Coloring Face Images Using Autoencoder

Coloração de Imagens de Faces Utilizando Autoencoder

Mateus Reis Santos[1], Raoni Florentino da Silva Teixeira[1], Joyce Marins[1], Gracyelli Santos Souza Guarienti[1], Alisson Cerutti[1], João Vitor Monteiro Souza[1]

[1]Departamento de Computação e Automação - Faculdade de Engenharia – Universidade Federal de Mato Grosso (UFMT)

Caixa Postal 78060-900 – Cuiabá – MT – Brasil

mateusreis.ufmt@gmail.com,
raoni.teixeira@ufmt.br,Joyce.marins@ufmt.br,
gracyeli.guarienti@ufmt.br,c3.cerutti@gmail.com,jvmtds@gmail.com

Abstract*. Until 1940s, most photographs were produced in black and white. Bringing a little more realism to these historical relics costs time and money with the non-automatic methods used for coloring photos. In order to improve these issues, this work proposes the use of a Neural Network Architecture known as Autoencoder U-NET to build a model to color these images. The model built in this work shows that it is possible to use the U-NET Autoencoder to color images of faces automatically.*
Keywords: *Images coloring. Autoencoder. celebA*

Resumo*. Até a década de 40 grande parte das fotografias foram produzidas em preto e branco. Trazer um pouco mais de realismo a essas relíquias históricas custa tempo e dinheiro com os métodos não automáticos usados para coloração de fotos. Visando melhorar essas questões, é proposto neste trabalho o uso de uma arquitetura de rede neural conhecida como Autoencoder U-NET para construção de um modelo para colorir imagens. O modelo construído neste trabalho mostra que é possível usar o Autoencoder U-NET para colorir imagens faciais de forma automática.*
Palavras-chave: *Coloração de imagens. Autoencoder. celebA*

1. Introdução

Desde o advento da fotografia em 1826 [Gernsheim 1986], com fotos em tons de cinza (famosas fotos preto e branco) se discute a coloração desses retratos para torná-los mais fiéis à realidade. A história da coloração de imagens se confunde com o começo da fotografia em si. As primeiras técnicas de coloração eram feitas de maneira artística, muitas vezes colorindo imagens inteiras em tons de azul ou vermelho, ou em tons de sépia. As técnicas mais elaboradas de colorir retratos eram aquelas em que todo o trabalho era feito a mão. Isso resultava em um processo demorado e caro e, pouquíssimas pessoas podiam pagar por esse trabalho. Até a década de 40 do século passado, a grande maioria das fotos foram registradas em preto e branco. Ricos momentos históricos como os acontecimentos da primeira e segunda guerra mundial foram em maior parte registrados em preto e branco [InstaRestoration 2020]. A maneira artística demorada e cara de coloração de imagens ainda existe. Este trabalho busca apresentar uma maneira simples

e eficiente de poder colorir fotografias antigas usando a arquitetura de rede neural convolucional conhecida como Autoencoder U-NET. Os resultados obtidos neste trabalho mostram que é possível utilizar essa arquitetura de rede neural para colorir de maneira automática imagens faciais e também imagens de pessoas.

Na literatura há trabalhos com abordagens similares a apresentada neste artigo, objetivando colorir imagens faciais em tons de cinza como por exemplo [Persch et al. 2017] e [Zaware et al. 2021]. No entanto, há diferenças nessas outras abordagens. Em [Persch et al. 2017] o que ocorre é a transferência de cores de uma imagem facial para outra imagem facial. Em [Zaware et al. 2021] foi usado um Autoencoder com camadas dropout no encoder e no decoder para colorir as imagens de forma automática. Neste trabalho foi escolhido o Autoencoder UNET para implementar o modelo de colorir. Em, [Zaware et al. 2021] foi usado somente uma base de dados para a efetivação dos testes. Neste trabalho foram utilizadas 3 bases de dados diferentes, tornando o modelo proposto neste artigo mais confiável evitando ter um modelo enviesado. Também vale destacar, que nos testes realizados foram utilizadas amostras que possuem parte do corpo das pessoas, não somente a face, ao contrário do trabalho previamente citado.

Este artigo está organizado da seguinte maneira: A seção 2 aborda a metodologia utilizada para o treinamento dos modelos usados para coloração. A seção 3 apresenta os experimentos realizados e os resultados obtidos. A seção 4 apresenta as conclusões obtidas com a realização deste trabalho e por fim expõem-se as referências bibliográficas.

2. Metodologia

Neste trabalho foi treinado um Autoencoder com a arquitetura U-NET para colorização automática de imagens em tons de cinza no espaço de cor YCBCR. As bases de dados utilizadas foram a CASIAWebFace, CelebA e Flickr-Faces-HQ (FFHQ).

2.1. Base de dados

A CASIAWebFace [Yi et al. 2014] é uma base de dados de imagens de faces. Ela geralmente é usada para tarefas de verificação e identificação de faces. A base de dados possui um total de 494.414 imagens de faces e 10.575 imagens de identidades reais coletadas através da internet. Essa base de dados possui muito de sua utilidade presente em pesquisas científicas, sendo permitido a sua utilização por pesquisadores somente em produtos que não possuem finalidade comercial.

CelebFaces Attributes Dataset [Liu et al. 2015], também conhecida CelebA, é uma outra base de dados utilizada para a implementação deste trabalho. Essa base de dados posui mais de 200 mil imagens originalmente e cada uma dessas imagens possui 40 anotações de atributos. Essa base de dados trata-se de uma base de dados muito usada para o treinamento de algoritmos que envolvem reconhecimento facial.

Finalmente, a última base de dados usada é Flickr-Faces-HQ [Karras et al. 2019], o conjunto de dados consiste em 70.000 imagens PNG de alta qualidade com resolução de 1024×1024 e contém variações consideráveis em termos de idade, etnia e plano de fundo da imagem. Ele também tem uma boa cobertura de acessórios como óculos, óculos de sol, chapéus, etc. Essa base de dados foi produzida a partir de imagens colhidas do Flickr. Vários filtros automáticos foram usados para podar o conjunto nesta base de dados.

2.2. Pré-Processamento e limpeza dos dados

Foram identificadas imagens das três bases de dados que estavam em tons de cinza e não foram utilizadas essas imagens para treinamento. A identificação das imagens para saber se eram ou não coloridas foi realizada pela identificação do espaço de cor e também verificando se as matrizes RGB da imagem eram iguais. Na base de dados CasiaWebFace foram encontradas 16.474 imagens em tons de cinza. Enquanto que na base de dados CelebA foram encontradas 44 imagens em tons de cinza e na base de dados FFHQ não foram encontradas nenhuma imagem em tons de cinza.

As imagens das bases de dados CasiaWebFace possuem resolução 250 x 250, da base de dados FFHQ possuem resolução 1024 x 1024, da base de dados celebA na versão alinhada possui dimensões 178 x 218 e na versão sem alinhar possui imagens de diferentes resoluções. Neste trabalho foi utilizada a versão alinhada inicialmente, mas depois foi utilizada a versão sem alinhamento com o objetivo de melhorar os resultados. As imagens das bases de dados FFHQ e CasiaWebFace foram redimensionadas para 256 x 256 porque esta é a dimensão de entrada da Arquitetura U-NET que foi utilizada neste trabalho. Nas imagens da base de dados celebA na versão alinhada, foi realizado o processo conhecido como padding no qual foram adicionados pixels com a cor preta em volta da imagem até atingir o tamanho 256 x 256. Na versão sem alinhamento da base celebA as imagens foram redimensionadas mantendo a proporção. Depois foi feito o padding. Juntando as imagens coloridas dessas três bases de dados foram obtidas 750.000 imagens das quais foram separadas em 95% para treinamento e 5% para teste. A figura 1 contém amostras das bases de dados utilizadas para treinamento após o pré-processamento.

Figura 1. Bases de dados utilizadas nos treinamentos.

2.3. Arquitetura Autoencoder U-NET

A arquitetura de rede neural utilizada para testes neste projeto é o Autoencoder U-NET [Ronneberger et al. 2015] que baseia-se no Autoencoder. Essa arquitetura surgiu inicialmente para segmentação de imagens médicas sendo posteriormente utilizada em outras atividades como para colorização [Treneska et al. 2022].

O Autoencoder é uma rede neural usada para codificação de dados de maneira não supervisionada, embora utilize métodos supervisionados no seu treinamento[Jason Brownlee 2020]. Essa rede é dividida em *encoder* e *decoder.* O encoder é a primeira parte

da rede neural onde a entrada é mapeada a um ponto no espaço latente e a partir desse ponto pode-se gerar uma saıda similar a entrada, mas com perdas de qualidade, passando esse ponto através do decoder, a segunda parte do Autoencoder. A rede é treinada para encontrar os melhores pesos para ambas as partes do Autoencoder, de tal maneira que a perda do dado de saída em relação ao dado de entrada seja a menor possível. O vetor de representação é uma compressão da imagem original em um espaço latente de dimensão inferior [Foster 2019]. O U-Net tem ambas as propriedades listadas acima, propriedade de codificação e decodificação. Mas ele permite que camadas sejam "puladas", isto é, nem sempre o dado de entrada passará por todas as camadas de decodificação e codificação presentes nas redes. Há uma desvantagem no Autoendoer clássico, e isso se dá porque ao passar o dado por todas as camadas, muito da informação original acaba se perdendo e isso não é bom para um algoritmo de coloração de imagens.

A arquitetura proposta é composta por 7 camadas sequenciais responsáveis pela compressão do dado de entrada, dado esse que consiste em imagens com resolução de 256 x 256 pixels em tons de cinza. A parte da decodificação é composta de 7 camadas sequenciais que visam restabelecer a dimensionalidade original das imagens. Mas junto a essas camadas há as camadas concatenate, que trabalham os pulos de alguns dados durante a sua codificação. Há também, como última camada da rede a Conv2DTranspose que retoma a dimensionalidade original da imagem, de acordo com a entrada. E como função de ativação temos em nosso trabalho a ReLU, definida por:

$$f(x) = \{x,\ se\ x\ \geq 0\ 0,\ caso\ contrário \tag{1}$$

2.4. Treinamento

Neste trabalho foi escolhido o espaço de cor YCBCR porque nesse espaço de cor o primeiro canal representa a imagem em tons de cinza e os outros dois canais da imagem são responsáveis pelas cores, ao contrário do espaço de cor RGB que possui três componentes de cor. Na colorização é comum a utilização também dos espaços Lab e YUV conforme afirma [Cao et al. 2017] pois esses espaços de cor também só possuem duas componentes que representam as cores, economizando assim largura de banda.

O treinamento do Autoencoder com Arquitetura U-NET foi realizado da seguinte maneira: A imagem colorida é convertida para o espaço de cor YCBCR. A matriz Y (primeiro canal no espaço de cor YCBCR) de dimensão 256 x 256 é submetida ao Autoencoder que tem como saída duas matrizes 256 x 256. O erro é estimado calculando a média da diferença absoluta (mean absolute error) entre os pixels das matrizes CB e CR da imagem original e os pixels das duas matrizes previstas pelo Autoencoder. O objetivo do treinamento é encontrar pesos para o Autoencoder que minimizem a diferença entre as matrizes CB e CR das imagens originais e as matrizes CB e CR calculadas pelo Autoencoder a partir de suas matrizes Y. O processo de otimização é feito a partir do gradiente do erro, utilizando o otimizador Adam por 300.000 épocas. A função de custo mean absolute error também conhecida como L1 pode ser descrita como:

$$Erro = \left(\frac{1}{n}\right) | \sum_{i=1}^{n} \quad Y_i - X_i \tag{2}$$

onde Y representa as matrizes CB e CR originais da imagem e X representa as matrizes previstas pelo Autoencoder.

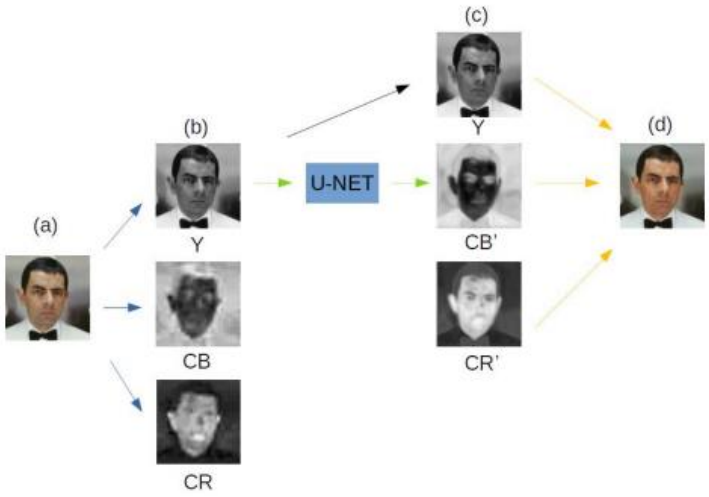

Figura 2. Diagrama treinamento e inferência do Autoencoder.

A figura 2 mostra um diagrama com os passos utilizados para o treinamento do Autoencoder e posterior utilização para colorização de imagens. Nesse diagrama em (a) temos a imagem original no espaço de cor RGB. Essa imagem é convertida do espaço de cor RGB para o espaço de cor YCBCR e seus canais Y, CB e CR são exibidos em(b). Depois o canal Y da imagem que representa a imagem em tons de cinza é submetido como entrada para o Autoencoder (U-NET) que calcula duas matrizes CB' e CR' conforme observamos em (c). Durante o processo de treinamento o erro é estimado calculando a diferença média absoluta entre as matrizes CB e CB' e CR e CR'. Sendo que CR e CB correspondem aos canais de cor da imagem original no espaço de cor YCBCR e CB' e CR' correspondem aos canais de cor calculados pelo modelo U-NET. Após o treinamento a imagem em tons de cinza que corresponde ao canal Y é submetida ao modelo U-NET que calcula as matrizes CB' e CR', juntando essas duas matrizes com a matriz Y obtemos os canais da imagem no espaço de cor YCBCR que são exibidos em (C) no diagrama. Para visualizar a imagem fazemos a conversão da mesma do espaço de cor YCBCR para o espaço de cor RGB cujo resultado é exibido em (d) no diagrama.

3. Experimentos e Resultados

3.1. Experimento I

O primeiro experimento realizado foi treinar o Autoencoder com as bases de dados CasiaWebFace, FFHQ e CelebA na versão alinhada. Após a realização do treinamento foi verificado que a colorização estava funcionando bem com as faces, mas outras partes da imagem não. E quando foi testada a colorização com imagens que possuíam mais de uma pessoa, a colorização por meio desse modelo mantinha tons de cinza para os indivíduos nos extremos da imagem. Outra limitação identificada a partir dos resultados foi que em retratos que apareciam outras partes do corpo da pessoa além do rosto como o pescoço e ombro por exemplo a face ficava colorida, mas o corpo permanecia em tons de cinza.

3.2. Experimento II

Com o objetivo de aumentar a generalização da colorização de modo que fosse possível colorir retratos de pessoas e não só a face, foi realizado um segundo experimento no qual

trocou-se a base de dados CelebA na versão alinhada para a mesma base de dados, mas em sua versão não alinhada e assim o Autoencoder foi treinado novamente. Foi utilizada essa abordagem pois essa versão da base dados contempla imagens de pessoas em diferentes contextos e ambientes e as imagens em muitos casos abrangem o corpo do indivíduo e o ambiente em que ele se encontra. Algo que não ocorre na versão alinhada da base de dados CelebA.

3.3. Experimento III

Visando melhorar os resultados foi realizado mais um treinamento do Autoencoder com a mesma base de dados utilizada no experimento II, mas aqui foram realizados alguns procedimentos para aumentar a base de dados que ocorreram durante o treinamento. Primeiro é estabelecido um ângulo aleatório entre [-15 , 15] graus no qual a imagem é rotacionada. Depois é feito o preenchimento em volta da imagem com pixels pretos até a imagem atingir o tamanho 480 x 480. Por fim é realizado um corte aleatório na imagem de modo que ela mantenha a resolução original de 256 x 256. Na figura 3 tem-se o processo usado para aumentar a base de dados sendo que em (a) tem-se a imagem original, em (b) tem-se a imagem rotacionada, em (c) tem-se a imagem após adição de bordas pretas para atingir a resolução 480 x 480 e em (d) tem-se uma nova imagem após um corte aleatório de tamanho 256 x 256.

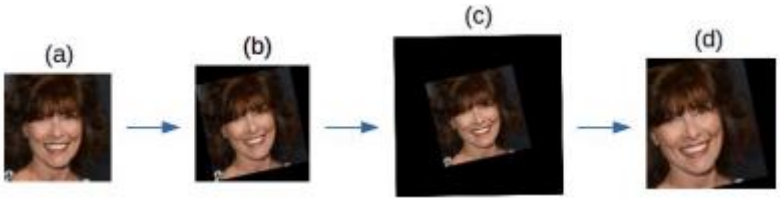

Figura 3. Processo realizado para aumentar a base de dados.

3.4. Resultados

A figura 4 mostra os resultados obtidos da colorização das imagens em tons de cinza com os modelos treinados nos 3 experimentos realizados neste trabalho. Nesta imagem na primeira coluna tem-se as imagens originais que foram convertidas em tons de cinza e foram utilizadas como entrada dos modelos U-NET. Na segunda coluna tem-se as imagens originais coloridas. Na terceira coluna tem-se as imagens coloridas pelo modelo treinado no experimento I. Na quarta coluna tem-se as imagens coloridas pelo modelo do experimento II e na quinta coluna tem-se as imagens coloridas pelo modelo do experimento III.

Na figura 4 vê-se que a colorização das faces ficou satisfatória pois as cores ficam próximas das originais. Porém, na imagem da segunda linha é perceptível que existe uma mão no lado direito da imagem que permaneceu em tons de cinza, o que dá um indicativo de que talvez os modelos aprenderam a colorir apenas o que está em primeiro plano ou talvez apenas as faces.

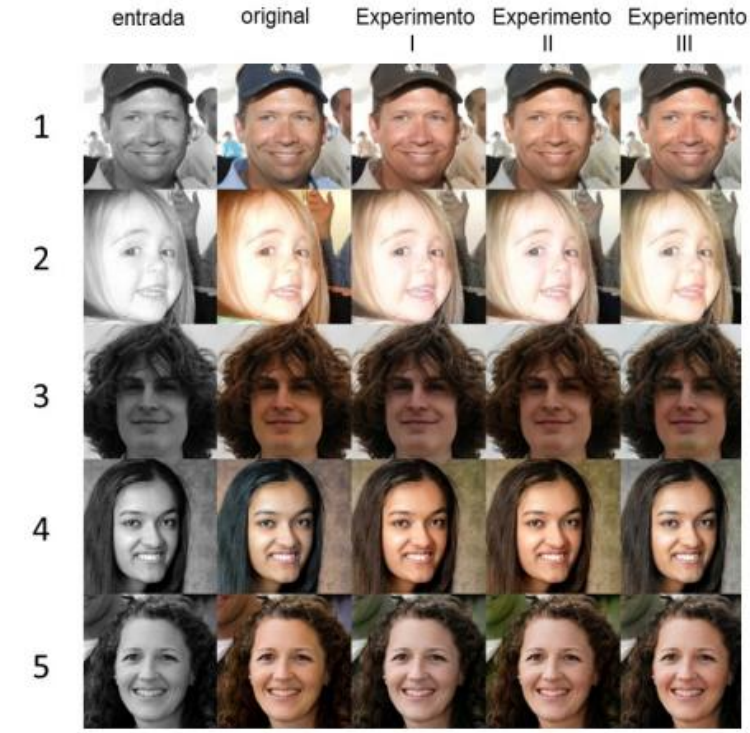

Figura 4. Resultados base de dados FFHQ.

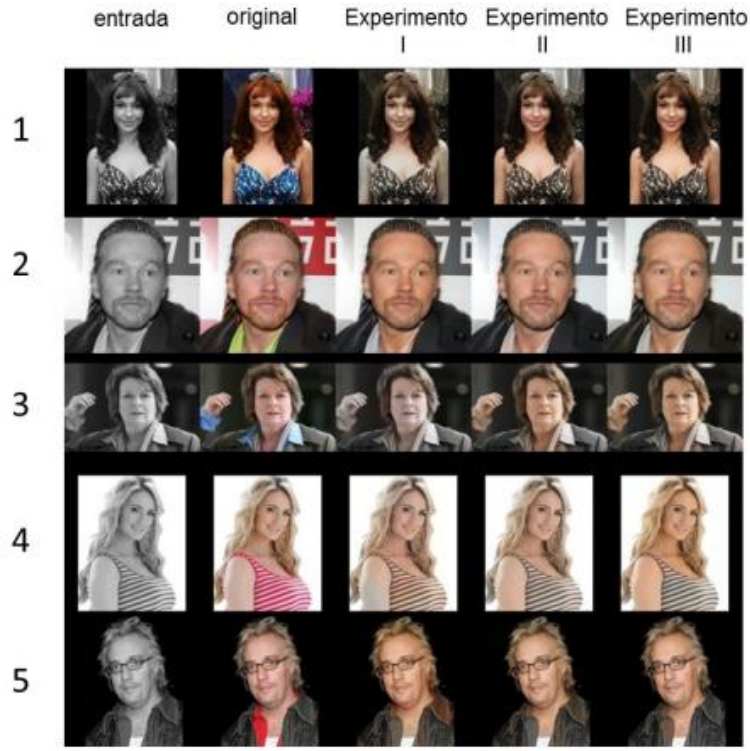

Figura 5. Resultados base de dados CelebA versão não alinhada.

Para investigar melhor o desempenho dos modelos em relação a generalização do modelo, foi testada a colorização com imagens da base de dados CelebA em sua versão não-linhada. A figura 5 mostra os resultados desse teste. Nesta figura percebe-se

diferenças na coloração do corpo das pessoas nas linhas 1, 3 e 4 da imagem. O modelo treinado no experimento II e o modelo treinado no experimento III tiveram um melhor resultado na coloração do corpo das pessoas do que o modelo treinado no experimento I. Com base nesses resultados acredita-se que a abordagem de usar a versão não alinhada da base de dados CelebA melhorou o aprendizado do modelo pois nessa base tem imagens que englobam mais o corpo das pessoas. Essa abordagem foi utilizada nos experimentos II e III.

Foi realizado um teste na base de dados CasiaWebFace cujos resultados estão expressos na figura 6. Nesta figura ressaltar a respeito das imagens da linha 3. Nessa linha, a saída do modelo do experimento II ficou melhor do que a saída do modelo do experimento I e também ficou melhor que a saída do modelo III, considerando a cor da pele do indivíduo.

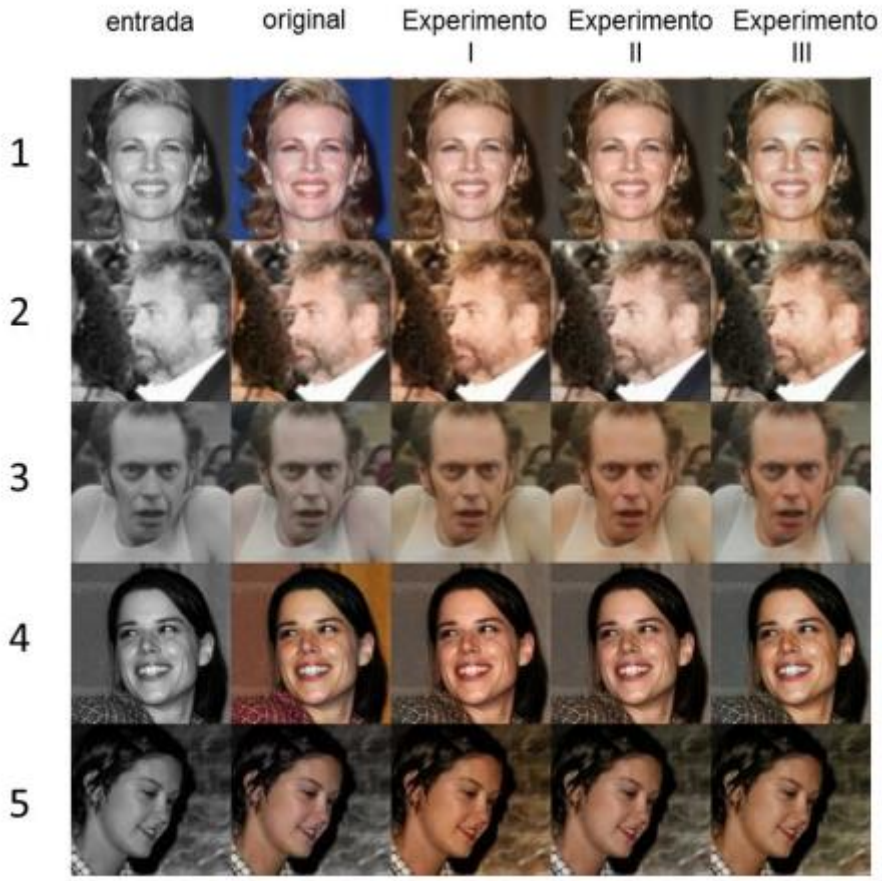

Figura 6. Resultados da base de dados CasiaWebFace.

Por fim foi realizado um teste na versão alinhada da base de dados CelebA. Na figura 7 apresenta-se os resultados desse experimento. Nessa imagem é interessante observar a segunda linha. Nessa linha, se observar a imagem colorida pelo modelo do experimento I é possível notar que a cor da camisa do sujeito está mais próxima da original do que nos demais modelos.

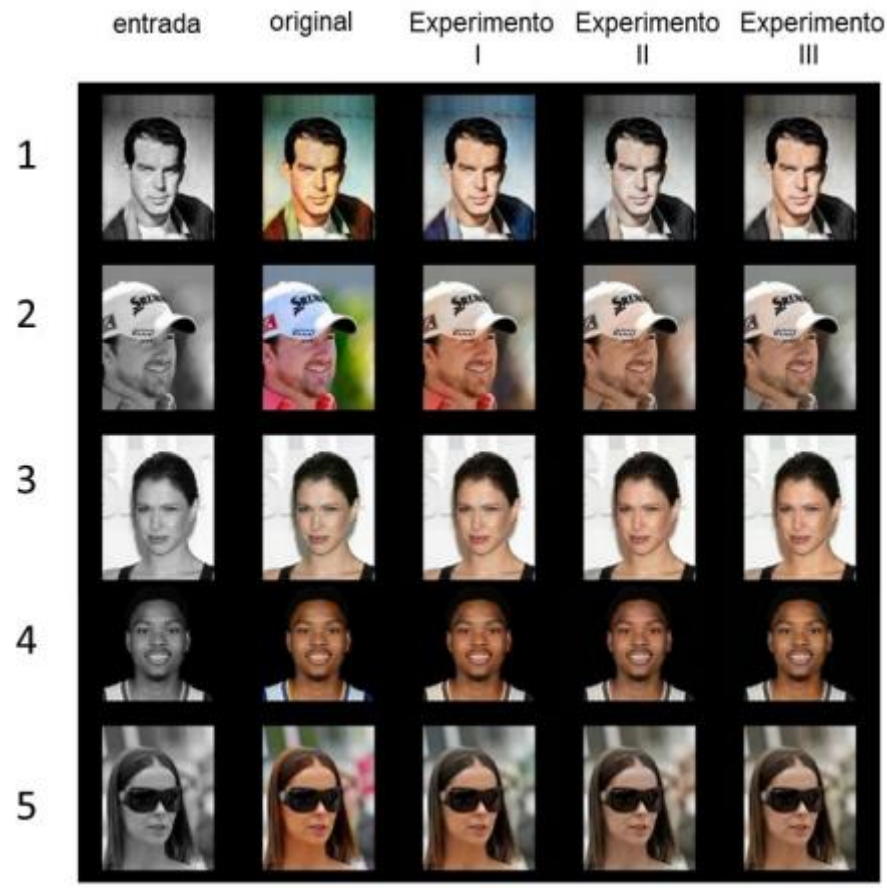

Figura 7. Resultados base de dados CelebA versão com Alinhamento.

4. Conclusão

Com base nos experimentos realizados e nos resultados obtidos, verificou-se que a colorização das faces foi bem sucedida. Porém é necessário destacar que há limitações. Elementos como roupa da pessoa, pessoas no fundo da imagem, corpo da pessoa não são coloridos de forma realista. Possivelmente nesses elementos o algoritmo só está aplicando uma cor média. Com relação aos modelos treinados, não foi identificado um que tem sempre melhor desempenho que os demais. Entretanto, acredita-se que os modelos dos experimentos II e III devem se sair melhor que o modelo I quando na imagem tiver o corpo da pessoa porque esses modelos foram treinados com imagens nessa configuração. Como trabalhos futuros pretende-se testar técnicas não supervisionadas de colorização, como por exemplo, o uso de redes generativas adversárias [Goodfellow et al. 2014] porque assim acredita-se que o algoritmo vai aprender a colorir de forma mais realista elementos como a roupa, corpo e ambiente.

Referências

Cao, Y., Zhou, Z., Zhang, W., and Yu, Y. (2017). *Unsupervised Diverse Colorization via Generative Adversarial Networks*, pages 151–166. doi: 10.1007/978-3-319-71249-9_10.

Foster, D. (2019). *Generative Deep Learning*. O'Reilly Media, Sebastopol, CA, USA. Gernsheim, H. (1986). A concise history of photography. Number 10. Courier Corporation.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks. *Advances in Neural Information Processing Systems*, 3. doi: 10.1145/3422622.

InstaRestoration (2020). History of photo colorization. Disponível em: *https://www.instarestoration.com/blog/history-of-photo-colorization.* Acesso em: 07 maio 2022.

Jason Brownlee (2020). Autoencoder feature extraction for classification. Disponível em: *https://machinelearningmastery.com/Autoencoder-for-classification/.* Acesso em: 07 maio 2022.

Karras, T., Laine, S., and 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*, pages 4396–4405. doi: 10.1109/CVPR.2019.00453.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). *Deep learning face attributes in the wild. In Proceedings of International Conference on Computer Vision (ICCV)*. doi: 10.1109/ICCV.2015.425.

Persch, J., Pierre, F., and Steidl, G. (2017). *Exemplar-based face colorization using image morphing. Journal of Imaging, 3(4).* doi: 10.3390/jimaging3040048.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *In Medical Image Computing and Computer-Assisted Intervention* – MICCAI 2015, pages 234–241, Cham. doi: 10.1007/978-3-319-24574-4 28.

Treneska, S., Zdravevski, E., Pires, I. M., Lameski, P., and Gievska, S. (2022). Gan-basedimage colorization for self-supervised visual feature learning. *Sensors*, 22(4). doi:10.3390/s22041599.

Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014). Learning face representation from scratch. *arXiv preprint arXiv*:1411.7923. doi: 10.48550/arXiv.1411.7923.

Zaware, S., Pathak, D., Patil, V., Sangale, G., and Gupta, V. (2021). Gray scale image colorization for human faces. *In 2021 International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C)*, pages 107–110. doi: 10.1109/ICDI3C53598.2021.00030.