OPEN ACCESS

Kashf Journal of Multidisciplinary Research
Vol: 03 - Issue 04 (2026)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

COLOR RECOGNITION UNDER ILLUMINATION SHIFT: A CNN STUDY WITH LOIO EVALUATION AND RELIABILITY ANALYSIS ON SADACOLORDATASET (SCD)

Paras Mangi[1], Sadaf Bibi[2], Sadia Bibi[3]

**[1]*Institute of Computer Science, Shah Abdul Latif University, Khairpur Mirs, Sindh, Pakistan.*
[2]Department of Artificial Intelligence, Aror University of Art, Architecture, Design & Heritage, Sukkur, Sindh, Pakistan.
[3]Department of Software Engineering, Comera LLC, Abu Dhabi, UAE.

***Corresponding Author:** (72scholar.paras110@gmail.com)*

***DOI:**(https://doi.org/10.71146/kjmr898)*

Article Info

Abstract

Color recognition looks easy on paper, yet it often breaks in practice because illumination spectra, auto white balance, exposure, and in-camera processing shift the observed color distributions. To study this failure mode in a controlled but realistic setting, we use SadaColorDataset (SCD), a color-paper dataset with 9 classes (Black, Blue, Gray, Orange, Pink, Purple, Sky Blue, White, Yellow) captured under four illuminations (Fluorescent_Light, Indoor, Indoor_Night, Sunlight), totaling 10,843 images. We benchmark lightweight CNNs (MobileNet/ResNet family) with two input variants: RAW images and a simple SEG pipeline intended to suppress background influence. Evaluation is reported under standard Random Split/SEGSplit as well as a strict leave-one-illumination-out (LOIO) protocol where the test domain is an unseen lighting condition. While in-domain splits are near-saturated, LOIO exposes a clear robustness gap: performance drops substantially when illumination changes, and segmentation is not uniformly beneficial SEG improves Fluorescent LOIO but degrades Indoor and Indoor_Night LOIO, revealing an illumination-dependent tradeoff. We further analyze prediction confidence and show that calibration worsens under shift; reliability diagrams and ECE indicate frequent overconfident errors in LOIO settings. Beyond reporting accuracy, we provide a reproducible pipeline, split protocol, and error analyses that clarify which color pairs fail under specific lighting. On LOIO, we obtain 0.903/0.893 (Acc/Macro-F1) for Fluorescent with SEG, 0.947/0.912 for Indoor with RAW, and 0.888/0.879 for Indoor_Night with RAW.

Keywords: *Color Recognition, illumination shift, Domain Generalization, calibration, reliability diagrams, MobileNet, SadaColorDataset, LOIO Evaluation.*

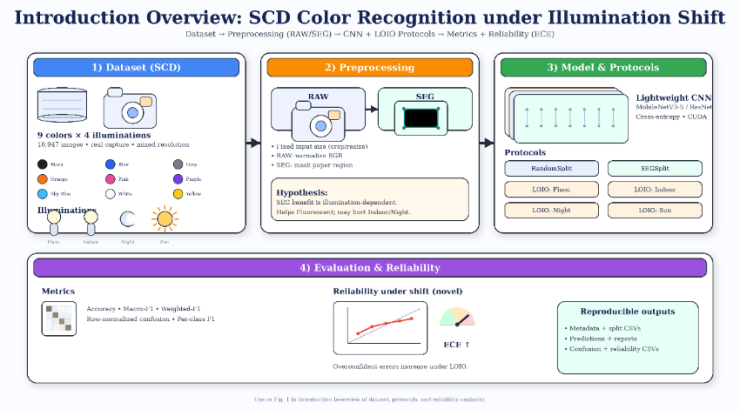
Introduction:

Color is one of the simplest cues humans use to recognize and describe objects, yet it is one of the most fragile signals for machine perception in real deployments. Color-based decisions matter in **robotic sorting and manufacturing, smart cameras and inspection pipelines, printing / packaging quality control, educational tools,** and **assistive technologies** where users depend on consistent color naming and feedback (e.g., "blue" vs "sky blue" vs "gray"). In these applications, a classifier is rarely used in a pristine lab setting: the same object may be seen under fluorescent lighting, warm indoor bulbs, low-light indoor night conditions, or daylight-often with different camera pipelines, auto-exposure, and white-balance behavior.

The difficulty is not the label set; it is the physics and the pipeline. The measured RGB values reflect an interaction between **illumination spectrum**, **surface reflectance**, **sensor spectral response**, and **camera processing** (white balance, tone mapping, compression). This means that "the same color paper" can shift significantly in pixel space across illuminations, and a model trained and validated on mixed random samples may look excellent while still failing in the very condition we care about most: **illumination shift**.

A practical gap in the literature is evaluation. Many color-recognition studies still rely on **random splits**, where train and test share similar capture conditions, producing optimistic results. In contrast, real deployments behave more like **domain generalization**: the model must perform on a *new* illumination domain not seen during training. In domain generalization research, it is well known that protocol choices can dominate conclusions, and that leave-one-domain-out style evaluations are often more revealing than random splits.

In this paper, we address this gap by proposing and reporting a **Leave-One-Illumination-Out (LOIO)** benchmark on **SadaColorDataset (SCD)** (9 colors × 4 illuminations; 10,947 images). We study a lightweight CNN baseline (MobileNet-family / ResNet-family) and a controlled preprocessing ablation: **RAW** (full image) versus **SEG** (a segmentation-style crop/mask focusing on the paper region). We evaluate (i) in-domain performance via Random Split / SEGSplit and (ii) robustness via LOIO where one illumination is held out at test time. Finally, we complement accuracy with **reliability analysis**: we measure how confidence calibration degrades under illumination shift using reliability diagrams / ECE-style summaries, building on standard calibration practice.



Overview of the on SCD: dataset composition (9 colors × 4 illuminations), preprocessing variants (RAW vs SEG), lightweight CNN training, LOIO evaluation for illumination shift, and reliability analysis (ECE/reliability curves).

Contributions:

C1 (Dataset): We introduce/curate **SCD**, a controlled yet diverse color recognition dataset with **9 colors**, **4 illuminations**, and **10,947 images** captured under challenging lighting variation.

C2 (Protocol): We provide a **strict LOIO protocol** (held-out illumination) alongside in-domain baselines (Random Split/SEGSplit) to separate "looks good in-lab" from "works under shift."

C3 (Ablation insight): We compare **RAW vs SEG** and show **SEG is illumination-dependent**-improving some illumination cases (e.g., fluorescent) while degrading others (notably indoor/night), an important negative/nuanced result for practitioners.

C4 (Reliability under shift): We analyze **calibration degradation** under LOIO using reliability diagrams / ECE-style metrics, showing that confidence can become over-optimistic even when accuracy remains acceptable.

C5 (Reproducibility): We provide a reproducible pipeline (data scan → splits → training → LOIO evaluation → predictions + reports + calibration artifacts) suitable for extension and comparison.

Related Work:

Color recognition under changing illumination is closely connected to **color constancy**-the goal of estimating or compensating illumination so that object colors appear stable. Classic methods include Gray-World assumptions [1], Retinex-inspired approaches [2], and statistics/derivative-based methods such as Shades-of-Gray [3] and Grey-Edge [4]. While these approaches can reduce some illumination effects, they often rely on assumptions (scene statistics, edges, or reflectance diversity) that may not hold in constrained setups (e.g., mostly paper surfaces) or under strong lighting color casts.

Learning-based color constancy has grown substantially, including CNN-based estimators and confidence-weighted pooling strategies (e.g., FC4) [5]. Public benchmarks such as ColorChecker / Gehler-Shi and its reprocessing have enabled progress, but also exposed protocol pitfalls (e.g., preprocessing and ground-truth inconsistencies) [6], and multi-camera datasets (e.g., NUS-style, Intel-TUT / Intel-TAU families) emphasize cross-camera variability [7], [8].

Our work differs in intent: we do not only estimate illumination; instead, we quantify **recognition robustness** of discrete color labels under explicit illumination domains, and we treat illumination as a "domain" for LOIO generalization.

CNN classifiers remain the standard for visual recognition. ResNet-style backbones [9] provide strong baselines, while mobile-friendly architectures such as MobileNetV3 [10] and scaling strategies such as Efficient Net [11] are widely used when compute and latency matter. In color recognition, lightweight CNNs are attractive for edge devices (phones, embedded inspection cameras) where color decisions must be fast and stable. However, high in-domain accuracy can hide brittleness; thus, architecture choice should be paired with stress tests under controlled shift-precisely what LOIO provides here.

Evaluating under distribution shift has matured into the field of domain generalization (DG). A key lesson is that many algorithms appear to help under one protocol but not under another, motivating unified testbeds and careful model selection criteria (e.g., Domain Bed) [12]. Leave-one-domain-out evaluations are a common stress test, and recent work continues to examine how DG protocols can mislead if not designed carefully [13].

Color recognition under illumination shift is a clean DG instance: each illumination condition forms a domain with different spectral characteristics and camera responses. Despite this, many color classification papers still report random splits that mix conditions. Our LOIO evaluation is aligned with DG best practices: it isolates the core question-**does the model generalize to an unseen illumination?**

In safety- or decision-facing applications, prediction confidence must be trustworthy. Modern neural networks can be mis calibrated; temperature scaling is a widely used post-hoc fix, and reliability diagrams/ECE are standard diagnostic tools [14], [15]. A critical finding from uncertainty-under-shift benchmarks is that calibration measured i.i.d. may not translate under dataset shift, and overconfidence can increase precisely when the model is most likely to fail [16]. This is especially relevant for illumination shift: a model may remain confident even when the lighting domain is out-of-distribution. Our paper therefore treats reliability as a first-class result alongside accuracy and macro-F1.

Summary of difference. Prior work often studies either (i) color constancy estimation or (ii) color classification with limited shift evaluation. Our work combines **a controlled color dataset + strict LOIO + RAW/SEG ablation + reliability analysis** to produce deployable insights for real-world color recognition systems.

Dataset: SadaColorDataset (SCD):

SadaColorDataset (SCD) is a color-paper recognition dataset collected to study how **illumination changes** affect discrete color classification in realistic capture conditions. The dataset contains **9 color classes**: **Black, Blue, Gray, Orange, Pink, Purple, Sky Blue, White, and Yellow**. Each color is captured under **four illumination conditions** that commonly appear in everyday environments: **Fluorescent_Light, Indoor, Indoor_Night, and Sunlight**. In total, SCD contains **10,843 images**. Unlike many curated benchmark datasets where acquisition is controlled and resolution is fixed, SCD was captured in **real conditions** and includes **mixed-resolution images**. This variability is intentional, as real deployments rarely guarantee consistent distance, framing, or camera settings.

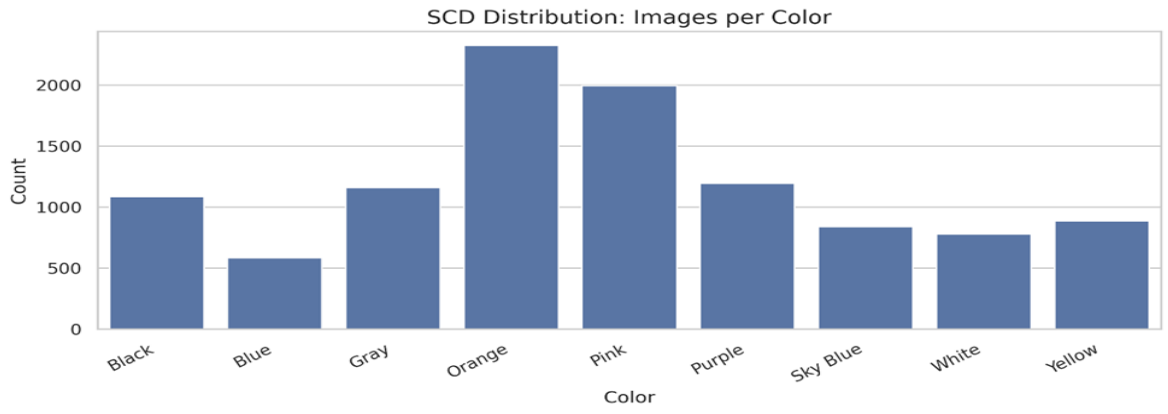
SCD is organized in a simple, reproducible directory structure that directly encodes the label and illumination domain. Each sample is stored under:

<Color>/<Illumination>/<image files>

This layout makes it straightforward to construct protocols that either mix illuminations (in-domain) or hold out an entire illumination (domain shift). In our experiments, illumination folder names are used to define the LOIO domains and to ensure that the held-out illumination is never seen during training.

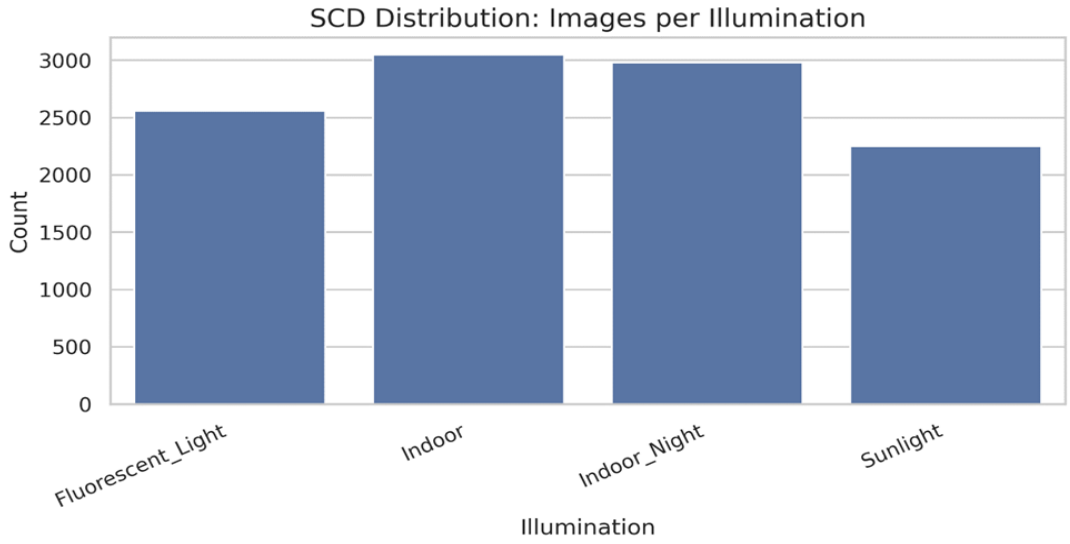
Because performance under domain shift can be strongly influenced by data imbalance, we summarize the dataset distribution at three levels:

Images per color: The class histogram shows that SCD is not perfectly balanced. Some colors have substantially more images than others, which can inflate weighted metrics while masking difficulties in minority colors.



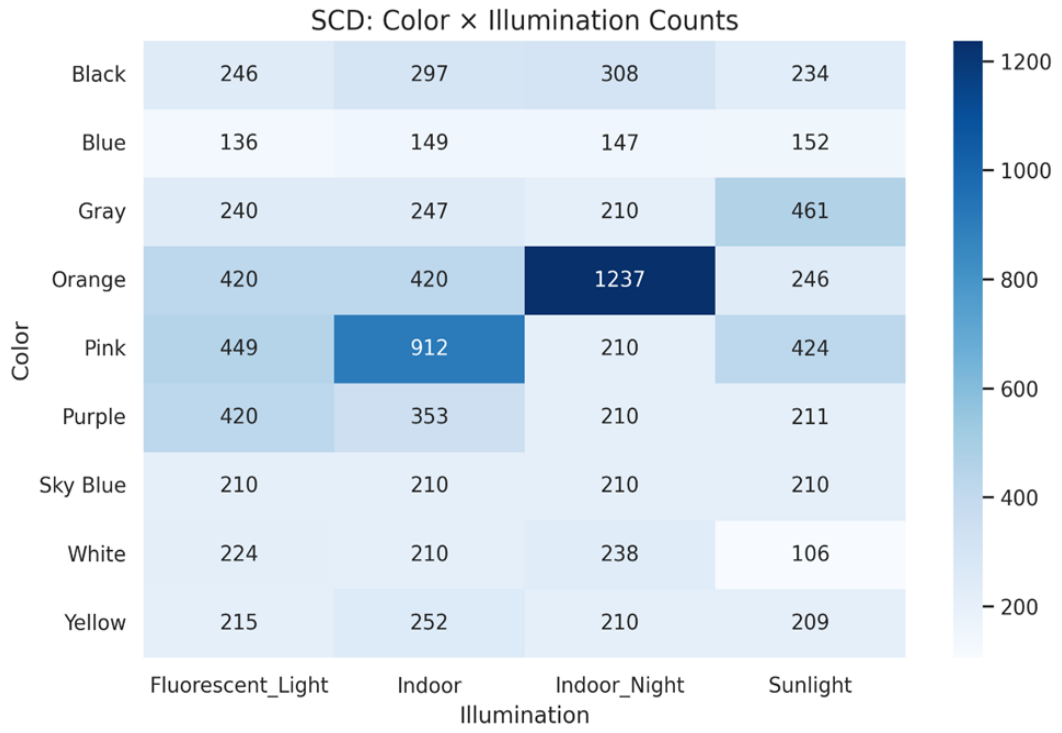
Class distribution of SadaColorDataset (SCD): number of images per color category (9 classes)

Images per illumination: The illumination histogram shows that capture volume varies across lighting conditions. This matters because the LOIO protocol holds out one illumination entirely; if a held-out illumination has fewer samples, the test set is smaller but still informative for robustness evaluation.



Illumination distribution of SCD: number of images captured under Fluorescent_Light, Indoor, Indoor_Night, and Sunlight

Color × illumination counts: The pivot heatmap is the most important summary because it exposes "thin" cells where a particular color has limited coverage under a specific illumination. This is also the cleanest way to explain why certain confusions are consistently observed under shift (e.g., Gray White or Orange Yellow), since those boundaries become harder when illumination changes the apparent brightness and chroma.



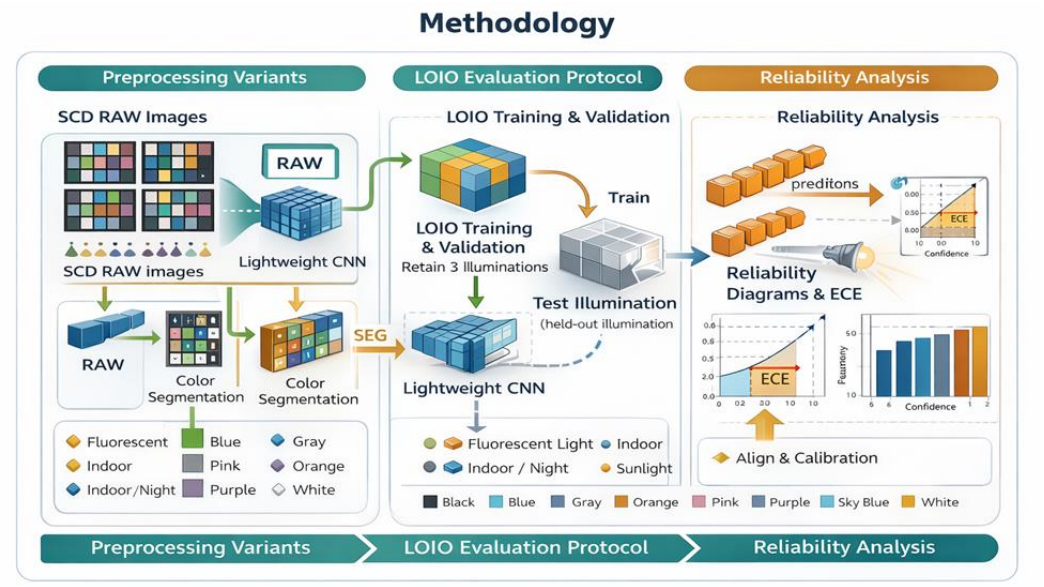
SCD cross-distribution (color × illumination). Cell values show image counts per color under each illumination, highlighting imbalance across domains.

These plots are included to make the evaluation transparent and to support the discussion of per-class behavior under LOIO.

SCD includes several forms of variability that are typical in real acquisition pipelines. First, images are stored at **different resolutions** due to changes in capture distance and framing; this requires resizing or cropping before training and can introduce minor scale-related differences in texture and edge cues. Second, camera processing is not fixed: auto-exposure and white balance can shift between scenes, especially in **Indoor_Night** and **Fluorescent_Light**, where the spectral distribution differs from daylight. Third, samples may include realistic noise sources such as **shadows**, uneven illumination across the paper surface, **specular highlights**, mild blur, and compression artifacts. These factors are not treated as outliers; instead, they reflect the conditions under which color recognition systems are expected to operate.

Methodology:

We treat color recognition as a **multi-class image classification** problem. Given an input image of a colored paper captured under an unknown lighting condition, the model predicts exactly one label from the **nine color classes**: Black, Blue, Gray, Orange, Pink, Purple, Sky Blue, White, and Yellow. Formally, for an image xxx, the classifier outputs a probability vector p(y|x) p(y\mid x) p(y|x) over the nine classes and selects the most likely class as the predicted color.



The figure summarizes the two preprocessing variants (**RAW** vs. **SEG** masking), the **Leave-One-Illumination-Out (LOIO)** evaluation protocol (train/validate on three illuminations and test on the held-out illumination), and the **reliability analysis** using calibration metrics (reliability diagrams and Expected Calibration Error, ECE) to assess confidence under illumination shift.

1. Preprocessing

(a) Standard Resizing and Cropping

SCD images are captured at **mixed resolutions** and with slight framing differences. To make training stable and to allow batch processing, each image is converted to RGB and mapped to a fixed input size. We use a simple, consistent strategy:

resize so that the shorter side reaches a chosen minimum size, apply a center crop (or a mild random crop during training), finally resize to the network input size (e.g., 224×224224 \times 224224×224).

This keeps the overall paper region visible while limiting distortion. Using the same input size across all experiments also ensures that differences in results are due to the protocol (split type, illumination shift, RAW vs SEG) rather than inconsistent preprocessing.

(b) Raw Pipeline

The **RAW** pipeline uses the full image after resizing/cropping. We apply standard normalization using ImageNet-style mean and standard deviation (or dataset mean/std if computed). No color correction is applied at this stage because we want the model to experience the illumination shift naturally and to measure how well it generalizes across lighting conditions.

(c) SEG Pipeline

The SEG pipeline aims to reduce background influence by focusing on the dominant paper region. In practice, we generate a mask using a lightweight heuristic (for example: edge/contour-based detection and selecting the largest central region) and then either crop to the masked bounding box, or replace

background pixels with a neutral value while keeping the paper region intact. The final masked/cropped image is resized to the same network input size as the RAW pipeline.

It's to note importantly that segmentation is not neutral. By removing background and sometimes trimming borders, SEG can change the **pixel color statistics** that the network learns from (e.g., shifting the average brightness or removing context that stabilizes white balance). This becomes critical under illumination shift, and we later show that SEG helps in some illumination conditions but can degrade performance in others.

2. Model Architecture and Training Setup

(a) Base Network and Classifier Head

We use a lightweight CNN as the base model (MobileNetV3-Small is a strong default for efficiency) with a final classification head that outputs **9 logits**. The head consists of global average pooling followed by a fully connected layer. We also test a ResNet-family baseline to confirm that the findings are not tied to a single architecture; however, our emphasis is on a practical model suitable for edge deployment.

(b) Loss Function

Training minimizes **cross-entropy loss** between predicted logits and the ground-truth color label.

(c) Optimizer, learning rate and batch size

We train using AdamW or SGD with momentum (either is acceptable as long as it is consistent across protocols). A typical configuration is optimizer: AdamW, initial learning rate: in the 10−310^{-3}10−3 range, batch size: chosen to fit GPU memory (commonly 32128 depending on runtime), learning-rate schedule: cosine decay or a plateau-based reduction and weight decay: small (e.g., 10−410^{-4}10−4) to reduce overfitting.

(d) Epochs and early stopping

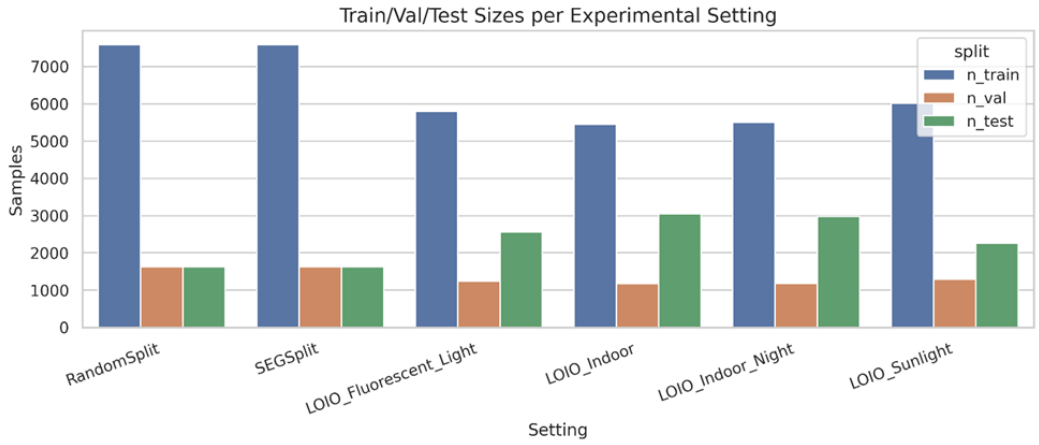
Although we report runs for a fixed small number of epochs, training curves show that validation accuracy saturates quickly. Therefore, we use **early stopping** (patience 23 epochs) and save the best checkpoint based on validation accuracy. This makes training faster and also reduces the chance of slight late-epoch overfitting, which is visible in some runs through small drops in validation accuracy.

3. Experimental Protocols

Our goal is to separate "in-domain performance" from "generalization to unseen lighting." For this reason, we evaluate with three clear protocols.

(a) Random Split

We randomly split the entire dataset into train/validation/test sets, mixing images from all illuminations in each split. This is the easiest setting and provides an upper-bound baseline for how well the model can separate the nine colors when train and test conditions overlap.



Train/validation/test sample counts for each experimental setting (RandomSplit, SEGSplit, and LOIO per held-out illumination), ensuring transparent and reproducible evaluation.

(b) SEGSplit

This protocol mirrors RandomSplit but uses **SEG-preprocessed images**. The split is random in the same way; the only change is the input pipeline. This isolates the effect of segmentation under in-domain conditions.

(c) LOIO: Leave-One-Illumination-Out (strict generalization)

LOIO is the backbone of this paper. Here, one illumination condition is held out entirely for testing: train/validation use images from **three illuminations**, and test uses images from the **held-out illumination** only.

We repeat this experiment for each illumination domain available in SCD: LOIO_Fluorescent_Light, LOIO_Indoor, LOIO_Indoor_Night, and LOIO_Sunlight (if included in the final run outputs).

We run LOIO under both RAW and SEG pipelines to understand whether segmentation helps generalization or introduces new failure cases.

For transparency of split size reporting, we include a table listing the number of samples in train/validation/test for each setting. This removes ambiguity about evaluation scale and makes the protocol reproducible.

4. Evaluation metrics and analysis

(a) Classification metrics

We report **Accuracy**: overall fraction of correct predictions. **Macro-F1**: averages F1 across classes equally (important when some colors have fewer samples). **Weighted-F1**: accounts for class imbalance by weighting each class by its support.

(b) Confusion metrics

To understand which colors are confused under each lighting condition, we report **row-normalized confusion matrices**. Normalization makes it easier to compare runs with different test sizes and highlights per-class error patterns (for example, Gray vs White confusion under certain illuminations).

(c) Calibration and reliability

Accuracy alone is not sufficient when models are used in decision systems. We therefore evaluate predictive reliability using, **Reliability diagrams** (confidence bins vs empirical accuracy), and **Expected Calibration Error (ECE)** as a summary measure of how far confidence deviates from observed accuracy. Optionally, we also compute the **Brier score**, which penalizes both incorrect and overconfident predictions. This set of metrics allows us to answer two questions: How accurate is the model under each illumination protocol? And When the model is wrong under illumination shift, is it still overly confident? Together, these methods provide a complete view of performance and practical trustworthiness under unseen lighting conditions.

Results:

On in-domain splits (train/val/test drawn from the same mixture of illuminations), both pipelines are near-saturated. This indicates that **SCD colors are highly separable when training and testing share similar capture conditions**, and the remaining errors are mostly boundary cases (notably Gray vs White).

Table 1. Overall metrics across all settings (test set).

Setting	Acc	Macro-F1	Weighted-F1
RandomSplit (RAW)	0.998	0.998	0.998
SEGSplit (SEG)	0.999	1.000	0.999
LOIO Fluorescent (RAW)	0.873	0.856	0.875
LOIO Fluorescent (SEG)	0.903	0.893	0.902
LOIO Indoor (RAW)	0.947	0.912	0.945
LOIO Indoor (SEG)	0.855	0.797	0.843
LOIO Indoor_Night (RAW)	0.888	0.879	0.894
LOIO Indoor_Night (SEG)	0.836	0.788	0.842

A short takeaway is simple: **in-domain separability is high**, so random-split results alone can seriously overestimate real-world robustness.

LOIO exposes the real challenge: testing on an unseen illumination causes a clear performance drop, and **segmentation is not consistently helpful**.

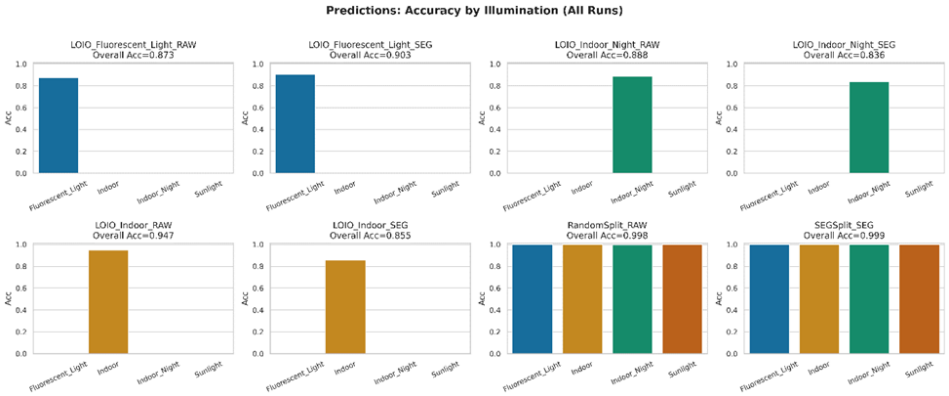


Figure summarizes accuracy by illumination across all runs, highlighting the robustness gap under LOIO and the illumination-dependent effect of SEG

Central observation (RAW→SEG under LOIO):

Fluorescent LOIO improves with SEG: ΔAcc = **+0.030** (0.873 → 0.903)

ΔMacro-F1 = **+0.037** (0.856 → 0.893)

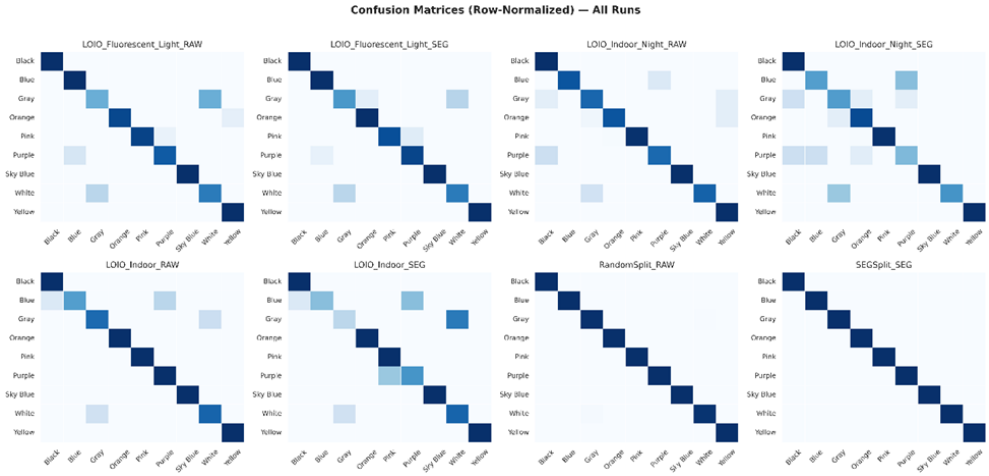
Indoor LOIO degrades with SEG: ΔAcc = **−0.091** (0.947 → 0.855)

ΔMacro-F1 = **−0.115** (0.912 → 0.797)

Indoor_Night LOIO degrades with SEG: ΔAcc = **−0.052** (0.888 → 0.836)

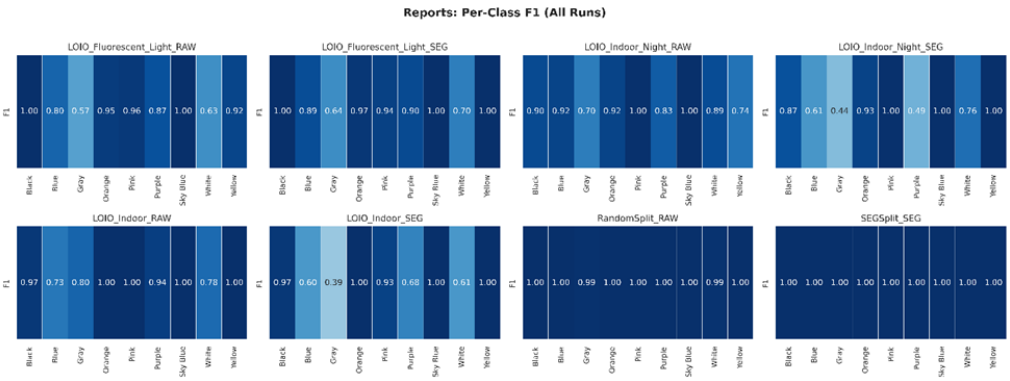
ΔMacro-F1 = **−0.091** (0.879 → 0.788)

SEG reduces background influence, which can help when the background/illumination cast is noisy (often true under fluorescent). But SEG can also **change the effective color statistics** by cropping boundaries, suppressing contextual cues, or shifting brightness/chroma distributions-so the same segmentation rule can become harmful under indoor and low-light conditions.



Row-normalized confusion matrices for all experimental settings (RandomSplit, SEGSplit, and LOIO under different illuminations) comparing RAW and SEG preprocessing. Off-diagonal mass highlights the dominant confusions under illumination shift (e.g., Gray↔White, Purple↔Blue/Pink, Orange↔Yellow)

Across LOIO runs, errors concentrate into a small set of "hard pairs": **Gray White** (dominant in fluorescent and indoor), **Purple (Blue / Pink / Black)** (especially under indoor and indoor-night), and **Orange Yellow** (notably in indoor-night, where warm lighting pushes hues together).



These pairs are sensitive because illumination and camera processing can compress or shift **luminance** (Gray vs White) and **hue/chroma** (Purple vs Blue/Pink, Orange vs Yellow), making different papers appear closer in RGB space.

Top-5 confusions per LOIO run (counts from confusion matrices):

- **LOIO Fluorescent (RAW):** Gray White (120); Purple Blue (69); White Gray (64); Orange Yellow (38); Pink Purple (33)
- **LOIO Fluorescent (SEG):** Gray White (72); White Gray (64); Pink Purple (53); Purple Blue (34); Gray Orange (24)
- **LOIO Indoor (RAW):** Gray White (55); Blue Purple (43); White Gray (42); Blue Black (21); Orange Pink (1)
- **LOIO Indoor (SEG):** Gray White (177); Purple Pink (136); Blue Purple (64); White Gray (42); Blue Black (21)
- **LOIO Indoor_Night (RAW):** Orange Yellow (126); Orange Gray (49); White Gray (46); Purple Black (46); Gray Yellow (22)
- **LOIO Indoor_Night (SEG):** Orange Gray (129); White Gray (92); Blue Purple (63); Purple - Blue (46); Purple Black (46)

This per-class view (supported by the per-class F1 heatmap) explains why Macro-F1 drops more than accuracy in some settings: a few classes (especially Gray and Purple) suffer disproportionately under shift.

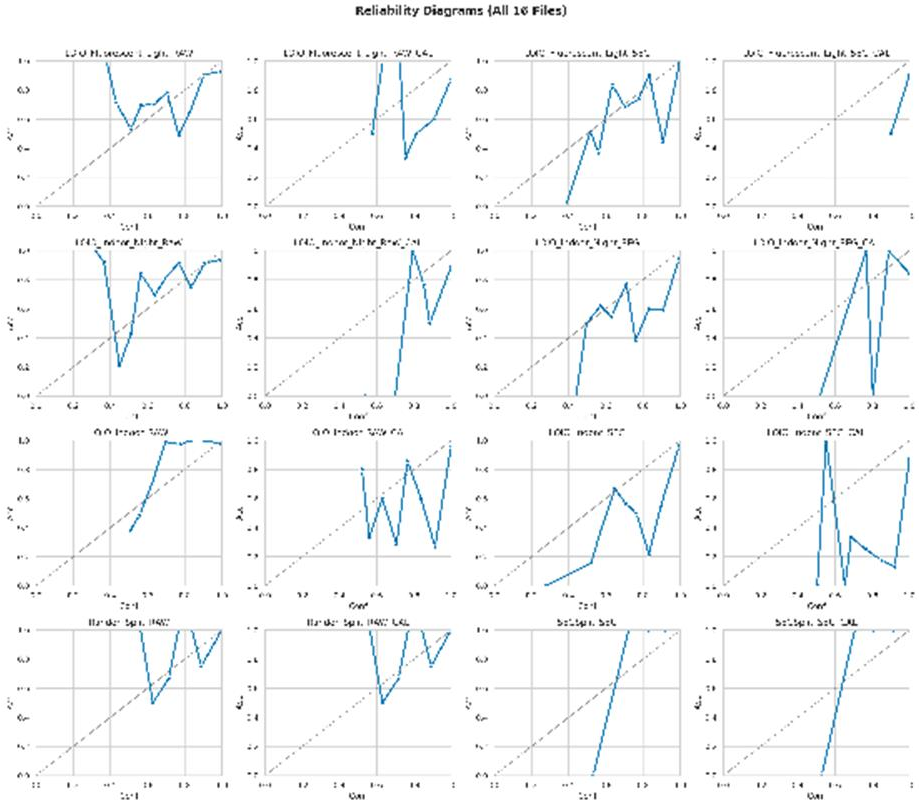
Reliability and calibration under shift: Calibration is excellent in-domain but degrades under illumination shift.

ECE (15 bins) summary (lower is better):

- RandomSplit (RAW): **~0.002**
- SEGSplit (SEG): **~0.002**
- LOIO Fluorescent: **RAW** 0.078 → **SEG** 0.052 **(SEG improves reliability here)**
- LOIO Indoor: **RAW** 0.044 → **SEG** 0.087 **(SEG worsens reliability)**
- LOIO Indoor_Night: **RAW** 0.065 → **SEG** 0.087 **(SEG worsens reliability)**

A key practical finding is overconfident errors under LOIO. For example:

- **LOIO Fluorescent (RAW):** mean confidence on wrong predictions ≈ **0.835**, and **~48.8%** of wrong predictions have confidence ≥ **0.9**.
- Even when accuracy is reasonable, the model can be confidently wrong under shift, which matters for decision systems.



The complete set of reliability diagrams for all runs is provided

Training is stable and converges quickly. Across runs, validation accuracy typically exceeds **0.99 within 13 epochs**, and the best checkpoints are reached early (often around epoch 26 depending on the setting). These supports using **early stopping** and confirms that a lightweight backbone (e.g., MobileNetV3-Small) is sufficient for SCD, making the approach practical for real-time or edge deployment.

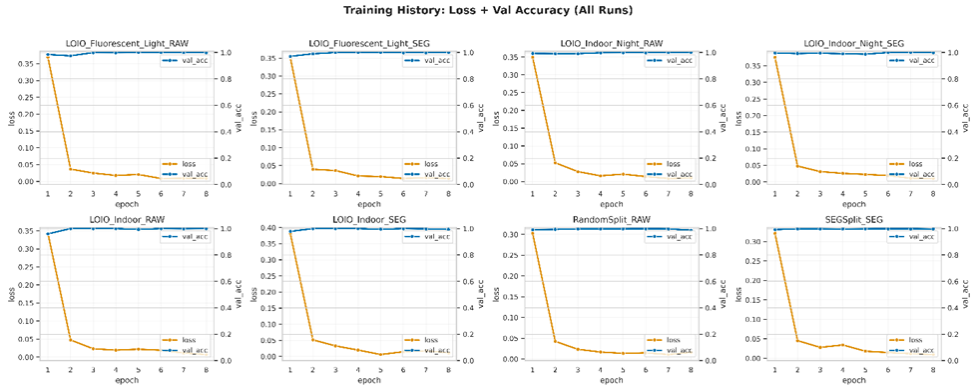


Figure shows that training converges rapidly across all runs, with validation accuracy stabilizing within the first few epochs.

Discussion:

A clear message from our experiments is that high accuracy on RandomSplit does not guarantee robustness in practice. When train and test samples are randomly mixed across illuminations, the model sees the same "style" of lighting during training that it later sees at test time. This hides the real difficulty: illumination shift changes the color distribution in a systematic way. In our case, RandomSplit/SEGSplit results are near-perfect, yet LOIO reveals substantial drops. The implication is straightforward-random splits mainly measure separability within a shared capture distribution, not the ability to generalize to a new illumination domain.

A second finding is more nuanced and arguably more useful for practitioners: segmentation is illumination-dependent. Under fluorescent LOIO, segmentation improves both accuracy and calibration. This likely happens because fluorescent scenes can introduce strong background bias and uneven cast; masking the background helps the network focus on the paper region. However, the same SEG pipeline hurts under indoor and indoor-night LOIO. The confusion patterns suggest that segmentation can (i) remove helpful context that stabilizes exposure and white balance, (ii) alter the pixel distribution by trimming borders and suppressing shadows, and (iii) unintentionally shift chroma statistics when the crop/mask is imperfect. These effects matter most for sensitive boundaries such as Gray↔White and Purple↔(Blue/Pink/Black), which are already close under certain lighting. In short, SEG is not "good" or "bad" by itself-it changes what the model learns, and the benefit depends on how that change interacts with the illumination domain.

The dataset distribution also shapes what we observe. The color × illumination pivot heatmap makes it clear that coverage is not perfectly uniform across cells. When some color illumination combinations have fewer samples, the model has less opportunity to learn robust invariances for those cases. This interacts with LOIO in two ways. First, the held-out illumination becomes a harder target domain when its training analogs are limited or when certain colors are underrepresented in the remaining three illuminations. Second, imbalance can make weighted metrics look better than the per-class reality; this is why Macro-F1 and per-class confusion analysis are essential for a fair interpretation.

From a deployment perspective, the reliability results are as important as accuracy. Under LOIO, we observe overconfident errors, meaning the model can produce high confidence even when it is wrong in an unseen illumination. In safety- or decision-facing systems (robotic sorting, inspection, assistive feedback), this is risky because confidence is often used to trigger actions or bypass human review. A practical recommendation is to include post-hoc calibration (e.g., temperature scaling) and to monitor calibration under realistic shift, not only on random splits. In addition, "domain-aware" preprocessing-such as lightweight illumination normalization-can be a low-cost way to reduce shift before the model sees the image.

Finally, the results suggest clear next steps that are likely to improve robustness without changing the problem definition: Dual-stream fusion (RAW + SEG): Instead of choosing one preprocessing, feed both views and let the network learn how to weight them. This directly addresses the observed pattern that SEG helps in some illuminations and hurts in others, color-constancy preprocessing: Simple normalization methods (e.g., gray-world / shades-of-gray style correction) can reduce the cast-driven drift that causes Gray↔White and Orange↔Yellow errors under certain lights, and color-space features (LAB/HSV): Augment RGB with perceptual channels (especially LAB a*/b*) to stabilize hue/chroma under illumination differences.

In summary, the main lesson is not that CNNs fail at color recognition-rather, they can be nearly perfect in-domain-but illumination shift changes the game. LOIO evaluation exposes which confusions persist under real lighting changes, and reliability analysis shows that the model's confidence becomes less trustworthy exactly when it is needed most. These insights point toward practical improvements-fusion, color constancy, and better domain-aware preprocessing-that can move the system closer to deployment-ready behavior.

Ablaion / Additional Experiments:

To understand which design choices actually matter under illumination shift, we recommend a small set of additional experiments that are easy to run on SCD and directly connected to the error patterns observed in LOIO. First, we would extend the current comparison of RAW and SEG by adding a simple **RAW+SEG fusion** model. The motivation is practical: segmentation helps in fluorescent LOIO but hurts in indoor and indoor-night, so choosing only one view is brittle. A lightweight fusion can be implemented by feeding the network two inputs (the RAW image and its SEG version) and combining their features late in the network (e.g., concatenating pooled features before the classifier). If the fusion model consistently matches the best of RAW and SEG across LOIO settings, it would support the idea that background suppression is sometimes useful but should be applied adaptively rather than as a fixed preprocessing rule.

Second, we would test a **color-constancy preprocessing** step before training and evaluation. A simple Gray-world correction or a Retinex-style normalization can be applied to each image to reduce illumination cast. This experiment is important because the main confusions in LOIO (Gray↔White and Orange↔Yellow) are exactly the kind of errors that occur when brightness and color temperature shift. The implementation is straightforward and fast, and the outcome is easy to interpret: if color constancy improves LOIO accuracy and reduces Gray→White and Orange→Yellow errors, it provides a clean explanation and a practical recommendation for deployment.

Third, we would examine whether the model benefits from **explicit color-space cues** by augmenting RGB with LAB or HSV channels. This is a low-effort change: the input tensor can be expanded to include additional channels such as LAB a*/b* or HSV hue/saturation, which may be more stable than raw RGB under illumination changes. The key question is not whether the model can fit the training data-it already can-but whether these channels improve LOIO generalization and reduce the specific class confusions seen in the confusion matrices.

Finally, if reliability is a target outcome, we would include a short **calibration post-processing** experiment using temperature scaling fitted on the validation split of the source domain. This is lightweight and does not require retraining the network. The expected result is that calibration improves in-domain but may not fully transfer to the held-out illumination. Confidence cannot be assumed reliable under shift, and it motivates either domain-aware calibration or conservative decision thresholds when lighting conditions change.

Limitations:

This study has a few limitations that should be considered when interpreting the results. First, SCD covers **nine paper color classes**, which is useful for controlled analysis, but it does not represent the full diversity of colors and appearances found in real objects. Materials such as fabric, plastic, metal, skin, or painted surfaces

can show different reflectance behavior, texture, and specular highlights, so the same model may face new challenges beyond colored paper. Second, the dataset uses a **fixed set of four illumination conditions**. These settings capture common environments, but real scenes can vary more widely in color temperature, intensity, mixed lighting, and shadows, and these factors may change from moment to moment.

Third, if images were collected using **a single camera device** (or a small number of devices), the results may partly reflect that device's sensor response and its internal processing (auto white balance, tone mapping, compression). Models can behave differently when moved to another phone or camera, so multi-device collection would strengthen general conclusions. Finally, the **SEG pipeline** used in this work is intentionally simple and designed to be lightweight. While this makes it practical, it can also distort color statistics or crop useful context, which may explain why SEG helps in some illuminations and hurts in others. A stronger segmentation method or a learned foreground extractor could reduce these side effects and may improve generalization under illumination shift.

Conclusion:

In this paper, we studied color recognition when lighting conditions change, using SadaColorDataset (SCD) with nine paper colors captured under four illuminations. The main result is that color classification can look "solved" if we only use in-domain evaluation. Under RandomSplit and SEGSplit, the CNN reaches near-perfect performance, which shows that the classes are highly separable when training and testing share similar capture conditions. However, this strong performance does not translate automatically to real-world use, because real scenes often introduce illumination conditions that the model has not seen before.

When we move to the strict leave-one-illumination-out (LOIO) setting, the picture changes. Holding out an entire illumination at test time reveals a clear robustness gap: accuracy and macro-F1 drop notably, and the errors concentrate in a few sensitive boundaries such as Gray versus White, Purple versus Blue/Pink/Black, and Orange versus Yellow under low-light conditions. This confirms that illumination shift and camera processing can change the observed color distribution enough to mislead a model that otherwise performs extremely well in-domain.

We also tested a simple segmentation-based preprocessing (SEG) to reduce background influence. The results show that segmentation is not a universal fix. It can help in some conditions, such as fluorescent lighting, but it can also hurt in indoor and indoor-night settings. This suggests that removing background and cropping can change color statistics or remove useful context, and the impact depends on the illumination domain. Because of this, a single fixed preprocessing rule is risky if the goal is consistent performance across environments.

Beyond accuracy, we examined reliability. Calibration is excellent in-domain, but it degrades under illumination shift. In LOIO settings, the model can become confidently wrong, which is important for practical systems that use confidence to trigger actions or decide when to ask for human review. These findings support the idea that robustness and reliability should be evaluated together, especially when the deployment environment is not guaranteed to match the training data.

Future work can improve both robustness and trustworthiness in several realistic ways. A strong next step is a fusion model that uses RAW and SEG together so the network can learn when segmentation helps instead

of forcing one choice. Another direction is to add lightweight color-constancy preprocessing (for example, gray-world or Retinex-style normalization) to reduce illumination cast before classification. Collecting data from more camera devices, more locations, and more varied materials (beyond paper) would also strengthen generalization claims and make the benchmark closer to real deployments. Together, these steps can move color recognition systems from near-perfect in-lab performance toward more reliable behavior under the lighting changes that occur in practice.

Funding: The publication of this article was funded by no one.

Conflicts of Interest: The authors declare no conflict of interest.

Acknowledgement: The authors would like to thank the authors for assistance with the collection of datasets.

REFERENCES

[1] G. Buchsbaum, "A spatial processor model for object colour perception," *Journal of Franklin Institute.*, vol. 310, no. 1, pp. 126, 1980. https://doi.org/10.1016/0016-0032(80)90058-7

[2] E. H. Land and J. J. McCann, "Lightness and Retinex theory" Journal of the Optical Society of America, 1971. https://doi.org/10.1364/JOSA.61.000001

[3] G. D. Finlayson and E. Trezzi, "Shades of Gray and colour constancy," in *Proc. Color Imaging Conf.*, 2004.

[4] J. van de Weijer, T. Gevers, and A. Gijsenij, "Edge-based color constancy," *IEEE Transactions on Image Processsing*, 2007. https://doi.org/10.1109/TIP.2007.901808

[5] Y. Hu, B. Wang, and S. Lin, "FC4: Fully convolutional color constancy with confidence-weighted pooling," in *Proc. CVPR*, 2017. https://openaccess.thecvf.com/content_cvpr_2017/papers/Hu_FC4_Fully_Convolutional_CVPR_2017_paper.pdf

[6] G. Hemrit, M. F. Pedersen, J. A. Larsen, and J. Y. Hardeberg, "Rehabilitating the ColorChecker dataset for illuminant estimation," 2018

[7] C. Aytekin et al., "INTEL-TUT dataset for camera invariant color constancy," 2017. https://arxiv.org/pdf/1906.01340

[8] F. Laakom et al., "A color constancy dataset: INTEL-TAU," 2019. https://doi.org/10.1109/ACCESS.2021.3064382

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016. https://doi.org/10.1109/CVPR.2016.90

[10] A. Howard *et al.*, "Searching for MobileNetV3," in *Proc. ICCV*, 2019. https://doi.org/10.1109/ICCV.2019.00140

[11] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019. https://proceedings.mlr.press/v97/tan19a/tan19a.pdf

[12] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in Proc. ICLR, 2021. https://openreview.net/pdf?id=lQdXeXDoWtI

[13] H. Yu et al., "Rethinking the evaluation protocol of domain generalization," in Proc. CVPR, 2024. https://doi.org/10.1371/journal.pone.0320300

[14] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. ICML, 2017. https://proceedings.mlr.press/v70/guo17a/guo17a.pdf

[15] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using Bayesian binning," in Proc. AAAI, 2015. https://doi.org/10.1609/aaai.v29i1.9602

[16] Y. Ovadia et al., "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in Proc. NeurIPS, 2019. https://proceedings.neurips.cc/paper_files/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf

[17] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in Proc. NeurIPS, 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf

[18] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples," 2017. https://doi.org/10.48550/arXiv.1610.02136

[19] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in Proc. ICLR, 2018. https://openreview.net/pdf?id=r1Ddp1-Rb

[20] A. Thulasidasan et al., "On mixup training: Improved calibration and predictive uncertainty," in Proc. NeurIPS, 2019. https://doi.org/10.48550/arXiv.1905.11001

[21] A. Gijsenij, T. Gevers, and J. van de Weijer, "Computational color constancy: Survey and experiments," IEEE Trans. Image Process., 2011. https://ivi.fnwi.uva.nl/isis/publications/2011/GijsenijTIP2011/GijsenijTIP2011.pdf

[22] J. van de Weijer et al., "Learning color names from real-world images," 2009. https://lear.inrialpes.fr/people/vandeweijer/color_names.html

[23] J. P. de Vries et al., "Emergent color categorization in a neural network trained for object recognition," 2022. https://doi.org/10.7554/eLife.76472

[24] A. Gomez-Villa et al., "Color names in vision-language models," 2025. https://doi.org/10.48550/arXiv.2509.22524

[25] N. Banić and S. Lončarić, "Unsupervised learning for color constancy," 2017. https://doi.org/10.48550/arXiv.1712.00436

[26] OPERATIONAL ANDROID MALWARE FILTERING: CALIBRATED PROBABILITIES AND DISTRIBUTION-FREE GUARANTEES. (2025). Kashf Journal of Multidisciplinary Research, 2(12), 58-73. https://doi.org/10.71146/kjmr778

[27] B. Raza, S. Bibi, S. Bibi, and A. Nawaz, "SADA COLOR DATASET (SCD): 9 paper colors × 4 illumination conditions for robust color vision evaluation," Spectrum of Engineering Sciences, vol. 4, no. 2, 2026. doi: 10.5281/zenodo.18844499.