exemplos, pertencentes a categorias diferentes, que possam existir em um *corpus*. Logo, o mesmo modelo generativo pode ser usado para gerar exemplos de classes diversas.

Entre as abordagens mais recentes e promissoras de DL para a geração de novos exemplos sintéticos está a *Generative Adversarial Networks* (GAN). Ela tem sido utilizada para produzir imagens realistas, por exemplo: de objetos, de cômodos e de dígitos manuscritos (GURUMURTHY et al., 2017).

As pesquisas em GANs têm alcançado resultados impressionantes na geração de imagens de rosto, como pode ser visto em (KARRAS et al., 2019, DIAMANT et al., 2019, KARRAS et al., 2017).

Alguns exemplos de propostas de arquitetura GAN recentes, que alcançaram resultados impressionantes, são: *StyleGAN* (KARRAS et al., 2019), *BigGAN* (BROCK et al., 2018) e *Progressive Growing GAN* (PGGAN) (KARRAS et al., 2017). Alcançar resultados como os destas pesquisas envolve a utilização de um grande poder computacional: (i) no caso da PGGAN, os autores treinaram a GAN com 800 mil imagens em um computador configurado com 8 GPUs Tesla V100 por 4 dias e, somente após este período, não observaram diferenças qualitativas nos resultados; (ii) no caso da BigGAN, os autores utilizaram entre 2 e 4 vezes mais parâmetros nas redes neurais e um *batch size* 8 vezes maior em comparação a técnicas anteriores; e (iii) no caso da StyleGAN, no repositório *Git* do projeto (https://github.com/NVlabs/stylegan), os autores informam que utilizaram o computador NVIDIA DGX-1, que tem 8 Tesla V100 GPUs. Ter acesso a um

poder computacional equivalente pode ser um obstáculo para pesquisadores e praticantes e, consequentemente, um entrave para o desenvolvimento dessa área.

Treinar GANs para produzirem resultados tão impressionantes requer uma grande quantidade de dados de treinamento, geralmente maior do que em outras soluções de Inteligência Artificial (GURUMURTHY et al., 2017, BRIOT et al., 2017).

Por outro lado, treinar GANs com menos dados pode reduzir a variabilidade dos resultados ou dificultar a convergência do modelo levando a outros problemas (KARRAS et al., 2020).

Do ponto de vista da utilização da capacidade do modelo GAN e do tempo de treinamento, as abordagens recentes são ineficientes. Por exemplo, no caso da geração de imagens, parte da capacidade do modelo é perdida ao tentar representar as regiões esparsas dos dados de treinamento que, em geral, não são organizados para a tarefa de produção automática de imagens sintéticas. Uma das consequências é o comprometimento da qualidade de parte dos resultados. Atualmente, as técnicas buscam, em geral, eliminar os resultados de baixa qualidade após o modelo ser treinado (DEVRIES et al., 2020).

As duas principais abordagens para lidar com o problema de treinar GANs com um conjunto reduzido de dados são: (i) a manipulação do espaço latente, um exemplo é a pesquisa (GURUMURTHY et al., 2017); e (ii) a utilização de *data augmentation* para gerar novos exemplos durante o treinamento do modelo, como visto em (KARRAS et al., 2020).

Na arquitetura proposta por GURUMURTHY et al. (2017), para cenários contendo conjunto de dados diversos e limitados, o espaço latente foi modificado para obter amostras nas regiões de alta probabilidade e permitiu a diversidade nos resultados gerados.

NUHA e AFIAHAYATI (2018) investigaram o impacto da redução do tamanho do conjunto de treinamento, agrupando as imagens sem nenhuma forma de tratamento específica. Indicaram ter alcançado bons resultados treinando um modelo DCGAN (*Deep Convolutional Generative Adversarial Networks*) (RADFORD et al., 2015) com 50 mil imagens em comparação a modelos treinados com 2 mil e 200 mil.

Recentemente, DEVRIES et al. (2020) buscaram entre as técnicas de *Instance Selection* (OLVERA-LÓPEZ et al., 2010), uma forma de tratar o problema da qualidade irregular dos resultados dos modelos GAN, antes do início do treinamento do modelo. Eles obtiveram importantes achados: (i) produziram resultados melhores com menos

treino; (ii) conseguiram melhorar a qualidade geral dos resultados em troca de alguma redução na diversidade; (iii) conseguiram uma redução dos requisitos de capacidade do modelo; e (iv) conseguiram diminuir o tempo de treinamento.

A revisão de literatura sugere que o treinamento dos modelos GAN, com volume de dados menor, tendo em vista a geração de resultados viáveis, é um problema interessante a ser investigado. O desafio é maior ao considerar-se que, embora as pesquisas com GAN tenham alcançado considerável avanço na geração de novas imagens, avaliar e comparar os resultados de modelos de diferentes arquiteturas ainda é uma tarefa difícil.

LUCIC et al. (2018) sugerem três aspectos importantes sobre a comparação entre modelos de arquiteturas diferentes: (i) não existe um algoritmo que domine claramente os outros; (ii) para um intervalo interessante de valores da métrica FID (*Fréchet Inception Distance*) (HEUSEL et al., 2017), um modelo "ruim" treinado com um orçamento grande pode ter um desempenho melhor do que um modelo "bom" treinado com um orçamento pequeno; e (iii) quando o orçamento é limitado, qualquer comparação estatisticamente significativa dos modelos é inatingível.

As formas para avaliar a qualidade dos resultados da GAN ainda são majoritariamente objetivas e poucos trabalhos têm demonstrado a sua consistência com a percepção humana (WANG et al., 2020).

Recentemente, algumas pesquisas têm focado em formas diferentes das tradicionais para avaliar a qualidade das imagens geradas pelas GANs, indicando a importância da percepção humana como critério para a avaliação dos resultados gerados. Alguns exemplos podem ser vistos em (WANG et al., 2020, FUJII et al., 2020, ZHOU et al., 2019).

WANG et al. (2020), por exemplo, buscaram avaliar a qualidade das imagens a partir dos sinais neurais dos participantes dos experimentos, captados por meio de exames de eletroencefalograma. Eles criaram uma pontuação chamada *Neuroscore* e compararam com três métricas objetivas: a *Inception Score* (IS) (SALIMANS et al., 2016), a *Kernel Maximum Mean Discrepancy* e a *Fréchet Inception Distance* (FID). O trabalho sugere que a *Neuroscore* é mais consistente com o julgamento humano do que as demais métricas. Destacaram também que a *Neuroscore* teve por objetivo avaliar a qualidade das imagens geradas, mas que essa métrica não é capaz de lidar com os problemas inerentes da GAN como o *Mode Collapse*.

A partir da revisão de literatura, observou-se que pouco foi investigado sobre os efeitos do agrupamento das imagens de rosto, por meio dos atributos faciais, no problema

do treinamento dos modelos GAN com poucos dados e baixo poder computacional, considerando-se a variabilidade dos resultados e o critério de avaliação da percepção humana.

Agrupar o conjunto de dados de treinamento com base em seus atributos é particularmente interessante no domínio de imagens de rostos, pois estes dados apresentam diversas características multimodais que podem ser exploradas de várias formas, por exemplo, para reduzir o tempo de treinamento dos modelos e para produzir imagens contendo rostos com alguma característica específica.

As características faciais são inúmeras, o que possibilita a anotação de atributos de diferentes categorias, por exemplo: (i) gênero; (ii) idade; (iii) etnia; (iv) atributos faciais geométricos; (v) atributos calculados a partir da distância entre os pontos de interesse do rosto (*landmarks*); e (vi) atributos de fotografia como a iluminação e a posição do rosto na imagem.

Conforme sugerem os autores NUHA e AFIAHAYATI (2018), trata-se de um interessante desafio para a nossa capacidade de manipular e representar distribuições de probabilidade de alta dimensão.

1.3 **Problema de Pesquisa**

Geração de imagens sintéticas de rostos utilizando GANs com qualidade perceptível ao ser humano requer uma grande quantidade de imagens devidamente rotuladas e muito poder computacional. Por exemplo, KARRAS et al. (2017) treinaram a sua arquitetura GAN com 800 mil imagens em um computador configurado com 8 GPUs Tesla V100 por 4 dias. Existem poucas bases de dados com número de imagens suficientes para suprir tal treinamento e o poder computacional necessário pode ser um limitador.

O problema que esta pesquisa procura resolver é o da dependência de grandes bases de treinamento para gerar uma imagem que contenha um rosto humano perceptível por pessoas.

O problema pode ser formulado da seguinte forma: é possível gerar uma imagem de rosto humano com GAN que seja a síntese de um conjunto de dados, por meio de um volume menor de dados, mas que ainda seja perceptível pelas pessoas como um rosto humano?

Assim, este problema possui dois aspectos importantes a serem consideradas: (i) dependência na avaliação dos resultados de boas bases de imagens faciais para que

possamos testar; (ii) dependência de uma coletividade de pessoas, não relacionadas com a nossa pesquisa, que possam avaliar as imagens sintéticas geradas de acordo com a percepção humana.

1.4 **Hipótese**

Após uma revisão de literatura que apontou soluções ainda ineficientes, formulamos nossa hipótese de pesquisa como: o uso de certas características faciais como filtro para selecionar imagens usadas no treinamento de uma DCGAN afeta significativamente o tamanho da base de treinamento necessário para a síntese de imagens faciais que gerem imagens com qualidade percebida por humanos, permitindo a produção de imagens com qualidade compatível à de modelos que receberam mais imagens porém sem esse tratamento. Logo, nossa hipótese nula é que nossa DCGAN treinada com menos dados, mas selecionados com certas características faciais, vai necessitar do mesmo tamanho de base para dar um resultado humanamente aceitável.

Além disso, queremos investigar se existe um *trade-off* no treinamento da DCGAN entre o tamanho do conjunto de dados de treinamento e as diversas características faciais em uma imagem.

Nossa hipótese se baseia em: acreditamos que certos atributos faciais são fundamentais para diminuir a variância entre as imagens e com isso diminuir a necessidade de grandes volumes de treinamento.

1.5 **Objetivos**

O objetivo principal do presente estudo foi investigar a geração de imagens de rosto que sintetizem um conjunto de dados, produzidas através de modelos de DCGAN, e o impacto na qualidade dos resultados, quando reduzido o conjunto de treinamento e agrupados os dados por meio dos seguintes atributos faciais: maçãs do rosto salientes, sobrancelhas arqueadas, rosto oval e lábios grandes.

A hipótese traz consigo a ideia de *trade-off* entre a diminuição da variabilidade das imagens de treinamento e a qualidade e a variabilidade dos exemplos sintéticos gerados, por isso, foi utilizado o critério da percepção humana para a avaliação das imagens produzidas.

Investigou-se também o *trade-off* entre o poder computacional, a configuração dos modelos DCGAN e a qualidade das imagens geradas do ponto de vista objetivo através da métrica FID.

Alguns questionamentos foram levantados para serem analisados ao longo do presente estudo:

- Todos os atributos faciais geométricos utilizados nos agrupamentos ajudaram no aprendizado dos modelos DCGAN?
- As imagens sintéticas geradas apresentaram os atributos utilizados nos agrupamentos?
- As características demográficas das pessoas que participaram nos experimentos influenciaram na avaliação das imagens sintéticas geradas?
- Quão próximo de imagens reais ficaram as imagens sintéticas produzidas, conforme a percepção das pessoas?

1.6 **Metodologia**

Para o presente estudo foi realizada uma pesquisa quantitativa com a aplicação de experimentos em laboratório e de questionários *online* por meio da plataforma de *crowdsourcing* Appen (Appen, 2020).

Foram treinados diversos modelos DCGAN, com configurações diferentes utilizando-se imagens agrupadas com e sem os atributos faciais geométricos.

Para cada modelo DCGAN treinado foi aplicada a métrica FID nos resultados de cada *epoch* a fim de selecionar-se a melhor versão do modelo.

As imagens geradas na melhor versão de cada modelo DCGAN foram apresentadas para as pessoas através de questionários. Foi construído um questionário para cada modelo treinado e, em cada experimento, participaram 100 pessoas.

Os resultados foram analisados de forma exploratória e com a aplicação de teste de hipótese.

O presente estudo buscou um *dataset* de imagens faciais em posição frontal, contendo um único rosto por imagem e anotações de atributos faciais.

O *dataset* ImageNet (DENG et al., 2009), apesar do volume de imagens, não serviria porque tem por objetivo catalogar objetos que pertençam a um mesmo *synset*: um grupo de dados considerados semanticamente equivalentes e, portanto, não apresenta anotações de atributos faciais.

Datasets utilizados em predição de beleza facial, como o SCUT-FBP5500 (LIANG et a., 2018), geralmente trazem atributos faciais anotados, mas com um baixo volume de imagens e podem apresentar vieses de seleção, como por exemplo, somente imagens de pessoas asiáticas e caucasianas.

O *dataset* CIFAR-100 (KRIZHEVSKY, 2009), possui imagens com baixa resolução, 32 x 32 *pixels*, o que poderia dificultar a percepção humana.

O *dataset* utilizado foi o CelebA (LIU et al., 2015). Trata-se de um *dataset* muito comum em pesquisas que envolvem a área de visão computacional, como por exemplo: reconhecimento de atributos faciais e detecção, edição e síntese de rostos.

O *dataset* CelebA, contém mais de 200 mil imagens de rostos de celebridades masculinas e femininas e 40 atributos faciais anotados para cada imagem. Utilizou-se um subconjunto deste *dataset,* composto por imagens de mulheres jovens, sem óculos e em imagens não desfocadas. Todas as imagens passaram por um pré-processamento que envolveu convertê-las para a escala de cinza e centralizar o rosto da imagem através da aplicação da técnica *Histogram of Oriented Gradients* (HOG) (DALAL e TRIGGS, 2005).

1.7 **Principais Contribuições**

Foram treinados modelos DCGAN com volumes de imagens diferentes: (i) Tipo 1, modelos que serviram de *baseline*, treinados com 7,5 mil, 10 mil e 12,5 mil imagens sem serem agrupadas por atributos faciais geométricos e; (ii) Tipo 2, treinados com 5 mil imagens filtradas por atributos faciais geométricos. Os modelos do Tipo 2 apresentaram resultados melhores com menos imagens de treinamento. Logo, a principal contribuição do presente estudo foi a demonstração de que é possível gerar imagens de rosto melhores a partir de uma melhor escolha de imagens, selecionadas por meio de atributos faciais.

Entre os atributos faciais utilizados no agrupamento das imagens do Tipo 2, o atributo que representa o formato do rosto (rosto oval) foi o que obteve os melhores resultados tanto na avaliação objetiva (métrica FID) quanto da subjetiva (percepção humana).

Outras contribuições podem ser observadas:

- Resultados dos modelos DCGAN com a indicação das configurações utilizadas;
- Resultados dos questionários;
- Um novo método proposto e aplicado;
- Utilização da métrica FID para a seleção da melhor versão do modelo com base no resultado de cada *epoch*;
- Demonstração da viabilidade da utilização da plataforma Appen para realização de experimentos que envolvam *crowdsourcing*;
- Os resultados não sugerem correlação entre a métrica FID e a percepção humana;

- Sobre o impacto das características demográficas dos avaliadores nos resultados, observou-se diferença nos resultados agrupados por faixa etária e por país de origem;
- Disponibilização de repositório *Git* (https://github.com/danieldasilvacosta/dissertacao-2020) contendo os *notebooks* utilizados bem como as imagens geradas selecionadas.

1.8 **Estrutura da Dissertação**

Este trabalho está organizado da seguinte forma: no Capítulo 2 (Fundamentação Teórica) são trazidos conceitos fundamentais necessários à realização desse estudo. Em seguida, no Capítulo 3 (Metodologia), é apresentada a metodologia utilizada e a descrição do método proposto. No capítulo 4 (Experimentos), foram descritas as configurações dos modelos treinados conforme o método proposto e é apresentada a estrutura dos questionários aplicados nos experimentos. Então, no Capítulo 5 (Resultados e Discussão), são apresentados e analisados os resultados obtidos. O capítulo seguinte, Capítulo 6 (Trabalhos Relacionados), apresenta pesquisas que envolveram assuntos tratados no presente estudo. E, finalmente, as conclusões são apresentadas no Capítulo 7 (Considerações Finais).

2. Fundamentação Teórica

2.1 *Generative Adversarial Networks* (GANs)

Durante a última década, ocorreram muitos avanços em *Deep Learning* principalmente a partir da pesquisa seminal de (KRIZHEVSKY et al., 2012), cujo modelo proposto superou outras técnicas na competição de classificação *ImageNet*.

As *Generative Adversarial Networks* (GANs) foram propostas originalmente em (GOODFELLOW et al., 2014) e pertencem ao domínio dos modelos de *Deep Learning*. GANs pertencem à subcategoria dos modelos generativos e têm sido estudadas, amplamente, com o objetivo de produzir novos exemplos sintéticos convincentes a partir de exemplos reais. Podem empregar técnicas utilizadas em outras arquiteturas de redes profundas (GOODFELLOW, 2016).

Este tipo de modelo aprende as distribuições de probabilidade sobre múltiplas variáveis dos exemplos reais permitindo que a função de distribuição de probabilidade do modelo seja avaliada explicitamente ou não (GOODFELLOW et al., 2016). Além da GAN, outros exemplos de modelos generativos são as arquiteturas *Variational Autoencoder* (KINGMA e WELLING, 2014) e *Generative Latent Optimization* (BOJANOWSKI et al., 2017).

Na GAN, duas redes neurais profundas participam de um jogo *minimax* para dois jogadores: a primeira rede, denominada geradora, aprende a mapear uma nova distribuição a partir de vetores de ruído aleatórios de entrada, cria exemplos novos e os envia para a segunda na forma de exemplos sintéticos; a segunda, denominada discriminadora, estima a probabilidade de um dado exemplo ser real ou falso (GOODFELLOW et al., 2014). Dessa forma, ambas as redes aprendem juntas e simultaneamente.

Portanto, a GAN é treinada de uma forma não supervisionada, mas utilizando-se de duas redes que aprendem de forma supervisionada: a geradora busca construir imagens reais e ajusta os pesos dos seus neurônios com base no erro reportado pela discriminadora que, por sua vez, indica a probabilidade de as imagens serem reais ou falsas.

A Equação 2.1 define a relação *minimax* entre uma rede geradora G e uma discriminadora D (GOODFELLOW et al., 2014):

$$\min_{G} \max_{D} V(D, G) = E_{x \sim P_{Data}}[\log D(x)] + E_{z \sim P_z(z)}[\log (1 - D(G(z)))] \quad (2.1)$$

onde:

- $D(x)$ representa a probabilidade de que x pertence ao conjunto de dados reais;
- $E_{x \sim P_{Data}}[log\ D(x)]$ é a expectativa de $log\ D(x)$ com relação a x ter sido sorteado a partir do conjunto de dados reais;
- $D(G(z))$ representa a probabilidade de $G(z)$ pertencer ao conjunto dos dados reais. Logo, $1 - D(G(z))$ representa a probabilidade de $G(z)$ não pertencer ao conjunto dos dados reais;
- A expectativa de $log\ (1 - D(G(z)))$ com relação a $G(z)$ ter sido produzido por G, a partir de um ruído aleatório z, é representada por $E_{Z \sim P_z(z)}\ [log\ (1 - D(G(z)))]$.

A Figura 2.1 ilustra a estrutura geral da GAN. É possível ver as redes geradora e discriminadora. A entrada para a geradora é um vetor de ruído aleatório (z) a partir do qual ela irá produzir uma imagem de rosto sintética (falsa), enquanto a discriminadora recebe uma imagem e determina se esta é de um rosto real ou falso. Os elementos *Loss* (perda), se referem ao valor calculado pelas funções de custo.

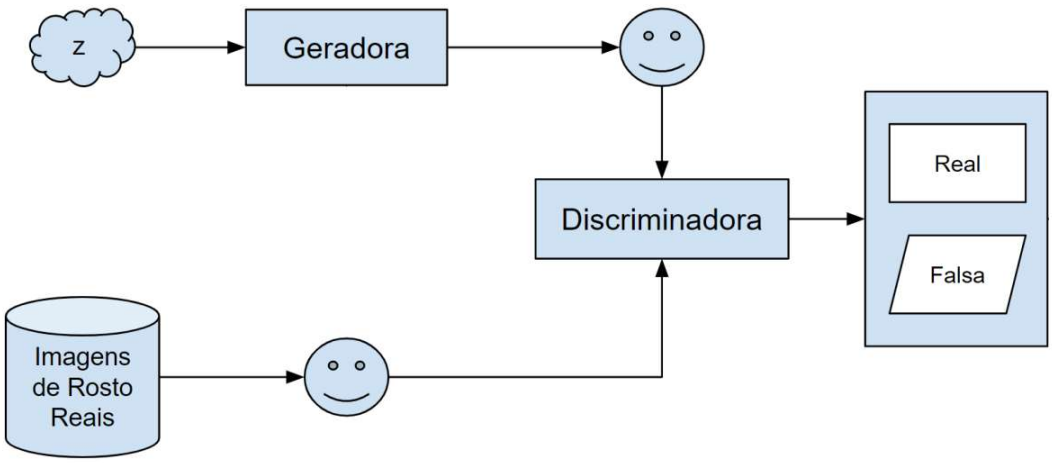

Figura 2.1: Arquitetura da GAN. Fonte: adaptada de (RAMSUNDAR e ZADEH, 2018), página 14.

Os vetores de ruídos aleatórios que são passados para a geradora como entrada são comumente extraídos de uma distribuição Gaussiana e não têm significado a princípio. Conforme o treinamento avança, os pontos neste espaço vetorial multidimensional corresponderão a pontos no domínio do problema, formando uma representação compactada da distribuição dos dados (BROWNLEE, 2020). Esse espaço vetorial é chamado de espaço latente e representa a distribuição dos dados de entrada como conceitos de alto nível.

As pesquisas que envolvem manipulações do espaço latente têm buscado compreender e permitir maior controle sobre os atributos presentes nas imagens sintéticas geradas. Entre as pesquisas recentes, que buscam entender melhor a estrutura do espaço latente, destacamos: MUKHERJEE et al. (2019), LIPTON e TRIPATHI (2017) e CHEN et al. (2016).

A estabilização do processo de aprendizagem da GAN ainda é um problema em aberto, embora seja possível obter-se um bom desempenho quando a arquitetura do modelo e os hiperparâmetros são cuidadosamente selecionados (GOODFELLOW et al., 2016). Nos últimos anos, alguns avanços foram alcançados a fim de melhorar esta estabilização, como a arquitetura DCGAN, que será explicada em detalhes na Seção 2.2 (Deep Convolutional Generative Adversarial Networks (DCGAN)).

Tipicamente, treinar um modelo GAN envolve lidar com dois problemas principais que têm sido discutidos na literatura: convergência e *Mode Collapse*.

Tendo em vista que as duas redes (discriminadora e geradora) aprendem simultaneamente, alcançar um equilíbrio é uma tarefa difícil pois a cada *epoch,* ambas são atualizadas e o problema de otimização em questão se modifica.

O problema de convergência (GOODFELLOW, 2016) pode ser identificado visualmente quando o valor da função de perda (*loss function*) do modelo não segue um padrão observável durante o treinamento. Isso reflete a dificuldade da GAN em encontrar um equilíbrio entre a aprendizagem da rede geradora e da discriminadora.

Eventualmente, as duas redes encontrarão um equilíbrio, mas, tipicamente, não há garantia de que o algoritmo de otimização irá manter um movimento de descida na busca pelo menor valor da função de custo, o que é um reflexo da mudança constante de cenário, resultado de cada *epoch* do treinamento.

Sendo assim, a maneira mais comum de identificar se a GAN continua aprendendo é observando os valores de *loss* da rede discriminadora e da GAN propriamente.

Na Figura 2.2, é apresentado um trecho do *log* de execução de um programa durante o treinamento de um modelo DCGAN que não alcançou a convergência. Esse programa foi executado previamente e adaptações foram aplicadas na arquitetura de DCGAN para permitir uma melhor estabilidade do modelo. Essas adaptações serão indicadas na próxima seção. Na imagem é possível ver a variação abrupta de *loss* na rede discriminadora (D_loss) e na GAN (G_loss): observam-se os valores negativos na discriminadora e os valores muito altos na GAN.

```
D_loss: 0,0061 -- G_loss: 21,0764
D_loss: 0,0374 -- G_loss: 18,6088
D_loss: 0,4525 -- G_loss: 12,1049
D_loss: 0,1536 -- G_loss: 10,0561
D_loss: 0,3693 -- G_loss: 25,1995
D_loss: 0,0408 -- G_loss: 93,1584
D_loss: 20,8025 -- G_loss: 155,5976
D_loss: 0,1422 -- G_loss: 41,8587
D_loss: 0,6127 -- G_loss: 90,2770
D_loss: -7,3694 -- G_loss: 2561,0210
D_loss: -3,9506 -- G_loss: 690,3988
```

Figura 2.2: Exemplo de *log* de execução de não convergência.

E na Figura 2.3, um trecho de *log* do programa final já com os ajustes na arquitetura DCGAN e com as heurísticas que serão apresentadas na seção 3.4: (Configuração da GAN). Esse programa foi usado para treinar os modelos definitivos deste estudo. Observa-se que não há uma mudança abrupta nos valores de *loss* da discriminadora (d1, quando treinada com os exemplos reais e d2, quando treinada com exemplos sintéticos) e da GAN (g).

```
d1=0,464, d2=0,531, g=1,693
d1=0,433, d2=0,633, g=1,627
d1=0,374, d2=0,586, g=2,212
d1=0,581, d2=0,617, g=1,619
d1=0,441, d2=0,533, g=1,581
d1=0,437, d2=0,597, g=1,954
d1=0,452, d2=0,567, g=1,805
d1=0,362, d2=0,597, g=1,404
d1=0,350, d2=0,580, g=1,804
d1=0,404, d2=0,549, g=1,724
d1=0,436, d2=0,587, g=1,701
```

Figura 2.3: Exemplo de *log* de execução sem o problema de não convergência.

O *Mode Collapse* ou o cenário *Helvetica* (GOODFELLOW, 2016), é um problema no qual a GAN, ao longo do treinamento, aprende a mapear diferentes valores de *z* para

uma mesma distribuição de saída. É, provavelmente, a forma mais comum de não convergência (GOODFELLOW, 2016).

No contexto deste estudo, um exemplo seria dizer que o modelo aprendeu a produzir o mesmo rosto para diferentes valores de z. Visualmente, este problema pode ser observado na Figura 2.4.

Figura 2.4: Exemplo do problema de *Mode Collapse*.

O programa utilizado para gerar os rostos da Figura 2.4, foi um dos construídos durante os estudos preliminares. No modelo treinado com este programa, a cada *epoch*, eram geradas 30 imagens pela geradora da GAN. Neste caso, são mostrados todos os rostos gerados em uma determinada *epoch*. Pode-se observar que, apesar de o modelo ter produzido boas imagens do ponto de vista da percepção humana, ele parece ter gerado sempre a mesma pessoa para pontos diferentes do espaço latente, o que indica a possível ocorrência do problema de *Mode Collapse*, pois diferentes pontos do espaço latente parecem ter sido mapeados para um mesmo conjunto de atributos faciais.

2.2 *Deep Convolutional Generative Adversarial Networks* (DCGAN)

Aproveitando o sucesso das arquiteturas de Redes Neurais Convolucionais (CNN) (LECUN et al., 1989) da literatura de aprendizado supervisionado, desde os surpreendentes resultados obtidos por KRIZHEVSKY et al. (2012), RADFORD et al. (2015) propuseram o uso desta arquitetura no contexto de GAN. Para tanto, adotaram as seguintes principais mudanças das CNNs tradicionais:

(i) substituir as camadas de *pooling* por camadas de *strided convolutions* para a rede discriminadora e *fractional-strided convolutions* para a geradora;

(ii) remover camadas ocultas totalmente conectadas em arquiteturas mais profundas;

(iii) usar *batch normalization* em ambas as redes, o que ajuda a lidar com as consequências de inicialização pobres e com fluxo gradiente;

(iv) na geradora, usar a ativação ReLU em todas as camadas, exceto na saída que deve ser *Hyperbolic Tangent* (Tanh);

(v) e na discriminadora, usar a ativação LeakyReLU para todas as camadas.

A Figura 2.5 exibe um exemplo de arquitetura para o modelo de rede geradora conforme proposta pelos autores da DCGAN (RADFORD et al., 2015). Pode-se observar que um vetor de ruído aleatório z, com 100 dimensões, é a entrada desta rede. Ao longo da rede, são aplicados *feature maps* para a extração de características pelos neurônios da rede. Observa-se também que, ao aplicar as quatro convoluções *fractional-strided*, o vetor de ruído foi convertido em uma imagem de 64 × 64 *pixels*.

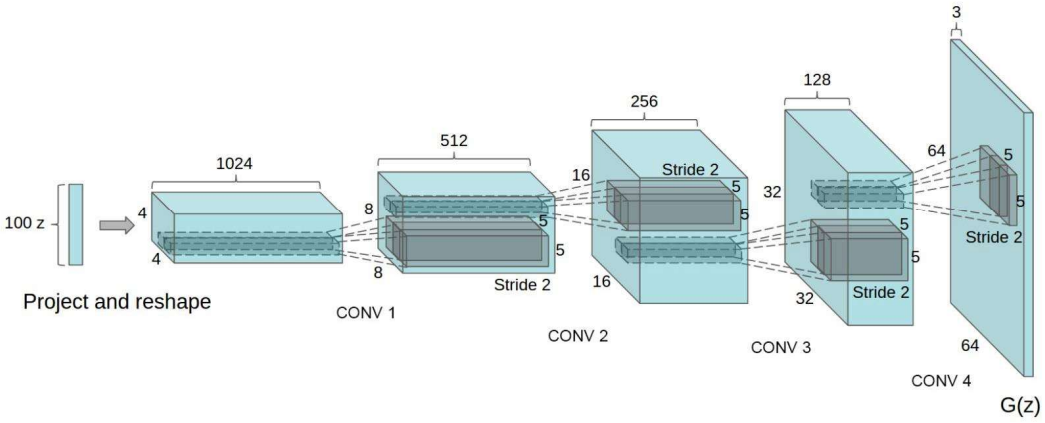

Figura 2.5: Um exemplo de rede geradora proposta para a arquitetura DCGAN. Extraída de (RADFORD et al., 2015), página 4.

2.3 **Appen: Plataforma de *Crowdsourcing***

Foi utilizada a plataforma Appen (Appen, 2020), uma plataforma de *crowdsourcing* que permite a realização de diversos experimentos com milhares de pessoas de vários países por meio da *Web*.

A Appen pode também ser chamada de plataforma de micro tarefas. Neste tipo de plataforma, as pessoas recebem uma quantia arbitrária em dinheiro para responderem aos questionários. Estas plataformas têm sido bastante utilizadas em pesquisas de Inteligência Artificial, por causa, entre outros fatores, da facilidade e da velocidade de se conseguir

um grande volume de pessoas para responder ao questionário em um espaço curto de tempo.

Na Appen, os experimentos são pagos em dólar e podem ser construídos a partir de *templates* predefinidos. Como exemplo de categorias de *templates*, pode-se destacar: (i) análise de sentimentos; (ii) categorização de dados; (iii) coleta de dados; (iv) validação de dados; (v) anotações em imagens e (vi) transcrição de conteúdo. Para este estudo, foi construído um questionário no formato de página da *Web* no qual as pessoas puderam avaliar as imagens sintéticas geradas pelos modelos de DCGAN treinados e selecionados conforme será visto no próximo capítulo.

3. Metodologia

3.1 Método Proposto

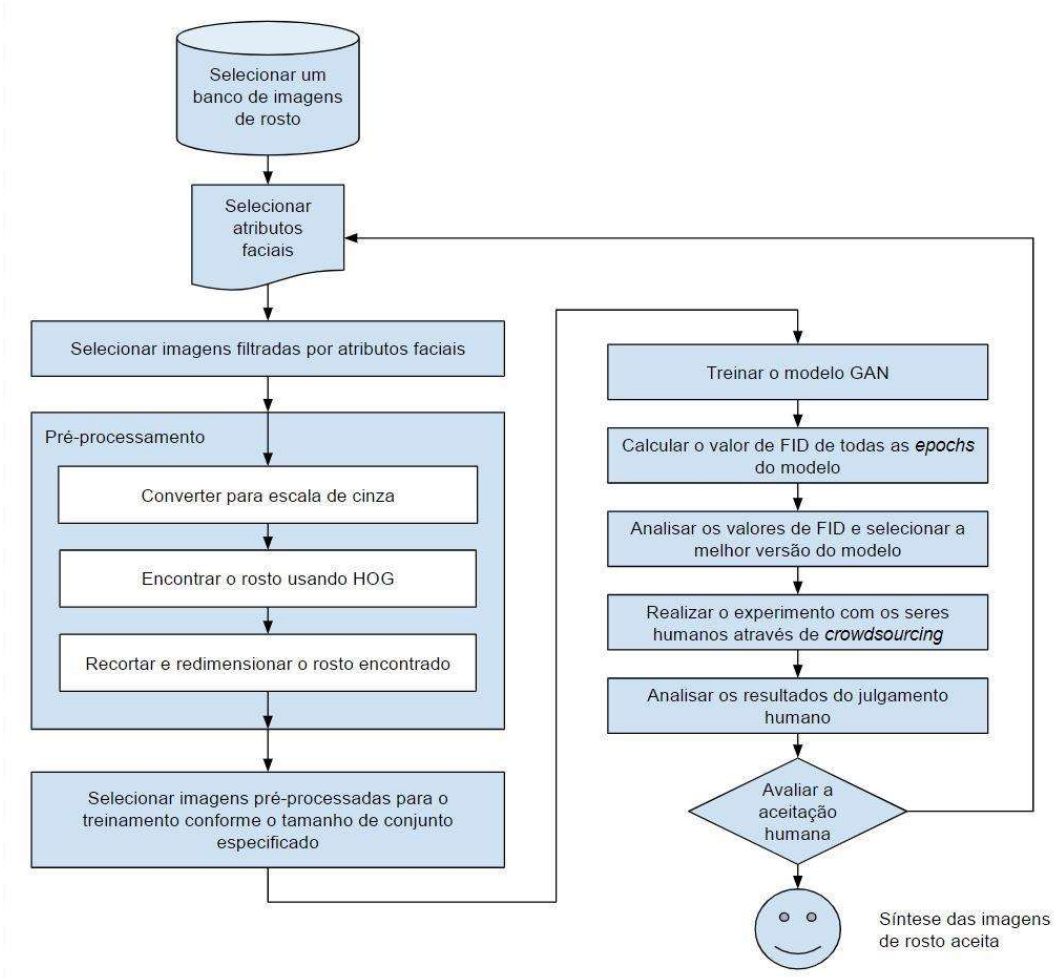

Figura 3.1: Método proposto e instanciado.

Levando-se em consideração que GAN é uma área emergente e a natureza de engenharia dos treinamentos dos modelos, o método (Figura 3.1) proposto e aplicado pelo presente estudo buscou permitir uma pesquisa experimental sobre se e quais atributos faciais interferem no aprendizado da GAN e, consequentemente, na produção de melhores resultados. Todas as etapas até o treinamento dos modelos GAN podem ser

feitas em laboratório e os experimentos com pessoas, em ferramentas *online* de *crowdsourcing*.

Pode ser aplicado em qualquer conjunto de imagens de rostos humanos contendo arquivos em quaisquer dimensões. As dimensões das imagens podem interferir nos resultados gerados pelos modelos e, consequentemente, na percepção das pessoas. Neste estudo, foram utilizadas imagens com 128 x 128 *pixels*, que é um formato comumente utilizado na literatura, e os resultados não foram comprometidos.

O método proposto, pode ser executado em qualquer configuração de computador com qualquer arquitetura de GAN. Além do tempo de processamento, a configuração do computador pode interferir no *batch size* dos modelos devido a restrições de memória. De maneira geral, pode-se entender o *batch size* como sendo a quantidade de exemplos de imagens reais e falsas que serão mostradas para a rede discriminadora durante o treinamento e em cada *epoch*, ou seja, a rede discriminadora não necessariamente verá todas as imagens reais e falsas de uma vez em cada *epoch*, mas um subconjunto.

Primeiramente, deve-se escolher e organizar um conjunto de imagens de rostos, que serão filtradas posteriormente para o treinamento dos modelos GAN. Para o presente estudo, foi utilizado o *dataset* CelebA (LIU et al., 2015) que é bastante comum em pesquisas de temas relacionados.

Após a organização deste conjunto, as imagens devem ser filtradas ou agrupadas por meio de um atributo facial. Podem ser utilizados mais de um atributo facial para filtrar um único conjunto, mas, no presente estudo, foram organizados grupos de dados conforme um único atributo facial por grupo.

Os atributos utilizados neste estudo, constavam no próprio *dataset* CelebA e foram anotados por seres humanos. Para pesquisas futuras, pode-se utilizar alguma técnica, por exemplo, baseada em CNNs, para filtrar de forma automática essas imagens. Também pode-se buscar formas não-supervisionadas para a realização dos agrupamentos.

A seguir, está a etapa de pré-processamento que consiste em melhorar a padronização das imagens por meio de 3 subetapas: (i) converter as imagens para escala de cinza; (ii) encontrar o rosto na imagem usando a técnica *Histogram of Oriented Gradients* (HOG) (DALAL e TRIGGS, 2005) e; (iii) recortar e redimensionar as imagens conforme o *bounding box* encontrado pela HOG.

A partir do conjunto de imagens pré-processadas, procede-se com a escolha do conjunto de imagens de treinamento, selecionando-se, de forma aleatória e sem repetição, um quantitativo de imagens necessário ao treinamento do modelo GAN.

Tendo definido uma arquitetura GAN previamente, treina-se o modelo com as imagens pré-processadas. Esta etapa pode ser realizada várias vezes, selecionando-se conjuntos de treinamento de tamanhos variados e/ou alterando outros hiperparâmetros do

modelo para avaliar-se o impacto nos resultados. Os hiperparâmetros deste estudo serão discriminados ainda na seção 4.1 (Tipos de Modelos Treinados Conforme os Dados de Entrada).

Todos os modelos deste estudo foram treinados durante 200 *epochs*. Em testes realizados anteriormente, a quantidade de 200 *epochs* foi suficiente para a arquitetura definida para a geração de bons resultados, considerando-se imagens de 128 x 128 *pixels* e um experimento-piloto realizado com pessoas.

Entre as métricas mais utilizadas recentemente, para a avalição objetiva dos resultados dos modelos GAN, está a *Fréchet Inception Distance* (FID) (HEUSEL et al., 2017). No caso do presente estudo, a FID foi utilizada para a seleção da melhor versão do modelo. Para isso, para cada modelo treinado, foi calculada a FID dos resultados produzidos em cada *epoch* desse modelo. Dessa forma, foi possível identificar-se a melhor versão de cada modelo e comparar as configurações utilizando diferentes valores de *batch size* e tamanho do espaço latente. Também foi possível utilizar as imagens sintéticas geradas na melhor *epoch* nos experimentos com humanos para a análise e a discussão dos resultados.

Após os experimentos com as pessoas, pode-se efetuar novo agrupamento de imagens de rostos, a partir do conjunto original, com base em novos atributos faciais e, assim, executar novamente o método para avaliar-se quais atributos facilitaram ou não o treinamento e o aprendizado dos modelos GAN. Isto foi realizado no presente estudo como será visto ao longo do trabalho.

3.2 **O Conjunto de Dados de Treinamento**

Para o presente estudo, a seleção do *dataset* baseou-se nos seguintes critérios: (i) *dataset* atual contendo um rosto humano em cada imagem, pois imagens com muitos rostos aumentariam o trabalho e não trariam benefícios para a discussão deste estudo; (ii) *dataset* com atributos faciais previamente anotados; (iii) contendo imagens com menos variação de luz e sombra; (iv) imagens de rosto sem oclusões e; (v) contendo um grande volume de dados para permitir experimentação com vários subconjuntos diferentes.

Entre os *datasets* encontrados na literatura, são citados alguns, a seguir, e as limitações que levaram a não utilização destes dados.

- ImageNet (DENG et al., 2009): não contém anotações de atributos faciais. Tem por objetivo catalogar objetos que pertençam a um mesmo *synset*: um grupo de dados considerados semanticamente equivalentes.

- SCUT-FBP5500 (LIANG et al., 2018): utilizado em pesquisas de predição de beleza facial. Apresenta atributos faciais anotados, mas com baixo volume de imagens. Possui somente imagens de pessoas asiáticas e caucasianas.
- CIFAR-100 (KRIZHEVSKY, 2009): as imagens têm baixa resolução (32 x 32) o que poderia dificultar a percepção humana.

Neste estudo, foi utilizado o *dataset* CelebA (LIU et al., 2015) que é um conjunto de dados recente, muito utilizado em pesquisas que envolvem tarefas de visão computacional, como por exemplo: reconhecimento de atributos faciais e detecção, edição e síntese de rostos. Este conjunto, tem 202599 imagens de rostos de celebridades em posição frontal. Cada imagem possui um único rosto humano.

Este *dataset* também foi selecionado por apresentar a maioria das imagens na mesma posição, além de frontal, na vertical. O entendimento é que isto poderia tornar o aprendizado dos modelos mais rápido e fácil, e levou-se em consideração que o objetivo do estudo era avaliar o impacto dos agrupamentos por atributos faciais em modelos com configurações diferentes.

O CelebA tem uma rica quantidade de atributos. Ao todo, são 40 atributos anotados para cada imagem. Além do identificador da imagem, os atributos são: *five o'clock shadow*, sobrancelhas arqueadas, atraente, olheiras, careca, franja, lábios grandes, nariz grande, cabelo preto, cabelo loiro, desfocado, cabelo castanho, sobrancelhas espessas, bochechudo, queixo duplo, óculos, cavanhaque, cabelo grisalho, maquiagem pesada, maçãs do rosto salientes, masculino, boca ligeiramente aberta, bigode, olhos estreitos, sem barba, rosto oval, pele pálida, nariz pontudo, recuo da linha do cabelo, bochechas rosadas, costeletas, sorrindo, cabelo liso, cabelo ondulado, usando brincos, usando chapéu, usando batom, usando colar, usando gravata e jovem.

Cada anotação contém um valor 1 (presença do atributo) ou -1 (ausência do atributo). Estas anotações podem permitir uma rica investigação e entendimento sobre a percepção humana sobre as imagens sintéticas geradas, bem como a sua relação com os atributos faciais.

A Figura 3.2, traz alguns exemplos de imagens existentes neste *dataset*.

Figura 3.2: Exemplos de imagens de rosto do *dataset* CelebA. Fonte: imagens extraídas de (LIU et al., 2015).

Os critérios utilizados para a seleção dos atributos faciais foram: (i) atributos que permitissem a seleção do maior número de imagens; (ii) exclusão de atributos masculinos, como cavanhaque e costeletas; (iii) exclusão de atributos que não estariam presentes nas imagens após o pré-processamento realizado com HOG, como cabelo preto e cabelo loiro e; (iv) exclusão dos atributos relacionados à maquiagem.

Foram selecionadas apenas imagens de mulheres para reduzir a variabilidade das imagens de treinamento e permitir a posterior análise sobre se os modelos DCGAN seriam capazes de produzir, do ponto de vista da percepção humana, imagens com características mais femininas do que masculinas.

A Tabela 3.1 apresenta os atributos que se enquadram nos critérios citados, bem como o volume de imagens que existem em cada grupo, se filtradas pelo atributo em questão, e a indicação sobre se ele foi selecionado para utilização nos treinamentos dos modelos DCGAN do presente estudo.

Tabela 3.1: Atributos do CelebA que poderiam ser utilizados para o agrupamento das imagens do presente estudo.

Atributo	Quantidade de Imagens	Selecionado?
Atraente	73910	Não selecionado
Maçãs do rosto salientes	**53497**	**Selecionado**
Sorrindo	51258	Não selecionado
Boca ligeiramente aberta	49856	Não selecionado
Sobrancelhas arqueadas	**42949**	**Selecionado**
Nariz pontudo	36540	Não selecionado
Rosto oval	**34963**	**Selecionado**
Lábios grandes	**30430**	**Selecionado**
Olhos estreitos	10334	Não selecionado.
Olheiras	8941	Não selecionado.
Nariz grande	8516	Não selecionado.
Sobrancelhas espessas	7858	Não selecionado.
Bochechudo	688	Não selecionado.
Queixo duplo	329	Não selecionado.

Procurou-se utilizar atributos de áreas diferentes do rosto e que resultassem em um volume razoável de imagens tendo em vista os volumes de imagens necessários nos treinamentos.

Dessa forma, foi selecionado um atributo de cada área do rosto conforme indicado no trabalho (CAO et al., 2018). Tendo em vista que no presente estudo foi utilizada uma arquitetura GAN baseada em convoluções, procurou-se na literatura de CNN artigos que pudessem ajudar a selecionar os melhores atributos. Vale ressaltar que CAO et al. (2018) utilizaram o CelebA.

O atributo *atraente* não foi selecionado por ser muito subjetivo. Para representar o grupo *Whole Image Group*, foi selecionado o *rosto oval*. Foi escolhido o atributo *maçãs do rosto salientes*, em detrimento do *nariz pontudo*, para o *Middle Group*, pois o primeiro apresentou mais imagens. Nos outros dois grupos, *Upper Group* e *Lower Group* foram selecionados, respectivamente, *sobrancelhas arqueadas* e *lábios grandes*.

Os atributos *sorrindo* e *boca ligeiramente aberta* não foram selecionados pois poderiam tornar mais difícil o aprendizado dos modelos do que os atributos anteriores.

Os demais atributos não foram utilizados por apresentarem baixo volume de imagens para os treinamentos e comparações entre os modelos.

3.3 **Pré-Processamento dos Dados**

Primeiramente, foi construído um subconjunto com imagens do CelebA, com o intuito de facilitar o treinamento da DCGAN. Este subconjunto foi organizado com todas as imagens contendo mulheres jovens, sem óculos e em imagens não desfocadas, de acordo com os atributos anotados e existentes no próprio CelebA. Desta forma, foram selecionadas 97609 imagens para a etapa de pré-processamento.

A Figura 3.3 exibe exemplos de imagens de mulheres filtradas dessa forma.

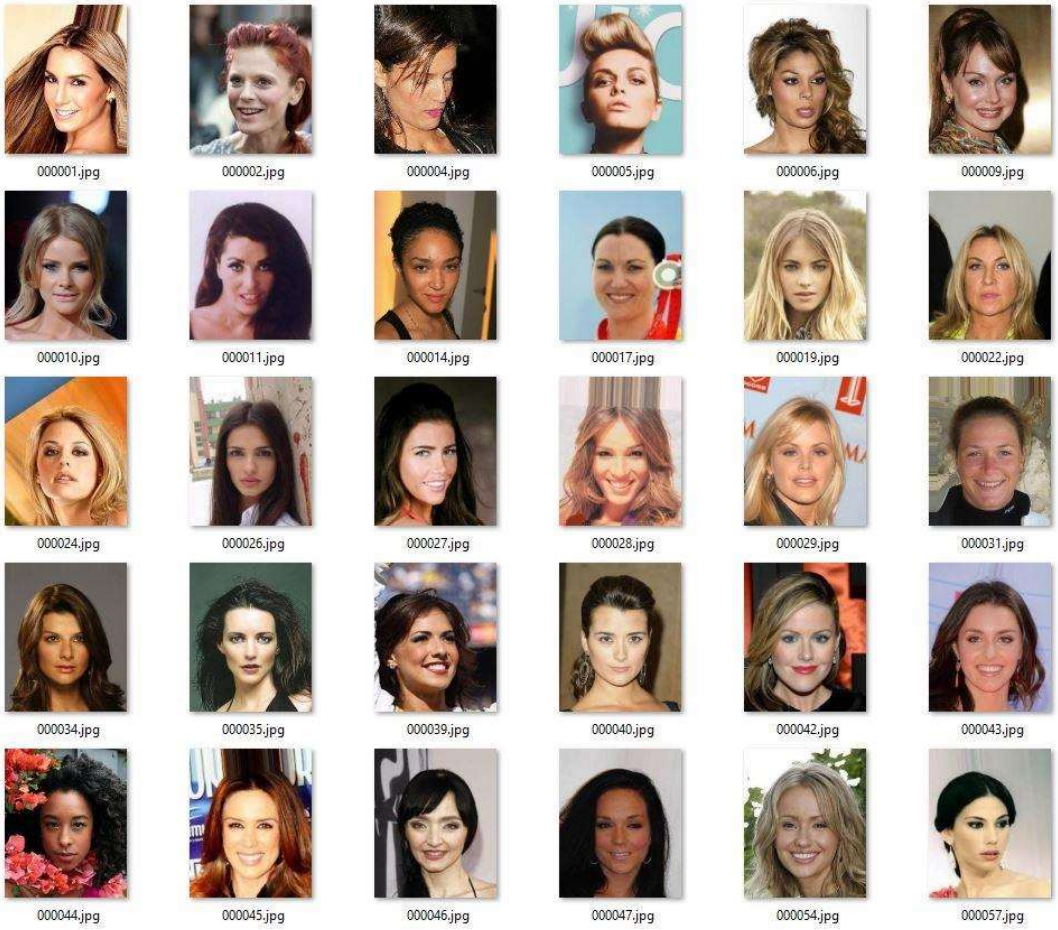

Figura 3.3: Exemplos de imagens de rosto de mulheres jovens, sem óculos e em imagens não desfocadas do *dataset* CelebA. Fonte: imagens extraídas de (LIU et al., 2015).

A escolha de um único gênero permitiu, posteriormente, perguntar às pessoas se as imagens geradas pelos modelos DCGAN traziam mais atributos masculinos ou femininos possibilitando a avaliação sobre se os modelos foram capazes de produzir imagens respeitando os atributos de gênero implícitos das imagens reais.

Imagens borradas não foram selecionadas, com o objetivo de melhorar a definição das imagens geradas e diminuir a interferência na percepção das pessoas.

Imagens contendo óculos também foram excluídas, porque a oclusão poderia dificultar o aprendizado da DCGAN e afetar a geração das novas imagens. Por exemplo, no caso de imagens agrupadas por um atributo específico como sobrancelhas arqueadas.

As imagens de idosos também foram deixadas de lado a fim de reduzir-se a variabilidade das imagens reais e facilitar o aprendizado dos modelos DCGAN.

Todas as imagens reais que foram utilizadas nos treinamentos dos modelos DCGAN foram pré-processadas.

Cada imagem foi convertida para a escala de cinza para reduzir o tempo de treinamento dos modelos. Imagens coloridas que, por exemplo, utilizem o sistema de cores RGB, possuem três canais de cor: vermelho, verde e azul e, portanto, mais informações do que uma imagem em escala de cinza, que possui apenas uma camada. Esta camada traz valores de luminosidade de cada *pixel* que representam o quão claro ou escuro é aquele *pixel*.

Como cada imagem apresenta um rosto que pode estar mais ou menos distante do observador ou ligeiramente inclinado, isso poderia trazer ainda mais dificuldades no treinamento das DCGANs e na percepção das pessoas sobre os resultados. Então, em cada imagem foi aplicada a técnica *Histogram of Oriented Gradients* (HOG) (DALAL e TRIGGS, 2005) para encontrar um rosto na imagem. Encontrado o rosto, a imagem foi recortada conforme o *bounding box* identificado pela HOG. Este recorte foi redimensionado para as dimensões de 128 x 128 *pixels* e salvo em disco.

Inicialmente, havia 97609 imagens. Após o pré-processamento, restaram 96241. A redução se deve ao fato de que a HOG não necessariamente encontrou um rosto em cada imagem, talvez por causa da luminosidade ou da posição da cabeça da pessoa, por exemplo. Ainda assim, a quantidade de imagens foi suficiente para a realização dos experimentos do presente estudo.

A Figura 3.4, exibe algumas imagens após a etapa de pré-processamento. Observa-se que os rostos ficaram bem enquadrados.

Figura 3.4: Exemplos de imagens de rostos após a etapa de pré-processamento.

3.4 **Configuração da GAN**

Para o presente estudo, foi desenvolvido um programa na linguagem de programação Python para treinar os modelos numa arquitetura baseada em DCGAN (RADFORD et al., 2015). Não foram utilizadas as camadas de *batch normalization*. Também foram acrescentadas algumas heurísticas indicadas em (GOODFELLOW, 2016, SALIMANS et al., 2016, CHINTALA et al., 2020), que são: (i) normalização dos *pixels* das imagens de treinamento para valores entre -1 e 1; (ii) utilização de camadas de ativação LeakyReLU na rede geradora e na discriminadora; (iii) utilização de *label smoothing* dos rótulos das imagens reais e das imagens sintéticas antes de treinar a discriminadora; (iv) utilização de *noisy labels* para as imagens reais e para as sintéticas antes do treinamento da discriminadora; e (v) utilização de otimizador Adam na discriminadora e na GAN propriamente, com *learning rate* igual a 0,0002 e *decay rate* exponencial para as

estimativas do primeiro momento igual a 0,5. Essas heurísticas ajudaram a estabilizar o programa diminuindo as chances de ocorrer o problema *Mode Collapse*. A *loss function* utilizada na discriminadora e na DCGAN foi a *Binary Crossentropy*.

Tipicamente, as redes são treinadas com dados rotulados numericamente com valores inteiros. No caso das GANs, geralmente se utiliza o valor 1 para imagens reais e 0 para imagens falsas. Durante o treinamento, a tendência é que a rede se torne muito confiante na sua avaliação, identificando os exemplos como pertencentes a uma ou exclusivamente a outra classe, o que pode comprometer a sua capacidade de aprender a partir dos novos exemplos. Além disso, grandes *datasets* frequentemente apresentam dados rotulados erroneamente e neste caso a alteração em partes dos rótulos dos dados de treinamento evitaria que a rede se especializasse em dados mal rotulados.

As heurísticas *label smoothing* e *noisy labels*, servem para deixar a rede menos confiante nos seus resultados durante o treinamento, melhorando a sua capacidade de generalização para dados futuros. Ambas modificam os rótulos das imagens reais e das falsas antes delas serem entregues à rede discriminadora, durante o seu treinamento. A *label smoothing* acrescenta um pequeno valor entre -0,3 e 0,2 ao rótulo das imagens reais e um valor entre 0,0 e 0,3 aos rótulos das imagens falsas (CHINTALA et al., 2020). A *noisy labels*, por sua vez, inverte parte dos rótulos das imagens reais e das falsas (CHINTALA et al., 2020). No caso do presente estudo, a cada *epoch*, 5% dos rótulos das imagens reais e das falsas eram trocadas e os rótulos eram modificados conforme a heurística *label smoothing*.

Foram realizados testes, previamente, para se chegar aos valores usados nesta pesquisa, como número de *epochs*, tamanho do espaço latente e *batch size*. Por exemplo, no computador utilizado no presente estudo não foi possível executar-se o programa para um *batch size* maior que 100. Mais detalhes sobre a configuração desse computador podem ser vistos na seção 3.6 (Configuração do Computador Utilizado no Treinamento dos Modelos DCGAN).

Nas tabelas 3.2 e 3.3 é possível ver a arquitetura e as camadas da rede discriminadora e da geradora. Nas camadas de Conv2D da discriminadora e nas camadas de Conv2DTranspose e de Conv2D da geradora, foi utilizado *padding*: *same*. Ao lado dos valores de *params*, quando indicado, aparecem entre parênteses os valores de espaço latente (EL). As camadas Conv2DTranspose (*Transposed Convolution Layer*) funcionam como uma convolução inversa, ampliando as dimensões dos dados. Em literaturas mais antigas é chamada de *Deconvolution Layer*.

Tabela 3.2: Arquitetura da rede discriminadora.

Layer	*Activation*	*Filter*	*Kernel*	*Stride*	*Output Size*	*Params*
Conv2D	LeakyReLU	128	(6, 6)	-	128×128×128	4 736
Conv2D	LeakyReLU	128	(6, 6)	(2, 2)	64×64×128	589 952
Conv2D	LeakyReLU	128	(6, 6)	(2, 2)	32×32×128	589 952
Conv2D	LeakyReLU	128	(6, 6)	(2, 2)	16×16×128	589 952
Conv2D	LeakyReLU	128	(6, 6)	(2, 2)	8×8×128	589 952
Flatten	-	-	-	-	8 192	0
Dropout	-	-	-	-	8 192	0
Dense	Sigmoid	-	-	-	1	8 193
Total de parâmetros treináveis:						2 372 737

Tabela 3.3: Arquitetura da rede geradora.

Layer	*Activation*	*Filter*	*Kernel*	*Stride*	*Output Size*	*Params*
Dense	LeakyReLU	-	-	-	8 192	827 392 (EL: 100) 2 056 192 (EL: 250) 4 104 192 (EL: 500)
Reshape	-	-	-	-	8×8×128	0
Conv2DTranspose	LeakyReLU	128	(4, 4)	(2, 2)	16×16×128	262 272 (EL: 100, 250 e 500)
Conv2DTranspose	LeakyReLU	128	(4, 4)	(2, 2)	32×32×128	
Conv2DTranspose	LeakyReLU	128	(4, 4)	(2, 2)	64×64×128	
Conv2DTranspose	LeakyReLU	128	(4, 4)	(2, 2)	128×128×128	
Conv2D	Hyperbolic tangent	1	(5, 5)	-	128×128×1	3 201
Total de parâmetros treináveis:						1 879 681 (EL: 100) 3 108 481 (EL: 250) 5 156 481 (EL: 500)

É importante ressaltar que a cada *epoch*, a discriminadora treina separadamente e duas vezes: uma com as imagens reais e outra com as imagens sintéticas produzidas pela geradora. Quando a GAN treina, ainda na mesma *epoch* que a discriminadora, o

aprendizado da discriminadora é desabilitado. Esse processo também é uma das heurísticas estudadas e testadas anteriormente.

3.5 **FID: Avaliação da Melhor Versão do Modelo**

Em (BORJI, 2019), várias métricas objetivas são discutidas para avaliar os resultados dos modelos generativos sob diversos aspectos. Para este estudo, decidiu-se utilizar a *Fréchet Inception Distance* (FID) (HEUSEL et al., 2017). Na revisão de literatura, ela apareceu como uma das principais métricas utilizadas atualmente. No caso da FID, quanto menor o valor calculado, melhor é o resultado do modelo.

A FID foi selecionada porque apresenta um bom desempenho em termos de distinção, robustez e eficiência computacional (BORJI, 2019). Além disso, conforme (BORJI, 2019), a FID parece permitir uma classificação mais próxima da percepção humana e tende a ser mais robusta a ruídos nas imagens reais.

Cada modelo DCGAN foi treinado durante 200 *epochs*. Em cada *epoch* foram produzidas 30 imagens sintéticas.

Selecionou-se a *epoch* com os melhores resultados em cada modelo, aplicando-se a FID para comparar o conjunto de imagens geradas em cada *epoch,* com o conjunto de imagens reais utilizado no treinamento daquele modelo. Dessa forma, foi possível calcular-se a FID de cada *epoch* de um mesmo modelo. As pontuações de FID de cada *epoch* foram ordenadas e aquela com menor valor indicou em que momento aquele modelo obteve os melhores resultados, isto é, o momento em que aquele modelo foi capaz de gerar as imagens sintéticas mais próximas das imagens reais.

3.6 **Configuração do Computador Utilizado no Treinamento dos Modelos DCGAN**

A Tabela 3.4 apresenta a configuração do computador utilizado no treinamento de todos os modelos DCGAN do presente estudo. Não foram utilizadas outras configurações de computador citadas em outras pesquisas devido a limitações de recursos.

Os programas foram escritos em linguagem de programação Python e utilizou-se as *Application Programming Interfaces* (APIs) chamadas TensorFlow e Keras, comumente utilizadas em projetos de *Deep Learning* por permitirem maior facilidade para a construção de redes neurais artificiais.

A configuração do computador utilizado foi colocada aqui para permitir a reprodutibilidade do presente estudo. Esta configuração é mais acessível do que as configurações utilizadas nos trabalhos recentes como em (KARRAS et al., 2019, KARRAS et al., 2017). Por exemplo, a placa de vídeo utilizada no presente estudo: Nvidia GeForce RTX 2060 Super, custava cerca de R$ 3 mil à época, enquanto a placa de vídeo utilizada por KARRAS et al. (2019, 2017), Tesla V100 GPUs, cerca de R$ 28 mil. Assim, a configuração do computador do presente estudo, possibilitou inclusive uma discussão sobre se seria possível alcançar bons resultados com os modelos DCGAN, sob o ponto de vista da percepção humana, com um computador mais acessível.

Tabela 3.4: Configuração do computador utilizado nos treinamentos dos modelos DCGAN deste estudo.

Linguagem de Programação	Python (versão: 3.7.7)
APIs	Keras (versão: 2.3.1) e TensorFlow (versão: 2.1.0)
GPU	Nvidia GeForce RTX 2060 Super
CPU	AMD Ryzen 5 1600X Six-Core 3.60 GHz
RAM	32 GB
Sistema Operacional	Windows 10 Home Single Language, 64 bits

Além do treinamento dos modelos, também foram realizados os cálculos de FID e a organização dos conjuntos de dados de treinamento neste computador.

Vale ressaltar que as atualizações deste computador, como as do sistema operacional e as da GPU, foram desabilitadas durante os treinamentos dos modelos para permitir uma comparação mais realista entre os resultados dos modelos e evitar assim interferências, por exemplo, nos tempos de processamento.

4. Experimentos

Este capítulo apresenta as configurações dos modelos DCGAN treinados, assim como os tempos dos treinamentos e os valores de FID obtidos nas *epochs* com melhores resultados de cada modelo.

Todos os modelos foram treinados utilizando-se o método proposto no capítulo anterior e as imagens geradas pelos modelos foram utilizadas nos experimentos.

Neste capítulo também será apresentada a estrutura e as perguntas dos questionários aplicados por meio da plataforma Appen.

4.1 Tipos de Modelos Treinados Conforme os Dados de Entrada

Foram treinados modelos DCGAN em duas categorias diferentes conforme o conjunto de dados utilizado: modelos do Tipo 1 e do Tipo 2.

4.1.1 Modelos do Tipo 1

Os modelos do Tipo 1 foram treinados com imagens organizadas e pré-processadas conforme definido no capítulo anterior: mulheres jovens, sem óculos e em imagens não desfocadas.

Cada modelo recebeu seu próprio conjunto de imagens de treinamento, selecionadas aleatoriamente.

As tabelas 4.1 a 4.7, exibem os modelos do Tipo 1. Os modelos que alcançaram o melhor FID, considerando-se o mesmo tamanho de conjunto de dados, estão destacados em negrito. As imagens produzidas por estes modelos foram separadas para os experimentos com as pessoas, para a análise e a discussão sobre a percepção humana.

Tabela 4.1: Configuração dos modelos do Tipo 1 com tamanho do conjunto de imagens de treinamento igual a 1 mil.

Tamanho do conjunto de dados	*Batch size*	Espaço Latente	*Epoch* com melhor FID	FID	Tempo aproximado de processamento (em horas)
1000	10	100	109	84,46	0,70
1000	10	250	71	85,87	0,72
1000	10	500	68	88,56	0,72

1000	50	100	123	79,64	0,53
1000	**50**	**250**	**127**	**79,05**	**0,53**
1000	50	500	164	89,77	0,55
1000	100	100	178	89,11	0,51
1000	100	250	191	92,05	0,52
1000	100	500	200	88,32	0,51

Tabela 4.2: Configuração dos modelos do Tipo 1 com tamanho do conjunto de imagens de treinamento igual a 2,5 mil.

Tamanho do conjunto de dados	*Batch size*	Espaço Latente	*Epoch* com melhor FID	FID	Tempo aproximado de processamento (em horas)
2500	10	100	98	83,12	1,73
2500	10	250	161	81,60	1,75
2500	10	500	39	87,68	1,77
2500	50	100	167	76,74	1,30
2500	**50**	**250**	**197**	**76,29**	**1,30**
2500	50	500	115	79,14	1,31
2500	100	100	177	76,75	1,24
2500	100	250	173	77,89	1,26
2500	100	500	113	85,01	1,26

Tabela 4.3: Configuração dos modelos do Tipo 1 com tamanho do conjunto de imagens de treinamento igual a 5 mil.

Tamanho do conjunto de dados	*Batch size*	Espaço Latente	*Epoch* com melhor FID	FID	Tempo aproximado de processamento (em horas)
5000	10	100	94	77,08	3,59
5000	10	250	80	80,92	3,55
5000	10	500	43	81,78	3,60
5000	50	100	133	71,10	2,67
5000	**50**	**250**	**156**	**70,52**	**2,68**
5000	50	500	138	73,93	2,58
5000	100	100	178	73,25	2,51
5000	100	250	192	75,28	2,50
5000	100	500	193	80,01	2,51

Tabela 4.4: Configuração dos modelos do Tipo 1 com tamanho do conjunto de imagens de treinamento igual a 7,5 mil.

Tamanho do conjunto de dados	*Batch size*	Espaço Latente	*Epoch* com melhor FID	FID	Tempo aproximado de processamento (em horas)
7500	10	100	49	74,86	5,20
7500	10	250	33	79,87	5,28
7500	10	500	49	79,01	5,40
7500	50	100	186	70,37	3,93
7500	50	250	127	70,52	3,88
7500	50	500	129	70,24	3,94
7500	**100**	**100**	**199**	**65,42**	**3,78**
7500	100	250	158	75,87	3,79
7500	100	500	199	73,76	3,78

Tabela 4.5: Configuração dos modelos do Tipo 1 com tamanho do conjunto de imagens de treinamento igual a 10 mil.

Tamanho do conjunto de dados	*Batch size*	Espaço Latente	*Epoch* com melhor FID	FID	Tempo aproximado de processamento (em horas)
10000	10	100	33	74,80	7,22
10000	10	250	45	74,11	6,97
10000	10	500	35	77,64	7,06
10000	50	100	73	71,44	5,10
10000	50	250	91	69,38	5,19
10000	50	500	99	74,39	5,12
10000	**100**	**100**	**165**	**67,84**	**5,04**
10000	100	250	103	72,39	5,06
10000	100	500	147	71,36	5,05

Tabela 4.6: Configuração dos modelos do Tipo 1 com tamanho do conjunto de imagens de treinamento igual a 12,5 mil.

Tamanho do conjunto de dados	*Batch size*	Espaço Latente	*Epoch* com melhor FID	FID	Tempo aproximado de processamento (em horas)
12500	10	100	191	76,49	8,62
12500	10	250	22	78,01	8,67
12500	10	500	134	77,45	8,62
12500	50	100	133	68,91	6,32
12500	50	250	67	72,80	6,38
12500	50	500	84	74,56	6,35
12500	100	100	188	70,06	6,32
12500	**100**	**250**	**85**	**68,21**	**6,28**

12500	100	500	96	70,58	6,28

Tabela 4.7: Configuração dos modelos do Tipo 1 com tamanho do conjunto de imagens de treinamento igual a 15 mil.

Tamanho do conjunto de dados	*Batch size*	Espaço Latente	*Epoch* com melhor FID	FID	Tempo aproximado de processamento (em horas)
15000	10	100	34	78,55	10,23
15000	10	250	22	81,88	10,42
15000	10	500	42	83,09	10,37
15000	50	100	67	73,76	7,63
15000	**50**	**250**	**191**	**67,50**	**7,65**
15000	50	500	52	72,75	7,65
15000	100	100	156	71,90	7,49
15000	100	250	97	78,31	7,41
15000	100	500	106	72,77	7,47

4.1.2 **Modelos do Tipo 2**

Foram treinados quatro modelos DCGAN com imagens filtradas com atributos faciais geométricos rotulados existentes no *dataset* CelebA.

Primeiramente, o conjunto de dados de treinamento para cada um desses modelos foi organizado da mesma forma que os modelos do Tipo 1. Em seguida, as imagens foram filtradas novamente, usando-se um atributo facial diferente para cada um dos quatro modelos do Tipo 2. Dessa forma, cada modelo recebeu um conjunto único dados.

Os atributos utilizados foram: (i) maçãs do rosto salientes; (ii) sobrancelhas arqueadas; (iii) rosto oval e; (iv) lábios grandes.

Para investigar se a filtragem pelos novos atributos faciais facilitou o aprendizado da DCGAN, cada modelo do Tipo 2 recebeu um conjunto de 5 mil imagens de treinamento. Esse valor foi escolhido para, posteriormente, avaliar-se se foi possível produzir-se melhores resultados com uma quantidade menor de dados, tendo em vista que os melhores modelos do Tipo 1, em termos de FID, foram aqueles treinados com 7,5 mil, 10 mil e 12,5 mil imagens.

Esses modelos foram treinados com *batch size*: 100 e espaço latente: 100, conforme os melhores modelos do Tipo 1.

A Tabela 4.8, apresenta a configuração e os resultados de FID dos modelos do Tipo 2. Nela é possível ver que o modelo baseado no atributo rosto oval alcançou o melhor

valor de FID (59,59), inclusive sendo menor que o melhor modelo do Tipo 1, com tamanho de conjunto de dados igual a 7,5 mil, que obteve o valor de FID (65,42).

Tabela 4.8: Configuração e FID dos modelos do Tipo 2.

Atributo utilizado na filtragem	Tamanho do conjunto de dados	*Batch size*	Espaço Latente	*Epoch* com melhor FID	FID	Tempo aproximado de processamento (em horas)
Maçãs do rosto salientes	5000	100	100	199	63,69	2,49
Sobrancelhas arqueadas	5000	100	100	191	66,41	2,47
Rosto oval	**5000**	**100**	**100**	**197**	**59,59**	**2,45**
Lábios grandes	5000	100	100	192	73,14	2,48

4.2 **Questionário**

Em cada experimento com os seres humanos, foram mostradas, para cada pessoa, todas as 30 imagens sintéticas produzidas na *epoch* com melhor FID de cada modelo DCGAN selecionado. Foi solicitado às pessoas que avaliassem as imagens conforme as perguntas indicadas.

Cada experimento teve seu próprio conjunto de avaliadores: um total de 100 pessoas por experimento.

Antes da realização dos experimentos propriamente, foi realizado um experimento-piloto na plataforma Appen, com 5 pessoas, para verificar se o questionário estava funcionando corretamente e se as pessoas conseguiriam realizar o experimento até o final sem deixar nenhuma resposta em branco.

São apresentadas a seguir, as instruções que constavam em todos os experimentos:

1) Para cada experimento
    a) *Visão Geral*
        i) *Estamos estudando a geração de imagens de rostos de pessoas usando inteligência artificial. Portanto, apresentamos imagens geradas em nosso projeto para avaliar o quão perto de imagens de rosto reais estão as imagens que geramos; e*

ii) *Ajude-nos a determinar se há pessoas nas imagens que fornecemos e se elas se parecem mais com homens ou com mulheres.*

b) *Passos*

i) *Selecione seu gênero;*

ii) *Selecione sua faixa etária;*

iii) *Observe a imagem;*

iv) *Se houver uma pessoa na imagem, marque a opção apropriada;*

v) *Indique o quão perto de um rosto real está o rosto na imagem; e*

vi) *Indique se o rosto da imagem possui mais características femininas ou masculinas.*

c) *Dicas de Regras:*

i) *Olhe para uma imagem de cada vez e responda às perguntas indicadas ao lado de cada imagem.*

2) Para cada imagem

a) *Há uma pessoa nesta imagem? (campo obrigatório)*

i) *sim*

ii) *não*

b) Se a pessoa respondeu *sim* à primeira pergunta, a ferramenta exibia a seguinte pergunta: *Quão próxima de uma imagem real está essa imagem? Escolha um número na escala abaixo, onde 1 significa "muito distante de uma imagem real" e 5 significa "muito próximo de uma imagem real". (campo obrigatório)*

i) *1 (muito distante de uma imagem real)*

ii) *2*

iii) *3*

iv) *4*

v) *5 (muito próximo de uma imagem real)*

c) Se a pessoa respondeu *sim* à primeira pergunta, a ferramenta exibia a seguinte pergunta: *Esta imagem tem mais características masculinas ou femininas? (campo obrigatório)*

i) *masculino*

ii) *feminino*

Para os experimentos que envolviam algum outro atributo facial específico como o rosto oval e maçãs do rosto salientes nos modelos do Tipo 2, o experimento trazia ainda informações, por exemplo, como as que seguem:

1) No experimento
    a) *Passos*
        i) *Indique se o rosto na imagem tem um rosto oval*.
2) Para cada imagem
    a) *Você consegue ver um rosto oval nesta imagem? (campo obrigatório)*
        i) *sim*
        ii) *não*

Além das perguntas sobre as imagens, também foi perguntado à cada pessoa:

a) *Como você define seu gênero? (campo obrigatório)*
    i) *masculino*
    ii) *feminino*
    iii) *não binário*
    iv) *prefiro não dizer*
b) *Quantos anos você tem? (campo obrigatório)*
    i) *menos de 20 anos*
    ii) *21 a 30 anos*
    iii) *31 a 40 anos*
    iv) *41 a 50 anos*
    v) *51 a 60 anos*
    vi) *mais de 60 anos*
    vii) *prefiro não dizer*

5. Resultados e Discussão

5.1 Ajustes dos Modelos GAN

Treinar modelos GAN ainda é uma tarefa de engenharia. É comum na literatura a realização de testes com configurações diferentes para poder estabilizar-se a arquitetura a ser utilizada no treinamento do modelo.

As tabelas 4.1 a 4.7, do capítulo anterior, mostraram que encontrar os valores ideais para o *batch size*, o tamanho do espaço latente e o tamanho do conjunto de imagens de treinamento, não é uma tarefa trivial. Por exemplo, o pior valor de FID, alcançado pelos modelos do Tipo 1, foi 92,05 no modelo com tamanho de conjunto dados igual a 1 mil, mas com as mesmas configurações de *batch size* (100) e de espaço latente (250), foi possível obter-se o melhor resultado de FID (68,21) entre os modelos com tamanho do conjunto de dados igual a 12,5 mil.

Na Tabela 5.1 pode-se observar os melhores modelos do Tipo 1, em termos de FID, considerando-se o tamanho do conjunto dados. A linha destacada em negrito, indica o modelo que alcançou o melhor resultado (FID: 65,42). Observava-se que os quatro últimos modelos apresentaram valores de FID muito próximos e com configurações diferentes. Ao contrário do que poderia ser pensado previamente, aumentar o número de exemplos de treinamento não provocou uma melhoria sensível nos resultados dos modelos (diminuição maior dos valores de FID).

Tabela 5.1: Configurações dos melhores modelos do Tipo 1.

Tamanho do conjunto de dados	*Batch size*	Espaço Latente	*Epoch* com melhor FID	FID	Tempo aproximado de processamento (em horas)
1000	50	250	127	79,05	0,53
2500	50	250	197	76,29	1,30
5000	50	250	156	70,52	2,68
7500	**100**	**100**	**199**	**65,42**	**3,78**
10000	100	100	165	67,84	5,04
12500	100	250	85	68,21	6,28
15000	50	250	191	67,50	7,65

Na Tabela 5.2, pode-se ver as médias e as medianas dos valores de FID dos modelos do Tipo 1 agrupados por tamanho do conjunto de dados. Os modelos que foram treinados com 10 mil imagens apresentaram a melhor média, o menor desvio padrão e a melhor mediana.

Tabela 5.2: Médias e medianas dos valores de FID dos modelos do Tipo 1 por tamanho do conjunto de imagens de treinamento.

Tamanho do conjunto de dados	Média	Desvio padrão	Mediana
1000	86,31	4,25	88,32
2500	80,47	3,87	79,14
5000	75,99	3,96	75,28
7500	73,32	4,38	73,76
10000	**72,59**	**2,82**	**72,39**
12500	73,01	3,55	72,80
15000	75,61	4,84	73,76

Para os modelos que receberam 7,5 mil imagens, entre os quais está o modelo que alcançou o melhor FID (65,42), o desvio padrão (4,38) apresentou uma dispersão cerca de 55,32% maior do que a dos modelos que receberam 10 mil imagens. Estes resultados reforçam a literatura no que diz respeito à dificuldade de se encontrar a melhor configuração para um modelo GAN.

Na Figura 5.1, observa-se um gráfico com as dispersões desses valores. Pode-se ver que, tendo em vista a arquitetura DCGAN e as heurísticas utilizadas nesta pesquisa, os melhores modelos treinados para o Tipo 1 estão entre aqueles que utilizaram entre 7,5 mil e 12,5 mil imagens de treinamento.

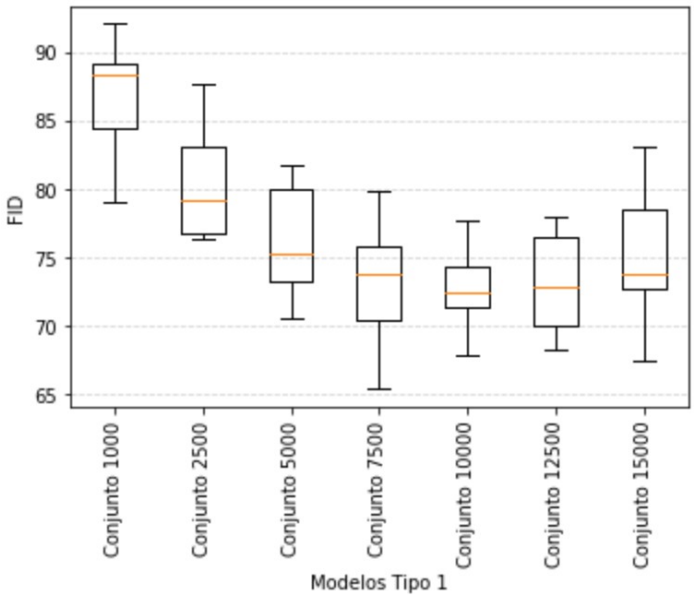

Figura 5.1: *Boxplot* dos valores de FID dos modelos do Tipo 1 agrupados pelo tamanho do conjunto de dados.

Na Figura 5.2, é apresentado um gráfico construído a partir dos valores de FID dos melhores modelos do Tipo 1 considerando-se o tamanho do conjunto de dados.

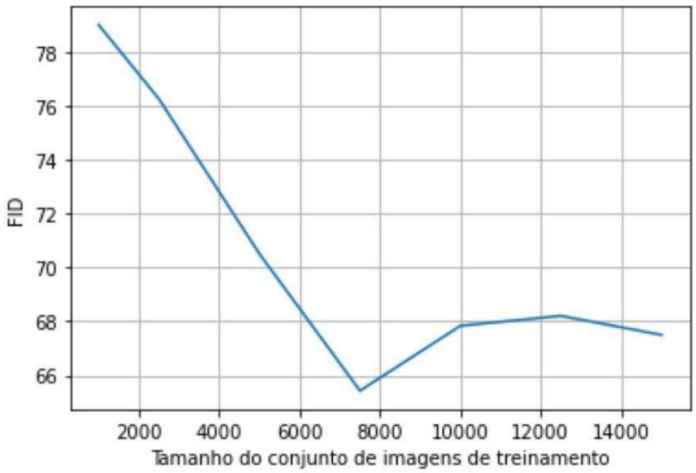

Figura 5.2: FID dos melhores modelos do Tipo 1 considerando-se o mesmo tamanho de conjunto de dados.

Os resultados aqui apontados sugerem que os modelos treinados com um volume de imagens a partir do tamanho 7,5 mil, apresentam resultados relativamente próximos, variando pouco. Para futuras pesquisas, seria interessante treinar modelos com maiores

quantidades de imagens para chegar-se a uma conclusão mais segura sobre se ao aumentar-se a quantidade de imagens de treinamento, para além de 15 mil, obter-se-iam resultados melhores de FID.

No contexto deste estudo, os melhores modelos do Tipo 1, foram aqueles treinados com uma quantidade de dados no intervalo entre 7,5 mil e 12,5 mil, e serviram como *baseline* para as discussões que se seguirão, onde serão apresentados os resultados dos modelos do Tipo 2, treinados com imagens filtradas por atributos faciais geométricos.

5.2 O *Trade-off*: Tamanho do Conjunto de Treinamento e a Variabilidade dos Exemplos

O treinamento dos modelos do Tipo 2 teve por objetivo explorar o impacto da diminuição da variabilidade das imagens de treinamento, filtradas por atributos faciais geométricos, no resultado da DCGAN e no tamanho necessário do conjunto de dados.

As tabelas que constam a partir desta seção irão apresentar a mediana dos votos dos avaliadores que participaram dos experimentos na plataforma Appen.

Foi utilizada a mediana para diminuir o impacto que as melhores ou as piores imagens sintéticas geradas em cada conjunto de 30 imagens, pudessem causar aos resultados. Um modelo GAN pode, por exemplo, gerar poucas imagens muito boas e o restante muito ruim enquanto aprende. Assim, no presente estudo, o melhor modelo foi aquele que apresentou o melhor resultado, seja no valor de FID seja na percepção humana, considerando-se todas as 30 imagens geradas por cada modelo.

Na Tabela 5.3, estão organizados todos os valores de FID dos melhores modelos do Tipo 1 e dos modelos treinados do Tipo 2. Foram destacados em negrito os 3 modelos de cada tipo que obtiveram os melhores valores de FID.

Tabela 5.3: Valores de FID dos modelos do Tipo 1 e do Tipo 2.

Tipo do Modelo	1							2			
Atributo utilizado na filtragem	-							**Maçãs do rosto salientes**	**Sobrancelhas arqueadas**	**Rosto oval**	Lábios grandes
Tamanho do conjunto de dados	1000	2500	5000	**7500**	**10000**	**12500**	15000	**5000**	**5000**	**5000**	5000
FID	79,05	76,29	70,52	**65,42**	**67,84**	**68,21**	67,50	**63,69**	**66,41**	**59,59**	73,14

Observa-se que o melhor valor de FID (59,59) alcançado entre todos os modelos foi do modelo do Tipo 2 que recebeu imagens de treinamento filtradas pelo atributo rosto oval.

Este modelo recebeu 5 mil imagens para o seu treinamento, menos do que os modelos do intervalo definido como *baseline*, que receberam entre 7,5 mil a 12,5 mil. Por exemplo, o FID do melhor modelo do Tipo 1, que recebeu 7,5 mil imagens, foi de 65,42.

Os modelos treinados com as imagens com maçãs do rosto salientes e sobrancelhas arqueadas também alcançaram bons resultados: o FID do primeiro (63,69) ficou abaixo do intervalo de valores da *baseline*: 65,42 a 68,21; e o valor do segundo (66,41), ficou dentro deste intervalo.

Para cada um dos modelos da Tabela 5.3, foi realizado um experimento na plataforma Appen, no qual foram exibidas as 30 imagens sintéticas geradas pelo modelo e 100 avaliadores puderam informar se havia um rosto humano em cada imagem.

Cada experimento recebeu um conjunto aleatório de avaliadores. Estes experimentos tiverem por objetivo permitir a análise da percepção humana sobre as imagens sintéticas geradas pelos modelos DCGAN treinados.

Na Figura 5.3 observa-se como foi a percepção dos avaliadores, nos experimentos, sobre a presença de um rosto humano nas imagens sintéticas geradas por cada modelo. Cada modelo gerou 30 imagens e cada *boxplot* representa a dispersão do total de votos que cada imagem recebeu. Esse gráfico foi construído a partir das respostas dadas à pergunta: *Há uma pessoa nesta imagem?*

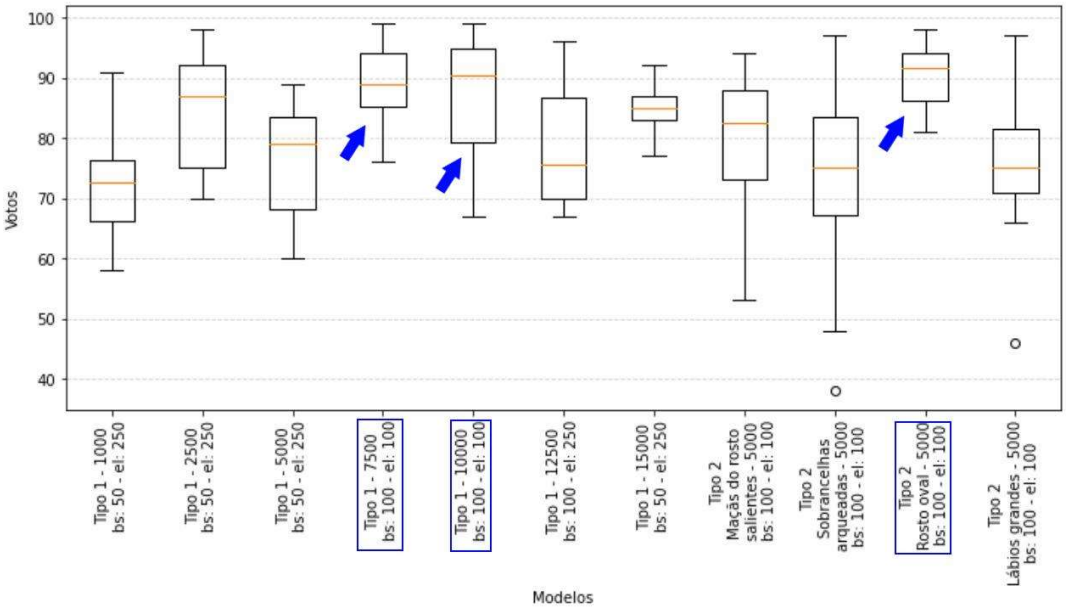

Figura 5.3: Presença de um rosto humano nas imagens sintéticas geradas, conforme indicado pelos avaliadores.

Percebe-se que o modelo treinado com rostos ovais obteve o maior número de votos, considerando-se o 50º percentil, e apresentou uma das menores distribuições.

A Tabela 5.4 apresenta as medidas de tendência central e de dispersão dos votos sobre a presença de um rosto humano nas imagens sintéticas geradas nos três melhores modelos conforme observado na Figura 5.3. Como foram geradas 30 imagens por cada modelo, primeiro foi calculado o total de votos recebidos por cada imagem e em seguida calculou-se as demais medidas como a média e a mediana entre o total de votos de cada imagem.

Tabela 5.4: Medidas de tendência central e de dispersão dos valores dos votos sobre a presença de um rosto humano nas imagens sintéticas geradas pelos modelos do Tipo 1 treinados com 7,5 mil e 10 mil imagens e do Tipo 2 treinado com rostos ovais.

Tipo do Modelo	1		2
Atributo utilizado na filtragem	-		**Rosto oval**
Tamanho do conjunto de dados	7500	10000	**5000**
FID	65,42	67,84	**59,59**
Média	89,43 ±6,10	87,13 ±8,71	**90,47 ±4,91**
Mediana	89,0	90,5	**91,5**
Menor valor	76	67	**81**
Maior valor	**99**	**99**	98
Amplitude	23	32	**17**

Observa-se como o modelo do Tipo 2 treinado com imagens contendo rostos ovais obteve os melhores resultados, exceto no item maior valor. Isto reforça a hipótese do presente estudo sobre a possibilidade de redução dos dados de treinamento, a partir da filtragem das imagens por um atributo facial específico, para a obtenção de resultados compatíveis com os dos modelos treinados com mais dados e sem este tratamento de filtragem, tendo em vista o critério de avaliação da percepção humana. Na Figura 5.4, são exibidas as imagens sintéticas geradas por este modelo e apresentadas aos avaliadores.

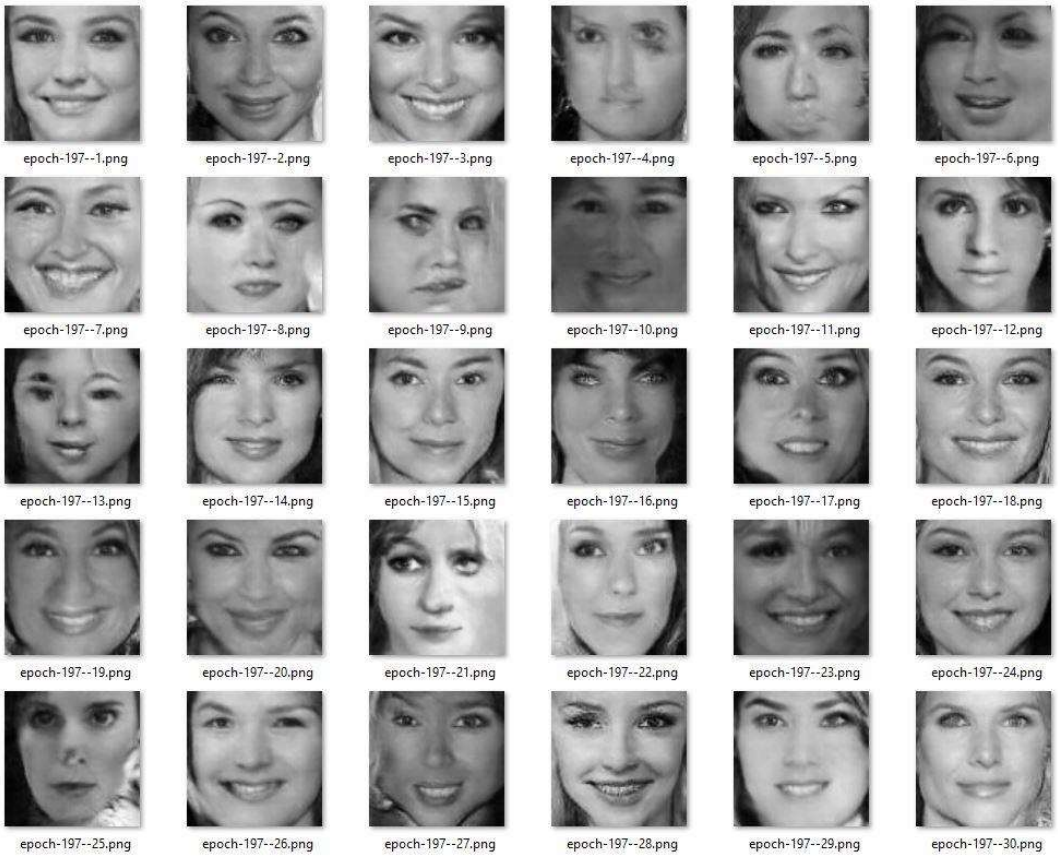

Figura 5.4: Imagens sintéticas geradas pelo modelo do Tipo 2 treinado com imagens filtradas pelo atributo rosto oval.

Na Figura 5.5, são exibidas as imagens sintéticas geradas pelo modelo treinado com 10 mil imagens. Entre os modelos do Tipo 1, este modelo obteve um dos melhores resultados FID (67,84) e o melhor resultado na percepção humana (90,5). Esta figura foi colocada aqui para permitir a visualização e a comparação com as imagens da Figura 5.4. A Figura 5.4 (modelo treinado com rostos ovais) parece apresentar rostos mais bem definidos.

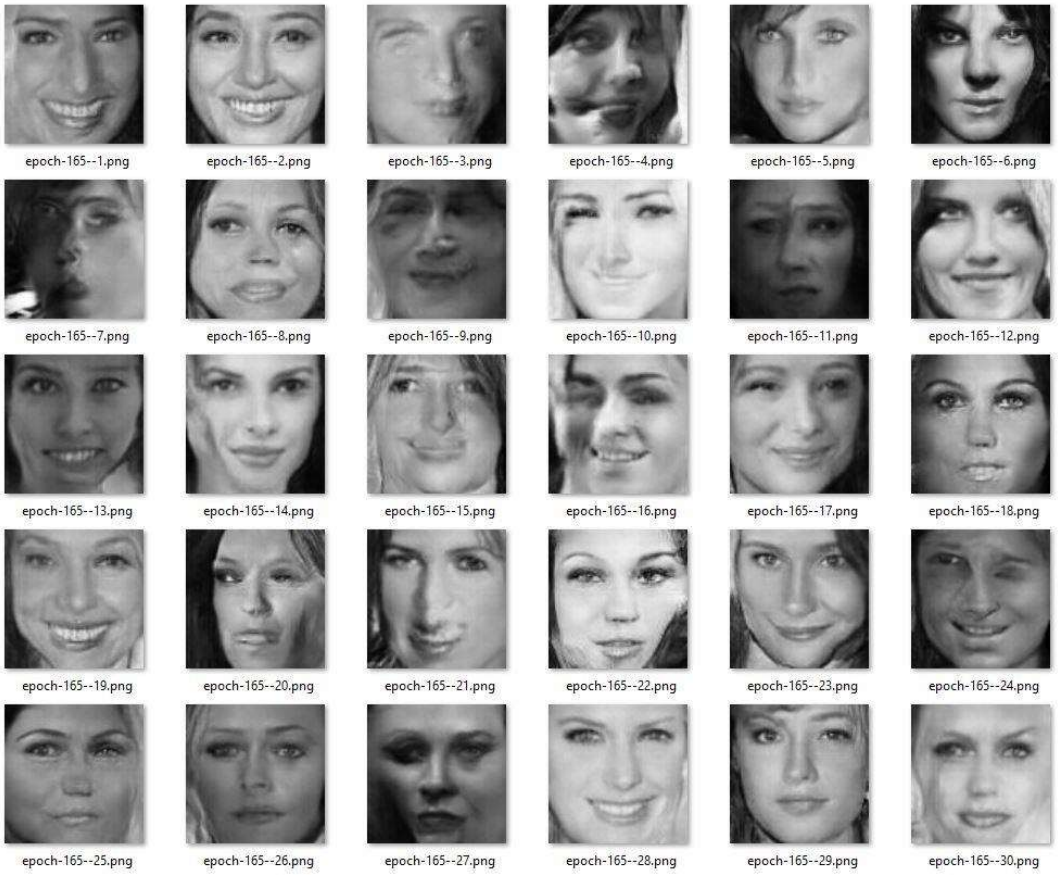

Figura 5.5: Imagens sintéticas geradas pelo modelo do Tipo 1 treinado com 10 mil imagens.

Com o objetivo de reforçar os resultados observados na Tabela 5.4 e o que foi discutido até este ponto, foram realizados dois testes de hipótese T de Student com os valores dos votos alcançados por cada modelo. Foram avaliados dois casos: (a) entre os resultados do modelo do Tipo 1 treinado com 7,5 mil imagens e os do modelo do Tipo 2 treinado com rostos ovais e; (b) entre os resultados do modelo do Tipo 1 treinado com 10 mil imagens e os do modelo do Tipo 2 treinado com rostos ovais. Em ambos os casos, a hipótese nula foi a de que os valores médios (esperados) seriam estatisticamente iguais, e a hipótese alternativa, diferentes.

Esta investigação ocorreu da seguinte forma:

(i) Primeiro, verificou-se se os conjuntos em cada caso (a) e (b), apresentavam distribuição normal.

(ii) Em seguida, avaliou-se se os conjuntos apresentavam variâncias iguais.

(iii) Então, foi aplicado o Teste de hipótese T de Student.

Os testes aplicados em cada etapa e os resultados foram colocados na Tabela 5.5 para facilitar a visualização.

Tabela 5.5: Testes de hipótese realizados para os casos (a) e (b) a fim de avaliar se, estatisticamente, os conjuntos são iguais.

	Caso (a)		Caso (b)	
Testes	Modelo Tipo 1 (7500)	Modelo Tipo 2 (rostos ovais)	Modelo Tipo 1 (10000)	Modelo Tipo 2 (rostos ovais)
Os conjuntos apresentam distribuição normal?				
Shapiro-Wilk	Sim	Sim	Não	Sim
D'Agostino's K-squared	Sim	Sim	Sim	Sim
Os conjuntos apresentam variâncias iguais				
Bartlett	Sim		Não	
Levene	-		Não	
As amostras são estatisticamente diferentes?				
T de Student	Não		Não	
p-value				
	0,480		0,078	

Após a aplicação do teste de hipótese T de Student, observa-se na Tabela 5.5 que o *p-value* obtido para o caso (a) foi de 0,480 e para o caso (b), 0,078, resultados que indicam que em ambos os casos as amostras são estatisticamente iguais, o que reforça a hipótese de que é possível reduzir o conjunto de treinamento pela filtragem dos dados através de um atributo facial específico.

O modelo treinado com imagens que possuem maçãs de rostos salientes também apresentou bom resultado na percepção humana quando comparado aos modelos da *baseline*.

O modelo treinado com imagens filtradas pelo atributo sobrancelhas arqueadas parece ter alcançado uma mediana similar à do modelo do Tipo 1 treinado com 12,5 mil imagens. Neste caso, a confirmação de que as medianas foram muito próximas, pode ser conferida na Tabela 5.6, onde a mediana do primeiro foi de 75,0 e a do segundo, 75,5. Esta tabela apresenta a mediana dos votos sobre a existência de um rosto humano nas imagens sintéticas geradas e os valores de FID dos modelos DCGAN do Tipo 1 e do Tipo 2.

Tabela 5.6: Mediana dos votos indicando a presença de um rosto humano nas imagens sintéticas geradas e os valores de FID dos modelos dos tipos 1 e 2.

Tipo do Modelo	1	2

Atributo utilizado na filtragem	-							**Maçãs do rosto salientes**	**Sobrancelhas arqueadas**	**Rosto oval**	Lábios grandes
Tamanho do conjunto de dados	1000	2500	5000	**7500**	**10000**	**12500**	15000	**5000**	**5000**	**5000**	5000
FID	79,05	76,29	70,52	**65,42**	**67,84**	**68,21**	67,50	**63,69**	**66,41**	**59,59**	73,14
Mediana de votos indicando presença de um rosto humano	72,5	87,0	79,0	89,0	90,5	75,5	85,0	82,5	75,0	**91,5**	75,0

Importante destacar o resultado do modelo do Tipo 1 treinado com 2,5 mil imagens, que apresentou FID igual a 76,29, mas teve um dos melhores valores de mediana de votos (87,0) na percepção humana. Neste caso, a percepção humana parece divergir da métrica FID. Em (KARRAS et al., 2020) há uma indicação de que a métrica FID não é uma métrica ideal quando o modelo é treinado com pequenos conjuntos de dados o que pode explicar este caso. Seria interessante realizar, em trabalhos futuros, uma investigação mais aprofundada para este caso, por exemplo, treinando modelos com volumes ainda menores de dados.

Pela Figura 5.3 observa-se que o modelo do Tipo 1 treinado com 15 mil imagens apresentou a menor distribuição dos votos dos avaliadores e uma das melhores medianas. Verificando-se as imagens geradas por este modelo, exibidas na Figura 5.6, supõe-se que se trata de um caso de *Mode Collapse*. Por exemplo, as imagens destacadas contêm rostos que parecem ter os mesmos olhos e boca. Isto sugere ainda que a métrica FID não foi capaz de identificar essa situação e que talvez as imagens sintéticas com rostos muito similares tenham interferido na percepção humana.

Figura 5.6: Imagens geradas pelo modelo do Tipo 1 treinado com 15 mil imagens. Esse modelo parece ter entrado em *Mode Collapse*. Observa-se que os olhos e a boca das imagens destacadas parecem os mesmos.

Uma pergunta que pode surgir naturalmente ao se observar a Tabela 5.6 é: os valores de FID e das medianas possuem correlação? Em outras palavras, conforme os valores de FID dos modelos diminuem, a mediana dos votos da percepção humana aumenta? Assim, foi realizado um teste de correlação, conforme descrito a seguir, onde a hipótese nula foi a de que os conjuntos são independentes e a hipótese alternativa, dependentes (correlacionados).

Essa questão foi investigada, neste estudo, da seguinte forma:

(i) Os FIDs e as medianas foram organizados em dois conjuntos diferentes para comporem os pontos do gráfico;

(ii) Ordenou-se de forma crescente o conjunto de FIDs;

(iii) Com o objetivo de realizar-se um teste de correlação, primeiro verificou-se:

a. Se os dois conjuntos, FIDs e medianas, apresentavam distribuição normal;
b. Se os dois conjuntos apresentavam a mesma variância;

(iv) Foram aplicados dois testes para verificar a normalidade dos dois conjuntos: Shapiro-Wilk e D'Agostino's K-squared. Os dois conjuntos apresentaram distribuição normal.

(v) Para verificar se as variâncias eram iguais, foi aplicado o teste de Bartlett. Os conjuntos apresentaram a mesma variância.

(vi) Com os resultados dos dois itens anteriores, aplicou-se o teste de correlação de Pearson. O *p-value* obtido foi de 0,106. Assim, o teste apontou que não há correlação.

Em (HEUSEL et al., 2017) é indicado que a FID tem um desempenho comprometido em imagens com rotações em parte da imagem. Talvez, as imagens sintéticas que não ficaram tão boas tenham apresentado distorções como as citadas em (HEUSEL et al., 2017) o que poderia explicar um mau desempenho da métrica FID e, consequentemente, a não correlação com a percepção humana.

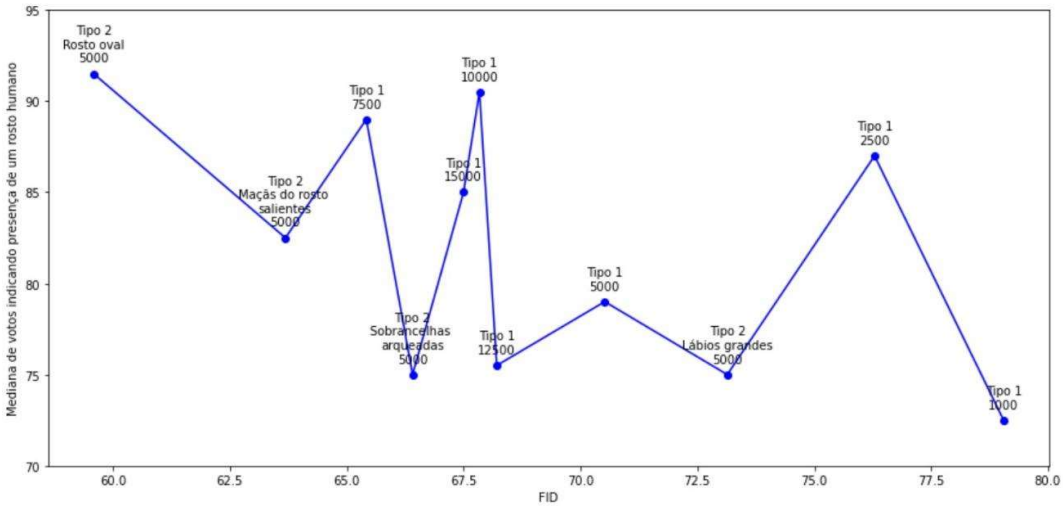

Figura 5.7: Gráfico que apresenta a relação entre o FID calculado para cada modelo e a percepção humana sobre a presença de um rosto humano nas imagens sintéticas geradas.

A relação de não linearidade entre os FIDs e as medianas também pode ser observada pelo gráfico da Figura 5.7. Neste gráfico, quanto mais à esquerda estiver situado o

modelo, melhor a sua avaliação em termos de FID. E quanto mais para cima, melhor na avaliação na percepção humana. Como exemplo, o melhor modelo em termos de FID e na percepção humana está situado no canto superior esquerdo: modelo treinado com imagens contendo rostos ovais.

5.3 **Observações Demográficas: o Gênero, a Idade e o País Impactam na Forma como os Indivíduos Percebem as Imagens Geradas?**

Nesta seção, foram organizadas tabelas com informações agrupadas por gênero, faixa etária e país de origem dos avaliadores, construídas a partir das perguntas do questionário: (i) *Como você define seu gênero?*; (ii) *Quantos anos você tem?*; e (iii) *Esta imagem tem mais características masculinas ou femininas?* A informação sobre o país de origem dos avaliadores é disponibilizada pela plataforma Appen após a realização do experimento.

Na Tabela 5.7, foram organizados, para cada modelo, a quantidade de votos dada por cada grupo de avaliadores considerando-se o gênero do avaliador conforme respondido no questionário.

Como cada experimento apresentou quantidades diferentes de avaliadores de cada grupo, além da mediana dos votos, foi calculada a razão entre a mediana e o total de avaliadores daquele grupo para permitir uma melhor comparação entre os resultados.

Tabela 5.7: Percepção humana sobre a presença de um rosto humano nas imagens sintéticas, geradas nos modelos do Tipo 1 e do Tipo 2, por gênero dos avaliadores.

Tipo do Modelo	1							2			
Atributo utilizado na filtragem	-							**Maçãs do rosto salientes**	**Sobrancelhas arqueadas**	**Rosto oval**	Lábios grandes
Tamanho do conjunto de dados	1000	2500	5000	**7500**	**10000**	**12500**	15000	**5000**	**5000**	**5000**	5000
FID	79,05	76,29	70,52	**65,42**	**67,84**	**68,21**	67,50	**63,69**	**66,41**	**59,59**	73,14
Gênero *masculino*											
Total de avaliadores	69	66	71	58	60	68	70	66	65	62	65
Mediana de votos indicando presença de um rosto humano	50,5	59,0	54,5	51,0	54,5	51,5	58,0	53,0	48,5	58,0	47,5
Razão entre a mediana de votos e o total de avaliadores	0,73	**0,89**	0,77	**0,88**	**0,91**	0,76	0,83	0,80	0,75	**0,94**	0,73
Gênero *feminino*											

Total de avaliadores	28	31	26	36	36	30	29	33	34	37	34
Mediana de votos indicando presença de um rosto humano	19,0	24,5	21,0	33,0	32,0	22,5	27,0	28,0	25,0	33,0	27,0
Razão entre a mediana de votos e o total de avaliadores	0,68	0,79	0,81	**0,92**	**0,89**	0,75	**0,93**	0,85	0,74	**0,89**	0,79
Gênero *não binário*											
Total de avaliadores	0	0	0	1	0	0	0	0	1	0	1
Mediana de votos indicando presença de um rosto humano	0	0	0	1,0	0	0	0	0	1,0	0	1,0
Razão entre a mediana de votos e o total de avaliadores	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	1,00	0,00	1,00
Gênero *prefiro não dizer*											
Total de avaliadores	3	3	3	5	4	2	1	1	0	1	0
Mediana de votos indicando presença de um rosto humano	2,0	2,0	3,0	5,0	3,0	2,0	0	1,0	0	1,0	0
Razão entre a mediana de votos e o total de avaliadores	0,67	0,67	1,00	1,00	0,75	1,00	0,00	1,00	0,00	1,00	0,00

As proporções de homens e mulheres, nos experimentos, foi relativamente a mesma, sendo a quantidade de homens em média: 65,45 ±3,92, e de mulheres, 32,18 ±3,46.

Tendo em vista que poucas pessoas indicaram como gênero as opções: *não binário* e *prefiro não dizer*, os votos desses gêneros não foram considerados.

Na percepção dos homens, o modelo com melhor resultado foi o modelo treinado com imagens contendo rostos ovais, cuja razão entre a mediana de votos e o total de avaliadores foi 0,94.

No caso das mulheres, a percepção foi parecida, mas o melhor modelo de todos foi o do Tipo 1 treinado com 7,5 mil imagens (0,92), considerando-se que o modelo treinado com 15 mil imagens, embora tenha alcançado uma mediana maior (0,93), parece ter entrado em *Mode Collapse*, conforme descrito na seção anterior. Ainda assim, o modelo do Tipo 2, treinado com imagens com rostos ovais, obteve um resultado tão bom quanto o modelo do Tipo 1 treinado com 10 mil imagens. Esses resultados parecem sugerir que não há uma diferença de percepção entre os gêneros masculino e feminino.

A Tabela 5.8 foi organizada da mesma forma que a anterior, porém os avaliadores foram agrupados em faixas etárias.

Tabela 5.8: Percepção humana sobre a presença de um rosto humano nas imagens sintéticas, geradas nos modelos do Tipo 1 e do Tipo 2, por faixa etária dos avaliadores.

Tipo do Modelo	1							2			
Atributo utilizado na filtragem	-							**Maçãs do rosto salientes**	**Sobrancelhas arqueadas**	**Rosto oval**	Lábios grandes
Tamanho do conjunto de dados	1000	2500	5000	**7500**	**10000**	**12500**	15000	**5000**	**5000**	**5000**	5000
FID	79,05	76,29	70,52	**65,42**	**67,84**	**68,21**	67,50	**63,69**	**66,41**	**59,59**	73,14
Faixa Etária *menos de 20 anos*											
Total de avaliadores	5	2	1	4	7	2	6	5	3	7	2
Mediana de votos indicando presença de um rosto humano	2,5	2,0	1,0	3,0	7,0	1,0	6,0	4,0	2,0	6,0	2,0
Razão entre a mediana de votos e o total de avaliadores	0,50	1,00	1,00	0,75	1,00	0,50	1,00	0,80	0,67	0,86	1,00
Faixa Etária *21 a 30 anos*											
Total de avaliadores	63	70	69	43	45	70	47	58	48	49	59
Mediana de votos indicando presença de um rosto humano	46,5	58,5	54,0	38,5	38,0	57,0	37,0	43,0	33,5	45,0	37,0
Razão entre a mediana de votos e o total de avaliadores	0,74	**0,84**	0,78	**0,90**	**0,84**	0,81	0,79	0,74	0,70	**0,92**	0,63
Faixa Etária *31 a 40 anos*											
Total de avaliadores	16	14	18	35	31	13	32	26	27	26	30
Mediana de votos indicando presença de um rosto humano	8,0	12,0	12,5	29,5	27,5	6,0	30,0	24,0	22,0	22,5	27,0
Razão entre a mediana de votos e o total de avaliadores	0,50	0,86	0,69	0,84	0,89	0,46	**0,94**	**0,92**	0,81	0,87	**0,90**
Faixa Etária *41 a 50 anos*											
Total de avaliadores	9	10	5	7	12	7	4	8	16	12	7
Mediana de votos indicando presença de um rosto humano	9,0	8,5	4,0	7,0	12,0	5,0	4,0	8,0	12,0	12,0	6,0
Razão entre a mediana de votos e o total de avaliadores	1,00	0,85	0,80	1,00	1,00	0,71	1,00	1,00	0,75	1,00	0,86
Faixa Etária *51 a 60 anos*											
Total de avaliadores	5	4	4	7	4	6	10	0	5	5	1
Mediana de votos indicando presença de um rosto humano	4,0	4,0	4,0	7,0	4,0	6,0	9,0	0	5,0	5,0	1,0

Razão entre a mediana de votos e o total de avaliadores	0,80	1,00	1,00	1,00	1,00	1,00	0,90	0,00	1,00	1,00	1,00
Faixa Etária *mais de 60 anos*											
Total de avaliadores	0	0	0	0	0	0	0	1	1	1	1
Mediana de votos indicando presença de um rosto humano	0	0	0	0	0	0	0	1,0	1,0	1,0	1,0
Razão entre a mediana de votos e o total de avaliadores	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	1,00	1,00	1,00
Faixa Etária *prefiro não dizer*											
Total de avaliadores	2	0	3	4	1	2	1	2	0	0	0
Mediana de votos indicando presença de um rosto humano	1,0	0	3,0	4,0	1,0	2,0	0	2,0	0	0	0
Razão entre a mediana de votos e o total de avaliadores	0,50	0,00	1,00	1,00	1,00	1,00	0,00	1,00	0,00	0,00	0,00

As faixas etárias *menos de 20 anos*, *mais de 60 anos* e *prefiro não dizer* foram desconsideradas por apresentarem poucos avaliadores e pouca variação nos resultados da razão entre a mediana e o total de avaliadores nos experimentos.

As faixas etárias *41 a 50 anos* e *51 a 60 anos* apresentaram pouca variação na razão entre a mediana e o total de avaliadores e foram desconsideradas nesta análise.

Na faixa etária *21 a 30 anos*, novamente o modelo treinado com imagens com rostos ovais obteve o melhor resultado entre todos os modelos. A faixa etária *31 a 40 anos* apresentou resultados diferentes da anterior, mas o modelo treinado com rostos ovais apresentou o resultado 0,87, ficando entre os modelos do Tipo 1 treinados com 7,5 mil (0,84) e 10 mil imagens (0,89).

No caso dessas duas faixas etárias, observa-se diferença na percepção das pessoas. Seria interessante a realização de novos experimentos com outras pessoas dessas duas faixas etárias, para se investigar possíveis variáveis da percepção humana que justifiquem essa diferença.

Na Tabela 5.9, foram organizados os votos sobre a presença de um rosto humano nas imagens sintéticas geradas, considerando-se os 3 países com mais avaliadores entre os experimentos: (i) Venezuela, com média de avaliadores entre os experimentos igual a 44,00 ±7,59; (ii) Estados Unidos, 30,64 ±7,95; e (iii) Egito, 15,82 ±5,57.

Tabela 5.9: Percepção humana sobre a presença de um rosto humano nas imagens sintéticas, geradas nos modelos do Tipo 1 e do Tipo 2, por país de origem do avaliador.

Tipo do Modelo	1							2			
Atributo utilizado na filtragem	-							**Maçãs do rosto salientes**	**Sobrancelhas arqueadas**	**Rosto oval**	Lábios grandes
Tamanho do conjunto de dados	1000	2500	5000	**7500**	**10000**	**12500**	15000	**5000**	**5000**	**5000**	5000
FID	79,05	76,29	70,52	**65,42**	**67,84**	**68,21**	67,50	**63,69**	**66,41**	**59,59**	73,14
País *Venezuela (VEN)*											
Total de avaliadores	44	44	34	49	48	29	39	51	41	57	48
Mediana de votos indicando presença de um rosto humano	29,0	37,0	27,5	45,5	43,0	25,5	34,0	45,0	32,0	51,0	39,5
Razão entre a mediana de votos e o total de avaliadores	0,66	0,84	0,81	**0,93**	**0,90**	0,88	0,87	0,88	0,78	**0,89**	0,82
País *Estados Unidos (USA)*											
Total de avaliadores	35	37	39	27	31	46	28	30	19	18	27
Mediana de votos indicando presença de um rosto humano	25,0	32,0	31,0	23,0	28,5	35,0	24,0	25,0	13,0	17,0	18,0
Razão entre a mediana de votos e o total de avaliadores	0,71	**0,86**	0,79	0,85	**0,92**	0,76	**0,86**	0,83	0,68	**0,94**	0,67
País *Egito (EGY)*											
Total de avaliadores	15	14	23	8	8	20	25	15	19	9	18
Mediana de votos indicando presença de um rosto humano	11,0	10,5	15,0	4,0	5,0	13,0	18,0	8,0	13,0	8,0	10,0
Razão entre a mediana de votos e o total de avaliadores	**0,73**	**0,75**	0,65	0,50	0,63	0,65	0,72	0,53	0,68	**0,89**	0,55

O modelo treinado com imagens de rostos ovais obteve o melhor resultado entre os avaliadores dos Estados Unidos e do Egito: 0,94 e 0,89, respectivamente. Embora este modelo não tenha alcançado o melhor resultado entre os avaliadores da Venezuela, ficou com o terceiro melhor resultado (0,89), ficando muito próximo do resultado do modelo do Tipo 1 treinado com 10 mil imagens (0,90).

As razões entre a mediana de votos e o total de avaliadores calculada para os modelos no caso dos avaliadores do Egito ficaram diferentes dos outros dois países. Isso parece indicar diferença na percepção humana sobre as imagens sintéticas para este povo.

A Tabela 5.10 foi construída para permitir a discussão sobre se as imagens sintéticas geradas pelos modelos apresentavam um rosto humano com características mais

femininas do que masculinas, na percepção dos avaliadores. Como cada modelo gerou 30 imagens, aqui também foi aplicada a mediana dos votos e no caso dos grupos organizados por gênero, foi calculada a razão entre as medianas de votos dos modelos e o total de avaliadores daquele gênero naquele experimento.

É importante relembrar que o questionário somente apresentava a pergunta *Esta imagem tem mais características masculinas ou femininas?* se o avaliador indicasse, primeiramente, haver um rosto humano na imagem sintética. Dessa forma, os números indicados nesta tabela, baseiam-se somente nas imagens onde foi percebido um rosto por pelo menos um avaliador.

Tabela 5.10: Percepção humana sobre a presença de um rosto feminino nas imagens sintéticas, geradas nos modelos do Tipo 1 e do Tipo 2, considerando-se o total de avaliadores e o agrupamento destes por meio dos gêneros masculino e feminino.

Tipo do Modelo	1							2			
Atributo utilizado na filtragem	-							**Maçãs do rosto salientes**	**Sobrancelhas arqueadas**	**Rosto oval**	Lábios grandes
Tamanho do conjunto de dados	1000	2500	5000	**7500**	**10000**	**12500**	15000	**5000**	**5000**	**5000**	5000
FID	79,05	76,29	70,52	**65,42**	**67,84**	**68,21**	67,50	**63,69**	**66,41**	**59,59**	73,14
Todos os avaliadores											
Mediana de votos indicando rosto feminino	67,0	84,5	78,0	**88,0**	**88,5**	75,0	84,0	81,0	75,0	**91,5**	81,5
Mediana de votos indicando rosto masculino	2,0	1,0	1,0	1,5	1,0	1,0	1,5	1,0	1,5	2,0	2,0
Gênero *masculino*											
Total de avaliadores	69	66	71	58	60	68	70	66	65	62	65
Mediana de votos indicando rosto feminino	48,0	56,5	54,0	50,0	54,0	51,0	57,0	53,0	47,0	57,0	54,5
Razão entre a mediana de votos e o total de avaliadores	0,70	0,86	0,76	0,86	**0,90**	0,75	0,81	**0,88**	0,72	**0,92**	0,84
Gênero *feminino*											
Total de avaliadores	28	31	26	36	36	30	29	33	34	37	34
Mediana de votos indicando rosto feminino	17,5	24,5	21,0	32,0	31,0	22,5	27,0	27,0	25,5	33,0	26,0

Razão entre a mediana de votos e o total de avaliadores	0,63	0,79	0,81	**0,89**	**0,86**	0,75	**0,93**	0,82	0,75	**0,89**	0,76

Considerando-se o total de avaliadores, os resultados de todos os modelos indicaram presença de um rosto feminino nas imagens sintéticas geradas, alcançando mais de 50% dos votos em todos os modelos, considerando a mediana. Isso indica que os modelos DCGAN foram capazes de aprender as características femininas das imagens de treinamento e que o resultado corresponde com a percepção humana.

O modelo treinado com rostos ovais apresentou a melhor mediana entre todos os modelos considerando todos os avaliadores.

A percepção entre homens e mulheres foi ligeiramente diferente. O melhor modelo avaliado no caso dos homens foi o treinado com rostos ovais. No caso das mulheres, o melhor modelo foi o do Tipo 1 treinado com 15 mil imagens. Se considerar-se que este modelo parece estar na situação de *Mode Collapse*, conforme visto na seção anterior, os melhores modelos passam a ser o do Tipo 1 treinado com 7,5 mil imagens e o do Tipo 2 treinado com 5 mil imagens de rostos com formato oval.

5.4 **Quão Próximo das Imagens Reais Estão as Imagens Sintéticas Geradas?**

Foi pedido aos avaliadores para indicarem uma nota, para cada imagem sintética exibida em cada experimento, em uma escala de 1 a 5, onde 1 equivalia a *muito distante de uma imagem real* e 5, *muito próximo de uma imagem real*.

É importante relembrar que o questionário somente apresentava a pergunta *Quão próxima de uma imagem real está essa imagem?* e as opções de 1 a 5, se o avaliador indicasse, primeiramente, haver um rosto humano na imagem sintética. Dessa forma, os números indicados nesta seção, baseiam-se somente nas imagens onde foi percebido um rosto.

A Tabelas 5.11 e 5.12, exibem as medianas das médias ponderadas das notas recebidas por cada imagem, para cada modelo do Tipo 1 e do 2. A primeira tabela apresenta a mediana de cada modelo e a segunda, as medianas agrupadas pelas opções de 1 a 5.

Em ambas as tabelas, as medianas foram calculadas da seguinte forma:

(i) Primeiro, para cada imagem, foi calculada a nota média ponderada pelos valores das opções de 1 a 5; e

(ii) em seguida, foi calculada a mediana entre as notas ponderadas das 30 imagens sintéticas geradas.

Tabela 5.11: Valores de medianas, das notas médias ponderadas de cada imagem (opções de 1 a 5), para cada modelo, sobre o quão próximo as imagens sintéticas estão das imagens reais.

Tipo do Modelo	1							2			
Atributo utilizado na filtragem	-							**Maçãs do rosto salientes**	**Sobrancelhas arqueadas**	**Rosto oval**	Lábios grandes
Tamanho do conjunto de dados	1000	2500	5000	**7500**	**10000**	**12500**	15000	**5000**	**5000**	**5000**	5000
FID	79,05	76,29	70,52	**65,42**	**67,84**	**68,21**	67,50	**63,69**	**66,41**	**59,59**	73,14
Mediana das notas médias ponderadas	10,60	**14,93**	12,97	13,50	**16,47**	12,27	12,70	13,77	13,03	**17,70**	11,43

Observa-se na Tabela 5.11 que o modelo treinado com 5 mil imagens de rostos ovais obteve a melhor nota, seguido do modelo treinado com 10 mil imagens. Isso reforça o entendimento de que o primeiro modelo foi capaz de alcançar um bom resultado, do ponto de vista da percepção humana, mesmo utilizando uma quantidade menor de imagens no seu treinamento.

Tabela 5.12: Valores das medianas dos votos por nota (opções de 1 a 5) recebido por cada imagem em cada modelo.

Tipo do Modelo	1							2			
Atributo utilizado na filtragem	-							**Maçãs do rosto salientes**	**Sobrancelhas arqueadas**	**Rosto oval**	Lábios grandes
Tamanho do conjunto de dados	1000	2500	5000	**7500**	**10000**	**12500**	15000	**5000**	**5000**	**5000**	5000
FID	79,05	76,29	70,52	**65,42**	**67,84**	**68,21**	67,50	**63,69**	**66,41**	**59,59**	73,14
Mediana dos votos *1*	**36,0**	14,0	16,0	**24,5**	16,5	**32,5**	24,0	22,5	16,0	15,5	15,0
Mediana dos votos *2*	27,5	**31,5**	30,0	24,5	20,0	19,5	**34,0**	19,0	22,0	17,0	**26,5**
Mediana dos votos *3*	8,0	15,0	17,0	20,5	**24,5**	8,5	**22,5**	20,5	18,0	**22,5**	18,5
Mediana dos votos *4*	3,0	**13,5**	9,0	9,5	**16,0**	3,0	4,0	11,0	12,5	**20,5**	6,0
Mediana dos votos *5*	2,0	4,0	2,0	4,0	**5,0**	**9,5**	2,0	5,0	3,0	**9,0**	1,0

Pela Tabela 5.12, observa-se que o modelo treinado com rostos ovais está entre os modelos que mais receberam notas 3, 4 e 5, o que sugere que as pessoas perceberam as imagens geradas por este modelo como sendo as que estiveram mais próximas do real.

5.5 As Imagens Geradas nos Modelos do Tipo 2, Treinados com Imagens Filtradas com Atributos Faciais Geométricos, Apresentaram esses Atributos conforme a Percepção Humana?

Foi analisado se as imagens sintéticas apresentaram os atributos faciais geométricos utilizados na filtragem das imagens de treinamento dos modelos do Tipo 2.

É importante relembrar que a pergunta referente à presença do atributo geométrico na imagem gerada somente foi feita nos casos em que o avaliador indicou ter percebido um rosto humano na imagem.

A Tabela 5.13 apresenta a mediana de votos dos avaliadores, sobre a presença dos atributos geométricos de rosto, nas imagens sintéticas geradas pelos modelos do Tipo 2.

Tabela 5.13: Percepção humana sobre a presença dos atributos faciais geométricos nas imagens sintéticas, geradas pelos modelos do Tipo 2.

Atributo utilizado na filtragem	Maçãs do rosto salientes	Sobrancelhas arqueadas	Rosto oval	Lábios grandes
Tamanho do conjunto de dados	5000	5000	5000	5000
FID	63,69	66,41	59,59	73,14
Mediana dos votos da presença do atributo nas imagens geradas	47,0	**54,5**	**52,5**	34,5

Pela Tabela 5.13, pode-se observar que os valores das medianas dos votos foram baixos considerando-se que 100 avaliadores participaram de cada experimento e, portanto, cada imagem poderia receber uma nota entre 0 e 100.

Talvez os resultados fossem melhores se os avaliadores tivessem sido treinados para identificar esses atributos e/ou se houvesse uma imagem de exemplo com a indicação do atributo. Neste estudo, evitou-se utilizar imagens de exemplo, tendo em vista que a qualidade dos resultados dos modelos foi diferente, o que poderia influenciar a avaliação das pessoas introduzindo algum viés.

É importante relembrar que os rótulos dos dados utilizados neste estudo já estavam presentes no *dataset* CelebA. O processo de anotação das imagens de treinamento pode

ser criticado para se buscar uma melhor rotulagem dos dados, o que por sua vez poderia afetar o treinamento dos modelos DCGAN e, consequentemente, a percepção humana sobre a presença dos atributos de filtragem nas imagens sintéticas produzidas.

5.6 O Treinamento dos Modelos GAN Realizado com Imagens com Rostos Ovais Realmente Impactou nos Resultados?

Neste ponto do estudo, o modelo do Tipo 2 treinado com imagens filtradas pelo atributo facial rosto oval apresentou resultados, tanto do ponto de vista objetivo (FID) quanto do subjetivo (percepção humana), que sugerem que é possível treinar-se um modelo com um volume menor de dados, a partir do agrupamento das imagens por meio de atributos faciais, e obter-se um resultado tão bom (e talvez melhor) do que modelos treinados com a mesma configuração mas sem o mesmo tratamento das imagens de treinamento.

Apesar dos resultados citados no parágrafo anterior e considerando-se a natureza de engenharia do treinamento das GANs, este estudo buscou realizar uma última análise dos valores de FID, treinando novos modelos, para verificar se, estatisticamente, os resultados dos modelos treinados com imagens filtradas por meio do atributo rosto oval são diferentes dos modelos treinados com imagens organizadas sem esta filtragem.

Para isso, foram treinados dois grupos novos de modelos DCGAN: (i) o Grupo A, que recebeu imagens contendo mulheres jovens, sem óculos e em imagens não desfocadas; e (ii) o Grupo B, que recebeu imagens nas mesmas condições, mas filtradas uma vez mais pelo atributo rosto oval. Foram treinados 30 modelos novos para cada grupo. A mesma configuração foi utilizada em ambos os grupos: tamanho do conjunto de treinamento (5 mil), *batch size* (100) e espaço latente (100).

Na Tabela 5.14, são exibidas as informações de tendência central e de dispersão dos valores de FID para os dois grupos.

Tabela 5.14: Medidas de tendência central e de dispersão dos valores de FID dos modelos dos Grupos A e B.

Medida	Grupo A	Grupo B
Média	**72,54 ±1,82**	**63,46 ±2,32**
Mediana	72,90	63,26
Menor valor	66,96	58,65
Maior valor	77,77	69,72

Amplitude	10,81	11,07

Na Figura 5.8, são exibidos os gráficos *boxplot* para melhor visualização da dispersão dos resultados dos FIDs dos modelos do Grupo A e do Grupo B.

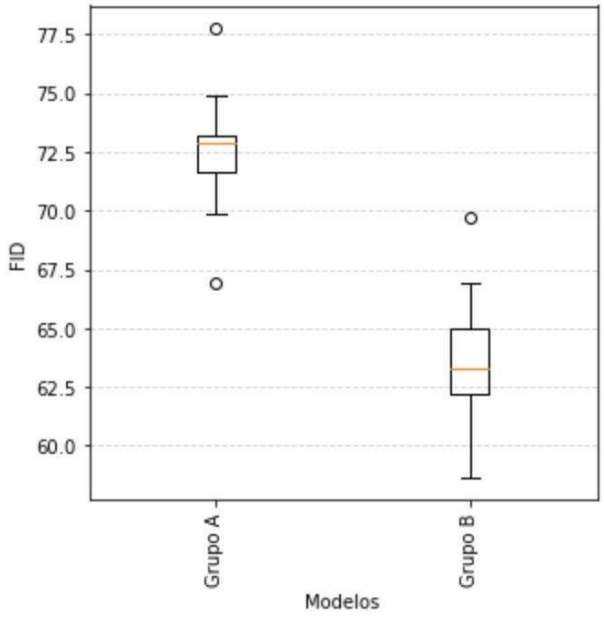

Figura 5.8: *Boxplots* das dispersões dos FIDs dos modelos dos Grupos A e B.

Como pode ser visto na Tabela 5.14 e na Figura 5.8, os modelos do Grupo B, treinados com imagens contendo rostos ovais, obtiveram os melhores resultados em termos de FID.

Com o objetivo de reforçar os resultados visualizados, foi realizado o teste de hipótese T de Student, no qual a hipótese nula era a de que os valores médios (esperados) de ambos os grupos eram estatisticamente iguais e a hipótese alternativa, diferentes.

Esta investigação ocorreu da seguinte forma:

(i) Primeiro, verificou-se:
   a. Se os conjuntos de valores de FID dos Grupo A e B apresentavam distribuição normal; e
   b. Se os dois conjuntos apresentavam a mesma variância;

(ii) Foram aplicados dois testes para verificar se os dois conjuntos apresentavam distribuição normal: Shapiro-Wilk e D'Agostino's K-squared. Em ambos os testes, o Grupo B apresentou distribuição normal e o Grupo A, não.

(iii) Para verificar se as variâncias eram iguais, como o Grupo A não apresentou distribuição normal, foi aplicado o teste de Levene. Os conjuntos apresentaram a mesma variância.

(iv) Tendo em vista os resultados dos dois itens anteriores, foi aplicado o teste de hipótese T de Student. O *p-value* obtido foi de 0,000, o que indicou que as amostras não são estatisticamente iguais.

Logo, a utilização do atributo rosto oval para filtrar as imagens de treinamento, efetivamente ajudou os modelos DCGAN do Grupo B a aprenderem melhor.

6. **Trabalhos Relacionados**

Este capítulo apresenta trabalhos relacionados ao problema da geração de uma imagem sintética por GAN que seja a síntese de um conjunto de dados, utilizando-se o menor conjunto de treinamento possível. Também são citados trabalhos que tratam da avaliação dos resultados dos modelos considerando-se a percepção humana.

Entre os trabalhos recentes, um dos que mais se aproxima da abordagem utilizada no presente estudo é (NUHA e AFIAHAYATI, 2018). Os autores investigaram qual seria o volume mínimo de dados para se obter resultados viáveis com os modelos GAN. Treinaram modelos com 3 tamanhos de conjunto de dados: 2 mil, 50 mil e 200 mil imagens. Chegaram à conclusão de que o modelo treinado com 50 mil imagens produziu os melhores resultados. Assim como o presente estudo, também utilizaram a arquitetura DCGAN nos modelos, mas não utilizaram as heurísticas encontradas em (GOODFELLOW, 2016, SALIMANS et al., 2016, CHINTALA et al., 2020) e não variaram os tamanhos de espaço latente para analisar os efeitos deste hiperparâmetro. O *batch size* foi configurado com o valor 100. As imagens de treinamento foram as do CelebA. Após os modelos serem treinados, foram geradas 15 imagens por cada modelo. Como critério de avaliação, desenvolveram uma fórmula própria para aplicar nos resultados da percepção humana. De forma diferente, e talvez mais robusta, o presente estudo buscou primeiramente avaliar os resultados gerados por meio da métrica objetiva FID e depois pela avaliação humana. Em (NUHA e AFIAHAYATI, 2018), apenas 5 pessoas responderam ao questionário construído. Por se tratar de uma avaliação subjetiva, o presente estudo apresentou resultados mais confiáveis, pois em cada experimento participaram 100 pessoas, não apenas de idades e gêneros diferentes, mas também de países diversos. Apresentaram um método parecido com o do presente estudo, com exceção das seguintes etapas: (i) a seleção das imagens não considerou o agrupamento das imagens por quaisquer meios; (ii) não houve uma etapa de pré-processamento e; (iii) não houve uma avaliação dos resultados por meio de uma métrica objetiva mais consolidada.

GURUMURTHY et al. (2017) também investigaram a geração de imagens a partir de conjuntos de imagens reduzidos. Propuseram uma arquitetura chamada DeLiGAN para

gerar imagens de diversas categorias. A abordagem deles é interessante, pois ao invés de aumentar a profundidade do modelo, eles buscaram aumentar o poder de modelagem sobre a distribuição dos dados através da seleção de exemplos nas regiões de alta probabilidade do espaço latente. Treinaram os modelos DeLiGAN em *toy data*, MNIST, CIFAR-10 e um *dataset* contendo esboços de objetos desenhados à mão. Estes *datasets*, não iriam servir ao presente estudo por não apresentarem imagens de rosto em posição frontal, com dimensões maiores e com anotações dos atributos faciais. Os modelos de *baseline* foram construídos com a arquitetura tradicional de GAN (GOODFELLOW et al., 2014). No presente estudo, todos os modelos foram treinados com a mesma arquitetura e heurísticas para permitir um melhor entendimento sobre o impacto da filtragem das imagens por meio de atributos faciais geométricos. Utilizaram uma versão modificada do *Inception Score* (SALIMANS et al., 2016) com o objetivo de permitir uma melhor avaliação da diversidade das imagens sintéticas geradas para uma categoria em particular. O presente estudo buscou uma ideia talvez mais simples, ao agrupar o conjunto de treinamento a partir dos atributos faciais para reduzir a variabilidade e o tamanho do conjunto de dados necessário e investigou se, na percepção humana, as imagens geradas apresentaram os atributos utilizados na filtragem. O presente estudo foi além ao buscar entender se a métrica objetiva era compatível com a avaliação humana.

Na literatura, observam-se pesquisas que buscam permitir uma representação desembaraçada dos dados. Como exemplo, CHEN et al. (2016) buscaram este objetivo adicionando alguma porção fixa de informação ao vetor de ruído da rede geradora, maximizando assim a informação mútua entre as imagens sintéticas geradas e o ruído. Foram capazes de gerar imagens pertencentes a classes diferentes, observáveis visualmente. Em outras palavras, a imagem sintética apresentava o atributo representativo daquela classe, sendo perceptível pelas pessoas. No presente estudo, a filtragem das imagens de treinamento por meio de atributos faciais pode ser vista como uma forma de induzir o aprendizado de representações desembaraçadas pelos modelos DCGAN. Foi analisado, por exemplo, se as imagens sintéticas apresentaram os mesmos atributos visuais, perceptíveis por humanos, das imagens de treinamento. A abordagem do presente estudo, trouxe ainda os seguintes benefícios: (i) permitiu a redução do volume de imagens de treinamento e; (ii) a discussão sobre se e quais atributos poderiam facilitar o aprendizado do modelo.

DE SOUZA e RUIZ (2018) usaram um modelo GAN para gerar imagens de rosto com variações de pose. Aplicaram um método de condicionamento para controlar a rotação

dos rostos sintéticos ao longo dos três eixos do espaço. Também utilizaram a CelebA. Apesar dos bons resultados, foi necessário treinar os modelos utilizando-se todo o *dataset*. Trata-se de uma pesquisa que poderia se beneficiar do agrupamento das imagens por meio de atributos faciais geométricos.

Em (SALIMANS et al., 2016), os autores realizaram um teste visual de Turing para avaliar se os seres humanos conseguiriam diferenciar as imagens sintéticas das reais. No presente estudo, foi aplicada uma abordagem diferente para permitir a análise sobre se as pessoas conseguiriam ver rostos humanos nas imagens sintéticas e o quão próximo estas imagens estariam das reais, com o objetivo de avaliar qual o melhor modelo treinado. Eles avaliaram em dois *datasets*: o MNIST, em que os seres humanos não conseguiram distinguir as imagens reais das sintéticas; e o CIFAR-10, onde reportaram uma taxa de erro humano de 21,3%. A imagem de um rosto humano tem características mais complexas do que as encontradas em imagens de algarismos manuscritos, como as do *dataset* MNIST. No caso do *dataset* CIFAR-10, trata-se de uma base com imagens de diferentes categorias e não apenas de rostos humanos. As imagens usadas em (SALIMANS et al., 2016) poderiam ter sido mais bem descritas e caracterizadas, pois não ficou claro, por exemplo, se elas continham rostos em foco e em posição frontal. Além disso, as imagens do CIFAR-10 possuem baixa resolução: 32 x 32 *pixels* o que pode dificultar ou interferir na percepção humana.

Alguns trabalhos mais recentes têm focado em formas diferentes das métricas objetivas tradicionais para avaliar a qualidade das imagens geradas por GAN. Entre esses trabalhos, destaca-se o de WANG et al. (2020) que buscaram avaliar a qualidade das imagens com base nos sinais neurais dos participantes dos experimentos captados por exames de eletroencefalograma. Os autores deste trabalho criaram uma pontuação chamada *Neuroscore* e compararam com três métricas objetivas: a *Inception Score* (IS) (SALIMANS et al., 2016), a *Kernel Maximum Mean Discrepancy* e a *Fréchet Inception Distance* (FID) (HEUSEL et al., 2017). Eles defenderam que esta pontuação é mais consistente com o julgamento humano do que as métricas convencionais e concluíram que os sinais neurais têm potencial para avaliação dos resultados dos modelos GAN. Mas os próprios autores indicam que o *Neuroscore* teve como objetivo avaliar a qualidade das imagens geradas, mas que essa métrica não é capaz de lidar com os problemas inerentes da GAN como o *Mode Collapse*.

Parece promissor o uso de sinais neurais para avaliar a qualidade das imagens conforme a percepção dos seres humanos, mas no caso do presente estudo, avaliamos o

julgamento visual humano sobre as imagens sintéticas geradas e se estas imagens carregaram os atributos faciais, além de observar se o treinamento dos modelos DCGAN tiraram proveito da filtragem das imagens de treinamento a partir desses atributos.

Um artigo recente, similar ao presente estudo, utilizou atributos das imagens para melhorar os resultados dos modelos GAN (DUAN et al., 2019). Nele, os autores buscaram melhorar o desempenho da GAN a partir de imagens anotadas por humanos. Foram organizadas 600 imagens sintéticas geradas por uma GAN às quais foram acrescidas 400 imagens reais do *dataset ImageNet*. Ao contrário do que consta neste estudo, os autores não utilizaram apenas imagens de rostos, e os 8 atributos anotados pertenciam ao domínio que talvez possa ser chamado de fotografia: cor, iluminância, objeto, pessoas, cena, textura, realismo e nível de estranheza. As 1 mil imagens foram anotadas através da plataforma *Amazon Mechanical Turk* por 10 avaliadores que julgaram 20 imagens cada um e a cada imagem deram notas entre 1 (*definitivamente não*) e 5 (*definitivamente sim*) para cada um dos 8 atributos. As imagens anotadas foram utilizadas para treinar três modelos de aprendizagem profunda (*VGG-16*, *ResNet50*, e *DenseNet169*) que foram chamados de *Attribute Nets*. Na arquitetura GAN proposta pelos autores, as imagens produzidas pela geradora são classificadas por uma *Attribute Net* e a saída desta rede é concatenada com a saída da discriminadora para alimentar uma camada *fully connected* para calcular a saída final da arquitetura. Usaram as métricas IS e FID para avaliar os resultados dos modelos GAN. Os autores sugeriram que os trabalhos futuros utilizassem um *dataset* maior com atributos mais estruturados e que os atributos anotados pudessem ser utilizados para treinar modelos GAN de forma mais condicionada ou para o aprendizado de atributos desembaraçados. No presente estudo, os atributos utilizados foram todos faciais e o *dataset* utilizado foi o CelebA que possui mais de 200 mil imagens de rostos e 40 atributos faciais anotados. Os experimentos foram realizados na plataforma Appen e a quantidade de avaliadores em cada experimento foi de 100 pessoas.

7. Considerações Finais

O problema investigado neste estudo foi a possibilidade da geração, de forma automática, de imagens de rostos sintéticos que tivessem uma boa avaliação na percepção humana, utilizando-se a técnica GAN (GOODFELLOW et al., 2014) e um conjunto de imagens e um poder computacional menores.

Através do método utilizado, dos experimentos realizados e da análise dos dados, conseguimos refutar a hipótese nula de que a nossa DCGAN treinada com menos dados, porém selecionados com certas características faciais, iria necessitar do mesmo tamanho de conjunto de dados de um modelo treinado sem este tratamento, para alcançar um resultado aceitável do ponto de vista humano.

Entre os resultados que reforçam esta conclusão, destacamos: (i) o modelo treinado com 5 mil imagens filtradas pelo atributo rosto oval obteve o melhor resultado entre todos os modelos treinados, na avaliação das pessoas nos experimentos, sobre a presença de um rosto humano nas imagens geradas, inclusive tendo alcançado melhores resultados que o modelo do Tipo 1 treinado com 10 mil imagens, cujos dados de treinamento não receberam o mesmo tratamento, como pode ser visto na Figura 5.3, na Tabela 5.5 e na Tabela 5.6; (ii) na Tabela 5.3, observa-se que os modelos treinados com imagens filtradas pelos atributos rosto oval, maçãs do rosto salientes e sobrancelhas arqueadas, obtiverem melhores resultados, em termos de FID, do que o modelo treinado com o dobro de imagens (10 mil), alcançando os valores: 59,59, 63,69 e 66,41, respectivamente, em comparação a 67,84 e; (iii) conforme discutido na seção 5.6 (O Treinamento dos Modelos GAN Realizado com Imagens com Rostos Ovais Realmente Impactou nos Resultados?), por meio da análise dos dados e da aplicação do teste de hipótese, o Grupo B, contendo 30 modelos treinados com imagens filtradas pelo atributo rosto oval, apresentou melhores resultados na métrica FID do que os do Grupo A, treinados sem esta filtragem.

Assim, o presente estudo apresentou uma investigação sobre o impacto do agrupamento de imagens de treinamento, do *dataset* CelebA (LIU et al., 2015), por meio dos atributos faciais geométricos: maçãs do rosto salientes, sobrancelhas arqueadas, rosto oval e lábios grandes, no treinamento e nos resultados de modelos DCGAN (RADFORD et al., 2015), na geração de imagens sintéticas.

Foram aplicadas duas formas de avaliação da qualidade das imagens geradas: (i) uma objetiva, por meio da aplicação da métrica FID (HEUSEL et al., 2017) e; (ii) uma subjetiva, por meio da análise exploratória dos resultados da percepção humana, coletados através da aplicação de questionários pela plataforma de *crowdsourcing* Appen (Appen, 2020).

Observou-se que a métrica FID não acompanhou a avaliação das pessoas em todos os casos, ao contrário do que é dito na literatura.

Foram realizados diversos experimentos para permitir a análise do *trade-off* entre a diminuição da variabilidade das imagens de treinamento e a qualidade e a variabilidade dos exemplos sintéticos gerados.

Os experimentos com as pessoas mostraram o potencial benefício da construção de novos modelos GAN, baseados em atributos faciais e que tenham uma melhor correspondência com a percepção humana.

Os resultados apresentados podem contribuir para um melhor entendimento da percepção humana sobre as imagens sintéticas geradas por GANs, bem como a identificação de possíveis categorias de atributos que podem facilitar o aprendizado dos modelos.

Um efeito que pode ocorrer por causa da diminuição da variabilidade dos dados de treinamento é a redução da variabilidade das imagens geradas (DEVRIES et al., 2020). Isto parece ter ocorrido no caso do modelo do Tipo 1, treinado com 15 mil imagens. Mas no caso dos modelos do Tipo 2, treinados com imagens filtradas por um atributo facial geométrico, isto não foi observado.

Dado que a arquitetura GAN usada neste trabalho é baseada em convoluções, uma possível interpretação sobre este modelo é que o formato do rosto pode facilitar o aprendizado de modelos DCGAN, pois o formato do rosto agiria como uma borda natural do rosto humano sendo aprendido mais rapidamente nas camadas de convolução.

Isto também poderia explicar a dificuldade do modelo treinado com imagens contendo lábios grandes, pois seria um atributo mais difícil de ser aprendido pelas camadas de convolução das redes.

Sobre a presença dos atributos faciais das imagens de treinamento nas imagens sintéticas (Tabela 5.13), esperava-se alcançar resultados mais altos. Os melhores modelos na percepção das pessoas, neste caso, foram aqueles treinados com imagens contendo sobrancelhas arqueadas (mediana dos votos: 54,5) e contendo rostos ovais (mediana dos votos: 52,5). Talvez, isto possa ser explicado, por exemplo, porque as pessoas não

receberam nenhuma instrução prévia sobre como identificar esses atributos ou que as pessoas que anotaram as imagens do CelebA não tiveram melhores instruções.

Os dados demográficos dos participantes dos experimentos foram organizados em 3 dimensões distintas: gênero, idade e país, com o objetivo de permitir a análise sobre a possibilidade das diferenças demográficas impactarem na percepção humana sobre as imagens geradas.

A partir dos resultados dos questionários não foi observada diferença de percepção entre os gêneros masculino e feminino.

As razões entre a mediana de votos dos modelos e o total de avaliadores para os participantes do Egito, ficaram diferentes dos outros dois países. Isto sugere, que há diferença na percepção das pessoas de povos diferentes.

Sobre a idade dos avaliadores, na comparação dos resultados das duas faixas etárias *21 a 30 anos* e *31 a 40 anos*, verificou-se diferença na percepção das pessoas sobre as melhores imagens sintéticas. Seria interessante realizar-se novos experimentos com outras pessoas dessas duas faixas etárias para investigar-se possíveis variáveis da percepção humana que justifiquem essa diferença.

O presente estudo também buscou investigar o *trade-off* entre o poder computacional, a configuração dos modelos DCGAN e a qualidade das imagens geradas. Como exemplo, destacamos a diferença de tempo de treinamento obtida entre o treinamento do modelo do Tipo 1 treinado com 10 mil imagens (5,04 horas, FID: 67,84 e mediana de votos da percepção humana: 90,5) e do modelo do Tipo 2 treinado com 5 mil imagens de rostos filtradas com base no atributo rosto oval (2,45 horas, FID: 59,59 e mediana de votos da percepção humana: 91,5): uma redução no tempo de treinamento de cerca de 2,59 horas mantendo os resultados do modelo do Tipo 2 compatíveis com os do Tipo 1. Em termos práticos, isso pode representar significativo ganho de tempo de pesquisa e redução em gastos com recursos como o de energia elétrica.

Além do ganho de tempo de treinamento devido ao agrupamento das imagens de treinamento, é importante salientar que o presente estudo alcançou bons resultados, do ponto de vista subjetivo, ao treinar modelos por algumas horas e em um computador mais acessível, em comparação a modelos recentes que obtiveram resultados impressionantes, como, por exemplo: *StyleGAN* (KARRAS et al., 2019), *BigGAN* (BROCK et al., 2018) e *Progressive Growing GAN* (PGGAN) (KARRAS et al., 2017), que treinaram seus modelos durante dias e utilizando um poder computacional superior ao do presente estudo.

Sobre isso, é destacado o seguinte trecho do trabalho de DEVRIES et al. (2020): “Ganhos na qualidade de geração de imagem são uma melhoria óbvia, mas talvez mais impactantes para a comunidade em geral sejam as reduções na capacidade do modelo e no tempo de treinamento que são oferecidas. Reduzir a barreira computacional de entrada para o treinamento de modelos generativos em grande escala fornece a muitos indivíduos, incluindo alunos, artistas de IA e entusiastas de ML, acesso a modelos que de outra forma seriam restritos apenas aos laboratórios com mais recursos” (DEVRIES et al., 2020).

Salientamos ainda um ponto importante que será oportuno para a comunidade científica e a profissional refletirem: ao passo em que pesquisas como a do presente estudo facilitam o desenvolvimento do conhecimento e o avançar da ciência, também facilitam a adoção desses conhecimentos e técnicas por pessoas de diversas áreas e indústrias que poderão, com cada vez menos obstáculos, utilizarem essas informações e ferramentas com finalidades diversas das científicas e, sendo assim, problemas de outras naturezas podem surgir, por exemplo: questões éticas relacionadas ao uso de imagens de pessoas famosas ou não para a criação de imagens de novos rostos semelhantes com finalidade comercial, mas sem a prévia autorização do dono da imagem, que poderão causar diversos danos, como: (i) danos morais; (ii) danos materiais e; (iii) danos psicológicos e emocionais e/ou ainda trazer impactos para as cadeias produtivas que envolvem a elaboração e a construção de imagens, como as relacionadas aos fotógrafos e aos editores de imagens.

7.1 **Contribuições**

A principal contribuição da investigação descrita no presente estudo foi demonstrar a possibilidade da geração de melhores imagens sintéticas de rosto, do ponto de vista da percepção das pessoas, que sintetizam as imagens reais, a partir de um conjunto reduzido de dados, considerando-se o treinamento de modelos DCGAN por meio da filtragem desses dados com base em atributos faciais geométricos.

As seguintes contribuições também foram destacadas: (i) registros dos resultados e das configurações dos modelos DCGAN treinados; (ii) resultados dos questionários, organizados por categorias demográficas, que foram apresentados e poderão servir a futuras pesquisas; (iii) foi proposto e aplicado um novo método para o treinamento das GANs e para a avaliação dos resultados; (iv) foi demonstrada a possibilidade de utilização da métrica FID para comparar os resultados produzidos em cada *epoch* de um modelo

treinado, para encontrar-se a melhor versão daquele modelo; (v) demonstração da possibilidade de utilização de uma plataforma de *crowdsourcing* chamada Appen, para a realização de experimentos de Inteligência Artificial com grande volume de pessoas e; (vi) disponibilização de repositório *Git* (https://github.com/danieldasilvacosta/dissertacao-2020) contendo os *notebooks* construídos para o treinamento dos modelos e as imagens geradas selecionadas.

7.2 **Limitações**

Ao utilizar a técnica (HOG) (DALAL e TRIGGS, 2005) no pré-processamento das imagens, parte da informação da imagem foi perdida. Essa informação equivale à parte superior do rosto das pessoas nas imagens (parte da testa e cabelo). Considerando-se a aplicação na qual serão utilizadas as imagens sintéticas, sugere-se utilizar outra técnica para identificação e posicionamento do rosto contido nas imagens.

Uma outra limitação do presente estudo foi a utilização de imagens contendo apenas um rosto e em posição frontal. Em um cenário mais realista, existe a possibilidade de as imagens possuírem mais rostos e em posições diferentes. No presente estudo, a escolha teve por objetivo reduzir a variabilidade das imagens e facilitar o aprendizado dos modelos. Imagens contendo vários rostos não iriam ajudar no propósito deste estudo, e por isso foram desconsideradas na investigação.

Outra decisão tomada para diminuir a variabilidade das imagens de treinamento foi a utilização de imagens de celebridades. Essas imagens podem apresentar vieses de seleção e organização, por exemplo, parte das celebridades pode estar maquiada, o que poderia facilitar o aprendizado da GAN em comparação a imagens de pessoas comuns.

Semelhante a limitação anterior, o presente estudo utilizou apenas imagens de mulheres buscando reduzir a variabilidade das imagens de treinamento e verificar se as imagens geradas pelos modelos DCGAN apresentariam aparência mais feminina do que masculina na percepção dos avaliadores. Buscando uma aplicação mais próxima do real, seria interessante, em pesquisas futuras, utilizar imagens de ambos os gêneros e avaliar como as pessoas percebem os resultados dos modelos. Novas perguntas poderão ser feitas para comparar-se os resultados com os do presente estudo, por exemplo: um modelo treinado com imagens de rostos no formato oval contendo tanto imagens de homens quanto de mulheres apresentaria resultados melhores ou piores do que os do modelo treinado no presente estudo?

A utilização de imagens de treinamento em escala de cinza também é uma limitação. Duas questões podem surgir desta limitação e poderão ser melhor investigadas no futuro: (i) se as imagens fossem coloridas, haveria algum impacto nos valores de FID? e; (ii) se as imagens fossem coloridas, qual seria o impacto no resultado da percepção humana? Por exemplo, o modelo do Tipo 2 treinado com imagens de rostos ovais ainda alcançaria os melhores resultados em termos de FID e da percepção das pessoas? No caso da primeira pergunta, a intuição diz que havendo mais informações para o modelo aprender, levar-se-ia mais tempo (número de *epochs*) para produzir-se resultados compatíveis com os do presente estudo. No caso da segunda pergunta, será mais seguro realizar novos experimentos para avaliar como as cores nas imagens irão interferir na percepção das pessoas.

Esta discussão leva a uma outra consideração importante sobre os experimentos realizados no presente estudo: as pessoas possuem acuidades visuais diferentes e podem apresentar condições específicas como o daltonismo. Tais condições podem interferir mais ou menos na percepção sobre as imagens sintéticas produzidas pelos modelos. Para uma análise mais exata sobre os resultados da percepção humana sobre as imagens sintéticas produzidas, poderia captar-se, no momento de realização do experimento, a condição visual da pessoa. Assim, para trabalhos futuros, seria interessante, por exemplo, colocar-se no questionário um campo onde a pessoa pudesse indicar se possui alguma condição visual específica que possa comprometer ou interferir na sua percepção sobre as imagens sintéticas geradas.

7.3 **Trabalhos Futuros**

Para permitir uma melhor compreensão da aplicabilidade do método proposto no presente estudo, seria interessante, em um trabalho futuro, aplicá-lo em outros *datasets* de imagens de rosto e utilizando outras arquiteturas de GAN.

Algumas questões poderiam ser levantadas em trabalhos futuros: (i) a utilização de imagens de rostos ovais de pessoas comuns para treinar os modelos DCGAN, trariam resultados compatíveis com o do presente estudo? e; (ii) os atributos ou categorias de atributos faciais contribuem da mesma forma para o aprendizado de GANs de arquiteturas diferentes.

Tendo em vista que os modelos DCGAN treinados com imagens de rostos ovais alcançaram os melhores resultados entre todos os modelos, uma questão que pode ser

investigada primeiramente é como o agrupamento de imagens de treinamento, a partir de outros formatos de rosto, pode afetar o aprendizado dos modelos GAN. Para o agrupamento com base no formato do rosto, podem ser lidas as pesquisas dos autores TIO (2019) e PASUPA et al. (2019).

Sobre o agrupamento das imagens, elas foram organizadas conforme os atributos faciais anotados previamente e existentes no *dataset* CelebA. Trabalhos futuros podem investigar formas automáticas de agrupar as imagens, por exemplo, usando-se modelos CNN (LECUN et al., 1989).

Uma outra abordagem a ser melhor explorada, apontada recentemente por DEVRIES et al. (2020), é a utilização de técnicas de *Instance Selection*, usadas tradicionalmente para treinar modelos de aprendizagem supervisionada, para a organização dos dados dos treinamentos das GANs. Uma revisão dos principais métodos pode ser vista em (OLVERA-LÓPEZ et al., 2010).

A diferença de percepção das pessoas do Egito em relação aos outros países pode ser um reflexo da diversidade cultural entre os povos. Pesquisas futuras poderão investigar melhor a influência do país de origem das pessoas nas avaliações dos resultados das GANs.

Outro ponto a se considerar é a investigação do efeito que o agrupamento dos dados de treinamento com base em etnias diferentes poderia causar na percepção das pessoas de diferentes grupos demográficos, para avaliar-se quais características culturais podem ser mais exploradas nos treinamentos de modelos GANs.

Ainda sobre o agrupamento dos dados de treinamento, um próximo passo seria observar o efeito da filtragem das imagens utilizando-se mais de um atributo facial. Por exemplo: modelos treinados com imagens que contenham, ao mesmo tempo, os atributos *rosto oval* e *maçãs do rosto salientes*, permitiriam resultados melhores dos modelos DCGAN dos que os observados neste estudo? Ou, ainda, existe um limite de atributos faciais que possam ser combinados para facilitar o aprendizado dos modelos, mas que, a partir deste limite, os resultados comecem a piorar ou aumente a frequência da ocorrência do problema do *Mode Collapse*?

Outro possível caminho de investigações é a combinação das técnicas de manipulação do espaço latente com a filtragem dos dados por meio de atributos faciais. Essa abordagem é interessante, por exemplo, por dois motivos: (i) a possibilidade de se alcançar resultados melhores dos modelos GAN e; (ii) a possibilidade da geração de imagens sintéticas que apresentem atributos específicos.

De forma semelhante, a aplicação da técnica *Adaptive Discriminator Augmentation* (ADA), deve melhorar consideravelmente o resultado do modelo GAN e permitir utilizar conjuntos de dados ainda menores se combinada com o agrupamento dos dados por meio de atributos faciais. A ADA foi proposta por KARRAS et al. (2020).

8. Referências

Appen, 2020, Appen Platform, disponível em: <https://appen.com/>, acessado em 04/07/2020.

BOJANOWSKI, P., JOULIN, A., LOPEZ-PAZ, D., et al., 2017, "Optimizing the Latent Space of Generative Networks", *arXiv preprint arXiv:1707.05776*, 2017.

BORJI, A., 2019, "Pros and Cons of GAN Evaluation Measures", In: *Computer Vision and Image Understanding*, pp. 41-65, 2019.

BRIOT, JP., HADJERES, G., PACHET, FD., 2017, "Deep Learning Techniques for Music Generation - A Survey", *arXiv preprint arXiv:1709.01620*, 2017.

BROCK, A., DONAHUE, J., SIMONYAN, K., 2018, "Large Scale GAN Training for High Fidelity Natural Image Synthesis", *arXiv preprint arXiv:1809.11096*, 2018.

BROWNLEE, J., 2020, *Generative Adversarial Networks with Python: Deep Learning Generative Models for Image Synthesis and Image Translation*. 1.5 ed., Machine Learning Mastery, 2020.

CAO, J., LI, Y., ZHANG, Z., 2018, "Partially Shared Multi-Task Convolutional Neural Network with Local Constraint for Face Attribute Learning", In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.

CHEN, X., DUAN, Y., HOUTHOOFT, R., et al., 2016, "InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets", In: *Advances in Neural Information Processing Systems*, pp. 2172-2180, 2016.

CHINTALA, S., DENTON, E., ARJOVSKY, M., et al., 2020, "How to Train a GAN? Tips and Tricks to Make GANs Work", disponível em: <https://github.com/soumith/ganhacks>, acessado em 05/07/2020.

DALAL, N., TRIGGS, B., 2005, "Histograms of Oriented Gradients for Human Detection". *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 20-25 June 2005.

DE SOUZA, D. M., RUIZ, D. D. A., 2018, "Gan-Based Realistic Face Pose Synthesis with Continuous Latent Code", In: "*Proceedings of the 31st Florida Artificial Intelligence Research Society Conference*", 2018.

DENG, J., DONG, W., SOCHER, R., et al., 2009, "ImageNet: A Large-Scale Hierarchical Image Database", *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248-255, Miami, FL, 2009.

DENTON, E. L., CHINTALA, S., SZLAM, A., et al., 2015, "Deep Generative Image Models Using a Laplacian Pyramid of Adversarial Networks". In: *Advances in Neural Information Processing Systems*, v. 28, pp. 1486-1494, 2015.

DEVRIES, T., DROZDZAL, M., TAYLOR, G. W., 2020, "Instance Selection for GANs", In: *Advances in Neural Information Processing Systems*, v. 33, 2020.

DI, X., PATEL, V. M., 2018, "Face Synthesis from Visual Attributes via Sketch using Conditional VAEs and GANs", *arXiv preprint arXiv:1801.00077*, 2018.

DIAMANT, N., ZADOK, D., BASKIN, C., et al., 2019, "Beholder-GAN: Generation and Beautification of Facial Images with Conditioning on Their Beauty Level", *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 739-743, Taipei, Taiwan, 22-25 September 2019.

DUAN, J., ONG, S. H., ZHAO, Q., 2019, "Human Annotations Improve GAN Performances", *arXiv preprint arXiv: 1911.06460*, 2019.

FUJII, K., SAITO, Y., TAKAMICHI, S., et al., 2020, "Humangan: Generative Adversarial Network With Human-Based Discriminator And Its Evaluation In Speech Perception Modeling", *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6239-6243, Barcelona, Spain, 2020.

GOODFELLOW, I., 2016, "NIPS 2016 Tutorial: Generative Adversarial Networks", *arXiv preprint arXiv:1701.00160*, 2016.

GOODFELLOW, I., BENGIO, Y., COURVILLE, A., 2016, *Deep Learning*, The MIT Press, disponível em: <https://www.deeplearningbook.org/>, acessado em 01/03/2020.

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., et al., 2014, "Generative Adversarial Nets", In: *Advances in Neural Information Processing Systems*, v. 27, pp. 2672-2680, 2014.

GURUMURTHY, S., KIRAN SARVADEVABHTLA, R., VENKATESH BABU, R., 2017, "DeLiGAN : Generative Adversarial Networks for Diverse and Limited Data", In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.

HEUSEL, M., RAMSAUER, H., UNTERTHINER, T., et al., 2017, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium", In: *Advances in Neural Information Processing Systems*, v. 30, pp. 6626-6637, 2017.

JIN, Y., ZHANG, J., LI, M., et al., 2017, "Towards the Automatic Anime Characters Creation with Generative Adversarial Networks", *arXiv preprint arXiv:1708.05509*, 2017.

KARRAS, T., AILA, T., LAINE., S., et al., 2017, "Progressive Growing of GANs for Improved Quality, Stability, and Variation", *arXiv preprint arXiv:1710.10196*, 2017.

KARRAS, T., AITTALA, M., HELLSTEN, J., et al., 2020, "Training Generative Adversarial Networks with Limited Data", *arXiv preprint arXiv:2006.06676*, 2020.

KARRAS, T., LAINE, S., AILA, T., 2019, "A Style-Based Generator Architecture for Generative Adversarial Networks", In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401-4410, Jun. 2019.

KAZEMI, H., IRANMANESH, M., DABOUEI, A., et al., 2018, "Facial Attributes Guided Deep Sketch-to-Photo Synthesis", *2018 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 1-8, Lake Tahoe, NV, 2018.

KINGMA, D. P., WELLING, M., 2014, "Auto-Encoding Variational Bayes" , *arXiv preprint arXiv:1312.6114*, 2014.

KRIZHEVSKY, A., 2009, *Learning Multiple Layers of Features from Tiny Images*, Citeseer, 2009.

KRIZHEVSKY, A., SUTSKEVER, I., HINTON, G. E., 2012, "ImageNet Classification with Deep Convolutional Neural Networks", In: *Advances in Neural Information Processing Systems*, v. 25, pp. 1097-1105, 2012.

LECUN, Y., BOSER, B., DENKER, J. S., et al., 1989, "Backpropagation Applied to Handwritten Zip Code Recognition", *Neural Computation*, v. 1, n. 2, 1989.

LIANG, L., LIN, L., JIN, L., et al., 2018, "SCUT-FBP5500: a Diverse Benchmark Dataset for Multi-Paradigm Facial Beauty Prediction", *24th International Conference on Pattern Recognition*, pp. 1598-1603, China, 20-24 August 2018.

LIPTON, Z., TRIPATHI, S., 2017, "Precise Recovery of Latent Vectors from Generative Adversarial Networks", *arXiv preprint arXiv: 1702.04782*, 2017.

LIU, Z., LUO, P., WANG, X., et al., 2015, "Deep Learning Face Attributes in the Wild". *IEEE International Conference on Computer Vision*, Santiago, Chile, 11-18 December 2015.

LUCIC, M., KURACH, K., MICHALSKI, M., et al., 2018, "*Are GANs Created Equal? A Large-Scale Study*", In: *Advances in Neural Information Processing Systems*, v. 31, pp. 700-709, 2018.

MUKHERJEE, S., ASNANI, H., LIN, E., et al., 2019, "ClusterGAN: Latent Space Clustering in Generative Adversarial Networks", In: *Proceedings of the AAAI Conference on Artificial Intelligence, 33(01)*, pp. 4610-4617, 2019.

NUHA, F. U., AFIAHAYATI, 2018, "Training Dataset Reduction on Generative Adversarial Network", *Procedia Computer Science*, v. 144, pp. 133-139, 2018.

OLVERA-LÓPEZ, J. A., CARRASCO-OCHOA, J. A., MARTÍNEZ-TRINIDAD, J. F. et al., 2010, "A Review of Instance Selection Methods", *Artificial Intelligence Review*, v. 34, n. 2, pp. 133-143, 2010.

PASUPA, K., SUNHEM, W., LOO, C. K., 2019, "A hybrid Approach to Building Face Shape Classifier for Hairstyle Recommender System", In: *Expert Systems with Applications*, v. 120, pp. 14-32, 2019.

RADFORD, A., METZ, L., CHINTALA, S., 2015, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". *arXiv preprint arXiv:1511.06434*, 2015.

RAMSUNDAR, B., ZADEH, R. B., 2018, *TensorFlow for Deep Learning*. 1 ed., O'Reilly Media, 2018.

SALIMANS, T., GOODFELLOW, I., ZAREMBA, W., et al., 2016, "Improved Techniques for Training GANs", In: *Advances in Neural Information Processing Systems*, v. 29, pp. 2234-2242, 2016.

TIO, A. E., 2019, "Face Shape Classification Using Inception v3", *arXiv preprint arXiv: 1911.07916*, 2019.

VOLZ, V., SCHRUM, J., LIU, J., et al., 2018, "Evolving Mario Levels in the Latent Space of a Deep Convolutional Generative Adversarial Network", In: *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 221-228, 2018.

WANG, Z., HEALY, G., SMEATON, A. F., et al., 2020, "Use of Neural Signals to Evaluate the Quality of Generative Adversarial Network Performance in Facial Image Generation", *Cognitive Computation*, v. 12, pp. 13-24, 2020.

ZALTRON, N., ZURLO, L., RISI, S., 2020, "CG-GAN: An Interactive Evolutionary GAN-Based Approach for Facial Composite Generation", *34th AAAI Conference on Artificial Intelligence*, pp. 2544-2551, New York, New York, USA, 7-12 February 2020.

ZHOU, S., GORDON, M., KRISHNA, R., et al., 2019, “HYPE: A Benchmark for Human eYe Perceptual Evaluation of Generative Models”, In: *Advances in Neural Information Processing Systems*, v. 32, pp. 3449-3461, 2019.