SCD: Statistical Color Distribution-based Objective Image Colorization Quality Assessment

Hongjin Lyu[1][0000−0002−2185−7754], Hareeharan Elangovan[1], Paul L. Rosin[1][0000−0002−4965−3884], and Yu-Kun Lai[1][0000−0002−2094−5680]

School of Computer Science and Informatics, Cardiff University, Wales, UK

Abstract. Colorization research has long been a focal point in computer vision and image processing. However, due to its inherently ill-posed nature, a reasonable assessment of the quality of their outcomes remains a challenge. Subjective evaluations are often restricted to a limited number of participants due to the high costs. This along with the existence of individual differences and subjective biases makes it difficult to derive convincing conclusions. Despite no need for participants in objective evaluation metrics, the currently widely applied objective metrics fail to accurately reflect the quality of colorization results, thereby impeding the attainment of consistency with subjective user opinions. Facing the above problems, we propose a novel Statistical Color Distribution-based Objective Evaluation Metric (SCD) for better consistency with human opinions. We first segment images into semantic regions. For each semantic type, a novel two-dimensional natural color distribution w.r.t. hue and saturation is collected to better align with human perceptual observations during image assessment. An adjacency weighted matrix considering surrounding neighboring regions smooths the color distribution table, enabling a more reliable quality assessment. The application of probability density eliminates the issue of frequency anomalies caused by human visual insensitivity, ensuring more accurate evaluation. Through extensive and comprehensive experiments involving two distinct datasets with the participation of 1321 volunteers, this paper demonstrates that the proposed SCD is more consistent with subjective user opinions compared with current objective metrics for evaluating colorization.

Keywords: colorization evaluation · objective metrics · statistical color distribution

1 Introduction

Image colorization techniques add colors to input images, which is well studied in computer graphics and computer vision, where multiple related problems have been studied. Among them, natural image colorization has garnered significant attention due to its wide-ranging application scenarios. The characteristic of natural image colorization tasks, where multiple plausible colors can be assigned to the same object, presents a typical ill-posed problem. This essence of natural image colorization tasks, with the inherent ambiguity and subjectivity, renders

the quality evaluation both crucial and challenging, which makes researchers prefer to adopt a combination of objective metrics, subjective assessments, and user studies to effectively provide a convincing evaluation.

Objective metrics [26, 30] commonly used for evaluating colorization results focus on the fidelity between the original and colorized images. However, the primary goal of colorization is to achieve visually pleasing results, where subjective satisfaction is influenced by natural color priors and personal aesthetics rather than similarity to the original image. Therefore, relying solely on similarity for evaluation may overlook perceptual differences and contradict the natural color distribution. Subjective evaluation involves human observers visually assessing colorization results for quality and satisfaction, providing direct feedback on visual quality, color accuracy, and perceptual satisfaction. However, individual differences and biases can cause inconsistent feedback. To mitigate this, a large number of observers are typically recruited for more comprehensive and consistent assessments. However, these user studies are resource-intensive and costly, making them challenging to implement in practice.

Given the existing research landscape, this paper proposes a novel statistical color distribution-based objective metric (SCD), which facilitates the accurate and cost-effective evaluation of colorization results. By leveraging statistical information of object colors, SCD assigns scores to the plausibility of colors at each pixel, providing a quantitative measure achieving high consistency with human opinions. Our main contributions are summarized below:

- This paper innovatively proposes an objective evaluation matrix (SCD) from a statistical perspective. A proper hue-saturation natural color distribution in HSV space is collected to approximate human perception.
- We further apply adjacency weight matrices and probability densities to handle different bin sizes in statistical analysis to ensure consistent prediction. This is necessary due to human perceptual insensitivity to hues when saturation is low, and such cases are treated with a bin that includes all hue values.
- Experimental results demonstrate that SCD achieves higher consistent with human subjective opinion compared with existing objective metrics.

2 Related Work

In computer graphics and computer vision, proper evaluation metrics are crucial for advancing technology. For example, using Random Forests to assess color transfer quality [7], employing Gaussian RBF kernels to evaluate image generation models [10] and using a binary classifier to measure texture tilability [19]. In natural image colorization research, common metrics roughly fall into two categories: Full-Reference (requiring a reference image) and No-Reference (not requiring one). In this section, the common metrics are introduced and their correlation with user perception is explored in Section 2.3.

2.1 Full-Reference Metrics

In this subsection, nine common colorization evaluation metrics that require input of a reference image or original image are introduced.

Mean Squared Error (MSE) widely applied in colorization methods [2, 17, 25] is commonly used to measure the degree of difference between colorized images and original image. Root Mean Squared Error (RMSE), which is the square root of MSE, is also widely applicable in practical scenarios. Mean Absolute Error (MAE) is less sensitive to outliers because it uses absolute values to measure errors rather than squaring them.

Structural Similarity Index (SSIM) [26] widely applied in colorization methods [14, 28] compares the structural information and pixel value distribution between images by simulating the way human eyes perceive images. It helps us understand differences in structure, brightness, and contrast between two images, but only capture the above differences based on single channel, e.g. grey. Multi-Scale Structural Similarity Index (MSSSIM) [27] applied in colorization methods [1, 17, 25] is an extended version of SSIM that takes into consideration the structural similarity of images at multiple scales (multiple resolution levels).

Peak Signal-to-Noise Ratio (PSNR) applied in colorization methods [1, 28, 25, 17] quantifies the relationship between the highest attainable power in a signal and the level of unwanted noise (measured as the difference between the colorization result and ground truth) that distorts its accurate representation.

Learned Perceptual Image Patch Similarity (LPIPS) [30] employs a pre-trained neural network (VGG [21]) to extract feature of images, capturing low-level textures and high-level semantic information.

Fréchet Inception Distance (FID) [6] widely applied in colorization methods [1, 3, 14, 28] utilizes a pre-trained deep neural network (Inception [24]) to extract high-level semantic features from images. It computes mean and covariance of Gaussian distributions for ground truth and colorized images, then calculates Fréchet distance to evaluate their difference.

CDR [11] (Cluster Discrepancy Ratio) discerns the similarity between original and recolorized images by examining the differences in super-pixel assignments.

2.2 No-Reference Metrics

This subsection introduces two commonly used metrics that do not require a reference image. Colorfulness [5] widely applied in colorization field [8, 28] represents colorfulness through a linear combination of statistical properties in the CIEab color space. Inception Score (IS) is computed from a pretrained Inceptionv3 and evaluates the diversity and classification accuracy of a dataset.

2.3 Objective Metrics vs. Human Perception

In natural image colorization research, a single object can have multiple plausible colors, causing possible inconsistencies with the original image. Therefore, the

goal of evaluating colorization algorithms is to assess the reasonableness and naturalness of the colors rather than matching the original. Given this fact, existing Full-Reference metrics may not accurately reflect actual performance, and No-Reference metrics fail to simulate human perception and judgment effectively, leading to inconsistent performance with subjective opinions.

Recent research, including the Human Evaluated Colourisation Dataset (HECD) [18] and the Subjective Evaluation of Colourized Images Dataset (SECID) [25], has shown that common objective metrics do not strongly align with user subjective opinions. This underscores the urgent need for an objective metric that better reflects user opinions. In response, this paper introduces SCD, which measures the naturalness from the color statistical perspective.

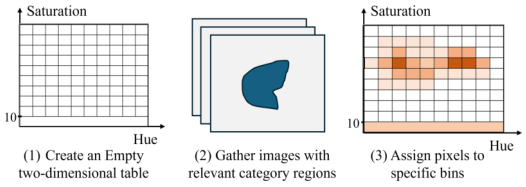


Fig. 1: Flowchart of collecting color distribution in Section 3.1.

Fig. 2: Flowchart of scoring input images in Section 3.2.

3 Methodology

3.1 Collections of Color Distributions

Color Space Among various color spaces, this paper opts to employ the HSV color space for describing and representing color attributes and features, primarily due to the following two reasons:

- In contrast to LAB, YUV, and XYZ color spaces, HSV intuitively decomposes colors into hue, saturation, and value, aligning more closely with human perception, which has also been mentioned in [3, 22] and been widely embraced in the research field of painting [4, 12, 15].
- The three components of HSV are independent, enabling accurate color recognition and selection for easier color analysis and processing.

In summary, opting for the HSV color space enhances the representation of color distributions, making it more effective for colorization evaluation tasks requiring careful attention to color aspects.

Color Distribution Dimensions In the HSV color space, hue represents basic colors, and saturation indicates color intensity. This paper constructs a two-dimensional color distribution using both hue and saturation to better capture color characteristics and differences, omitting luminance which is not relevant for evaluating colorization.

Data Selection Color distributions are object-specific, so we estimate them for each category using diverse natural images. Accurate color distributions require a substantial number of images with region label information, typically obtained via semantic segmentation. However, segmentation accuracy is crucial, as errors can skew the color distribution and affect evaluation criteria. Instead of relying on current segmentation methods, this paper uses the ADE20K dataset, which includes over 20,000 high-resolution images, 150 semantic categories, and precise pixel-level annotations, ensuring accurate color information for each category.

Overall Process As shown in Fig. 1, to collect the color distribution of a specific object category, we follow these steps: (1) an empty two-dimensional color distribution table is created for each category. Due to the human difficulty in accurately distinguishing different colors at low saturation, all colors with saturation ≤ 10 (out of 100) are placed within one specified bin. (2) All images that contain regions belonging to the category of interest are gathered. (3) For each image containing the category of interest, we extract pixels that belong to the semantic category and assign them to specific bins in the color distribution table based on their color information and bin size settings.

3.2 Statistical Distribution based Colorization Metric

We utilize the obtained statistical distributions to evaluate the quality of images, assigning scores based on the plausibility of each pixel and aggregating the results across the image. The overall flowchart of scoring input images is shown in Fig. 2.

Scoring One Image The overall score for an image is defined as the mean of the scores for all the pixels, as shown in the formula below:

$$SCD = \frac{1}{N}\sum_{i=1}^{N} p_i, \qquad p_i = \frac{S_i}{S_{\max}} \tag{1}$$

where N represents the total number of pixels, and p_i represents the score of the i-th pixel, which is calculated based on the above formula. S_i is the unnormalized score for the i-th pixel based on its color bin value, and $S_{\max}$ is the highest score among all bins for that category, representing the most common color. This ratio reflects how closely the pixel's color matches the typical color for that category. A higher score indicates a closer match, while a lower score suggests greater disparity. Details of S_i will be described in the next subsection. This approach normalizes the final score to a [0, 1] range, ensuring scores for different pixels or categories are on a similar scale.

Scoring One Pixel To work out the unnormalized plausibility score of a specific pixel S_i, we first utilize the color distribution of its corresponding semantic category and find out the corresponding bin $(\tilde{h}, \tilde{s})$ where $\tilde{h}$ and $\tilde{s}$ are the bin indexes for hue and saturation respectively. Let H and Q denote the window sizes for hue and saturation in terms of bins to obtain smoother plausibility score estimation. S_i is defined as:

$$S_i = \sum_{h=\tilde{h}-\frac{H-1}{2}}^{\tilde{h}+\frac{H-1}{2}} \sum_{s=\tilde{s}-\frac{Q-1}{2}}^{\tilde{s}+\frac{Q-1}{2}} W_{hs}^i \times PD_{hs}^i, \tag{2}$$

where the score of the allocated bin of the i-th pixel is the weighted average of all bins within $H \times Q$ regions. W_{hs}^i and PD_{hs}^i are the weight and probability density for hue h and saturation s.

To work out the weight for a bin, we first calculate the distance of the bin to the center bin $(\tilde{h}, \tilde{s})$ as

$$D_{hs}^i = \sqrt{\left(E_s^i \times C_s\right)^2 + \left(E_h^i \times C_h\right)^2} \tag{3}$$

where C_s and C_h respectively control the change rates in the saturation and hue dimensions, while E_s^i and E_h^i represent the distances in the saturation and hue dimensions to the central bin. For the saturation, the absolute difference $|s-\tilde{s}|$ is used as E_s, and for the hue, its circular characteristic is taken into account when working out E_h for hue values $h, \tilde{h} \in [0, 360)$, as $E_h = \min(|h-\tilde{h}|, 360-|h-\tilde{h}|)$.

With the distance to the central bin D_{hs}^i and let $D_{\max}^i$ be the maximum distance within the $H \times Q$ neighborhood, W_{hs}^i is defined as

$$W_{hs}^i = 1 - \frac{D_{hs}^i}{D_{\max}^i}. \tag{4}$$

For PD_{hs}^i, instead of directly using the frequency count $F_m(h, s)$ in the corresponding bin, we calculate the probability density

$$PD_m = \frac{F_m(h, s)}{B_s(h, s) \times B_h(h, s)} \tag{5}$$

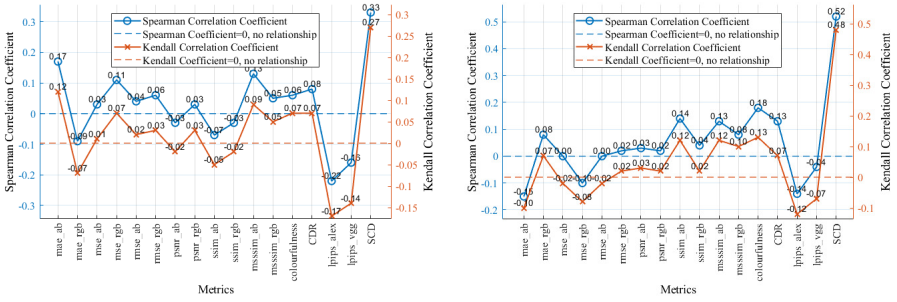
where B_s and B_h represent the saturation bin size and hue bin size, respectively.

4 Evaluation

To effectively reflect the performance of objective colorization evaluation metrics, the evaluation section of this paper is conducted on two datasets, Human Evaluated Colourisation Dataset (HECD) [18] and SECID [25] (to be introduced in Section 4.1). Then, Section 4.2 compares SCD with nine commonly used objective colorization evaluation metrics. In Section 4.3, this paper additionally explores the impact of key factors in the SCD computation process.

4.1 Datasets Details

HECD [18] used 20 benchmark images from the Berkeley Segmentation Dataset [16] and collected user ratings for different re-colorized versions. Each image had 65 variants: six from colorization algorithms (Adobe Photoshop, Colorize photos, and [13, 29, 31, 9]), and 59 from manual adjustments of chroma, hue, or overall image offset. This paper evaluates objective metrics using only the six algorithm-generated variants since they simulate typical colorization scenarios. SECID [25] used 20 benchmark images from the ImageNet validation set [20], generating four variants using different methods [13, 29, 9], including two versions of [9]. Human generally lean towards images that convey a sense of naturalness, a sentiment that aligns with the findings in SECID [25]. Therefore, the preference rating data of SECID [25] is selected as the validation data for this paper. To obtain reasonable semantic segmentation maps, Segmenter [23] known for its performance on the ADE20K dataset, is used to generate initial segmentation maps for both datasets. Given the limited number of benchmark images (20), we make minor manual corrections to enhance accuracy in boundary areas.



(a) HECD dataset [18] (b) SECID dataset [25]

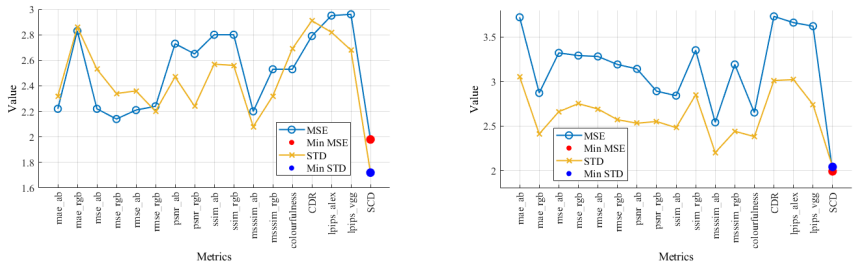
Fig. 3: Rank-based correlation analysis comparing with other metrics under HECD (a) and SECID (b) datasets in Section 4.2.

4.2 Comparison with Other Objective Metrics

Implementation Details This section compares SCD with nine widely used metrics: MAE, MSE, RMSE, PSNR, SSIM, MSSSIM, LPIPS, Colorfulness, and CDR. Evaluations for the first six metrics are conducted in LAB and RGB color spaces, with scores averaged across all color channels on a single test image. In LAB color space, the L channel is excluded since it remains consistent across different re-colorized images. SCD is not compared with IS and FID, as they focus on overall model performance rather than individual image assessment.

This paper initially uses Spearman and Kendall correlation coefficients to analyze the relationship between objective metric scores and user subjective scores. A high correlation between the ranking orders suggests strong similarity between these scores. Following this, mean squared error (MSE) and standard deviation

(STD) are employed for error analysis to measure difference and dispersion further. Results for MSE and STD are multiplied by 10 and displayed with two decimal places. All metrics and user ratings undergo min-max normalization to address score range discrepancies before analysis.



(a) HECD dataset [18] (b) SECID dataset [25]

Fig. 4: Error Analysis comparing with other metrics under HECD dataset (a) and SECID dataset (b) in Section 4.2.

Rank-based Correlation Analysis: Fig. 3(a) and Fig. 3(b) separately show the rank-based correlation analysis under HECD and SECID.

- RMSE and PSNR which use RGB, as well as MS-SSIM in LAB and RGB, exhibit consistent trends. However, the performance of RMSE in RGB is contrary to expectations; the other three metrics (PSNR in RGB, MS-SSIM in RGB and LAB) show weak correlation.
- MAE, MSE and SSIM show inconsistency across different evaluation datasets.
- Colorfulness exhibits an expected correlation but has 73.60% and 73.0% lower Spearman and Kendall correlations, respectively, compared to SCD.
- CDR shows consistent positive correlations on both datasets but averages only 0.09, which is much lower than SCD's correlation.
- Both versions of LPIPS exhibit consistent negative correlations on two datasets as anticipated, which are outperformed by SCD.
- SCD achieves the highest and positive (as expected) correlation strength in all testing scenarios, which indicates that SCD is most consistent with subjective user opinions.

Error Analysis: The error analysis results are shown in Fig. 4(a) and Fig. 4(b) for the HECD and SECID datasets, respectively. Despite some differences, both datasets show similar trends in error analysis.

- SCD consistently achieves the lowest MSE and STD values. This indicates both smallest average errors and least dispersion from user subjective ratings, suggesting closest alignment and greatest stability in SCD's predicted scores.

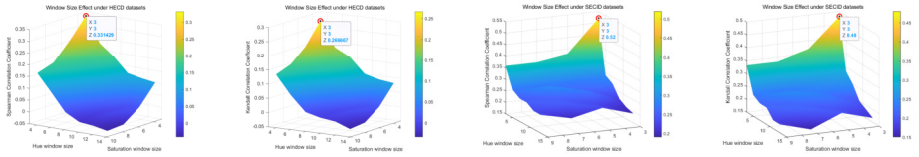
(a) Recolorized A (b) Recolorized B (c) Scores

Fig. 5: Intuitive display about the scores of real examples in Section 4.2. The blue and orange lines respectively represent the scores for Recolorization A and B. When a metric's scores for Recolorization A and Recolorization B are consistent with the user's subjective ratings and have similar ranges of variation, it indicates that the metric better reflects the user's subjective opinions.

- LPIPS and CDR achieve expected correlation coefficients, but show higher MSE and STD compared to SCD on both datasets, indicating instability in generating objective scores.
- MSSSIM in LAB does not surpass SCD in both error and correlation analysis, emphasizing SCD's advantage in using natural color statistics for accurate naturalness assessment.

Intuitive Display of Examples Fig. 5 shows three examples comparing objective metrics and subjective user ratings. Each row displays images of Recolorization A, Recolorization B, and a chart with metric scores and user ratings. Closer alignment in direction and variations between metric scores and user ratings indicates better reflection of subjective opinions and naturalness in the images. All scores are normalized and rounded to two decimal places for readability.

In the first row, Recolorization A is perceived by users as more natural than Recolorization B. SCD scores for both images align well with user ratings, showing a 0.5 difference that matches the 0.3 difference in user perceptions. PSNR scores in LAB also aligns with user opinions, but the 0.07 difference of PSNR in LAB suggests slight naturalness differences contrary to user perceptions (0.3 difference). In the second and third rows, users rate Recolorization A as less natural than Recolorization B. SCD scores show a minimal 0.08 difference, closely



(a) Spearman HECD (b) Kendall HECD (c) Spearman SECID (d) Kendall SECID

Fig. 6: Rank-based Correlation Analysis: Window size Effect under HECD ((a) and (b)) and SECID ((c) and (d)) Datasets in Section 4.3.

matching the human-scored difference, affirming SCD's accuracy in reflecting user opinions. PSNR in LAB performs similarly to SCD but shows a much larger 0.46 difference between images, indicating less precision in capturing naturalness variations. PSNR in RGB and SSIM in LAB align with user opinions but demonstrate inconsistent correlations across datasets, revealing challenges in evaluating structural changes and image uncertainties.

Table 1: Ablation study about Impact of Hue bin size in Section 4.3

HECD	6	10	18	24	36
Spearman	0.32	0.33	0.31	0.25	0.15
Kendall	0.25	0.27	0.23	0.20	0.12
SECID	6	10	18	24	36
Spearman	0.48	0.52	0.33	0.37	0.30
Kendall	0.43	0.48	0.32	0.33	0.27

Table 2: Ablation study about Impact of Saturation bin size in Section 4.3

HECD	6	10	15	30
Spearman	0.30	0.33	0.28	0.27
Kendall	0.24	0.27	0.22	0.22
SECID	6	10	15	30
Spearman	0.46	0.52	0.50	0.45
Kendall	0.43	0.48	0.45	0.40

4.3 Ablation Study

In this section, we will delve into an in-depth exploration of the impact of four key factors on the final performance during SCD computational process. The subsequent four sub-sections respectively address the influence of collected color distributions, the impact of window size, the function of probability density and the effect of change rates. Due to page limit, the corresponding error analysis results of the above four different ablation experiments and the detailed analysis of two more ablation experiments (involving the mean operation and color space selection) are provided in the supplementary material.

Color Distribution For collected color distributions, different settings for bin sizes will divide color space to distinct color bins. Thus, although the color information in images for each category in the ADE20K dataset is fixed, the color distributions with different bin size settings for the same pixel yield different scores, directly impacting the accuracy of the final image scores. Two ablation

Table 3: Ablation study under [1,1] and [3,3] (default) window sizes in Section 4.3

Dataset	Window Size	Spearman	Kendall
HECD	[3,3]	0.33	0.27
	[1,1]	0.27	0.2
SECID	[3,3]	0.52	0.48
	[1,1]	0.39	0.35

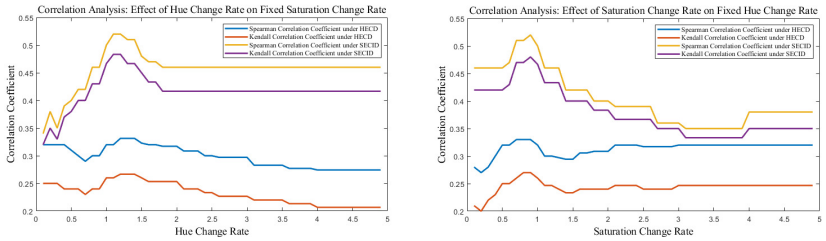
Table 4: Ablation study involving probability density in Section 4.3

Dataset	PD	Spearman	Kendall
HECD	enable	0.33	0.27
	disable	0.27	0.23
SECID	enable	0.52	0.48
	disable	-0.09	-0.03

experiments are conducted on the collected color distributions in this paper: in the first ablation experiment, the saturation bin size is fixed at 10, while the hue bin size is incrementally increased from 6 to 36 (see Table. 1); in the second ablation experiment, the hue bin size is fixed at 10, while the saturation bin size is incrementally increased from 6 to 30 (see Table. 2).

Table. 1 demonstrates the impact of hue bin size variations on the HECD and SECID datasets. Optimal performance are consistently achieved under all evaluation conditions when the hue bin size is set to 10. Although, in the last two rows of Table. 1, the SECID dataset obtained local optimal performance when the hue bin size is 24, its performance on Spearman Correlation and Kendall Correlation are still lower than that of the bin size set to 10, with percentages of 28.85% and 31.25%, respectively. Table. 2 illustrates the distinct effects of saturation bin size variations on the HECD and SECID datasets. Both tables exhibit a consistent trend of initially increasing and then decreasing. Furthermore, in all experiments, the optimal performance is attained when the saturation bin size is set to 10. Based on the above experimental analysis, this paper defaults to setting both the saturation bin size and hue bin size to 10.

Window Size Different window sizes determine the range of neighboring bins considered, aiming to reduce sensitivity to individual pixels and mitigate local noise or outliers' impact on pixel scores. Ablation experiments with varying window sizes (ranging from 3 to 9 for saturation and 3 to 13 for hue) are conducted to assess their impact on results. Fig. 6 shows correlation performances with different window sizes on the HECD and SECID datasets. Each subplot includes axes for hue window size, saturation window size, and correlation strength (Spearman or Kendall), demonstrating their relationships. Fig. 6 shows that increasing window sizes generally decrease consistency with subjective opinions. In Fig. 6 (c) and (d), a local peak occurs at saturation and hue window sizes of 7 for the SECID dataset, but performance remains below optimal levels. Across all conditions, the highest correlation with user opinions occurs with a window size of [3,3]. Further analysis with a [1,1] window size in Table. 3 confirms inferior performance compared to [3,3] window size, highlighting the importance of adjacent bins. Thus, this study adopts a default setting of [3,3] for both saturation and hue window sizes.



(a) Effect of Hue Change Rate (b) Effect of Saturation Change Rate

Fig. 7: Rank-based Correlation Analysis: Change Rate Effect under both Datasets in Section 4.3.

Probability Density This section validates the importance of probability density (PD) through substituting it with frequency. As detailed in Table. 4, results show that enabling PD consistently improves correlation with human subjective judgments compared to using frequency (disabling PD) across both HECD and SECID datasets. This highlights PD's role in normalizing data for fair comparisons between color bins.

Change Rate This paper conducted two change rate ablation experiments: one fixed the saturation change rate at 1 and adjusted the hue change rate from 0.1 to 4.9, while the other fixed the hue change rate at 1 and adjusted the saturation change rate similarly. Based on the Fig. 7, optimal performances across datasets are observed when saturation and hue change rates are set at [1:1.2] and [0.8:1], respectively. This indicates that slightly higher hue change rates contribute to superior performance for SCD. Based on these findings, the paper sets saturation and hue change rates at [1:1.2].

5 Conclusion

This paper proposes a novel objective metric SCD for evaluating colorized images quality based on statistical color information. This approach initiates from the HSV color space, which aligns closely with human perception of images, collecting two-dimensional color distributions for various objects. Subsequently, employing various normalization methods, including probability density and a locally weighted variant, the study conducts a systematic and reasonable scoring of the coloring images. Next steps may involve enhancing SCD by (1) integrating additional datasets with precise semantic segmentation information to improve accuracy in natural color distribution across object categories, (2) proposing extra mechanisms to measure color shifts, color bleeding, and biases in local regions to enhance overall performance and (3) extending SCD to High Dynamic Range imaging.

References

1. Cao, Y., Meng, X., Mok, P., Liu, X., Lee, T.Y., Li, P.: AnimeDiffusion: Anime Face Line Drawing Colorization via Diffusion Models. arXiv preprint arXiv:2303.11137 (2023)
2. Chen, S.Y., Zhang, J.Q., Gao, L., He, Y., Xia, S., Shi, M., Zhang, F.L.: Active Colorization for Cartoon Line Drawings. IEEE Transactions on Visualization and Computer Graphics **28**(2), 1198–1208 (2020)
3. Dou, Z., Wang, N., Li, B., Wang, Z., Li, H., Liu, B.: Dual Color Space Guided Sketch Colorization. IEEE Transactions on Image Processing **30**, 7292–7304 (2021)
4. Gurney, J.: Color and light: A guide for the realist painter, vol. 2. Andrews McMeel Publishing (2010)
5. Hasler, D., Suesstrunk, S.E.: Measuring Colorfulness in Natural Images. In: Human vision and electronic imaging VIII. vol. 5007, pp. 87–95. SPIE (2003)
6. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained By a Two Time-scale Update Rule Converge to a Local Nash Equilibrium. Advances in neural information processing systems **30** (2017)
7. Hristova, H., Le Meur, O., Cozot, R., Bouatouch, K.: Perceptual metric for color transfer methods. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 1237–1241. IEEE (2017)
8. Huang, Z., Zhao, N., Liao, J.: UniColor: A Unified Framework for Multi-modal Colorization with Transformer. ACM Transactions on Graphics (TOG) **41**(6), 1–16 (2022)
9. Iizuka, S., Simo-Serra, E., Ishikawa, H.: 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), 1–11 (2016)
10. Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., Kumar, S.: Rethinking FID: Towards a better evaluation metric for image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9307–9315 (2024)
11. Kim, E., Lee, S., Park, J., Choi, S., Seo, C., Choo, J.: Deep Edge-aware Interactive Colorization Against Color-bleeding Effects. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14667–14676 (2021)
12. Kim, E., Suk, H.J.: Hue extraction and tone match: Generating a theme color to enhance the emotional quality of an image. In: ACM SIGGRAPH 2015 Posters, pp. 1–1 (2015)
13. Larsson, G., Maire, M., Shakhnarovich, G.: Learning Representations for Automatic Colorization. In: Proceedings of the European Conference on Computer Vision. pp. 577–593. Springer (2016)
14. Li, Z., Geng, Z., Kang, Z., Chen, W., Yang, Y.: Eliminating Gradient Conflict in Reference-based Line-art Colorization. In: European Conference on Computer Vision. pp. 579–596. Springer (2022)
15. Lynch, D.K., Livingston, W.C.: Color and light in nature. Cambridge University P˙ress (2001)
16. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001. vol. 2, pp. 416–423. IEEE (2001)
17. Min, L., Li, Z., Jin, Z., Cui, Q.: Color edge preserving image colorization with a coupled natural vectorial total variation. Computer Vision and Image Understanding **196**, 102981 (2020)

18. Mullery, S., Whelan, P.F.: Human vs Objective Evaluation of Colourisation Performance. arXiv preprint arXiv:2204.05200 (2022)
19. Rodriguez-Pardo, C., Casas, D., Garces, E., Lopez-Moreno, J.: TexTile: A differentiable metric for texture tileability. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4439–4449 (2024)
20. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
21. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image recognition. arXiv preprint arXiv:1409.1556 (2014)
22. Smith, A.R.: Color Gamut Transform Pairs. ACM Siggraph Computer Graphics **12**(3), 12–19 (1978)
23. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for Semantic Segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7262–7272 (2021)
24. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going Deeper with Convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
25. Teng, X., Li, Z., Liu, Q., Pointer, M.R., Huang, Z., Sun, H.: Subjective Evaluation of Colourized Images with Different Colorization Models. Color Research & Application **46**(2), 319–331 (2021)
26. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
27. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale Structural Similarity for Image Quality Assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. vol. 2, pp. 1398–1402. IEEE (2003)
28. Weng, S., Zhang, P., Li, Y., Li, S., Shi, B., et al.: L-CAD: Language-based Colorization with Any-level Descriptions Using Diffusion Priors. Advances in Neural Information Processing Systems **36** (2024)
29. Zhang, R., Isola, P., Efros, A.A.: Colorful Image Colorization. In: Proceedings of the European Conference on Computer Vision. pp. 649–666. Springer (2016)
30. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features As a Perceptual Metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
31. Zhang, R., Zhu, J.Y., Isola, P., Geng, X., Lin, A.S., Yu, T., Efros, A.A.: Real-time User-guided Image Colorization with Learned Deep Priors. arXiv preprint arXiv:1705.02999 (2017)