Estudo de Sistemas para a Colorização Automática de Imagens Utilizando Aprendizagem Profunda

Leo Araújo e Luciana Veloso

***Resumo*— Métodos de Aprendizagem Profunda (*Deep Learning*) têm superado técnicas tradicionais em diversas aplicações de domínios da Visão Computacional e do Processamento de Sinais. Nesse contexto, esse trabalho realiza um estudo sobre a aplicabilidade de tais métodos para colorir imagens automaticamente. Três arquiteturas de Redes Neurais Convolucionais foram comparadas segundo três índices quantitativos, a raíz do erro médio quadrático (RMSE), a relação sinal ruído de pico (PSNR) e o índice de similaridade estrutural (SSIM), além de um teste de usabilidade. Foi verificado que os modelos avaliados são capazes de realizar a colorização automática, sendo capazes de produzir resultados fotorrealistas.**

***Palavras-Chave*— Aprendizagem Profunda, Redes Neurais Convolucionais, Colorização Automática.**

***Abstract*— Deep Learning methods have outperformed classic techniques in numerous applications from Computer Vision and Signal Processing. In this context, this project performs a study covering the applicability of said methods to colorize images automatically. Three Convolutional Neural Network architectures were evaluated according to three quantitative indexes, the root mean square error (RMSE), the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM), in addition to a user study. It was verified that the evaluated models are capable of performing automatic colorization, being able to produce photorrealistc results.**

***Keywords*— Deep Learning, Convolutional Neural Networks, Automatic Colorization.**

I. Introdução

A produção de um sistema capaz de colorir imagens automaticamente tem sido uma meta de pesquisas desde a década de 1980, quando estúdios cinematográficos começaram a lançar versões coloridas dos filmes em preto e branco. Colorir cada *frame* de um filme é um trabalho tedioso e de custo elevado, de maneira que diversos algoritmos foram desenvolvidos para esse fim e, com o passar do tempo, passaram a requisitar cada vez menos interação humana [4].

O problema supra citado apresenta certa complexidade devido à perda de informação observada ao converter uma imagem colorida em uma imagem em escala de cinza. Dois terços da informação são eliminados no processo [2], impossibilitando a determinação algébrica dos valores excluídos.

Todavia, ao observar uma imagem em escala de cinza ou preto e branco, é possível, para um humano, inferir qual seria a provável cor de alguns de seus elementos, como um morango, a grama, ou o céu. Estes fatores indicam a existência, dentro da própria imagem, de informações capazes de revelar sua matriz de cor. Para poder explorar essas informações e produzir colorizações realistas, um sistema deve ser capaz de interpretar a imagem, identificar e localizar seus objetos [3].

Modelos de aprendizado profundo (*Deep Learning*) como Redes Neurais Artificiais (RNAs) e Redes Neurais Convolucionais (CNNs) mapeiam entradas brutas nas saídas desejadas, realizando tarefas que exigem um alto nível de abstração. Diferentemente das técnicas tradicionais de aprendizado de máquina, que requerem a implementação de um extrator de características a partir dos dados de entrada, métodos de *Deep Learning* são capazes de aprender esse processo diretamente a partir dos dados brutos [5].

Essas características são as razões pelas quais técnicas de Aprendizagem Profunda são adequadas para a implementação de colorizadores automáticos. Embora esse tipo de sistema já exista, faz-se necessário avaliá-los e determinar suas contribuições utilizando um mesmo conjunto de métricas e uma mesma base de dados. Sendo assim, esse trabalho se propõe a analisar alguns dos principais colorizadores e compará-los.

A. Colorização Automática com Aprendizagem Profunda

Os primeiros trabalhos sobre colorizadores automáticos implementados com *Deep Learning* datam de 2015. A abordagem típica consiste em fornecer uma imagem em escala de cinza como entrada de uma CNN e obter, em sua saída, matrizes referentes aos seus canais de cores [3], [6], [1], [7]. Abordagens alternativas envolvem a utilização de CNNs para produzir distribuições probabilísticas da colorização para cada *pixel* [4], [2], [8], [9], [10], ou o uso de Redes Generativas Adversariais (GANs [11]) para produzir um colorizador a partir da competição entre duas redes neurais [12], [13].

Os canais de cores são usualmente representados nos espaços de cores CIE Lab [4], [2], [3], [1], [7], [9], [10], YUV [6], [8], [12], ou HCL [3]. A vantagem dessas alternativas é que nesses espaços a imagem em escala de cinza corresponde a um dos canais da imagem colorida, diferentemente do espaço RGB, não sendo necessário que a rede produza informações redundantes. Parte dos trabalhos considera um espaço de cores contínuo e a colorização é modelada como regressão, e, em outros, é modelada como classificação, sendo o espaço de cores subdividido em valores discretos.

Foram escolhidos três colorizadores que abordassem o problema de formas distintas. Deste modo, foi possível analisar sistemas que modelaram a colorização como classificação ou regressão, ou utilizaram espaços de cores diferentes, ou ainda utilizaram estratégias diferentes para prover supervisão semântica às redes. Segue-se uma descrição mais aprofundada das arquiteturas adotadas pela CNN de cada colorizador.

Leo Araújo, Departamento de Engenharia Elétrica, UFCG, Campina Grande-PB, e-mail: leo.araujo@ee.ufcg.edu.br; Luciana Veloso, Departamento de Engenharia Elétrica, UFCG, Campina Grande-PB, e-mail: luciana.veloso@dee.ufcg.edu.br.

1) ***Iizuka, Simo-Serra e Ishikawa 2016 [1]:*** Esse trabalho modela a colorização como uma regressão e utiliza uma CNN para produzir canais de cores no espaço CIE Lab a partir de uma imagem em escala de cinza. Camadas convolucionais e totalmente conectadas são utilizadas para extrair características, permitindo que a colorização realizada considere padrões locais e globais da imagem de entrada. O modelo foi treinado a partir de imagens do *Places Scene Dataset* [14], com 2,3 milhões de exemplos para treino e 19,5 mil para teste.

Conforme a Figura 1a, a CNN apresenta dois ramos de entrada. Um deles processa a imagem de entrada e, através de camadas convolucionais e totalmente conectadas, extrai padrões globais, enquanto o outro ramo utiliza exclusivamente camadas convolucionais para identificar padrões locais. As informações extraídas são concatenadas por uma camada de fusão e servem como entrada para a rede colorizadora.

A rede colorizadora alterna camadas convolucionais e operações de interpolação e tem como saída os canais de cores associados à imagem de entrada. Esses correspondem aos canais *a* e *b* do espaço CIE Lab que, concatenados à imagem de entrada, produzem uma imagem colorida. A CNN possui outra saída, esta responsável por classificar o contexto global da imagem a partir das características globais extraídas.

São consideradas as perdas de ambas as saídas, utilizando o erro médio quadrático para computar os erros na colorização e a função de entropia cruzada para os erros de classificação. Assim, a segunda saída influencia o processo de colorização porque durante o treinamento seus gradientes se propagam pelos extratores de caractetísticas globais e de baixo nível. Consequentemente a presença do classificador melhora a compreensão do colorizador sobre o contexto global da imagem.

2) ***Zhang, Isola e Efros, 2016 [2]:*** Diferentemente do modelo analisado anteriormente, a CNN proposta nesse trabalho considera o problema de colorização como uma tarefa de classificação. Segundo os autores, ambiguidades existentes quanto às colorizações possíveis fariam com que funções de perdas associadas a problemas de regressão, como o erro médio quadrático, levassem a rede a produzir tonalidades dessaturadas como forma de minimizar as perdas.

O banco de dados utilizado foi o ImageNet [15], com 1,3 milhão de imagens para treino e 10 mil para teste. Para poder modelar a colorização como tarefa de classificação, o plano *ab*, definido pelos valores das coordenadas *a* e *b* no espaço CIE Lab, foi quantizado com passo 10. Verificando quais pontos desse plano constavam nas imagens do banco de dados foram definidas 313 classes, cada uma associada a uma cor.

A CNN implementada (Figura 1b) recebe a imagem em escala de cinza e produz uma distribuição probabilística onde cada elemento representa a probabilidade de que um determinado *pixel* possua uma das 313 cores. A conversão para os canais *ab* se dá a partir do cálculo do valor esperado da distribuição em cada *pixel*.

As perdas são calculadas segundo uma versão rebalanceada da função de entropia cruzada. O rebalanceamento foi efetuado com objetivo de penalizar o uso de cores com baixa saturação, ou seja, teve por objetivo forçar a rede a produzir imagens mais coloridas e reduzir o uso de tonalidades acinzentadas. De modo a possibilitar o treinamento, uma função foi definida para mapear as cores da imagem original em uma distribuição probabilística como a produzida pela rede.

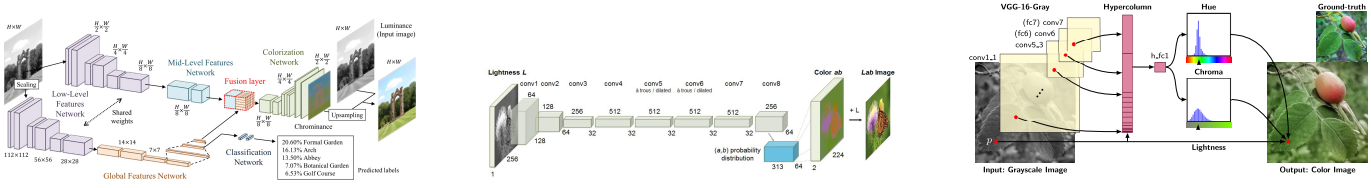
A CNN implementada não utiliza camadas totalmente conectadas ou de *pooling*. Na Figura 1b, cada bloco indica uma sequência de duas ou três camadas convolucionais seguidas de *batch normalization*. Mudanças na resolução foram produzidas por subamostragem ou sobreamostragem entre esses blocos.

3) ***Larsson, Maire e Shakhnarovich 2017 [3]:*** Novamente o modelo desenvolvido modelou a colorização como uma tarefa de classificação. Assim como no modelo analisado anteriormente, os canais de cores foram subdivididos em classes, no entanto, experimentos foram realizados utilizando tanto o espaço CIE Lab quanto o espaço HCL, tendo o segundo produzido melhores resultados. Os dados utilizados para treino foram obtidos do banco de imagens da ImageNet [15], tendo sido utilizadas 1,2 milhão de imagens para treino e 10 mil para teste.

Nesse trabalho, uma rede VGG-16 [16] pré-treinada para classificação com imagens do ImageNet é utilizada como base para o seu colorizador. Essa rede teve sua primeira camada convolucional substituída para receber como entrada uma imagem em escala de cinza, e não colorida. Após a adaptação a rede foi retreinada com imagens em escala de cinza para ajustar os demais pesos ao novo tipo de entrada.

A estratégia empregada está ilustrada na Figura 1c. Primeiramente a entrada é processada pela rede VGG-16, dados espacialmente relacionados a um *pixel* são extraídos das camadas dessa rede e concatenados, formando uma hipercoluna de 12.417 elementos. Essa estrutura age como um vetor de características do *pixel* e serve como entrada para uma camada totalmente conectada, que é seguida pelas saídas da rede que determinam as cores do *pixel*.

Embora o procedimento descrito acima permita a colorização do *pixel* utilizando um grande volume de informações



(a) (b) (c)

Fig. 1: Esquemáticos das redes convolucionais propostas por Iizuka et al [1] (a), Zhang et al [2] (b), Larsson et al [3] (c). Todas as imagens utilizadas nessa Figura foram retiradas dos respectivos artigos.

de diferentes níveis de abstração, hipercolunas apresentam um enorme custo computacional. Consequentemente, ao invés de realizar a colorização *pixel* a *pixel*, apenas 128 hipercolunas são extraídas de forma aleatória, sendo as demais aproximadas utilizando interpolação bilinear.

II. Metodologia

Com o objetivo de avaliar os colorizadores, os algoritmos foram testados com um banco de imagens diferente daqueles pelos quais foram treinados. Imagens coloridas do banco foram convertidas para escala de cinza e coloridas artificialmente pelos modelos, de modo a permitir comparações entre as imagens originais e as que foram coloridas artificialmente. Além disso, um experimento foi realizado para verificar o realismo das colorizações artificiais perante observadores humanos.

A. Banco de Dados

Foi escolhido o banco de imagens LabelMe [17], composto de 2688 imagens coloridas 256×256 divididas em 8 tipos de paisagem: costa, campo aberto, floresta, montanha, estrada, rua, centro de cidade e arranha-céu. Foram utilizadas 2000 imagens do banco, 250 de cada paisagem, redimensionadas para 128×128 para reduzir os requisitos de memória e tempo de processamento para a colorização das imagens.

A Figura 2b mostra um mapa de calor da distribuição das cores no banco de dados em escala logarítmica. Nele, o plano *ab*, do espaço CIE Lab, foi subdividido segundo uma grade 25×25 e as cores correspondentes a cada ponto da grade são ilustradas na Figura 2a para um valor fixo de intensidade luminos ($L = 60$). Valores mais altos de L produzem cores mais claras, e mais baixos produzem cores mais escuras.

Mapas de calor semelhantes foram produzidos para as versões do banco de dados recoloridas pelos sistemas avaliados, Figuras 2c, 2d, 2e. Nos quatro casos avaliados existe uma grande concentração de *pixels* de coloração acinzentada, localizados no centro dos mapas de calor. Esse resultado pode ser atribuído à existência frequente de elementos de coloração dessaturada (i.e. nuvens, edifícios, estradas) na base de dados utilizada. Comportamento semelhante é encontrado em [2], que utiliza a base de dados ImageNet [15].

Todavia, as Figuras 2c e 2e apresentam distribuições mais concentradas na região central, diferentemente das distribuições das Figuras 2b e 2d. Consequentemente, pode-se esperar que o modelo de Zhang [2] produza imagens mais coloridas e os modelos de Iizuka [1] e Larsson [3] produzam colorizações mais acinzentadas.

B. Avaliação Quantitativa

Avaliações quantitativas entre os modelos são realizadas a partir de métricas como o erro RMS ($RMSE$), a relação sinal ruído de pico ($PSNR$) e o índice de similaridade estrutural ($SSIM$). As métricas foram computadas segundo as Equações 2 a 4, para os canais de cores de duas imagens, denominadas I e C, de dimensões $H \times W$ e descritos no espaço de cores CIE Lab. Símbolos como μ_x, σ_x^2, σ_{xy} indicam, respectivamente, a média de x, a variância de x e a covariância entre x e y.

1) $RMSE$: Mensura o erro entre os canais de cores da imagem original e a imagem recolorida. O valor de $RMSE$ para uma imagem se refere à raiz quadrada do erro médio quadrático (Equação 1) calculado para todos os *pixels*, conforme a Equação 2. Seu valor ideal é 0, indicando que o algoritmo reproduziu as cores originais perfeitamente, e aumenta conforme as diferenças entre as imagens original e reconstruída aumentam.

2) $PSNR$: Representa a integridade da imagem reconstruída quando comparada à original. Quanto maior o valor da métrica, maior a qualidade da reconstrução, tendendo a infinito no caso ideal. Essa métrica é calculada a partir do MSE segundo a Equação 3, sendo o termo MAX_I referente ao maior valor permitido aos elementos de uma imagem, 255 para imagens de 8-*bits*.

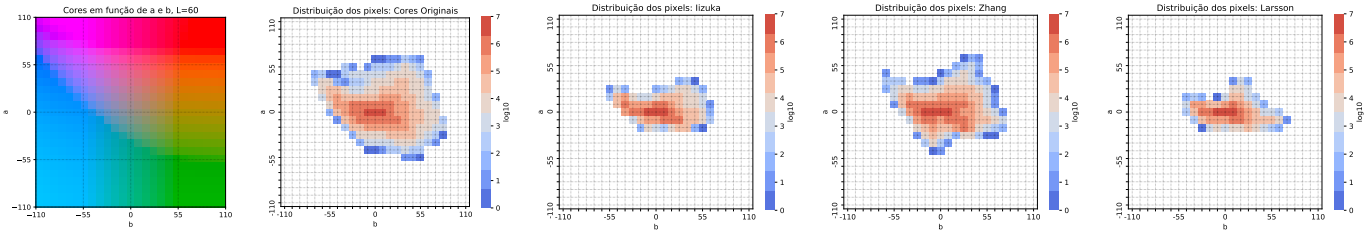
3) $SSIM$: Avalia a degradação da qualidade da imagem levando em conta sua perceptibilidade ao sistema visual humano [18], com $0 \leq SSIM \leq 1$, valores próximos da unidade indicam maior similaridade entre as imagens. Esse índice foi calculado localmente a partir de janelas 5×5 e o valor global ($SSIM$) de uma imagem corresponde à média de todos os valores computados localmente. As constantes $c_1 = 6,5025$ e $c_2 = 58,5225$ evitam divisões por zero no denominador e são proporcionais a MAX_I.

$$MSE = \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{[I(i,j) - C(i,j)]^2}{2.H.W} \tag{1}$$

$$RMSE = \sqrt{MSE} \tag{2}$$

$$PSNR = 20.log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right) \tag{3}$$

$$SSIM = \frac{(2.\mu_I.\mu_C + c_1).(2.\sigma_{IC} + c_2)}{(\mu_I^2 + \mu_C^2 + c_1).(\sigma_I^2 + \sigma_C^2 + c_2)} \tag{4}$$



(a) (b) (c) (d) (e)

Fig. 2: (a) Cores descritas pelo plano ab para $L = 60$, (b - e) distribuição de cores do banco de dados por colorizador.

C. Avaliação Qualitativa

Como o objetivo dos colorizadores é produzir colorizações convincentes e não necessariamente reproduzir as cores originais da imagem, o uso das métricas descritas na seção anterior é criticado em alguns trabalhos [2], [3], [8], [9]. Sendo assim, com inspiração nos trabalhos de Zhang et al [2] e Cao et al [12], foi realizado um experimento para avaliar o realismo das colorizações realizadas aos olhos de observadores humanos.

O experimento foi implementado em *Python* e, a cada etapa, uma imagem foi escolhida aleatoriamente. Cada imagem escolhida era apresentada na tela em escala de cinza até que o usuário pressionasse uma tecla, fazendo com que ela desaparecesse e duas versões da mesma imagem, uma com as cores originais e outra colorida artificialmente, aparecessem na tela durante 1 segundo cada. Uma vez que ambas desaparecessem, o participante selecionava qual das duas imagens apresentava cores artificiais.

Vinte participantes de idade entre 20 e 60 anos realizaram o experimento, sendo 53 imagens distintas sorteadas para cada um deles. Para familiarizá-los com a tarefa, as cinco primeiras imagens serviam como prática, seus resultados eram desconsiderados, as imagens com cores falsas utilizavam canais de cores escolhidos aleatoriamente do banco de dados e, após a resposta, o participante era informado se havia acertado ou não. As 48 imagens restantes eram divididas igualmente entre os modelos avaliados e não eram fornecidas informações sobre sua performance até o final do experimento.

III. Resultados e Discussões

Os valores de $RMSE$, $PSNR$ e $SSIM$ foram computados a partir das imagens do banco de dados e de suas versões reconstruídas por cada colorizador e em escala de cinza. Na Figura 3 são ilustrados os diagramas de caixa para cada métrica avaliada e na Figura 4 são mostradas colorizações realizadas pelos modelos avaliados em conjunto das respectivas métricas.

A. Avaliação Quantitativa

Analisando os diagramas de caixa, ilustrados na Figura 3, percebe-se que a CNN de Iizuka [1] apresenta os melhores resultados nas três métricas avaliadas e a CNN de Zhang [2] apresenta os piores. Devido à grande concentração de *pixels* dessaturados nas imagens, espera-se que colorizações mais acinzentadas produzam valores de MSE menores, implicando em $RMSE$ e $PSNR$ mais baixos para os colorizadores de Iizuka [1] e Larsson [3].

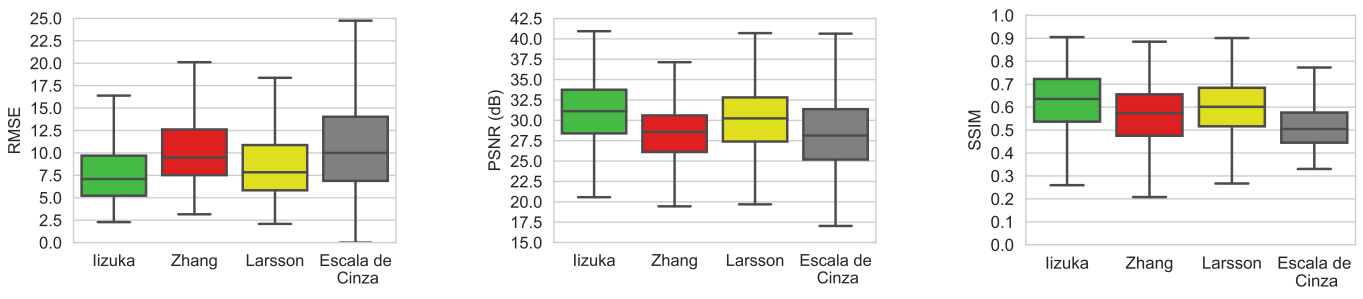
A distribuição das imagens em escala de cinza apresenta grande variabilidade na suas distribuições de $RMSE$ e $PSNR$ e resultados de $SSIM$ concentrados em torno de 0,5. Os dois primeiros resultados são consequência da diversidade do banco de dados, que possui imagens com cores saturadas (i.e. florestas coloridas, pôr do Sol) e dessaturadas (i.e. nuvens, edifícios, estradas), produzindo altos valores de MSE no primeiro caso e baixos valores no segundo. A concentração dos resultados de $SSIM$ em torno de 0,5 ocorre porque imagens em escala de cinza apresentam valores nulos de *a* e *b* para todo *pixel*, não apresentando similaridade aos das imagens originais a não ser que estas sejam naturalmente dessaturadas.

A Figura 4 mostra que resultados realistas podem ser encontrados mesmo em imagens que apresentam baixos valores nas métricas quantitativas. Analisando as colorizações de cada modelo confirma-se que, de fato, as imagens reconstruídas pelo colorizador de Zhang [2] apresentam maior diversidade de cores, e os demais modelos produzem imagens menos saturadas. Por outro lado, algumas imagens produzidas por esse colorizador apresentam manchas, comprometendo seu realismo, o que acontece em menor escala nos demais modelos.

Há certos padrões nas imagens que sugerem sobreaprendizagem (*overfitting*) nos modelos. A tendência de colorir o céu sempre de azul (Figura 4a, linha 1), vegetações sempre de verde (Figura 4a, linha 2) ou construções de cinza (Figura 4b, linhas 2 a 4, Figura 4c linhas 2 e 3) sugere que certos padrões são decorados pelas CNNs, o que apesar de produzir imagens realistas, leva a más avaliações quando esses elementos possuem cores atípicas na imagem original. Por outro lado, imagens como a linha 3 da Figura 4a e a linha 5 da Figura 4c apresentam manchas ou colorização inconsistente de elementos, falhas que comprometem seu realismo, mas não necessariamente impactam as métricas significativamente.

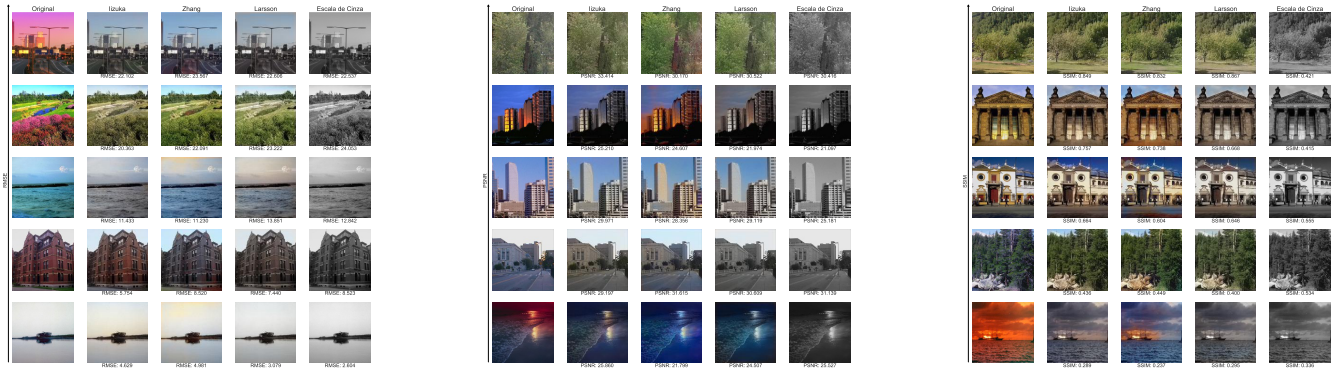
B. Avaliação Qualitativa

No experimento qualitativo, os voluntários apresentaram um erro médio de 41,46% na identificação das imagens coloridas artificialmente. Foram obtidos erros médios de 50,0%, 43,18% e 31,25% para os colorizadores propostos por Iizuka et al [1], Zhang et al [2] e Larsson et al [3], respectivamente.



(a) (b) (c)

Fig. 3: Diagramas de caixa dos resultados obtidos.



(a) (b) (c)

Fig. 4: Comparação de imagens e valores das respectivas métricas.

Ressalta-se que, nesse experimento, um maior erro indica melhor performance, sendo o caso ideal representado por um erro de 50%, mostrando que em média as colorizações produzidas foram tão convincentes quanto as próprias cores originais. Consequentemente, pode-se dizer que as colorizações produzidas pelo modelo de Iizuka [1] foram mais realistas aos olhos dos voluntários e as do modelo de Larsson [3] foram as menos realistas. Observa-se que o modelo de Zhang [2] obteve uma boa pontuação na avaliação qualitativa apesar de ter apresentado os piores valores nas métricas quantitativas.

IV. Considerações Finais

Os resultados obtidos permitem inferir que os colorizadores promovem algum restauro nas imagens em escala de cinza, frequentemente tornando-as mais semelhantes às originais. Contudo, as métricas empregadas não são apropriadas para avaliar a performance desses sistemas porque mensuram o grau de restauração da imagem reconstruída e não o realismo das colorizações produzidas pelos colorizadores.

Além disso, foram verificadas tendências à produção de imagens dessaturadas e/ou com inconsistências na coloração de objetos. Os modelos avaliados fizeram contribuições significativas ao introduzir estratégias para mitigar essas tendências, a exemplo da modelagem do problema como uma classificação [2], [3], o rebalanceamento da função de perdas para favorecer cores mais saturadas [2] e o fornecimento de informações sobre o contexto da imagem para minimizar manchas [1], [3].

Por fim, considera-se viável a utilização desse tipo de sistema em aplicações mais complexas como a colorização de vídeos. Como sugestão de trabalhos futuros, propõe-se o estudo dos impactos de utilizar sinais de vídeo diretamente como entrada das redes. Supõe-se que melhores resultados seriam produzidos pois a disponibilidade de múltiplos *frames* possibilitaria a extração de mais informações sobre cada objeto antes da colorização. Além disso, *frames* com os mesmos objetos em instantes distintos do vídeo estimulariam que eles fossem coloridos da mesma forma, impedindo mudanças de cor repentinas em um mesmo objeto durante o vídeo.

Referências

[1] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (TOG)*, 35(4):110, 2016.

[2] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[3] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. *arXiv preprint arXiv:1603.06668*, 2016.

[4] Amelie Royer, Alexander Kolesnikov, and Christoph H Lampert. Probabilistic image colorization. *arXiv preprint arXiv:1705.04258*, 2017.

[5] Gabriel Chartrand, Phillip M. Cheng, Eugene Vorontsov, Michal Drozdzal, Simon Turcotte, Christopher J. Pal, Samuel Kadoury, and An Tang. Deep learning: A primer for radiologists. *RadioGraphics*, 37(7):2113–2131, 2017. PMID: 29131760.

[6] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 415–423, 2015.

[7] Mariusz Chybicki, Wiktor Kozakiewicz, Dawid Sielski, and Anna Fabijańska. Deep cartoon colorizer: An automatic approach for colorization of vintage cartoons. *Engineering Applications of Artificial Intelligence*, 81:37–46, 2019.

[8] Sergio Guadarrama, Ryan Dahl, David Bieber, Mohammad Norouzi, Jonathon Shlens, and Kevin Murphy. Pixcolor: Pixel recursive colorization. *arXiv preprint arXiv:1705.07208*, 2017.

[9] Jiaojiao Zhao, Jungong Han, Ling Shao, and Cees GM Snoek. Pixelated semantic colorization. *arXiv preprint arXiv:1901.10889*, 2019.

[10] Aditya Deshpande, Jiajun Lu, Mao-Chuang Yeh, Min Jin Chong, and David Forsyth. Learning diverse image colorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6837–6845, 2017.

[11] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[12] Yun Cao, Zhiming Zhou, Weinan Zhang, and Yong Yu. Unsupervised diverse colorization via generative adversarial networks. *arXiv preprint arXiv:1702.06674*, 2017.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[14] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[17] Audi Oliva. Labelme. Computational Visual Cognition Laboratory, 2019.

[18] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.