

# BIAS IN AUTOMATED IMAGE COLORIZATION: METRICS AND ERROR TYPES

Frank Stapel    Floris Weers    Doina Bucur

University of Twente

## ABSTRACT

We measure the color shifts present in colorized images from the ADE20K dataset, when colorized by the automatic GAN-based DeOldify model. We introduce fine-grained local and regional bias measurements between the original and the colorized images, and observe many colorization effects. We confirm a general desaturation effect, and also provide novel observations: a shift towards the training average, a pervasive blue shift, different color shifts among image categories, and a manual categorization of colorization errors in three classes.

*Index Terms*— Image, colorization, bias, error

## 1. INTRODUCTION

The task of colorizing grayscale images is ambiguous, and difficult for a fully automated process without human input. We provide *systematic measurements of color bias* against the ground truth. We focus on a widely used deep-learning colorizer, DeOldify [1] (also a popular Twitter bot), whose colorization models have generative adversarial network (GAN) architectures and are pretrained on ImageNet [2]. To measure colorization bias objectively, we take measurements over ADE20K [3], a dataset different than the training dataset.

Colorizers have biases, often demonstrated with example images, not systematic measurements. Early models were accurate on landscape and home, but not on complex scenes [4, 5]. A frequent problem is *desaturation* [4, 6], due to loss functions (inherited from standard regression) that encourage conservative predictions. Colorizers aiming at vibrant colors can still fail to recover long-range *color consistency* [7, 6, 8, 1], make confusions between colors [7], and output *sepia* [9, 7, 10, 11] or *gray tones* [12]. Some models are biased towards frequent colors (e.g., red cars) [10], and can confuse the *context or boundaries of objects* (regions with fluctuations may be colored like grassland) [10, 13, 8, 1]. When an object is not in the GAN distribution (or GAN inversion fails) the output contains unnatural or incoherent colors [14].

To validate colorization, prior work reports global statistics: pixels-averaged MAE or RMSE [5, 6, 15, 10]), peak signal-to-noise ratio (PSNR) [4, 6, 11, 8, 16, 14], structural similarity (SSIM) [11, 8, 14], per-pixel accuracy [7, 12, 10], the Fréchet inception score (FID) [14], a colorfulness

score [14], histograms, or human preference. These metrics cannot capture fine-grained biases, such as errors that occur persistently in a certain region of an image category. We thus design not only *global*, but also *local* and *regional bias metrics*. We find a pronounced increase in neutral shades and shift towards the training-average colors, a shift towards blue (pronounced in the center of images), but also that image categories are affected differently (sky patches in nature, urban, and industrial scenes are counterintuitively stripped of blue). In a user study, 60% of inaccurate colorizations were found to be plausible, with the rest colorization failures.

## 2. METHOD

DeOldify [1] has two colorization models, with different architecture and training process. The *artistic model* creates vibrant, colorful results, but does less well in common scenarios: nature scenes and portraits. The *stable model* is best on nature scenes and portraits, with fewer unnatural miscolorations, but also less vibrancy. Both are trained on a fraction of ImageNet [2]. We test here on all (except 5 non-RGB) color images from the ADE20K dataset [3]. ADE20K provides 25,564 images, split into 10 categories (as in Table 1, column # images). They are diverse in size and content, are annotated with their semantic category, and the scenes are tagged with objects (such as sky) and object parts.

**Table 1. ADE20K image categories and % sky**

Category	# images	% sky
Urban	7239	82.33%
Home or hotel	6117	0.72%
Nature landscape	3332	75.16%
Unclassified	2536	58.49%
Workplace	1565	1.91%
Sports and leisure	1528	41.53%
Cultural	1115	3.23%
Shopping and dining	1089	1.74%
Transportation	693	4.17%
Industrial	350	81.77%

To compare original with colorized images, we grayscale (ITU-R 601-2 luma transform [17]) and colorize the ADE20K dataset. We then take bias measurements in two color spaces. The **RGB** (trichromatic and additive) color space remains the

The first two authors contributed equally.

most widely supported and understood system for the characterization and comparison of colors in digital images. This uses three monochromatic primaries at standardized wavelengths (defined in standard CIE 1931 [18]), and is perceptually non-uniform: an equal distance in the color space may not correspond to equal differences in color. **CIELAB** (also  $L^*a^*b^*$ ), defined in CIE 1976 [18], expresses color as three values  $L^*$  (perceptual lightness),  $a^*$  (red to green) and  $b^*$  (blue to yellow). This was intended as a perceptually uniform space: a numerical change in color corresponds to a consistent perceived change in color, so Euclidean distances can be used to compare color transformations in all directions.

We systematically take three types of color bias measurements: global, local, and regional, between the original and the colorized images. **Global color bias** measurements show the bias in RGB and CIELAB channel values, treating equally all pixels in all test images. The method for this is shown in Fig. 1, schematically: the overall distributions of channel values, taken independently per channel, are compared across all images, and we report the channel shifts  $\Delta$ . This may show, for example, that the pixels in colorized images have high blue-channel values more frequently than the original images.

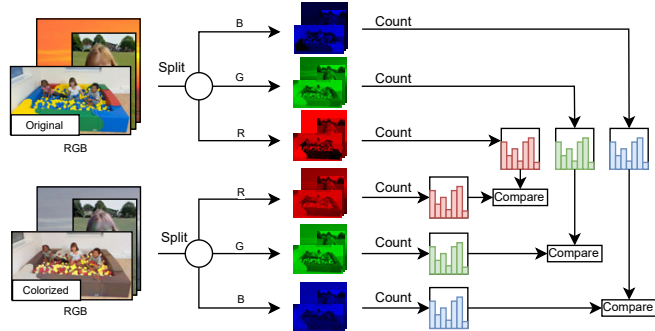


Fig. 1. Global color bias (method; similar for CIELAB)

**Local color bias** measurements are instead fine-grained. Their aim is to show a two-dimensional color shift between the colorized and the original, on average both for the entire dataset, and per image category. To achieve this, since the ADE20K images are diverse in size, each image, regardless of aspect ratio and resolution, is size-normalized into an aggregated 64x64 image. (This was chosen because it is smaller than the smallest original image size in the dataset.) To avoid confusion between the pixels in the original image and the normalized pixels, we call the latter “cells”. The color in each cell is the average color of the original pixels. For both RGB and CIELAB, we measure the average color shift per cell. The methodological pipeline for this is shown in Fig. 2.

We also take a second local bias measurement, which tests the hypothesis whether the colorization strips away some of the vibrant colors and replaces them with dull, muddy shades. For this, we first calculate, per cell in the normalized image size, the average color of the training dataset (2% of Ima-

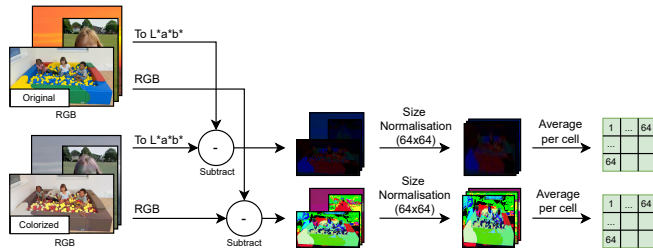


Fig. 2. Local color bias: color shift (method)

geNet). We call this the “mud” image. The cells in the mud image need not contain the exact same color. Then, per individual image and per cell, we measure the color distance to the mud color of that cell, and report whether, on average, this distance becomes *smaller* across colorized images. The methodological pipeline is shown in Fig. 3. Since this measurement relies on Euclidean distances between colors, we only perform it in the perceptually uniform CIELAB space.

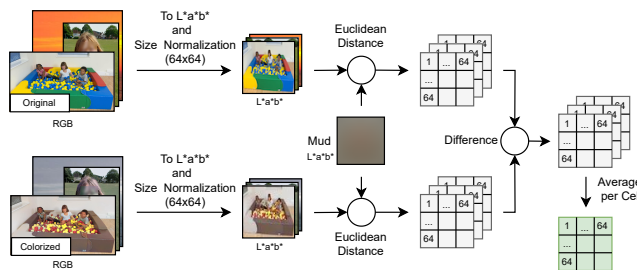


Fig. 3. Local color bias: distance to mud (method)

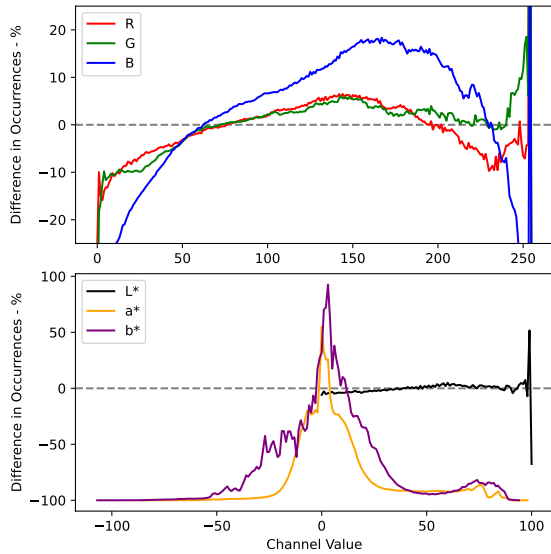
**Regional color bias** measurements are based on the local color shift measurements, but are done separately for specific special regions within the image, such as the center, or the top third. These special regions are image patches defined by popular composition rules in photography [19], such as the rule to fill the center of the frame with a subject, the *rule of thirds* which defines 3 equal vertical or horizontal patches and their intersection, and the *golden-rule grid*, which does the same in the ratio 1:0.618:1. They are applied as masks to the local color shift results, to draw regional conclusions.

### 3. RESULTS

We present the measurements, and insights gained from them. All are for DeOldify artistic, unless otherwise specified.

**Global color bias.** We show RGB and CIELAB channel shifts  $\Delta$  in Fig. 4. A positive  $\Delta$  means a more frequent occurrence of that channel value in the colorized images. (Maximum channel values—R, G, B near 255,  $L^*$ ,  $a^*$ ,  $b^*$  near 100—are rarely present in the data [12], leading to extreme or noisy shifts at those bounds.) We observe a shift in the

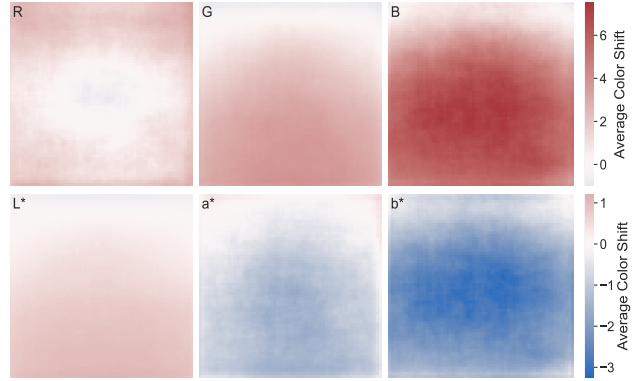
distribution of B channel values: an *increased frequency in mid-to-high blue components* among colorized images. This is also present, but is much less significant, for the R, G, and L\* channels. On the other hand, the shifts in the distributions of the a\* and b\* channels show a *pronounced increase in frequency for neutral shades* (between green and red for a\*, and blue and yellow for b\*), and thus a *pronounced decrease in frequency for saturated colors*. Many a\* channel values with an absolute value over 25 disappear from the distribution. In summary, we observe **global bias towards neutral shades, and global bias towards mid-to-high-range blues**.



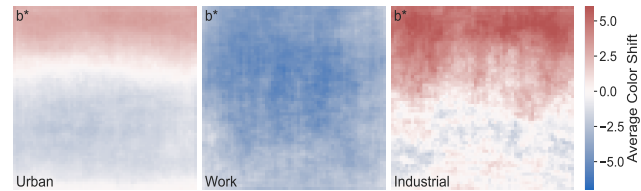
**Fig. 4. Global color bias:** channel  $\Delta$  (positive means higher frequency in colorized images)

**Local color bias: color shift.** We show per-cell average channel shifts (across the entire dataset) for the R, G, and B (top) and L\*, a\*, and b\* channels (bottom), in Fig. 5. A positive per-cell shift means a higher average value in the colorized images. While we had previously (in Fig. 4) observed similar global shifts between the red and green channels, the local measurements now show that these color shifts have a different spatial distribution: the colorization process *red-shifts slightly the periphery* (but not the center), but *green-shifts slightly the bottom two thirds* of the images. The latter is explained by the more frequent occurrence of natural landscapes at the bottom of the images, with these areas further shifted to green in colorization. The former implies that the (sometimes colorful) objects in the center of the original images are stripped of some of their red, with the opposite occurring for the image background.

The colorization also on average **blue-shifts almost every cell** of the images, with double the shift amplitude of the green. The most pronounced shift is in the center of the images. This implies that the blue shift is not a further deepening of the sky blue; it is instead a pervasive effect throughout the



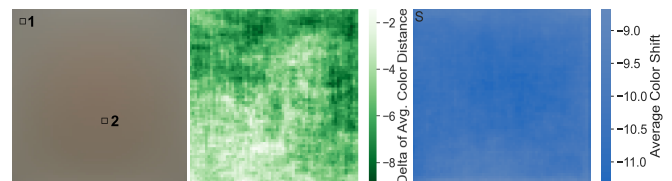
**Fig. 5. Local color bias:** average channel shift for R, G, and B (top) and L\*, a\*, and b\* (bottom), over the entire dataset (positive means higher average value in colorized images)



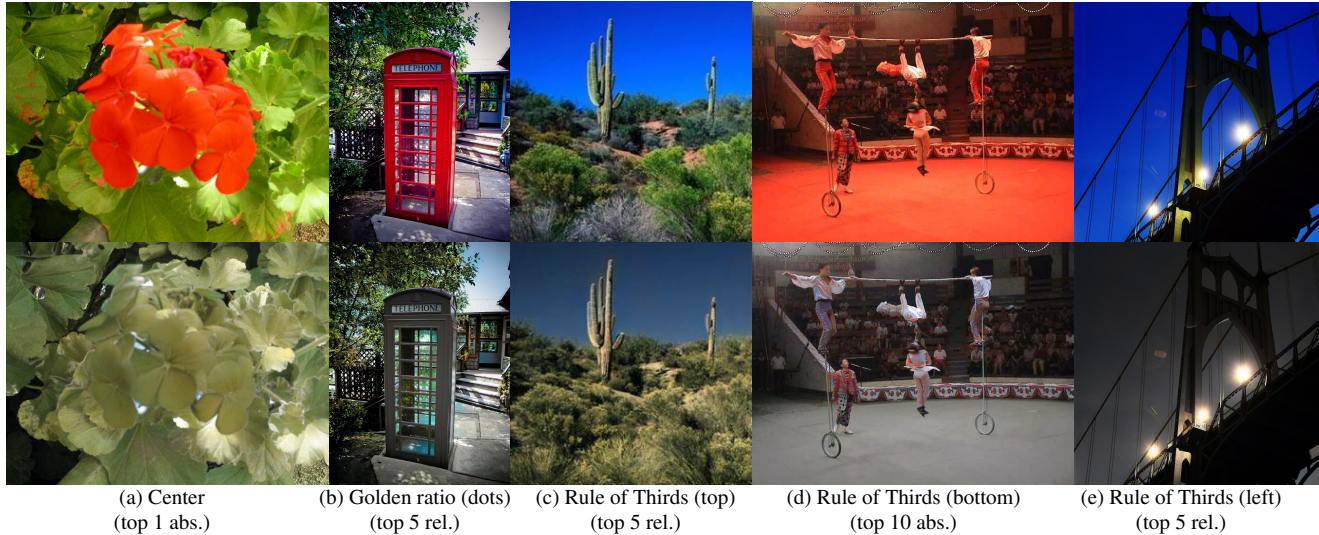
**Fig. 6. Local color bias:** b\* (blue) channel shift is not uniform per image category (urban, nature, and industrial scenes have top regions shifted away from blue)

images. This is seen in both the B and b\* channels: the same blue shift manifests as a positive shift in the B channel, but a negative one (away from yellow and towards blue) in b\*. We then verify that this blue shift is **not uniform across image categories**: Fig. 6 shows the breakdown of the b\* channel per image category, for three of the largest categories. (All categories not shown behave like Work, except Nature, which is like Urban.) While 7 out of 10 image categories are almost uniformly blue-shifted, urban, nature, and industrial scenes have their top regions colorized with a shift away from blue. Since these are the image categories with the most sky (present in 75+% of the images, from Table 1), this means that the colorizer **strips blue from patches of sky**.

The average Euclidean distance across cells in CIELAB (between colorized and original images) also varies per image



**Fig. 7. Local color bias:** (left) mud; (center) difference in distance to mud; (right) average S (saturation) channel shift



**Fig. 8. Regional color bias:** examples of most inaccurate colorizations (original image at the top)

category. With lower average distance meaning more accurate colorization, the top are: urban (average distance 1.885), unclassified (2.146), and nature (2.234). The bottom are: work (4.128), shopping (4.143), and cultural (4.331).

**Local color bias: distance to mud.** The mud image (the average color of the training dataset) is shown in Fig. 7 (left). It contains a gradient between cell 1 (RGB 131, 128, 119) and cell 2 (RGB 126, 111, 95). Using this image as a baseline of comparison, we then observe (in Fig. 7, center) that *the distance image-to-mud decreases* for the artistic model, on average, after colorization: a negative difference in the figure means that the distance is lower on colorized images. The result is very similar for the stable model, but with an even higher amplitude (reaching -10). This confirms our hypothesis and insights in the related work [4, 6, 12] that the colorization strips away some of the colors, and **shifts the colors towards the training average**. This is also confirmed by a supplementary measurement of the average shift in the S channel from the HSV color space (shown in Fig. 7, right, where the range of channel values is  $[0, 100]$ ). The colorized images are almost **uniformly and heavily desaturated** on average, in comparison to the originals.

**Regional color bias.** For each region defined by the rule of thirds and the golden ratio, we extract (1) the top  $n$  images by the *absolute* color shift in each region, and (2) the top  $n$  images by the *relative* color shift between the region and the rest of the image. The latter allows to capture examples where a region is colorized very differently from the remaining image. Fig. 8 shows examples of both. In (a) and (b), the colorization of the dominant object failed, either in absolute or relative terms: (a) the flower was colorized like its leaves, and (b) an object of characteristic color that a human would guess correctly was not colorized. In (c)-(e), the colorization is muted, but plausible. In all, there is overall desaturation.

**Manual categorization of errors.** Finally, we add a user study<sup>1</sup>: we selected the top 400 images by (1) the average and standard deviation in color shifts between original and colorized, and (2) regional color shifts of the relative type. After removing duplicate images in this set, the rest were manually categorised by the type of failure observed in colorization. A minority of the failure modes were not clear-cut. Only 5% of the images completely failed to colorize and were essentially grayscale. In 23% of the cases, one dominant object failed to colorize (examples include (a)-(b) in Fig. 8). However, in the majority (60%) of the cases, the results were judged still plausible (examples include (c)-(e) in Fig. 8), and akin to a “mood change” in the image.

#### 4. DISCUSSION AND CONCLUSIONS

We presented insights on the color shifts in images colorized by the GAN-based DeOldify model. We introduced local and regional bias measurements between the original and the colorized datasets, and obtained quantitative and qualitative results showing many colorization effects. We observe desaturation (confirming prior knowledge [4, 6, 12]), but also provide *novel observations*: a shift towards the training average, a pervasive blue shift, different shifts among image categories, and a manual classification of the errors. This study has limitations: the measurements included only the two public colorizers available with DeOldify, and only one image dataset—but we conclude that pervasive biases remain present in advanced colorization models. Our results may guide the development of automated AI colorizers, which could, for example, use semantic input to resolve some of the regional color shifts per image category.

<sup>1</sup>We provide more image examples at <https://github.com/WeersProductions/colorization-bias>.

## 5. REFERENCES

- [1] Jason Antic and contributors, “DeOldify,” <https://deoldify.ai/>, Accessed Jan 2022.
- [2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, 2009, pp. 248–255.
- [3] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba, “Scene parsing through ADE20K dataset,” in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017, pp. 633–641.
- [4] Zezhou Cheng, Qingxiong Yang, and Bin Sheng, “Deep colorization,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 415–423.
- [5] Aditya Deshpande, Jason Rock, and David Forsyth, “Learning large-scale automatic image colorization,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 567–575.
- [6] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich, “Learning representations for automatic colorization,” in *European conference on computer vision*. Springer, 2016, pp. 577–593.
- [7] Richard Zhang, Phillip Isola, and Alexei A Efros, “Colorful image colorization,” in *European conference on computer vision*. Springer, 2016, pp. 649–666.
- [8] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang, “Instance-aware image colorization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7968–7977.
- [9] 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)*, vol. 35, no. 4, pp. 1–11, 2016.
- [10] Kamyar Nazeri, Eric Ng, and Mehran Ebrahimi, “Image colorization using generative adversarial networks,” in *International conference on articulated motion and deformable objects*. Springer, 2018, pp. 85–94.
- [11] Gokhan Ozbulak, “Image colorization by capsule networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.
- [12] 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*, 2017, pp. 1125–1134.
- [13] Thomas Mouzon, Fabien Pierre, and Marie-Odile Berger, “Joint CNN and variational model for fully-automatic image colorization,” in *International Conference on Scale Space and Variational Methods in Computer Vision*. Springer, 2019, pp. 535–546.
- [14] Yanze Wu, Xintao Wang, Yu Li, Honglun Zhang, Xun Zhao, and Ying Shan, “Towards vivid and diverse image colorization with generative color prior,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14377–14386.
- [15] 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*, 2017, pp. 6837–6845.
- [16] Patricia Vitoria, Lara Raad, and Coloma Ballester, “Chromagan: Adversarial picture colorization with semantic class distribution,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2445–2454.
- [17] NTSC, “NTSC: ITU-R 601-2 luma transform,” 1982.
- [18] CIE, “Colorimetry, 3rd edition,” Standard, International Commission on Illumination (CIE), 2004.
- [19] Eftichia Mavridaki and Vasileios Mezaris, “A comprehensive aesthetic quality assessment method for natural images using basic rules of photography,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 887–891.