Data Descriptor

SadaColorDataset (SCD): 9 Paper Colors × 4 Illumination Conditions for Robust Color Vision Evaluation

Basit Raza *, Sadaf Bibi , Sadia Bibi , Ali Nawaz

Posted Date: 9 March 2026

doi: 10.20944/preprints202603.0599.v1

Keywords: SadaColorDataset(SCD); color constancy; illumination shift; Robust Color Vision

 doi:10.20944/preprints202603.0599.v1

Data Descriptor

SadaColorDataset (SCD): 9 Paper Colors × 4 Illumination Conditions for Robust Color Vision Evaluation

Basit Raza [1,*], Sadaf Bibi [2], Sadia Bibi [3] and Ali Nawaz [4]

[1] Institute of Computer Science, Shah Abdul Latif University, Khairpur Mirs, Pakistan
[2] Department of Artificial Intelligence, Aror University of Art, Architecture, Design & Heritage, Sukkur, Pakistan
[3] Department of Computer Science, Sukkur IBA University, Pakistan
[4] Department of Electrical Engineering Technology, The Benazir Bhutto Shaheed University of Technology & Skill Development, Khairpur Mirs, Pakistan
* Correspondence: basitraza.computerscience@gmail.com

Abstract

SadaColorDataset (SCD) is a publicly available image dataset designed to support research on robust color recognition and illumination-related color variation in real mobile captures. The dataset contains 10,843 photographs of nine physical color papers (Black, Blue, Gray, Orange, Pink, Purple, Sky Blue, White, and Yellow) recorded under four everyday lighting conditions: Fluorescent, Indoor, Indoor Night, and Sunlight. All images were captured using an Infinix NOTE 40 smartphone camera (108 MP) with a simple, repeatable setup intended to reflect practical conditions rather than laboratory calibration. For each color–illumination setting, multiple images were collected to cover natural variability due to exposure, white balance, shadows, and reflections. During acquisition, the paper was placed on a ground surface and the phone was mounted on a tripod; the viewpoint was varied by moving the tripod to different positions and orientations. However, because explicit angle labels were not recorded or reliably recoverable from the released file structure/metadata, SCD does not provide calibrated or discrete "angle IDs," and it should be treated as a dataset with unlabeled viewpoint variation. Along with the images, we release machine-readable metadata and summary files that describe image counts across colors and illuminations and provide basic color statistics (e.g., RGB/CIELAB-derived measures) to facilitate reproducible analysis. SCD is distributed under a public license and is intended for benchmarking illumination robustness, dataset shift, and color stability in mobile vision pipelines.

Keywords: SadaColorDataset(SCD); color constancy; illumination shift; Robust Color Vision

Introduction

Color is one of the most useful visual cues for recognition, tracking, inspection, and quality control. However, the recorded color in an image is not only a property of the surface—it is also shaped by the illumination spectrum, camera sensor and ISP (white balance, tone mapping), and the viewing geometry. This is why the same "blue" paper can look noticeably different under fluorescent office light, indoor warm light, indoor night lighting, and direct/indirect sunlight. These variations create a practical gap between "clean" color examples and real deployments, where models must remain stable under illumination change and camera auto-processing. Color constancy and automatic white balance research exists to reduce this sensitivity, but progress is tightly coupled to the availability of datasets with clear capture conditions and reliable metadata [13,16,18,22].

Many popular color/illumination datasets were built for illuminant estimation (global or local) and white-balance evaluation, often using a color checker, calibrated rigs, or multiple cameras. They

 doi:10.20944/preprints202603.0599.v1

are invaluable for benchmarking, but they may not match the needs of downstream tasks that explicitly require repeated captures of the same nominal color across multiple everyday lighting conditions and viewpoints (e.g., robust color recognition, color-based retrieval, and illumination-aware augmentation). Recent analyses also show that dataset composition can bias methods (for example, toward scenes containing particular semantic cues), which increases the value of datasets that make their capture factors explicit and easy to audit [23]. In parallel, the community has increasingly emphasized open, well-documented dataset releases that follow reproducible "data descriptor" practices: clear scope, transparent collection protocol, structured metadata, technical validation, and reusable code [1–3].

To support research that focuses on color appearance variation under common real-world illuminations, we introduce SadaColorDataset (SCD): a public dataset of nine commercially available color papers captured under four illumination conditions (Fluorescent, Indoor, Indoor Night, and Sunlight). Images were captured using a mobile camera, and multiple viewpoints were collected by repositioning the phone mounted on a tripod around a fixed paper placed on a flat ground surface. Importantly, the viewpoints are not geometrically calibrated (i.e., no measured degrees), so they are treated as view IDs rather than physical angles. Alongside the images, SCD provides machine-readable metadata and summary CSVs, plus visualization outputs and basic color statistics to help users understand coverage, imbalance, and illumination-driven shifts.

In the remainder of this paper, we describe the collection protocol and dataset structure, provide the metadata schema, and report technical validation results that summarize coverage across color and illumination, show per-condition appearance changes, and quantify a simple "color stability" proxy based on CIELAB drift statistics. We also release the code used to generate the metadata tables and figures to enable fully reproducible dataset reporting, consistent with modern data descriptor expectations [1–3].

Dataset Contributions

- Everyday illumination categories, real capture conditions: SCD targets common lighting environments (fluorescent, indoor, indoor night, sunlight) where color shifts naturally due to illumination and mobile camera processing.
- Simple, repeatable subjects: Using nine color papers makes the visual content easy to interpret and reduces confusion from complex scene semantics, which helps when diagnosing failures under domain shift.
- Metadata + reproducible reporting assets: The release includes structured CSV metadata and automatically generated tables/figures that support transparent dataset reporting and fast reuse.
- Evaluation-friendly design for robustness studies: SCD supports clear testing scenarios such as training on one illumination and evaluating on another, making it suitable for studying illumination robustness and generalization.
- Open access and practical usability: The dataset and code are publicly available, making it straightforward for others to replicate reported statistics and build baseline pipelines.

Related Work

Color Constancy, White Balance, and Illumination Robustness

Computational color constancy aims to estimate and correct the scene illuminant so that object colors appear more consistent under changing light [18,22]. Earlier approaches relied on statistical assumptions (e.g., gray-world variants, gamut methods), while modern methods often use learning-based models that map image evidence to an illuminant estimate or directly learn to distinguish correct vs. incorrect white balance [18,22]. Deep learning has become a strong baseline for illuminant estimation and AWB, including patch-based CNNs and fully convolutional models [17], and methods that improve robustness or editing quality under practical camera pipelines [20]. Because the problem is under constrained, evaluation is highly dataset-dependent, and several works explicitly discuss

 doi:10.20944/preprints202603.0599.v1

biases and hidden shortcuts that can appear in commonly used AWB/color constancy benchmarks [23].

Benchmark Datasets for Illumination and Color Constancy

A large fraction of the literature evaluates on a small set of established datasets. The ColorChecker/Gehler family is widely used, with reprocessing efforts emphasizing the importance of correct ground truth and careful handling of camera processing artifacts [11,12]. The Gray Ball dataset provided large-scale imagery with a known reference object for ground truth estimation [14]. The NUS 8-camera dataset was designed to study cross-camera behavior and generalization across sensors [15]. More recent datasets push toward *camera invariance* and richer capture factors: Intel-TUT explicitly targets camera-invariant evaluation and includes lab scenes under multiple illuminations, plus field scenes and a mobile camera component [7]. INTEL-TAU further expands coverage and is frequently used for modern learning-based benchmarking [6]. Cube+ / Cube++ provide controlled illuminant estimation settings using a reference object designed for ground-truth extraction, and Cube++ scales this idea with a larger collection [8,9]. CC-NORD contributes camera-invariant benchmarking with a focus on diverse illuminants beyond the typical color-temperature curve, including a synthetic benchmark component [10].

Multi-Illumination and "In-the-Wild" Appearance Variation Datasets

Beyond single-illuminant assumptions, several datasets capture objects or surfaces under multiple illuminations to study relighting, inverse rendering, or illumination-dependent appearance. Murmann et al. introduced a dataset of indoor surfaces under varying illumination designed to be collected at scale in real environments [19]. Open Illumination provides a large multi-illumination benchmark for inverse rendering with many views and many light configurations [21]. These datasets motivate a key idea that SCD also targets in a simpler setting: systematic sampling of common lighting conditions can expose appearance shifts that are invisible when only one illumination is present.

Where SCD Fits

SCD is not meant to replace classic illuminant-estimation benchmarks; instead, it complements them by focusing on repeatable, label-clean color surfaces (nine known paper colors) under everyday illuminations that non-expert users frequently encounter. This design supports (i) robust color classification and retrieval under illumination shift, (ii) stress-testing color features and augmentation strategies, and (iii) fast baselines for illumination-aware modeling without requiring specialized calibration hardware. Because SCD's viewpoints are uncalibrated, we intentionally treat them as discrete view IDs and avoid claims that require measured angles. This makes the dataset honest about what it controls (nominal color, illumination category, repeated captures) and what it does not (exact geometry, absolute spectral calibration), aligning with best-practice dataset reporting and bias-aware evaluation guidance [2,3,23].

Background & Summary

Color looks "stable" to people, but in images it often behaves like a moving target. The same surface can appear warmer, cooler, darker, or more saturated depending on the light source (sunlight vs fluorescent), the time of day, and the camera's automatic processing (auto white balance and exposure). Because of this, color-based methods that work well in one setting can fail badly when the illumination changes. This is a common reason why color recognition, color-based retrieval, and simple threshold rules break when moving from indoor scenes to sunlight, or from daytime to indoor night lighting.

Many existing color-related datasets were built under controlled conditions or rely on calibration targets (for example, standard color charts) to estimate the illuminant. Those benchmarks

 doi:10.20944/preprints202603.0599.v1

are very useful for measuring illuminant estimation accuracy, but they do not always reflect the kind of variability that appears in everyday mobile captures—where lighting is mixed, shadows are present, the camera adapts automatically, and the same object is photographed multiple times with small viewpoint changes. For practical applications, it is often more helpful to have a dataset where the subject is simple and repeatable (a known set of color papers), while the environment remains realistic (common indoor and outdoor lighting). This setting makes it easier to study *how much color shifts in practice* and to evaluate whether a method stays stable under illumination-driven domain shift.

To address this gap, we present SadaColorDataset (SCD), a public dataset of nine physical color papers (Black, Blue, Gray, Orange, Pink, Purple, Sky Blue, White, and Yellow) captured under four everyday illumination conditions**:** Fluorescent, Indoor, Indoor Night, and Sunlight. All images were captured using an Infinix NOTE 40 mobile camera (108 MP). During collection, the paper was placed on a flat ground surface and the phone was mounted on a tripod. The viewpoint was varied by moving and reorienting the tripod across captures. However, these viewpoint changes were not saved as consistent, machine-readable "angle IDs," and therefore the released dataset should be treated as having unlabeled viewpoint variation rather than calibrated angles.

SCD is released with supporting files that help readers and users inspect the dataset quickly and reproduce key statistics. Along with the images, we provide machine-readable metadata tables and summary files (CSV) as well as high-quality figures (PNG) generated through a public notebook. These resources describe the dataset structure, quantify how samples are distributed across colors and illumination conditions, and provide simple color statistics (e.g., RGB- and CIELAB-derived measures) that can be used for technical validation, sanity checks, and baseline experiments.

Relation to Existing Datasets

SCD complements widely used color constancy and white-balance benchmarks rather than replacing them. Many established datasets are designed around illuminant estimation and may include calibration objects or more complex scene content, which is valuable for estimating ground-truth illumination and comparing specialized algorithms. In contrast, SCD focuses on a simpler subject class (color papers) and emphasizes repeated captures under everyday lighting categories, which is convenient for studying illumination-driven appearance change, building robust color features, and running domain shift evaluations in a controlled label space. SCD does not provide spectral calibration or calibrated viewing geometry; instead, it offers a practical middle ground: repeatable color targets captured under realistic lighting variability with reproducible metadata and reporting artifacts.

Typical Use-Cases Enabled by SCD

- Color constancy / white balance evaluation (as a stress test): Analyze whether a correction strategy reduces cross-illumination color drift.
- Illumination classification: Predict illumination category from images or derived statistics.
- Robust color recognition: Train color classifiers and evaluate under held-out illumination (domain shift).
- Dataset shift analysis: Quantify how feature distributions change across lighting conditions and how that impacts model performance.

Methods

Materials

Color papers: SadaColorDataset (SCD) was created using nine physical color papers purchased from the local market. The color categories in the dataset are: Black, Blue, Gray, Orange, Pink, Purple, Sky Blue, White, and Yellow. The papers were used as the main target object so that the label ("color")

 doi:10.20944/preprints202603.0599.v1

is clear and easy to interpret. The exact brand and paper finish can vary by availability; in our capture setup the papers behaved like typical colored sheets used for crafts and printing. If a paper's surface finish is known (for example, matte or glossy), it can be recorded as optional metadata; otherwise, the dataset treats the paper type as a practical, real-world material.

Camera device: All images were captured using a mobile phone camera: Infinix NOTE 40 with a 108-megapixel rear camera. The goal was to reflect what users commonly experience with a modern smartphone pipeline, including automatic exposure and automatic white balance.

Illumination environments: Images were collected under four everyday lighting conditions that commonly produce visible color shifts:

- **Fluorescent:** indoor lighting dominated by fluorescent tubes (typical office/classroom light).
- **Indoor:** indoor ambient lighting (room light) during normal conditions.
- **Indoor Night:** indoor lighting at night, often warmer and dimmer with stronger shadows and noise.
- **Sunlight:** outdoor natural light, including bright conditions where shadows and reflections may appear.

Camera settings and capture setup: The phone was mounted on a tripod to reduce motion blur and make repeated captures easier. The paper was placed on a flat ground surface. The camera typically operated using the phone's default camera settings, which means auto exposure and auto white balance were active (as in normal use). The flash was kept off. The approximate distance between the camera and the paper was kept reasonably consistent within each session (close enough to clearly fill most of the frame with the paper), but small differences naturally occurred due to viewpoint changes and framing.

Setting	Value / Description
Capture device	Infinix NOTE 40 (rear camera)
Sensor / image resolution	108 MP (native camera capability); images saved at the phone's default output resolution for the selected mode
Capture mode	Standard phone camera mode (default pipeline)
Flash	Off
White balance (AWB)	Auto white balance enabled (default)
Exposure	Auto exposure enabled (default)
Stabilization	Phone mounted on a tripod to reduce motion blur
Target placement	Color paper placed flat on a ground surface
Camera–target distance	Approximately kept consistent within each session; small variations occurred due to viewpoint changes (report as an approximate range if measured)
Viewpoint variation	Introduced by moving/reorienting the tripod; no calibrated angles or angle IDs released
Illumination conditions	Fluorescent, Indoor, Indoor Night, Sunlight
Capture environment	Real-world settings (not laboratory-calibrated); natural variation in shadows and reflections may occur
File format	Images stored as standard compressed image files (e.g., JPG/PNG as released)
Dataset labels	Color label (9 classes) + illumination label (4 classes); viewpoint variation is unlabeled

Capture Procedure

General procedure: For each of the nine colors, we captured multiple images under each illumination condition. The main idea was to keep the target (paper) simple and stable, and let the lighting and camera pipeline create realistic variation. The paper was placed on the ground and

 doi:10.20944/preprints202603.0599.v1

oriented so that it was clearly visible. The phone, fixed on a tripod, was positioned to capture the paper with sufficient resolution and minimal shake.

Per-illumination capture notes: Because lighting conditions differ in real environments, the dataset was collected in separate sessions for each illumination:

- **Fluorescent:** captured indoors under fluorescent light.
- **Indoor:** captured indoors under standard room lighting.
- **Indoor Night:** captured indoors at night under artificial light, where exposure and noise effects are stronger.
- **Sunlight:** captured in outdoor daylight, where brightness, shadows, and reflections can change across time.

For each session, we tried to keep the background simple and the paper clearly visible; however, small changes in shadows and surrounding surfaces are natural and were not strictly controlled, as the dataset aims to reflect practical conditions.

Viewpoint variation (angles): During capture, the viewpoint was varied by moving and reorienting the tripod-mounted phone around the paper (for example, slightly changing left/right position and tilt). This was done to introduce realistic changes in viewing direction and framing. However, explicit angle labels were not stored in a consistent machine-readable form, and angle values in degrees were not calibrated. For this reason, the released dataset does not provide official "Angle ID 1–7" labels, and viewpoint changes should be considered unlabeled variation.

Number of images per setting: The dataset contains multiple images for each color–illumination combination. The counts are not perfectly balanced across all combinations because real capture sessions naturally produce different numbers of usable images (for example, due to time, lighting stability, and quality filtering). We therefore report the exact counts in the accompanying summary tables (CSV) and include heatmaps/plots that show the distribution of samples across colors and illuminations.

Quality Control

Image screening: After capture, images were checked for basic usability. Images with severe issues—such as strong motion blur, heavy out-of-focus regions, or extreme under/over-exposure that hides the paper's appearance—can be excluded. In practice, the dataset prioritizes real-world variability, but it avoids images where the color paper is not meaningfully visible.

Duplicate and near-duplicate handling: Smartphone captures may include near-duplicates, especially when burst-like capturing is used. To reduce unnecessary repetition, a simple duplicate check can be applied. A practical approach is to compute a hash-based signature (or perceptual hash) for each image and flag cases where two files are almost identical. Whether duplicates are removed or simply reported, the dataset release includes metadata generation code that makes it easy to audit repetition and build balanced subsets for fair evaluation.

File naming and directory structure: Images are organized using a directory structure that reflects the dataset labels. At minimum, the folder hierarchy should clearly represent color and illumination (and optionally angle/view ID if it is added later). Consistent naming ensures that metadata can be extracted automatically and that users can reproduce summary tables and figures directly from the raw image folders. Along with the dataset, we provide code that crawls the directory structure, generates per-image metadata, and exports paper-ready CSVs and figures.

Data Records

This section describes the files that are released with SadaColorDataset (SCD) and how they are organized. The dataset is hosted publicly on Kaggle and can be accessed through the dataset page. The released package includes the original images and supporting metadata files (CSV/JSON) that summarize the dataset and make it easier to reproduce the figures and statistics reported in this paper.

 doi:10.20944/preprints202603.0599.v1

Folder Structure and Naming Conventions:

SCD is organized using a simple directory structure so that the main labels can be understood directly from the folder path. At minimum, the folder names represent the color label and the illumination condition. Viewpoint variation is present in the images, but explicit angle IDs are not provided in the released metadata; therefore, the structure does not require an angle folder.

A typical layout follows the pattern:

SCD/
Black/
Fluorescent/
*.jpg (or *.png)
Indoor/
*.jpg (or *.png)
Indoor_Night/
*.jpg (or *.png)
Sunlight/
*.jpg (or *.png)
Blue/
Fluorescent/
Indoor/
Indoor_Night/
Sunlight/
...
Yellow/
Fluorescent/
Indoor/
Indoor_Night/
Sunlight/

Naming Conventions

- Color label is taken from the top-level folder name (e.g., Black, Blue, Gray, Orange, Pink, Purple, Sky Blue, White, and Yellow).
- Illumination label is taken from the second-level folder name (e.g., Fluorescent, Indoor, Indoor_Night, and Sunlight).
- The file name itself may be any camera-generated name (e.g., IMG_001.jpg), because the label is derived from the folder path.
- If angle labels are added in a future version, they would be introduced as an additional folder level (e.g., `Angle_1` … `Angle_7`) or as a mapping table; however, the current release treats viewpoint variation as unlabeled.

File Formats, Color Space, and License

Image files: The dataset contains standard image files saved in common formats (e.g., JPG and/or PNG, depending on the released version). Images are stored in the typical smartphone output color space (sRGB). Resolution varies by capture and camera processing mode; we recommend reporting the minimum, maximum, and mean resolution in the dataset statistics table (or in Technical Validation).

License and access.

SCD is released publicly under a **public license** through Kaggle. The dataset can be accessed at:

- **Kaggle dataset:** https://www.kaggle.com/datasets/basitaliharejo/sadacolordataset-9colors-4illums-final

 doi:10.20944/preprints202603.0599.v1

We also provide the exact Kaggle notebook used to generate the metadata tables and paper-ready figures:

- **Kaggle code notebook:** https://www.kaggle.com/code/basitaliharejo/sadacolordatasetscd-paper/edit

Metadata Tables and Supporting Files

In addition to the images, SCD is released with machine-readable metadata that supports transparent reporting and reproducible dataset inspection. These files are generated by the accompanying Kaggle notebook and are intended to help users quickly understand dataset coverage and to reproduce the figures shown in this paper.

Released Tables and Files

- metadata_images.csv — per-image metadata and basic color statistics.
- metadata_images_with_drift.csv — adds a color variability proxy (lab_drift).
- summary_counts_color_x_illum.csv — counts for each color–illumination combination.
- group_color_stats.csv — aggregated statistics per (color, illumination) and optional groupings.
- exif_dump.json — EXIF fields extracted when available (if present in the images).

Notes on Labels

- color_label and illumination are derived from the folder path.
- angle_id may appear in some metadata outputs, but in the current release it is not reliably available and should be treated as **missing** (i.e., the dataset does not publish angle labels).

Metadata Schema (Main Columns)

Below is a compact schema for the most important columns used in metadata_images.csv and metadata_images_with_drift.csv. Columns may be extended in future releases, but these form the core metadata needed to reproduce our summary statistics and figures.

Column name	Type	Description	Example
rel_path	string	Relative path of the image inside the dataset	Blue/Indoor/IMG_0123.jpg
filename	string	File name only	IMG_0123.jpg
filesize_bytes	integer	Size of the image file in bytes	2543891
width	integer	Image width in pixels	3000
height	integer	Image height in pixels	4000
color_label	string	Color class label (from folder name)	Blue
illumination	string	Illumination condition label (from folder name)	Indoor Night
mean_r, mean_g, mean_b	float	Mean RGB values (normalized 0–1) computed from the image	0.12, 0.18, 0.62
std_r, std_g, std_b	float	Standard deviation of RGB values (normalized 0–1)	0.04, 0.05, 0.07
mean_h, mean_s, mean_v	float	Mean HSV values (normalized)	0.60, 0.75, 0.62
mean_L, mean_a, mean_b_lab	float	Mean CIELAB values (L*, a*, b*)	42.1, 12.0, -38.5
lab_drift	float	Distance in CIELAB space from the median of its (color, illumination) group;	5.63

 doi:10.20944/preprints202603.0599.v1

		used as a stability proxy (only in metadata_images_with_drift.csv)
error	string (optional)	Read error message if an image could not be processed *(empty)*

Definition of lab_drift: For each image, we compute a simple drift value in CIELAB space: the Euclidean distance between the image's mean Lab vector and the median Lab vector of its corresponding (color, illumination) group. This produces an interpretable proxy for how much the observed color varies under a given illumination condition.

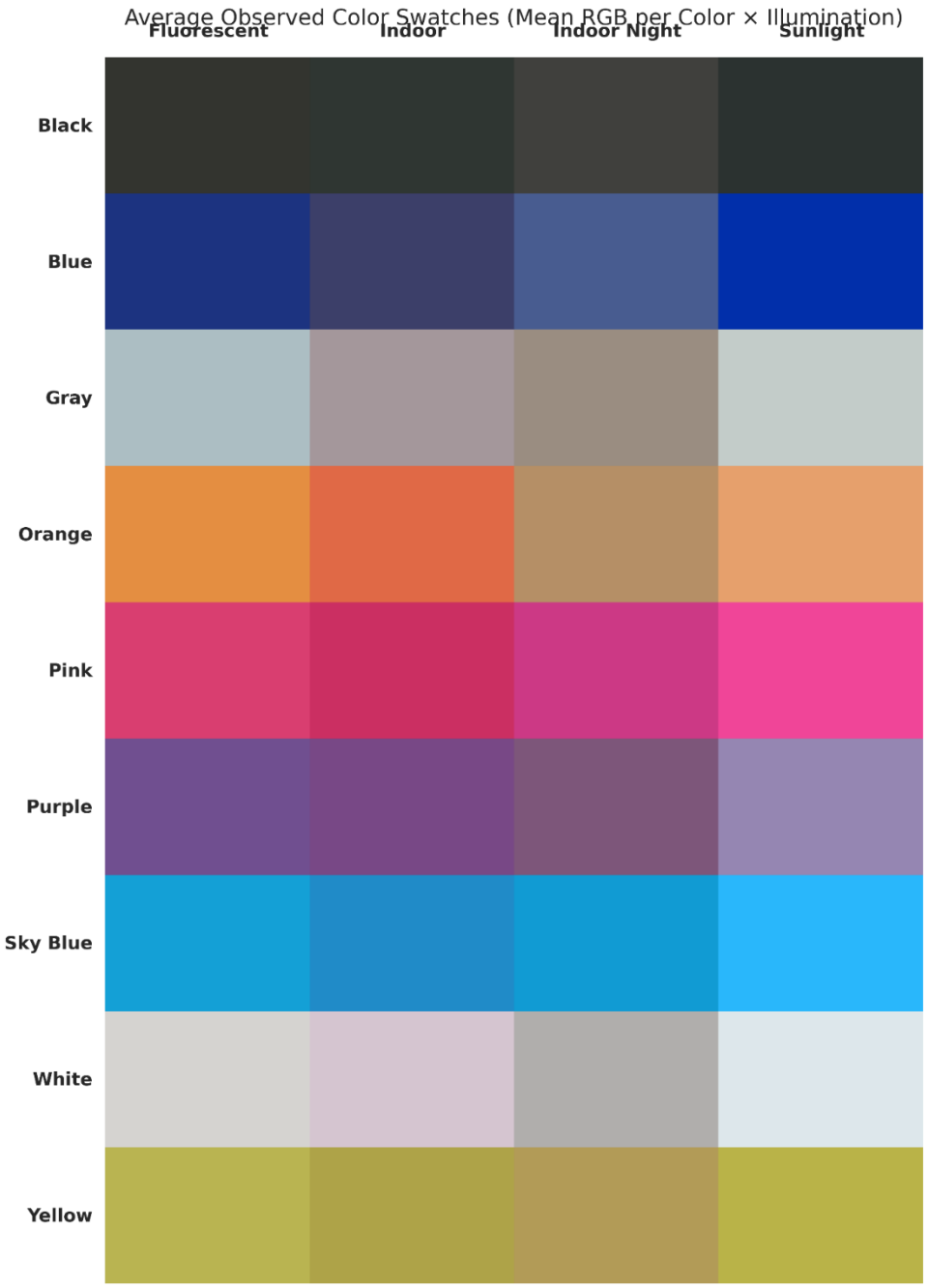

Figure 1. Average observed color appearance across illuminations. Grid of mean RGB swatches computed per **color × illumination**, summarizing how the same paper color shifts under Fluorescent, Indoor, Indoor Night, and Sunlight conditions. Swatches are intended as a descriptive dataset summary rather than a colorimetric ground truth.

 doi:10.20944/preprints202603.0599.v1

Technical Validation

This section checks whether the dataset is complete, internally consistent, and useful for evaluation under illumination change. We focus on (i) coverage across labels, (ii) basic balance properties, and (iii) whether illumination produces measurable and meaningful color variation.

Coverage and Balance

Overall size and completeness: The released metadata contains 10,843 images distributed across 9 color classes and 4 illumination conditions, giving 36 color–illumination cells. The Color × Illumination heatmap shows that all expected combinations are present (no missing cells), which is important for controlled cross-condition evaluation.

Imbalance (expected in real capture, but must be reported): The dataset is not perfectly balanced across cells. The smallest cell is White × Sunlight (106 images), while the largest cell is Orange × Indoor Night (1237 images). At the class level, Orange (2323) and Pink (1995) contribute the most images, while Blue (584) contributes the fewest. Across illuminations, totals are more even: Indoor (3050), Indoor Night (2980), Fluorescent (2560), and Sunlight (2253).

Recommended balanced subset protocol (for fair benchmarks): To support fair comparisons, we recommend a simple, reproducible balanced subset:

- **SCD-Balanced-106**: sample 106 images per (color, illumination) cell (the minimum cell size).
- This produces 106 × 36 = 3816 images, perfectly balanced across the full label grid.
- Sampling should be random but seeded (fixed seed in code) and ideally stored as an index file (e.g., balanced_index.csv) so that results are comparable across papers.

For training larger models, researchers can also use the full dataset with stratified sampling or class weights, but should always report whether balancing was applied.

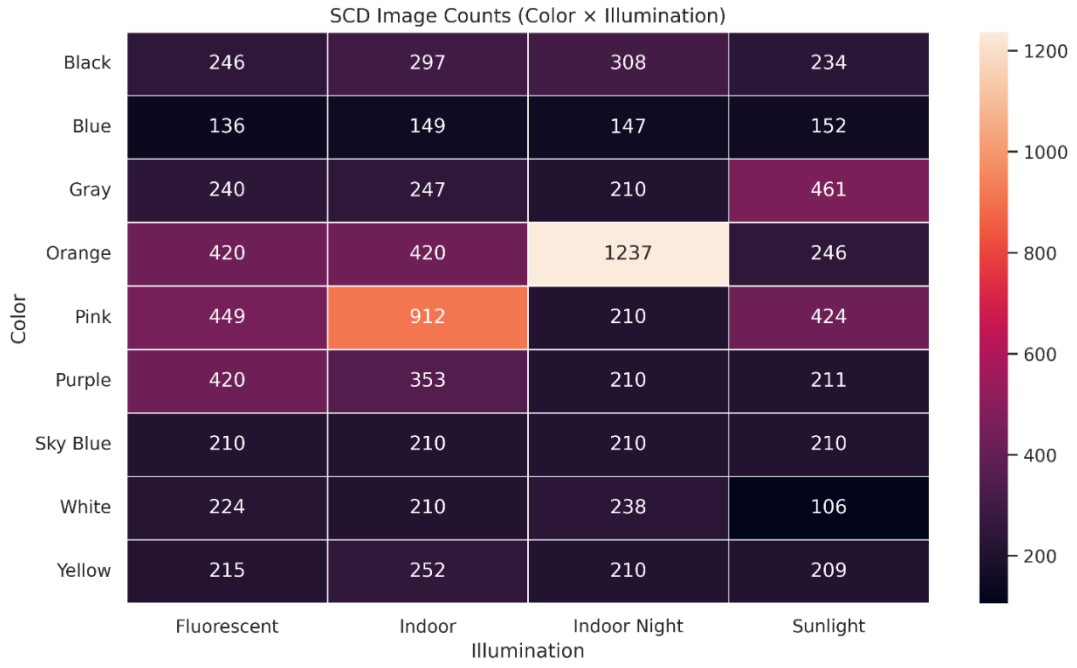

Figure 2. Dataset coverage across classes and conditions. Heatmap of image counts for each **color × illumination** combination in SadaColorDataset (SCD). The plot highlights coverage as well as class/condition imbalance that should be considered when building balanced training subsets.

 doi:10.20944/preprints202603.0599.v1

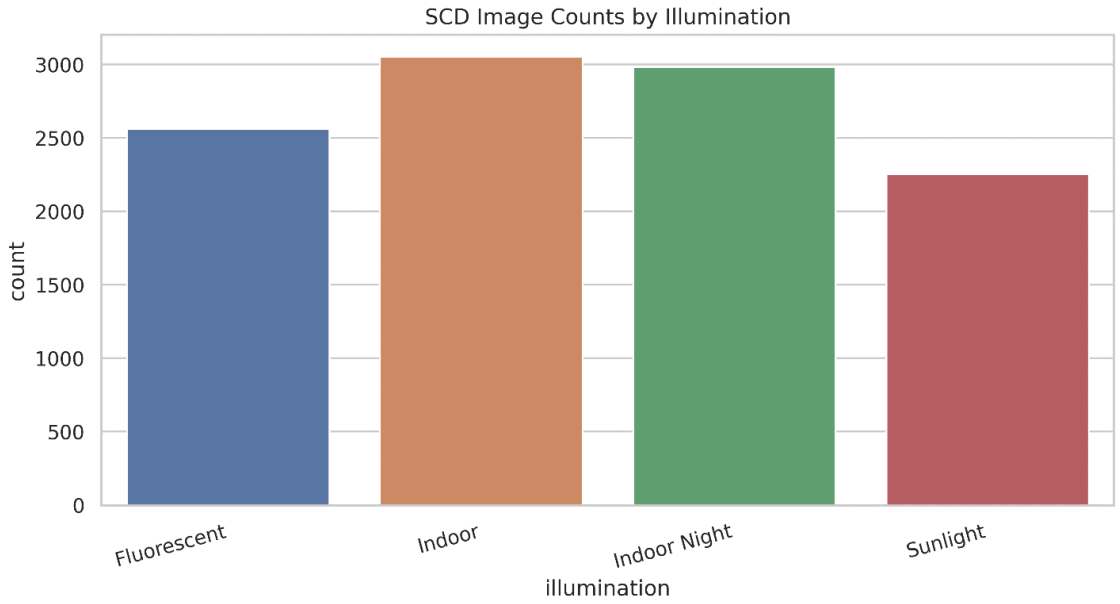

Figure 3. Overall distribution by illumination. Total number of images captured under each illumination condition. Differences reflect real capture constraints and can be handled using stratified sampling or a balanced subset protocol.

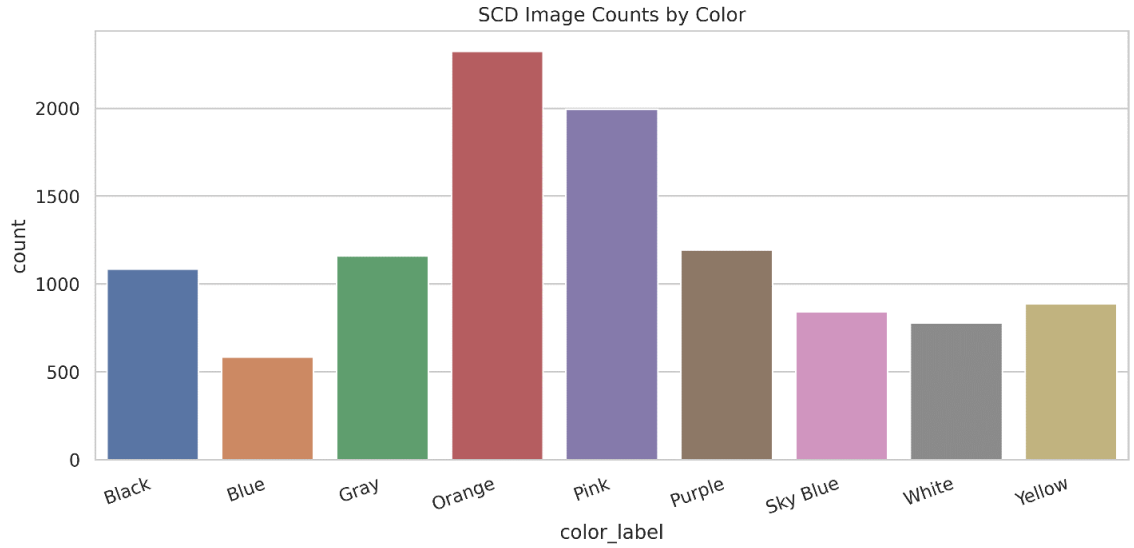

Figure 4. Overall distribution by color class. Total number of images per color category in SCD. The distribution shows which colors have more/less coverage and supports reporting of a balanced evaluation subset.

Illumination-Induced Variation (Color Stability Proxy)

To quantify how much the observed color varies under each illumination, we use a simple stability proxy called Lab drift (lab_drift). For each image, Lab drift is computed as the Euclidean distance between the image's mean CIELAB values and the median Lab of its corresponding (color, illumination) group. This gives an intuitive "how far from typical" measure within each condition.

What the violin plot shows: Lab drift differs clearly by illumination:

- Indoor Night shows the highest typical drift (median ≈ 5.63, n=2980), indicating stronger variability under low-light conditions.
- Fluorescent and Sunlight have similar medians (≈ 4.17 and ≈ 4.07), but Sunlight has the largest tail (maximum drift ≈ 32.76), consistent with harsh shadows, highlights, and changing daylight.

 doi:10.20944/preprints202603.0599.v1

- Indoor has the lowest median (≈ 3.29, n=3050), suggesting relatively more stable appearance in that setting.

These trends match practical expectations: indoor night captures often suffer from higher sensor noise and stronger automatic adjustments, while sunlight can produce extreme cases due to strong contrast and reflections.

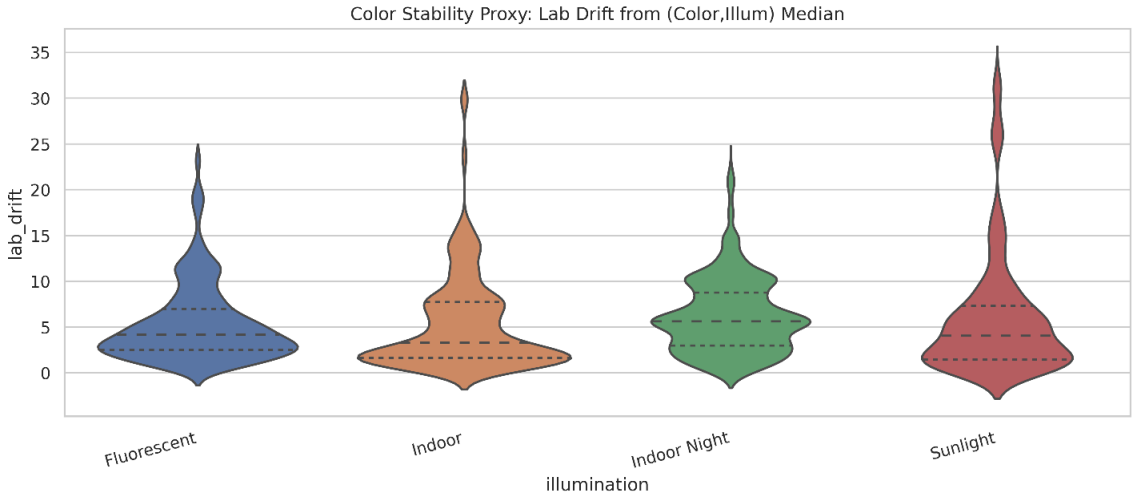

Figure 5. Illumination-driven variability using a Lab drift proxy. Violin plots of per-image **Lab drift** values grouped by illumination, where drift is computed relative to the median appearance within each **(color, illumination)** group. Wider tails indicate larger within-condition appearance variation, consistent with challenging lighting and smartphone auto-adjustments.

Color Separability (Optional But Strong)

A useful sanity check is whether the nine color classes remain reasonably separable in a perceptual color space such as CIELAB. A simple and informative figure is:

- A 2D scatter plot of a* vs b* for all images, colored by the ground-truth color label, and faceted by illumination (four small panels).

This plot typically shows that colors form clusters, while the cluster positions shift across illuminations due to white balance and exposure. Including this figure strengthens the dataset paper because it visually confirms that the labels correspond to consistent regions in color space, while also showing the expected cross-illumination drift.

Inter-Illumination Shift per Color (Practical and Publishable)

Beyond per-illumination variability, it is important to show *which colors are most sensitive* to illumination change. We summarize this by computing the mean Lab vector for each (color, illumination) group, then measuring the pairwise Lab distances between the four illuminations for each color (six pairs per color). The mean of these pairwise distances provides a compact "illumination sensitivity" score.

Observed pattern: Illumination sensitivity is not uniform across colors:

- Blue shows the strongest shift (mean pairwise distance ≈ 31.53, max ≈ 50.87).
- Orange is also highly sensitive (≈ 25.61, max ≈ 37.78).
- Black is the most stable (≈ 4.74, max ≈ 7.57).
- Other colors fall in between (e.g., Purple/Gray moderate; Sky Blue and Yellow lower than Blue/Orange).

 doi:10.20944/preprints202603.0599.v1

This is a useful result for users: it suggests that "hard colors" (like Blue and Orange) are good stress tests for illumination robustness, while Black is less sensitive and may be easier to classify across conditions.

A simple bar chart (one bar per color) is recommended for the paper, using the mean (or max) pairwise Lab distance as the value.

Angle-Based Validation (Only If Angles Are Labeled)

The current dataset release contains viewpoint variation introduced by moving and reorienting the tripod, but explicit angle IDs were not recorded or reliably recoverable, and the extracted metadata indicates that angle labels are missing. For this reason:

- We do not include angle coverage plots in the main validation results.
- We do not claim calibrated angles or official "Angle ID 1–7" in this version.

If a future version adds consistent angle_id labels (via folder naming, filenames, or an angles.csv mapping), then two additional validations should be included:

1. Color × Angle coverage heatmap (checks completeness across viewpoints)
2. Mean L* vs Angle ID line plot (shows viewpoint-dependent lightness changes, optionally separated by illumination)

These would extend SCD from "illumination robustness" into "illumination + viewpoint robustness" in a fully supported way.

Usage Notes

This section gives practical guidance on how to use SadaColorDataset (SCD) in a consistent and fair way. Because SCD is captured with a smartphone under everyday lighting, users should treat it as a dataset that reflects real camera behavior (auto white balance, auto exposure, shadows, and small viewpoint changes). The notes below help reduce avoidable noise and make results comparable across studies.

Recommended Preprocessing

Crop a region of interest (ROI) around the paper: Many images contain background pixels (ground surface, shadows, surrounding objects). If you compute color statistics from the full image, the background can bias the measured color and inflate variability. For most experiments, we recommend either:

- A center-crop that keeps the paper mostly in view, or
- A simple paper ROI (manual box, or automatic segmentation using thresholding/contours if the paper edges are clear).

A consistent ROI policy should be documented in the experiment description (crop size, position, and whether the same crop is applied to all images).

Resize consistently: For learning-based baselines, resize images to a standard resolution (e.g., 224×224 or 256×256). Keep the resizing method consistent (bilinear is typical). If you use an ROI first, crop then resize.

Keep the color pipeline simple and explicit: Images are stored in standard sRGB. When converting to other spaces (HSV, Lab), apply the same conversion method across all experiments. Avoid mixing different libraries without noting it, because small differences can slightly change numeric values.

Recommended Evaluation Protocols

SCD is most useful when experiments explicitly test how performance changes across illumination conditions. We recommend reporting at least one of the following protocols.

Protocol P1: Held-out illumination (domain shift): This is the main "robustness" protocol.

 doi:10.20944/preprints202603.0599.v1

- Choose one illumination as the test domain (e.g., Sunlight).
- Train the model using the other three illuminations (Fluorescent + Indoor + Indoor Night).
- Evaluate only on the held-out illumination.
- Repeat this four times (each illumination becomes the test set once).

This protocol answers a clear question: *If a model is trained without seeing a lighting condition, does it still work when that lighting appears in deployment?*

Protocol P2: Balanced subset across color × illumination: Because SCD is not perfectly balanced, it is easy for a model to gain accuracy by learning the majority patterns. To avoid that, we recommend building a balanced subset:

- Let k be the minimum sample count among all (color, illumination) cells.
- Randomly sample k images from each cell using a fixed random seed.
- Train and evaluate using only this balanced subset (or use it as a standardized benchmark split).

This protocol improves fairness and makes comparisons across methods more reliable. It also helps when computing macro-averaged metrics.

Protocol P3 (only if angle labels exist in a future version): Held-out angles: The current release includes viewpoint variation but does not provide official angle IDs. If a later release adds angle_id labels, then a useful geometry robustness test is:

- Train on angles 1–5, test on angles 6–7, while keeping illumination coverage consistent.
- Optionally repeat with different train/test angle partitions.

This protocol measures whether a method generalizes to new viewpoints, not just new illuminations.

Baseline Tasks Supported by SCD

Task 1: Illumination classification: Goal: predict the illumination label (Fluorescent / Indoor / Indoor Night / Sunlight). Why it matters: illumination classification can be used as a pre-step for illumination-aware correction or routing (choose a specialized model per lighting condition).
Recommended inputs: either the full image (CNN) or simple features such as mean Lab/HSV plus texture statistics.

Task 2: Color classification under domain shift: Goal: predict the color label (9 classes) while testing on a held-out illumination (Protocol P1). Why it matters: this is the most direct "robust color recognition" benchmark and exposes real failure modes caused by lighting change.
Recommended reporting: macro-F1, per-class accuracy, and confusion matrices for each test illumination.

Task 3: Color constancy calibration (future extension): SCD currently provides practical color samples but does not include a physical reference target (like a gray card) for ground-truth illuminant estimation. If a future version adds a gray reference (or a standard color checker), SCD could also support stronger color constancy tasks:

- Estimate illumination chromaticity,
- Apply correction,
- Quantify improvement in cross-illumination stability (e.g., reduced Lab drift).

Reporting Tips (To Make Your Results Reproducible)

- Always state whether you used full images or ROI crops.
- Report whether you used balanced sampling or class weights.
- Fix random seeds and, if possible, release split files (CSV/JSON) so others can reproduce your exact protocol.
- For Protocol P1, report results for each held-out illumination separately, not only an overall average.

 doi:10.20944/preprints202603.0599.v1

Limitations

This dataset was captured using a smartphone camera with the default imaging pipeline, which means auto white balance and auto exposure were active. This makes the images realistic and closer to what users see in practice, but it also reduces experimental control because the camera may adjust colors differently from one shot to the next, especially under difficult lighting such as indoor night or strong sunlight. The dataset is also not perfectly balanced across all color–illumination combinations; some colors and conditions contain many more images than others, which can bias simple training setups unless users apply stratified sampling, class weighting, or a balanced subset. Although the images include viewpoint variation because the tripod-mounted phone was moved and reoriented during capture, the current release does not provide consistent angle labels or calibrated angle measurements, so it should not be used for experiments that require precise geometry or known viewing angles. Finally, as the color papers were photographed in real environments, some images may contain background pixels, shadows, and reflections; when global color statistics are computed from the full image, this background variation can influence the measured values. For this reason, users may obtain more stable measurements by cropping a consistent region of interest around the paper or applying a simple paper segmentation step before extracting color features.

Author Contributions: Basit Raza (basitraza.computerscience@gmail.com): Conceptualization, dataset design, data acquisition, metadata generation, analysis, and manuscript drafting. Sadaf Bibi (f24ari184@aror.edu.pk, sadaf.bibi.dev@gmail.com): Data curation support, verification of labels and folder organization, and review/editing of the manuscript. Sadia Bibi (sadiabibi.bcsf18@iba-suk.edu.pk, sadiashah059@gmail.com): Literature support, validation of reporting content, and review/editing of the manuscript. Ali Nawaz (engralitechno@gmail.com): Technical support for data processing and reproducibility, code review, and manuscript review/editing.

Institutional Review Board Statement: This dataset contains photographs of colored paper sheets captured under different lighting conditions. It does not include human participants, personal information, faces, voices, or any other identifiable human data. No sensitive content is present, and the images were collected for research and educational purposes only.

Acknowledgments: The authors thank their respective institutions and colleagues for informal support and feedback during dataset preparation, organization, and documentation. We also acknowledge the Kaggle platform for hosting the dataset and code release.

Conflicts of Interest: The authors declare that there are no competing interests related to the creation or release of this dataset. The authors declare no funding from any source.

References

1. M. D. Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship,” *Scientific Data*, vol. 3, 2016.
2. *Scientific Data* (Nature Portfolio), “Data Descriptor (article type) / author guidance,” Nature Portfolio, accessed 2026.
3. FORCE11, “Joint Declaration of Data Citation Principles,” 2014.
4. A. Gh. Akbarinia and C. A. Parraga, “Colour constancy: Biologically-inspired contrast-variant pooling mechanism,” *arXiv preprint* arXiv:1711.10968, 2017.
5. S. Banić and S. Lončarić, “Unsupervised Learning for Color Constancy,” *arXiv preprint* arXiv:1712.00436, 2017.
6. A. Laakom, V. Raitoharju, A. Iosifidis, and M. Gabbouj, “INTEL-TAU: A Color Constancy Dataset,” *arXiv preprint* arXiv:1910.10404, 2019.
7. C. Aytekin, J. Nikkanen, M. Gabbouj, and M. A. Ghanem, “A Dataset for Camera-Invariant Color Constancy Research,” *arXiv preprint* arXiv:1703.09778, 2017.
8. S. Banić and S. Lončarić, “Cube+ dataset for illuminant estimation (SpyderCube-based),” 2018.

 doi:10.20944/preprints202603.0599.v1

9. E. Ershov, D. Gusak, S. Banić, and S. Lončarić, "The Cube++ Illumination Estimation Dataset," *arXiv preprint* arXiv:2011.10028, 2020.
10. O. Ulucan et al., "CC-NORD: A camera-invariant global color constancy dataset," in *Proc. EUSIPCO*, 2023.
11. E. Shi, "Re-processing of Gehler's ColorChecker dataset (notes and corrected data)," Simon Fraser University resource page, accessed 2026.
12. G. Hemrit, R. Luque-Baena, and A. P. de la Blanca, "Rehabilitating the ColorChecker dataset for illuminant estimation," 2018.
13. J. T. Barron, "Convolutional Color Constancy," in *Proc. ICCV*, 2015.
14. F. Ciurea and B. Funt, "A large image database for color constancy research," 2003.
15. D. Cheng, B.V.K.V. Kumar, and M. S. Brown, "Illuminant Estimation for Color Constancy: Why Spatial-Domain Methods Work and the Role of the Color Distribution," *IEEE Trans. Image Process.*, 2014.
16. Y.-C. Lo et al., "CLCC: Contrastive Learning for Color Constancy," in *Proc. CVPR*, 2021.
17. S. Bianco, C. Cusano, and R. Schettini, "Color Constancy Using CNNs," in *Proc. CVPR Workshops*, 2015.
18. G. Sharma, W. Wu, and E. N. Dalal, "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," *Color Research & Application*, vol. 30, no. 1, pp. 21–30, 2005.
19. L. Murmann, M. Gharbi, M. Aittala, and F. Durand, "A Dataset of Multi-Illumination Images in the Wild," in *Proc. ICCV*, 2019.
20. M. Afifi and M. S. Brown, "Deep White-Balance Editing," in *Proc. CVPR*, 2020.
21. I. Liu et al., "OpenIllumination: A Multi-Illumination Dataset for Inverse Rendering Evaluation on Real Objects," in *NeurIPS (Datasets and Benchmarks Track)*, 2023.
22. A. Gijsenij, T. Gevers, and J. van de Weijer, "Computational Color Constancy: Survey and Experiments," 2011.
23. M. Buzzelli, S. Zini, S. Bianco, G. Ciocca, R. Schettini, and M. K. Tchobanou, "Analysis of biases in automatic white balance datasets and methods," *Color Research & Application*, vol. 48, no. 1, pp. 40–62, 2023.
24. X. Xing et al., "Point Cloud Color Constancy," in *Proc. CVPR*, 2022.
25. International Commission on Illumination (CIE), "Colorimetry — Part 4: CIE 1976 L*a*b* Colour space," CIE publication, 1976 (and later standard editions).
26. IEC, "IEC 61966-2-1: Multimedia systems and equipment — Colour measurement and management — Part 2-1: Default RGB colour space — sRGB," 1999.