MODELING THE LIGHTING AS STYLE FACTOR VIA NEURAL NETWORKS FOR WHITE BALANCE CORRECTION

OSMAN FURKAN KINLI

ÖZYEĞİN UNIVERSITY

MAY, 2025

MODELING THE LIGHTING AS STYLE FACTOR VIA NEURAL NETWORKS FOR WHITE BALANCE CORRECTION

by
Osman Furkan Kınlı

Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy

in
Computer Science

Advisor: Asst. Prof. M. Furkan Kıraç

Graduate School of Science and Engineering
Özyeğin University
İstanbul

May, 2025

MODELING THE LIGHTING AS STYLE FACTOR VIA NEURAL NETWORKS FOR WHITE BALANCE CORRECTION

Approved by:

Asst. Prof. M. Furkan Kıraç, Advisor
Department of Computer Science
Özyeğin University

Prof. Lale Akarun
Department of Computer Engineering
Boğaziçi University

Prof. H. Fatih Uğurdağ
Department of Electrical and Electronics Engineering
Özyeğin University

Assoc. Prof. Aykut Erdem
Department of Computer Science and Engineering
Koç University

Prof. Erhan Öztop
Department of Artificial Intelligence and Data Engineering
Özyeğin University

Approval Date: May 12, 2025

To dead butterflies...

DECLARATION OF ORIGINALITY

I hereby declare that I am the sole author of this thesis and that this is the true copy of my thesis, including the final revisions, approved by my thesis committee. All data and information have been obtained, produced, and presented in accordance with the rules of research ethics and principles of academic honesty. As required by these rules, to the best of my knowledge I have acknowledged ideas, thoughts, and any copyrighted material in accordance with the standard referencing rules. I certify that any part of this thesis has not been submitted for a degree or diploma in another educational institution.

Osman Furkan Kınlı

ABSTRACT

This thesis explores White Balance (WB) correction by modeling lighting as a style factor through distribution-based approaches in both architectural design and optimization frameworks. Three novel methods are proposed to address the challenges of complex illumination scenarios. The first approach, Style WB, employs a UNet-like architecture with style modulation to effectively remove illumination-related style information, which achieves robust correction with enhanced spatial consistency. The second approach, FDM WB, introduces feature distribution matching within the Uformer architecture, which enables precise alignment of global and local illumination features for WB correction. Both approaches are evaluated on the Cube+ dataset and a synthetic multi-illuminant benchmark, and they demonstrate substantial improvements in WB correction across diverse lighting conditions. The third approach, FDM Loss, defines an optimization framework leveraging the [CLS] token of Vision Transformers to achieve exact matching of all moments between the predicted and ground truth images, capturing higher-order statistics essential for managing intricate lighting variations. This approach delivers reduced Mean Angular Error (MAE) and consistent illumination correction on the LSMI dataset across three camera setups. While these methods advance WB correction, integrating deterministic mapping mechanisms, such as DeNIM, in resource-constrained environments or leveraging diffusion-based models and neural ODEs could further enhance performance, particularly in handling complex lighting scenarios. This work redefines the role of distribution-based modeling in addressing illumination challenges, setting a foundation for future innovations in image restoration.

Keywords: White Balance correction, Style factor, Feature statistics, Illumination, EFDM

ÖZET

Bu tez, ışıklandırmayı bir stil faktörü olarak modelleyerek, hem mimari tasarım hem de optimizasyon çerçeveleri açısından dağılım tabanlı yaklaşımlar aracılığıyla Beyaz Dengesi (BD) düzeltmesini incelemektedir. Karmaşık ışıklandırma senaryolarındaki zorlukları ele almak için üç yeni yöntem önerilmektedir. İlk yöntem, Style WB, ışıklandırmaya bağlı stil bilgisini kaldırmak için stil modülasyonuna sahip UNet benzeri bir mimari kullanmakta ve geliştirilmiş mekânsal tutarlılığı artırırken güçlü bir düzeltme sağlar. İkinci yöntem, FDM WB, Uformer mimarisine ışıklandırmanın bütüncül ve yerel özniteliklerinin hassas şekilde hizalayan bir dağılım eşleme mekanizması entegre ederek, BD düzeltmesi için önemli iyileştirmeler sunar. Her iki yöntem de Cube+ veri seti ve sentetik çoklu ışıklandırma içeren değerlendirme kümesi üzerinde test edilmiş ve çeşitli ışıklandırma koşulları altında başarı göstermiştir. Üçüncü yöntem, FDM Loss, Vision Transformers mimarisinde yer alan [CLS] token'ı kullanarak, tahmin edilen ve gerçek görüntüler arasındaki tüm dağılım momentlerini tam eşleştiren bir optimizasyon çerçevesi tanımlar ve karmaşık ışıklandırma varyasyonlarını modellemek için gerekli olan yüksek mertebeden istatistikleri yakalar. Bu yöntem, üç farklı kamera ile çekilmiş resimler içeren LSMI veri seti üzerinde daha düşük ortalama açısal hata değerleri ve tutarlı BD düzeltmesi sağlamaktadır. Bu yöntemler BD düzeltmesi performansını geliştirirken, kaynak kısıtlı ortamlarda DeNIM gibi deterministik piksel eşleme mekanizmalarının entegrasyonu ya da karmaşık ışıklandırma senaryolarını ele almak için difüzyon tabanlı modeller ve nöral ADD'lerin kullanımıyla performansın artırılabileceği öngörülmektedir. Bu çalışma, zorlu ışıklandırma senaryolarında dağılım tabanlı modellemenin rolünü yeniden tanımlamakta ve görüntü iyileştirme alanında gelecekteki yeniliklere sağlam bir temel oluşturmaktadır.

Anahtar Kelimeler: Beyaz Dengesi, Stil faktörü, Öznitelik istatistiği, Işıklandırma

ACKNOWLEDGEMENTS

Foremost, I express my sincere gratitude to my advisor, Prof. M. Furkan Kıraç, for his unwavering support and invaluable knowledge. His patience and enthusiasm have been instrumental in shaping my research, and I deeply appreciate the intellectual freedom he provided, allowing me to explore and pursue ideas that truly inspired me. I extend my appreciation to my thesis committee for their forthcoming insights and guidance, which will refine my research and contribute significantly to its quality.

My gratitude also goes to my colleagues at the Vision and Graphics Lab (VGL) at Özyeğin University: Dr. Barış Özcan, Sami Menteş, Doğa Yılmaz, Mert Erkol, and Üveys Akbaş. Their camaraderie, discussions, and shared struggles have enriched this journey. Special thanks to Dr. Barış Özcan for his invaluable advice and continued support, to Sami for always being there during the darkest times, and to Doğa for his inspiring perspectives influenced my research approach. I am also grateful to Emir Arditi for his insightful assistance and encouragement throughout my research.

I am profoundly thankful to my family—my mother, Fatma Hanım, my father, Muammer, and my brother, Efecan—for their unwavering belief in me. Their endless encouragement and unconditional support have been my anchor, keeping me motivated and determined through the highs and lows of this journey.

Finally, I owe immense gratitude to the music that carried me through the darkest moments of my life and academic journey. The raw energy and depth of metalcore bands like Bad Omens, Motionless In White, VOLA, Architects, Katatonia, Bring Me The Horizon, Draconian, and Imminence fueled my resilience, turning even the hardest days into something bearable, and at times, even inspiring. And lastly, to the final boss, Alex—thank you for being a wonderful jumbo-sized friend. Your presence brought immeasurable comfort and joy, making this journey far more bearable and meaningful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

AdaIN	Adaptive Instance Normalization
AWB	Auto White-Balance
CDF	Cumulative Distribution Function
CRF	Conditional Random Field
CCTF	Color Component Transfer Function
CFA	Color Filter Array
CIE	International Commission on Illumination
CNN	Convolutional Neural Network
DeNIM	Deterministic Neural Illumination Mapping
DNCM	Deterministic Neural Color Mapping
DNN	Deep Neural Network
DRM	Dichromatic Reflection Model
eCDF	Empirical Cumulative Distribution Functions
EFDM	Exact Feature Distribution Matching
FDM WB	Feature Distribution Matching White Balancing
GeLU	Gaussian Error Linear Unit
GI	Grayness Index
HDR	High Dynamic Range
IFRNet	Instagram Filter Removal Network
ISP	Image Signal Processor

LSMI	Large-scale Multi-Illuminant
LDR	Low Dynamic Range
LeFF	Locally-enhanced Feed-Forward Network
LeWin	Locally-enhanced Window Transformer Block
LNRE	Locally Normalized Reflectance Estimation
MAE	Mean Angular Error
MLP	Multi-Layer Perceptron
MSR	Multi-to-Single Ratio
MSE	Mean Squared Error
MSA	Multi-Head Self-Attention
NLP	Natural Language Processing
PCA	Principal Component Analysis
RNN	Recurrent Neural Network
ViT	Vision Transformer
WB	White Balance

1. INTRODUCTION

1.1 Context and Motivation

Accurate reproduction of colors under varying lighting conditions remains a key challenge in digital image processing, especially when dealing with real-world scenes where lighting sources can vary widely [10]. Auto White-Balance (AWB) correction plays a critical role in ensuring that the colors are perceived correctly under different illuminants, making it one of the most important tasks in the Image Signal Processor (ISP) pipeline. This correction is essential to obtain images that are perceived accurately by the human eye, maintaining natural color representation in diverse lighting conditions [11]. However, while effective under controlled conditions, traditional AWB methods often fail to deliver consistent results in multi-illuminant environments or low-light conditions, leading to unnatural color casts and poor visual quality.

The foundation of White Balance (WB) correction lies in adjusting the illuminant color mapping or distributions in an image so that white objects appear white, regardless of ambient lighting. However, the challenge lies in how to model and correct the influence of different light sources on an image's illuminant color balance. Each light source, whether natural or artificial, introduces a different spectral signature, which can distort colors in complex and unpredictable ways [12]. This distortion often appears as unwanted color casts that must be removed to achieve a natural and balanced image. The challenge is further compounded when dealing with multi-illuminant environments [2], where various light sources may light different regions of an image, each with its own color temperature. For example, a room with both natural sunlight and artificial incandescent lighting can create a mixture of warm and cool reflects in different parts of the scene.

Figure 1.1: *Different spectral signatures of the light sources in a sample scene.*

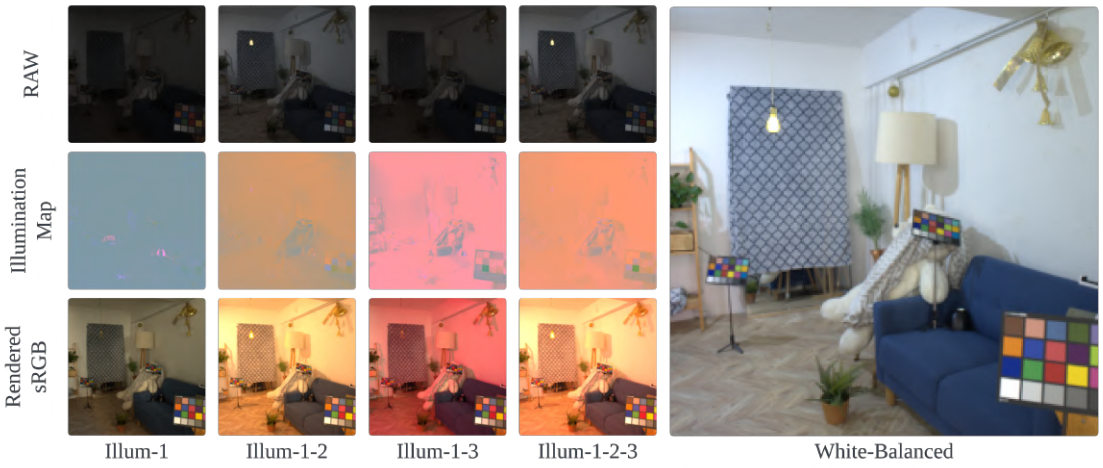


Figure 1.1 demonstrates how different spectral signatures of the light sources can alter the overall appearance of a scene. White balance correction methods aim to neutralize the color casts introduced by varying illumination sources, resulting in a balanced and visually accurate color representation under different lighting conditions. The complexity increases as the intensity and direction of these light sources vary. Traditional AWB methods often apply a global correction, adjusting the entire image based on a single averaged lighting estimate, which assumes that illumination is uniform. However, this assumption rarely holds in real-world scenarios. These methods typically rely on simple statistical measures, such as the Gray-World assumption [13], which assumes that the average color of a scene should be gray, or histogram-based approaches [14, 15]. Although computationally efficient, these methods lack the flexibility to adapt to the nuanced variations in lighting, often resulting in only partially corrected color casts or, in some cases, making them worse.

1.1.1 Deep Learning Era for White Balance Correction

Recent advances in deep learning have introduced more sophisticated approaches to WB correction [2, 16, 17, 1, 18, 19, 20, 21, 22, 23, 24]. These methods leverage large datasets [12, 1, 25, 26, 8] to learn complex mappings between input images and the desired white-balanced output, offering flexibility and adaptability beyond traditional methods. However, despite their successes, one critical limitation is that many of these models do not explicitly account for the underlying distribution characteristics, which are essential for accurate color correction.

Deep learning models, typically based on Convolutional Neural Network (CNN) [27] or attention-based architectures like Transformers [3], can learn to identify the effects of lighting on color distributions by directly mapping input images to corrected outputs. These models handle multi-illuminant environments and complex lighting variations by learning both global and local features, enabling them to apply corrections based on the conditions of each image.

A key advantage of deep learning models is their ability to perform end-to-end learning. These models do not require manual adjustment of parameters; instead, they adjust the color balance across the image using large labeled datasets. This allows for more accurate corrections, especially in cases where traditional methods struggle [24]. However, a significant limitation of current deep learning approaches is that they typically rely on learned features extracted implicitly through convolutions or attention mechanisms, without direct reference to the distributional properties of color, which play a crucial role in maintaining perceptual consistency across lighting conditions.

For example, when dealing with multi-illuminant environments, a model that does not explicitly consider feature distributions may apply corrections based solely on learned patterns from the training data, without understanding the statistical shifts introduced by different light sources. This can result in overcorrection or undercorrection, especially

Figure 1.2: *Simulated rendering results with different color temperatures (i.e., Shade and Tungsten) and the white-balance corrected version.*

when the training data predominantly features single illumination scenarios. Without an awareness of how multiple illuminants distort the color distributions, the model may struggle to generalize in more complex lighting conditions.

1.1.2 Feature Distributions as Style Factors

In the context of illumination and white balance, feature distributions refer to the statistical properties of an image's color channels or higher-level features that describe how illuminant colors are represented. These distributions are influenced by the lighting conditions in which an image is captured, and adjusting them can significantly improve the perceived color balance. This concept introduces the main idea of this study: feature distributions can represent or model lighting as a style factor [24, 28, 29].

Lighting is a dominant style factor because it directly affects the distribution of color values across an image. The type, intensity, and color temperature of the light source can drastically shift the appearance of objects, making whites appear warm (*e.g.*, orange) or cool (*e.g.*, blue) and distorting other colors accordingly. Figure 1.2 demonstrates the effect of different lighting conditions on a scene, showing how lighting can introduce unwanted color casts that need correction. Furthermore, different lighting conditions can alter the way camera sensors perceive colors [22], requiring device-dependent algorithms for image acquisition and post-processing algorithms to ensure accurate color perception.

To correct these lighting-induced distortions, WB correction aims to align the color distributions with a neutral reference, ensuring that the colors in the scene are represented accurately. This requires understanding how lighting conditions affect the color distribution and adjusting them accordingly. By treating feature distributions as style factors, it becomes possible to model lighting and illumination environments, allowing for their manipulation or mitigation of unwanted casts using neural networks.

1.1.3 Proposed Approach: FDM for WB Correction

To address the limitations of both traditional and recent AWB methods, this dissertation introduces a novel approach that models lighting as a style factor and corrects white balance through feature distribution matching. Rather than relying on global illumination estimates or heuristic-based assumptions, this method formulates WB correction as a feature alignment problem, ensuring that color distributions are adaptively corrected based on learned representations of illuminant styles.

The proposed framework leverages deep learning architectures to explicitly model the statistical properties of feature distributions influenced by lighting conditions. In particular, it employs Exact Feature Distribution Matching (EFDM) [30] to match the empirical distributions of image features, aligning them with a white-balanced reference. This approach allows for precise color correction by capturing higher-order statistics beyond simple mean and variance adjustments, addressing complex illuminant shifts that traditional methods struggle with.

A key advantage of this methodology is its robustness in multi-illuminant environments, where different light sources interact within the scene. Unlike prior methods that rely on a single illuminant assumption, the proposed approach dynamically learns and applies per-region corrections, preserving spatial consistency and reducing overcorrection artifacts. This is achieved through a combination of architectural improvements,

such as style modulation within a UNet-like framework (Style WB) and feature distribution matching integrated into a Transformer-based architecture (FDM WB). Furthermore, this study introduces a distribution-based loss function (FDM Loss), which ensures that the predicted white-balanced image adheres closely to the ground-truth distributions in various datasets, including Cube+ and Large-scale Multi-Illuminant (LSMI).

By adopting a feature distribution matching perspective, this dissertation not only enhances the adaptability of WB correction but also contributes to a broader understanding of how lighting conditions can be represented and corrected within deep learning frameworks. The proposed methods set a new benchmark for robust and perceptually accurate WB correction, which offers a scalable and efficient solution for real-world imaging applications.

1.2 Research Objectives

This dissertation aims to establish a robust and adaptable framework for WB correction by modeling lighting conditions as style factors and employing deep learning architectures that integrate feature distribution matching. Traditional WB correction techniques often struggle with complex illumination scenarios, particularly in multi-illuminant environments where standard statistical assumptions break down. This study proposes an alternative paradigm by treating lighting as a style factor and developing a novel feature alignment methodology that ensures perceptually consistent WB correction across diverse imaging conditions.

The objectives of this dissertation are structured to address both theoretical and practical challenges in WB correction, focusing on enhancing accuracy, adaptability, and computational efficiency. The research unfolds through the following core objectives:

i. **Establishing a Style Factor Learning Strategy for WB Correction**

The primary objective of this study is to develop a learning framework that models

lighting as a style factor by leveraging EFDM. This framework is designed to learn and infer the underlying statistical properties of different illuminants, allowing for adaptive WB correction. By incorporating feature alignment techniques, such as adaptive normalization mechanisms and deep feature statistics, the model can effectively neutralize undesired illumination-induced color distortions while preserving the natural structure of the scene.

ii. **Addressing the Limitations of Existing WB Correction Approaches**

A fundamental challenge in existing deep learning-based AWB methods is their reliance on implicit feature learning without explicitly considering feature distribution shifts caused by complex lighting variations. This dissertation aims to systematically investigate the limitations and proposes a more structured approach by explicitly incorporating distribution matching into the optimization framework.

iii. **Enhancing Multi-Illuminant Robustness through Feature Distribution Matching-Based Optimization**

Traditional methods often fail in multi-illuminant scenarios, where multiple light sources with different spectral properties create regionally varying color casts. The proposed framework aims to address this limitation by incorporating a feature distribution matching-based objective function into the optimization process. By enforcing distribution-level consistency, the framework can reduce the risk of overcorrection, undercorrection, or spatial inconsistencies in complex lighting conditions.

iv. **Conducting Extensive Experimental Validation for Performance Assessment**

This study aims to rigorously evaluate the effectiveness of the proposed WB correction framework through extensive experiments on benchmark datasets such as Cube+ and LSMI, covering single- and multi-illuminant conditions. In addition, qualitative assessments will be conducted to analyze perceptual consistency and

color fidelity in various lighting scenarios. The objective is to demonstrate the advantages of feature distribution matching-based WB correction over both traditional and recent deep learning-based approaches.

v. **Developing a Computationally Efficient and Resolution-Independent Version**
This study also aims to design a WB correction framework that maintains robust performance across varying image resolutions while ensuring computational efficiency for real-time applications. To achieve this, we explore resolution-independent feature learning strategies and integrate deterministic neural illumination mapping to optimize feature representations while minimizing computational overhead. The objective is to improve scalability to enable deployment in resource-constrained environments such as edge devices and mobile imaging applications.

1.3 Thesis Contribution

This dissertation presents significant contributions to the field of WB correction by introducing a novel paradigm that models lighting as a style factor, influencing the color distribution of a scene. By incorporating feature distribution matching into the optimization process, this approach enhances the robustness, accuracy, and adaptability of WB correction, particularly in complex and multi-illuminant environments. The key contributions of this work are as follows.

i. **Style Factor Learning for WB Correction**
This dissertation establishes a novel approach to WB correction by treating lighting as a style factor that influences the color distribution of the entire scene. By leveraging this perspective, the proposed method enables adaptive color correction in diverse illumination conditions, significantly improving robustness in complex lighting environments.

ii. **Feature Distribution Matching-Based WB Correction**

A key contribution of this work is the integration of EFDM into the WB correction pipeline. This approach ensures that the color distribution of an input image is exactly aligned with an ideal white-balanced reference at multiple stages of the neural network architecture and optimization process to improve the overall accuracy and perceptual consistency of WB correction.

iii. **A Robust Framework for Multi-Illuminant Environments**

This dissertation introduces a feature distribution-based objective function that explicitly accounts for complex lighting interactions to effectively handle variations in illuminant spectral properties.

iv. **Enhanced Accuracy and Real-World Adaptability**

The proposed approach surpasses state-of-the-art WB correction methods by integrating deep learning architectures with style-based illumination modeling. Experimental validation demonstrates that the proposed methods achieve superior accuracy and perceptual consistency across multiple benchmarks, making them applicable to imaging scenarios where existing WB correction methods fall short.

v. **Deterministic Neural Illumination Mapping for Efficient Feature Learning**

To improve computational efficiency and enable deployment in resource-constrained environments, this dissertation explores the application of deterministic pixel mapping for WB correction. This approach optimizes feature learning by ensuring resolution-independent processing, reducing computational overhead while maintaining robust correction across diverse imaging conditions.

vi. **New Paradigms for Style Representation in Image Restoration**

By formulating lighting as a style factor and integrating feature distribution matching into neural network architectures and the optimization process, this dissertation

introduces a new paradigm for WB correction. This contribution extends beyond WB correction, paving the way for future research in image restoration and enhancement, where feature distribution modeling can be applied to a broader range of imaging tasks.

The remainder of the dissertation is structured as follows. Section 2 provides an overview of the ISP pipeline, focusing on WB correction, feature distributions, and distribution matching. Section 3 reviews traditional AWB methods and explores the evolution of style-based learning and deep learning approaches in image processing. Section 4 presents the proposed methodologies, namely Style WB and Feature Distribution Matching White Balancing (FDM WB), along with the extended version of FDM WB, including improvements in architecture and optimization. Section 5 details the training setup, including datasets, training strategies, and hyperparameters, as well as the experimental setup and evaluation metrics. It also presents the experimental results, comparing the proposed methods to deep learning-based WB correction methods on various datasets, with quantitative and qualitative evaluations. Section 6 covers additional applications, including DeNIM for efficient feature learning and the impact of the proposed methods on night photography, and the potential future of the main idea using advanced methods, such as diffusion-based techniques and flow matching.

2. BACKGROUND

2.1 Style Factors and Representation

In perceptual systems, style factors refer to extrinsic attributes that alter the appearance of an image without modifying its core content. These factors can include elements such as lighting, texture, and color tone, which are often separated from the content that represents structural or intrinsic information in the scene. Human perceptual systems, and increasingly artificial models, are designed to distinguish between these style and content factors when interpreting visual information [31].

Examples of this distinction are found in daily life: words spoken with an unfamiliar accent (*i.e.*, style) still convey the same meaning (*i.e.*, content), letters written in different handwriting styles retain their textual meaning, and objects illuminated under varying lighting conditions are perceived consistently despite changes in their visual appearance. These illustrate how style factors are integrated into the content of audio, text, and images. In visual perception, lighting is a key style factor that modifies how objects are perceived, although the objects themselves remain unchanged.

Earlier studies have approached the separation of style from content using computational models that provide expressive representations of these factors [32, 33, 34, 31]. These models describe factors as well-defined representations of observations. In the context of natural images, separating content from style is particularly challenging. Convolution-based neural architectures are effective in producing generic feature representations that allow content and style to be processed independently. This ability has been applied to various tasks, such as texture recognition and synthesis [35, 36, 37, 38],

artistic style classification [39, 40, 41], style transfer [42, 43, 44, 30], and generative image synthesis [45, 46, 47]. These works demonstrate that style representation can be distilled by forming specific feature spaces for images through learning objectives designed to isolate these extrinsic factors.

The concept of style can be interpreted in different ways depending on the domain. For example, in face images, style might represent attributes such as age, type of haircut, or whether the person is wearing glasses. These style factors can be extracted as affine parameters within a feature space, packed together to represent different visual attributes. The mapping network extracts these parameters using random vectors or features from pretrained networks (*e.g.*, VGG [48], ViT [4]) to manipulate the image style. In another example, style may refer to the painting style of an artist [42] or filters applied to a photograph [41], where the style factor is captured through the correlation between features, allowing direct manipulation of style or the removal of applied filters.

Based on these insights, any disruptive or modifying factor that influences the entire image can be modeled as a style factor. This notion is central to this study, where it extends the concept of style to the domain of lighting and illumination. This study conducted on AWB correction proposes treating lighting as a style factor that affects the color distribution of an image. By modeling lighting in this way, the AWB correction process becomes one of identifying and neutralizing this style factor to correct the image color balance without affecting the content [24].

This approach enables deep learning models to handle color distortions by recognizing and adjusting lighting as a style factor and then aligning the color distributions of the image with a neutral reference. By addressing lighting as a stylistic attribute, we can achieve robust AWB correction that handles complex, multi-illuminant environments—an area where traditional AWB methods often struggle. This approach not only ensures accurate color correction but also maintains perceptual consistency, aligning the output with

human visual expectations. In essence, modeling lighting as a style factor creates a powerful framework for robust AWB correction, forming the foundation for this dissertation's contributions to real-world image processing challenges.

2.2 Distribution Alignment

In image processing, effectively handling variations in color distribution caused by different lighting conditions is critical to achieve perceptual consistency. Techniques like feature distribution alignment and matching have been utilized in various domains within deep learning; however, their usage in image processing-related tasks (*e.g.*, AWB correction) remains relatively unexplored. This dissertation introduces a novel application of these methods for AWB correction, which demonstrates how matching the feature distributions of an image during the optimization of learning in deep architectures can significantly improve color correction under varying illumination conditions. By treating lighting as a style factor that influences the overall color distribution, we leverage feature distribution matching to ensure robust corrections that align with human perceptual expectations.

2.2.1 Deep Feature Statistics

Deep feature statistics refer to the quantitative measures of the feature representations extracted by pre-trained deep learning models, which describe the distribution and variability of features represented in the high-dimensional latent space. These statistics encompass various metrics, such as mean, variance, skewness, kurtosis, and higher-order moments, along with the correlation and co-occurrence patterns among features. These measures offer a deep understanding of how visual features are distributed and how they vary under different conditions.

For a feature map $F \in \mathbb{R}^{C\times H\times W}$, where C represents the number of channels,

and $H \times W$ is the spatial resolution, the channel-wise mean μ_c and variance σ_c for each channel c are computed as

$$\mu_c = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} F_{c,h,w} \tag{2.1}$$

$$\sigma_c = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} (F_{c,h,w} - \mu_c)^2 \tag{2.2}$$

where μ_c and σ_c represent the first and second-order statistics (*i.e.*, mean and variance), respectively, which describe the distribution of feature values in a given channel.

In addition to mean and variance, skewness γ_1 and kurtosis γ_2 provide higher-order insights into the feature distributions. The former describes the asymmetry of the distribution, while the latter measures the tailedness or extremity of values.

The skewness γ_1 of a feature map F can be calculated as

$$\gamma_1 = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} \left(\frac{F_{c,h,w} - \mu_c}{\sigma_c} \right)^3 \tag{2.3}$$

The kurtosis γ_2, which measures the tailedness of the distribution, is defined as

$$\gamma_2 = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} \left(\frac{F_{c,h,w} - \mu_c}{\sigma_c} \right)^4 - 3 \tag{2.4}$$

Here, the subtraction of 3 normalizes the kurtosis so that the kurtosis of a normal distribution equals zero, which makes it easier to compare different distributions. Skewness and kurtosis allow for a more detailed understanding of the distribution of feature maps beyond what mean and variance provide. This makes them valuable for fine-grained adjustments in tasks related to the color space distributions.

Using these low-order feature statistics, deep learning models can capture how style factors such as artistic painting style or dominant pattern affect the overall appearance of an image [42, 37]. Additionally, higher-order moments such as skewness and kurtosis can provide further insight into the asymmetry and extremity of the tails of the

feature distributions, which make them useful for fine-grained adjustments in learning the representation of the style factor [30].

These statistics serve as the foundation for methods like feature alignment and distribution matching, which aim to correct shifts in any distribution introduced by varying conditions for different factors. As discussed in the following sections, these statistics also provide valuable insight into how lighting conditions, as style factors, affect the color distribution in an image. Deep learning models rely on these statistics to represent the color characteristics of the scene, and any shifts in the distribution can be attributed to lighting effects, which must be corrected for accurate AWB correction.

2.2.2 Feature Alignment

Feature alignment is widely used in deep learning methods to reduce disparities between feature distributions caused by input variations. Techniques such as batch normalization [49], layer normalization [50], and instance normalization [51] are common methods for aligning features in different domains and ensure that feature distributions remain consistent during training and inference. A specialized technique, Adaptive Instance Normalization (AdaIN) [44], goes one further step by separating and aligning the mean and standard deviation in the feature maps between the inputs of content and style. This approach allows for the adaptive transformation of specific features in the latent space while maintaining other aspects of the representation.

AdaIN is mathematically defined as

$$\text{AdaIN}(x, y) = \sigma(y) \left(\frac{x - \mu(x)}{\sigma(x)} \right) + \mu(y) \tag{2.5}$$

where x and y are the feature maps of the content image and the style input, respectively. This normalization technique aligns the channel-wise mean μ and variance σ of the content feature maps to those of the style feature maps.

Our foundational work [41] introduces the reverse style transfer technique, in

which the style information injected into the content is systematically removed or modified. This preliminary idea, which forms the basis of the current thesis, incorporates Adaptive Instance Normalization (AdaIN) across all layers of the feature encoder while adjusting the learning objectives to treat internal style factors, such as Instagram filters, as external elements to be discarded or corrected. This method proved to be particularly useful in tasks such as filter removal or image restoration, where style factors, such as Instagram filters, can be neutralized to restore content in the original scene.

The development of this idea highlighted the limitations of the Gaussian assumptions investigated in our first prior work in the WB correction [24], which often oversimplify real-world scenarios. These assumptions fail to capture the fine-grained variations introduced by the distinctive style factors. To address these challenges, the thesis extends these foundational works by introducing the use of EFDM to AWB correction, which is a more advanced technique that relaxes the Gaussian assumption and matches the entire empirical distribution of feature maps [28, 29], as detailed in the following section.

2.2.3 Exact Feature Distribution Matching (EFDM)

To address the limitations of Gaussian-based assumptions made by methods such as [44], EFDM [30] was proposed as a more advanced technique for feature alignment. EFDM focuses on directly matching Empirical Cumulative Distribution Functions (eCDF) of the feature maps, thus ensuring that not only the first- and second-order statistics (*that is*, mean and variance) but also higher-order statistics such as skewness and kurtosis are inevitably aligned. This approach provides a more comprehensive method for aligning feature distributions, particularly in scenarios where style factors inject more complex non-linear distortions.

EFDM is grounded in the Glivenko-Cantelli theorem [52], which states that the

Algorithm 1 PyTorch-like pseudo-code for EFDM.

$\mathbf{X}$: input vector, $\mathbf{Y}$: target vector
_, IndexX = torch.sort($\mathbf{X}$)
SortedY, _ = torch.sort($\mathbf{Y}$)
InverseIndex = IndexX.argsort(-1)
return $\mathbf{X}$+ SortedY.gather(-1, InverseIndex) $-\mathbf{X}$.detach()

eCDF of a random variable converges uniformly to the true Cumulative Distribution Function (CDF) as the sample size approaches infinity.

$$\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0 \text{ as } n \to \infty \tag{2.6}$$

where $\hat{F}_n(x)$ is the empirical CDF, $F(x)$ is the true CDF, and the convergence is uniform as the sample size increases. EFDM applies this concept to ensure that feature maps are matched with precision, which addresses recognizing the variations in various style factors.

The practical implementation of EFDM involves point-wise transformations of the feature maps based on the sorting of their values, thus ensuring that the empirical distributions are aligned. A PyTorch-like pseudocode for EFDM can be written as in Algorithm 1.

This algorithm matches the feature distributions by sorting the values in both the input and target vectors and then applying a pointwise transformation based on the sorted indices. The result is a precise alignment of feature distributions that goes beyond the simpler Gaussian assumptions used by previous methods.

In this thesis, EFDM plays a critical role in correcting color distributions in images affected by complex lighting conditions, ensuring that AWB correction is handled with better accuracy and perceptual consistency. By matching the exact distributions of the feature maps, EFDM enables a more robust adjustment of the lighting effects, leading to perceptually consistent results even in challenging scenarios.

2.3 Transformers Architecture

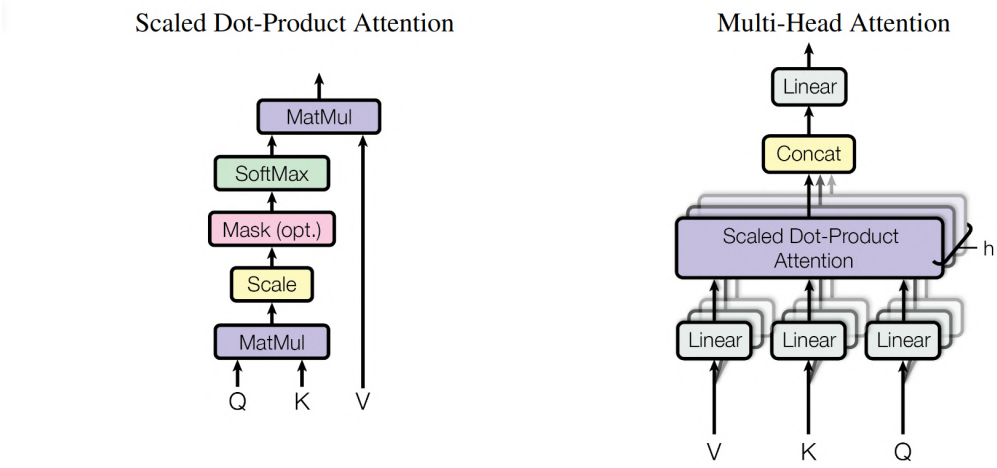
The Transformer architecture, first introduced by Vaswani *et al.* [3], is a highly influential model in deep learning, particularly in Natural Language Processing (NLP) tasks. Unlike earlier models such as the Recurrent Neural Network (RNN) [53, 54] or CNN architectures, Transformers use the self-attention mechanism to process input data more effectively by capturing both long-range dependencies and global context. The architecture is unique in its ability to weigh the importance of different elements in the input sequence, regardless of their positions. With this ability, this mechanism can be more powerful for modeling relationships in various types of sequences than the previous methods.

2.3.1 Multi-Head Self-Attention (MSA)

At the core of the Transformer architecture is the Multi-Head Self-Attention (MSA) mechanism, which calculates the relevance of one token in a sequence to all others, allowing the model to capture contextual relationships. The self-attention mechanism uses three key components: query (Q), key (K), and value (V) vectors. These vectors are linear transformations of the input and their relationships are crucial to determining how much influence one token should have over other tokens. The attention score between two tokens is calculated by taking the dot product of the Q and K vectors, scaling by the square root of the dimension of the K vectors, and then passing the result through a softmax function before multiplying it with the V vectors. This multiplication gives the weighted importance of each token to the others in the sequence. The formula for the self-attention mechanism can be seen as follows

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.7}$$

Figure 2.1: *(Left) Scaled Dot-Product Attention mechanism. (Right) Multi-head attention comprises multiple attention layers operating in parallel. The figure is obtained from [3].*



where where Q represents the query matrix, K the key matrix, V the value matrix, and d_k is the dimension of the key vectors. The final output of the self-attention mechanism is a weighted sum of the value vectors, where the weights are determined by the similarity between the queries and keys.

The MSA mechanism enhances this process by allowing multiple self-attention heads to operate in parallel. Each head focuses on different subspaces of the feature representations, which can capture different relationships in the data. The outputs of all the heads are concatenated and projected through a linear transformation, and this can be formally defined as

$$\text{MSA}(Q, K, V) = \text{Concat}(h_1, h_2, \ldots, h_h)W^O \tag{2.8}$$

where h_i represents the attention output of the i_{th} head, and W^O is a learned weight matrix. This mechanism allows the Transformer to capture both fine-grained and global dependencies in the data. The structure of the MSA mechanism can be seen in Figure 2.1, which illustrates both the scaled dot-product attention and the multi-head attention operating in parallel.

2.3.2 Vision Transformer (ViT)

The Vision Transformer (ViT) architecture [4] treats an image as a sequence of patches, which is similar to the way words are handled in NLP tasks. Instead of relying on convolutional operations, ViT applies the self-attention mechanism to capture long-range dependencies in images. The core innovation of this architecture lies in dividing the input image into patches and treating each one as a token for the Transformer model.

Initially, the image $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ is split into N patches of size $P \times P$. Each patch $\mathbf{X}_p \in \mathbb{R}^{P \times P \times C}$ is flattened into a vector $X_p \in \mathbb{R}^{P^2 \times C}$ and then linearly projected into a higher-dimensional space to create a patch embedding. These patch embeddings, each of size D, form a sequence of tokens, with the number of patches given by $N = \frac{H \cdot W}{P^2}$. Since Transformers do not have the ability to inherently capture positional information, positional embeddings are added to the patch embeddings to retain the spatial relationships between patches. These embeddings allow the model to understand the structure and layout of the image.

A key feature in ViT is the classification token (*i.e.*, denoted as [CLS]), which is appended to the sequence of patch embeddings. The [CLS] token, which provides a global representation of the image, aggregates information from all patches. It not only captures appearance and texture information, but also encodes more global style information such as object parts and high-level features that are pivotal for downstream tasks. In deeper layers, the [CLS] token gradually accumulates more abstract information, which makes it especially useful for tasks like style transfer or appearance-based classification [55].

The set of tokens, including the [CLS] token, is passed through the Transformer encoder, which consists of alternating layers of MSA and Multi-Layer Perceptron (MLP) [56]. The self-attention mechanism in the Transformer enables the model to learn both local and global dependencies across the entire image without any bias toward local structures. This flexibility allows the model to aggregate information from distant parts of the

Figure 2.2: *Vision Transformer (ViT) architecture. The input image is divided into non-overlapping patches, each treated as a token. These patch embeddings, along with a learnable [CLS] token, are processed by the Transformer encoder. Positional embeddings are added to retain spatial information, and the [CLS] token aggregates global information for downstream tasks such as classification. The figure is obtained from [4].*

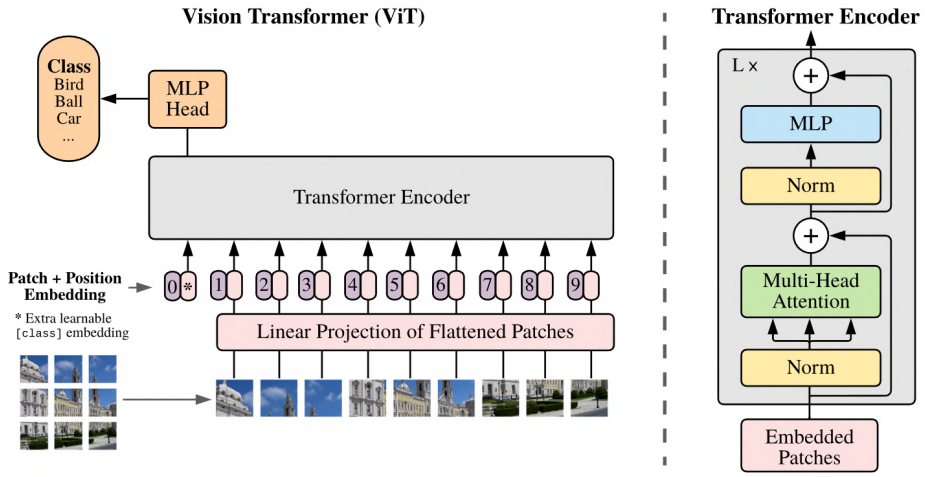


image. Figure 2.2 demonstrates the overall mechanism of ViT architecture.

ViT has demonstrated strong performance on large datasets, such as ImageNet [57] and JFT-300M [58], by leveraging its ability to capture long-range dependencies within images. However, due to the lack of inductive biases such as *translation equivariance* and *locality*, which are inherent in CNN, ViT models typically require substantial training data to achieve competitive results compared to CNN-based architectures.

2.3.2.1 Style Information in ViT

The [CLS] token has shown flexibility in representing 'appearance' across different scenes and images, which mainly captures key visual elements while being invariant to spatial configuration. It has been used effectively in tasks such as semantic appearance transfer, where the appearance information of one image is transferred to another while maintaining structural integrity [55]. Due to its ability to generalize and preserve global

appearance properties while disregarding specific poses or structural configurations, the [CLS] token has proven to be highly effective for modeling style factors in ViT.

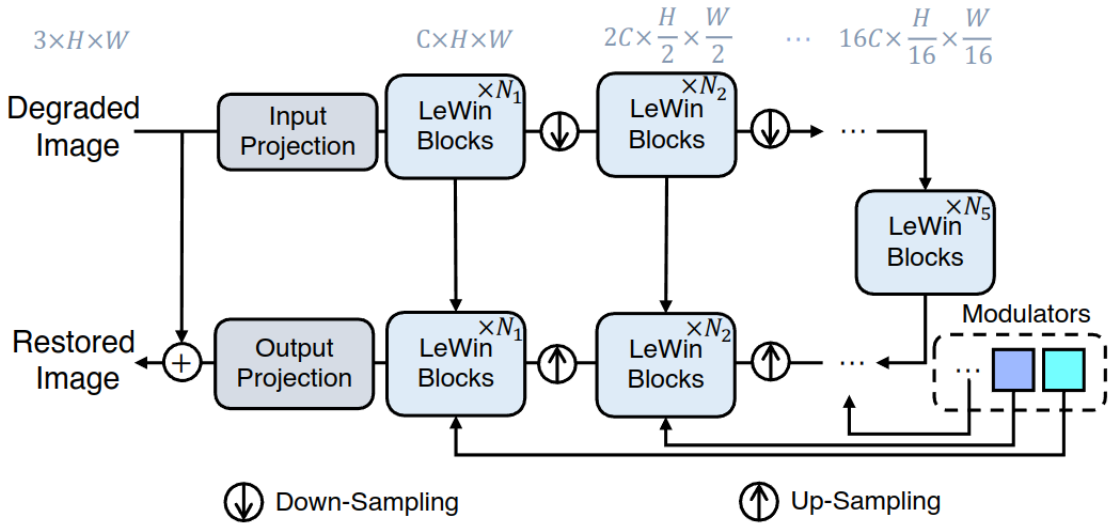
This flexibility of the [CLS] token can be extended to model lighting conditions in images. Since lighting can be considered a style factor, the ability of the [CLS] token to capture global appearance properties makes it ideal for encoding the effects of various illuminants. Using this token during the optimization of style factor learning, our approach can more effectively represent and correct lighting variations across different scenes. This approach allows us to model lighting as a style factor, which not only contributes to accurate WB correction, but also makes our study novel by utilizing ViT, instead of VGG, for the task of modeling complex illumination scenarios. The integration of the [CLS] token for representing the lighting enables the model to handle global and multiple illumination effects while ignoring contextual or structural information in the scene.

2.3.3 Uformer: U-Shaped Transformer

Uformer [5] is a U-shaped Transformer-based architecture designed for image restoration tasks. It integrates the hierarchical structure of U-Net [59] with Transformer blocks, which allows the model to capture both local and global dependencies in the data. Specifically, Uformer consists of an encoder and decoder with skip connections between them, which facilitates the preservation of significant features during downsampling and upsampling operations. The general structure of Uformer is illustrated in Figure 2.3.

The architecture of Uformer introduces two key innovations: the Locally-enhanced Window Transformer Block (LeWin) and the multi-scale restoration modulator.

Figure 2.3: *Uformer architecture, a U-shaped Transformer-based network for image restoration. The figure is obtained from [5].*



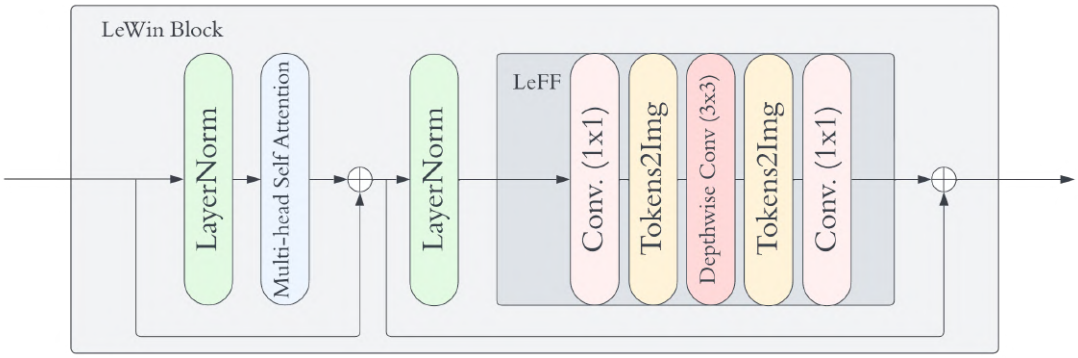
2.3.3.1 Locally-enhanced Window Transformer Block (LeWin)

The LeWin addresses the computational complexity associated with global self-attention, especially in high-resolution images. Instead of global self-attention, it performs window-based self-attention on nonoverlapping windows, which significantly reduces the computational cost while still capturing long-range dependencies. Given the feature maps $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, they are split into non-overlapping windows of size $M \times M$, and self-attention is performed within each window. The self-attention for each window i is computed as

$$\text{Attention}_i(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}} + B_i\right) V_i \tag{2.9}$$

where Q_i, K_i, and V_i represent the queries, keys, and values for the i_th window, d_k is the dimension of K_i, and B_i is the relative position bias specific to window i. The outputs of each window are concatenated and passed through a linear layer. This window-based self-attention reduces the computational complexity from $O(H^2W^2C)$ to $O(M^2HWC)$,

Figure 2.4: *Illustration of LeWin Transformer block, redrawn from [5], by combining LeWin block and Locally-enhanced feed-forward layer (LeFF) in a single figure.*



making it more efficient for high-resolution inputs.

LeWin incorporates a Locally-enhanced Feed-Forward Network (LeFF), as shown in the dark gray part of Figure 2.4, which uses depth-wise convolutions to capture local context to address the limitation of standard Transformers in modeling local dependencies. The output of the LeWin block is computed as

$$\mathbf{X}'_l = \text{W-MSA}(\text{LN}(\mathbf{X}_{l-1})) + \mathbf{X}_{l-1} \tag{2.10}$$

$$\mathbf{X}_l = \text{LeFF}(\text{LN}(\mathbf{X}'_l)) + \mathbf{X}'_l \tag{2.11}$$

where $\mathbf{X}_{l-1}$ and $\mathbf{X}_l$ refer to the feature map in layer $l-1$ and l, LN denotes layer normalization, W-MSA is the window-based multi-head self-attention, and LeFF is the Locally-enhanced Feed-Forward Network.

2.3.3.2 Multi-scale Restoration Modulator

The multi-scale restoration modulator is designed to adjust the feature representations at different scales in the decoder, which aims to facilitate the restoration of finer details in the image. This module introduces learnable bias terms on multiple scales, which are added to the feature maps in each LeWin. These modulators allow the network to adapt its feature representations for different image restoration tasks.

The modulator is a lightweight addition that operates with minimal computational overhead, improving the overall performance of the model on tasks such as denoising, deblurring, and deraining. It enables flexible adjustments of feature maps, which makes Uformer highly effective for image restoration.

2.4 Image Signal Processing (ISP) Pipeline

In digital imaging systems, the ISP is a critical component responsible for transforming the RAW sensor data into a visually coherent and color-accurate image. This pipeline consists of multiple sequential operations, including noise reduction, demosaicing, color space conversion, WB correction, tone mapping, and sharpening, each of which refines the image while preserving perceptual consistency. Given that the RAW sensor data lacks intrinsic color balance, WB correction plays a fundamental role in establishing accurate color perception before subsequent processing steps manipulate contrast, tone, and detail.

To fully understand the role of WB correction within an imaging system, it is essential to analyze its interaction with key components of the ISP pipeline. The ISP applies a sequence of transformations to RAW sensor data, where each stage depends on the accuracy of the preceding operations. Since WB correction is one of the earliest steps, its accuracy directly affects downstream tasks such as color space conversion, tone mapping, and noise reduction. Errors in WB estimation introduce systematic color shifts that propagate through the pipeline, leading to false color artifacts, perceptual inconsistencies, and potential color clipping in the final image.

Furthermore, different ISP modules process chromatic and luminance information differently, which makes the color consistency highly dependent on proper WB correction. For example, non-linear contrast adjustments in tone mapping may amplify color inaccuracies if the WB correction step fails to neutralize unwanted illumination effects.

Similarly, denoising algorithms, which often rely on statistical assumptions about color distributions, may not perform optimally when color channels retain residual illumination biases. These interdependencies highlight the critical role of WB correction in preserving color fidelity across all ISP stages.

The following sections present a detailed overview of key ISP components, describing their role in image formation and explaining how WB correction interacts with these processing stages.

2.4.1 RAW Image Capture and Initial Processing

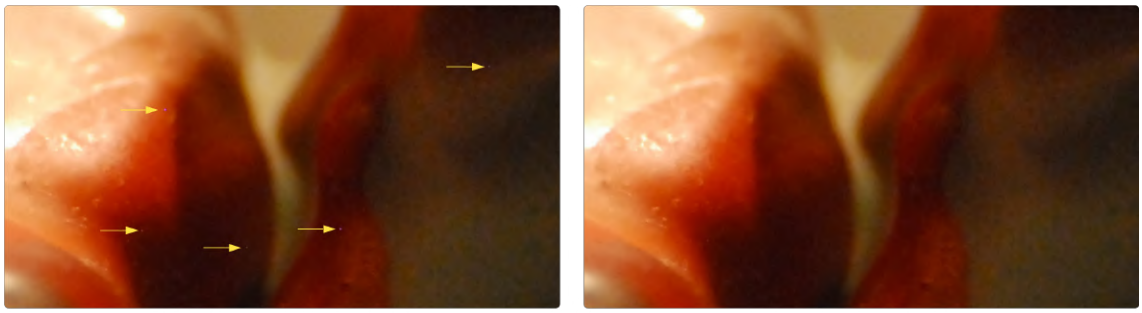
When a digital camera captures an image, the sensor records the light hitting each pixel as a raw intensity value. This raw data, called the RAW image, represents the sensor's most direct output. It contains all the light information without any in-camera processing, such as white balance, noise reduction, or sharpening. RAW images offer high dynamic range and flexibility for post-processing, but they require multiple correction steps before being converted into the final *RGB* image.

A typical RAW image contains imperfections due to the nature of the sensor and its processing environment. To transform this data into a usable form, the following corrections are applied.

2.4.2 Black-Level Correction

The "black level" refers to the sensor's baseline response to zero light exposure. Even in complete darkness, sensors can register nonzero values due to thermal noise or some factors related to the physical properties (*e.g.*, electricity) of the sensor. Black-level correction is the process of subtracting this baseline noise from each pixel to ensure that areas without light are represented as true black in the final image. The black level is usually determined during sensor calibration, and correcting it ensures that the darkest

Figure 2.5: *Visualization of hot pixel correction.*



(a) Before correction (b) After correction

areas in the image are accurately represented.

$$I_{\text{corrected}}(x, y) = I_{\text{RAW}}(x, y) - C_{BL} \tag{2.12}$$

where $I_{\text{RAW}}(x, y)$ represents the raw pixel value where x and y are the spatial locations of the corresponding pixel, and C_{BL} is the black level offset determined after calibration.

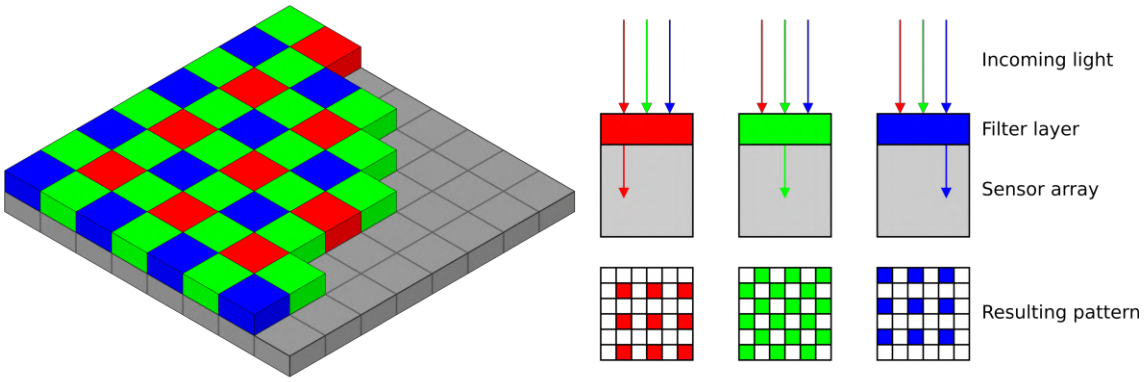
2.4.3 Hot-Pixel Correction

"Hot pixels" are defective pixels that appear brighter than they should, which often show up as isolated white or colored spots in the image, particularly in long exposures or high-temperature conditions. These anomalies may be caused by electrical imperfections or oversensitivity in certain pixels of the sensor. Hot-pixel correction identifies and replaces these outlier pixel values by interpolating from neighboring pixel values to minimize their impact on the image quality. The formula for this correction operation in the horizontal direction can be seen as follows

$$I_{\text{hot-pixel corrected}}(x, y) = \frac{I_{\text{corrected}}(x - \Delta_x, y) + I_{\text{corrected}}(x + \Delta_x, y)}{2} \tag{2.13}$$

where Δ_x is the offset for neighboring pixels in the horizontal direction. This process ensures that outlier pixels do not distort the final image. As seen in Figure 2.5, before the correction, hot pixels appear as bright outliers, which are highlighted by yellow arrows,

Figure 2.6: *Visualization of the Bayer filter and cross-sections of sensor sensitive to different color bands. Credit: Colin M. L. Burnett (CC BY-SA 3.0).*



(a) The Bayer filter

(b) Profile of sensor sensitive to different colors

due to sensor defects or long exposure times. After correction, the outliers are removed, and the pixel values are smoothed using neighboring pixel information, restoring a more natural appearance to the image.

2.4.4 Demosaicing

After black-level and hot-pixel corrections, the next critical step is demosaicing, known as debayering in the literature. Since most camera sensors use a Bayer filter (*i.e.*, a Color Filter Array (CFA)), as shown in Figure 2.6, each pixel in the RAW image only captures one of the three primary colors: red (R), green (G), or blue (B). The sensor layout typically follows a Bayer pattern, where each 2×2 block of pixels contains two green pixels, one red pixel, and one blue pixel. This layout means that each pixel lacks full-color information, which requires interpolation to reconstruct the missing color values at each pixel.

Demosaicing is the process of estimating the missing color values based on the known color values of surrounding pixels. Simple methods, such as bilinear interpolation, can be used, but more advanced techniques, such as gradient-based linear filtering [60] or directional filtering [61], produce better results by preserving fine details like edges and

textures while minimizing color artifacts (*i.e.*, false colors or moiré patterns).

Camera manufacturers have developed specialized color filter arrays that incorporate various improvements tailored to different scenarios, such as enhanced light absorption characteristics of color bands, reduced susceptibility to color moiré, and increased sensitivity to light. These specialized filter arrays often require sensor-specific conversion mapping for the demosaicing process. This means that using the demosaicing method provided by the camera manufacturer *generally* yields better results due to its optimization for that particular sensor and filter design.

2.4.5 Color Space Transformation

In the ISP pipeline, transforming the RAW image data into a standard color space is critical to ensure color consistency across various devices. The sensor data from the camera, which is in a manufacturer-specific native color space, needs to be mapped to a device-independent, standardized color space, (*i.e.*, *XYZ*), before further processing and color correction.

Transformation of RAW image data into a canonical color space is essential to achieve consistent and accurate color reproduction across different devices. The *XYZ* color space, defined by the International Commission on Illumination (CIE), is device-independent and designed to approximate human vision, making it a neutral reference point for color processing [62]. This standardization ensures that colors are represented uniformly on a variety of displays and printers. In addition, many key image processing tasks, such as white balance correction, color grading, and tone mapping, are more consistent when applied in a standardized color space, as it reduces the variability introduced by the unique characteristics of camera sensors and color filter arrays. By converting sensor data to *XYZ*, cameras can more accurately manage color by correcting sensor-specific

variations and enabling perceptually accurate color rendering. This transformation improves the color accuracy before the data is converted into an output color space, such as *sRGB*. Therefore, it ensures that the image aligns with human visual perception and maintains fidelity across different output devices.

Different camera manufacturers utilize unique CFAs and sensors, which leads to variations in the native color space captured by the sensor. For instance, *Sony* sensors may have different sensitivities to light, compared to *Canon* or *Nikon* sensors, while *Fuji*'s *X-Trans* sensors use a distinctive CFA pattern to reduce moiré. As a result, each manufacturer typically provides a custom transformation matrix that maps the native color space of camera to *XYZ*. This matrix accounts for the specific sensor properties, including the filter array and sensor sensitivities. A general transformation from the native camera color space to *XYZ* is represented by

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} R_{\mathrm{n}} \\ G_{\mathrm{n}} \\ B_{\mathrm{n}} \end{pmatrix} \tag{2.14}$$

where $R_{\mathrm{n}}, G_{\mathrm{n}}, B_{\mathrm{n}}$ are the values in the camera's native color space, and the matrix coefficients a_{ij} are specific to the camera model and sensor characteristics.

Once the image data has been transformed into the *XYZ* space, it can easily be mapped to other color spaces, such as *sRGB*, *Adobe RGB*, or *ProPhoto RGB*, depending on the display or output requirements [63]. The transformation into *XYZ* ensures that the color information is correctly represented and ready for further processing or projection into the final output space.

2.4.6 White Balance (WB) Correction

WB correction is one of the most critical stages in the ISP pipeline to ensure that the colors of an image are rendered accurately under various lighting conditions. The main

goal of WB correction is to neutralize any color cast introduced by the light source, so that objects that should appear white in reality also appear white in the image, regardless of the light source's color temperature. Proper WB correction is essential to produce natural images in different environments, such as daylight, incandescent, or fluorescent light.

Light sources have different color temperatures, typically measured in Kelvin (**K**). As demonstrated in Figure 2.7, the color temperature influences the hue of the light, where lower temperatures (*e.g.*, around 3000**K**) produce a warm yellow-orange tint, and higher temperatures (*e.g.*, 6000**K** and above) produce cooler and blueish tones. Without WB correction, images taken under these different lighting conditions would have color casts that distort their appearance. WB correction compensates for these color changes in the image by adjusting the relative intensities of the red, green, and blue channels.

In the RAW image capturing process, the camera sensor does not inherently know the color of the light source. Therefore, it captures color information as-is, which often leads to images with incorrect color casts. WB correction is typically applied in the *XYZ* or *native RGB* color space before transforming the image into a display-ready color space like *sRGB*. This correction ensures that colors are faithfully represented when viewed on screens or printed. The general approach to WB correction is to adjust the color channels so that objects that should be white are neutralized to a uniform gray [13]. This is achieved by scaling *RGB* values on the basis of the color temperature of the estimated illuminant. WB correction relying on Gray-World assumption can be formalized as follows

$$R' = \frac{R}{\frac{1}{N}\sum_{i=1}^{N} R_i}, \quad G' = \frac{G}{\frac{1}{N}\sum_{i=1}^{N} G_i}, \quad B' = \frac{B}{\frac{1}{N}\sum_{i=1}^{N} B_i} \tag{2.15}$$

where R, G, B are the original color values, R', G', B' are the white-balanced red, green, and blue values, and N is the number of pixels in the image.

Figure 2.7: *Color temperatures influence the hue of the light.*

(a) Tungsten (b) Shade (c) Original

2.4.7 Gamma Correction

Gamma correction is a non-linear transformation applied to the image to adjust the luminance levels. Camera sensors capture light linearly, and this means that the brightness values are proportional to the amount of light hitting the sensor. However, human vision is non-linear, thus being more sensitive to changes in dark areas than in bright areas [64]. Gamma correction compensates for this by applying a nonlinear transformation to the brightness values, which makes the image more perceptually accurate. The transformation can be described as

$$I_{\text{gamma}} = I_{\text{WB}}^{\frac{1}{\gamma}} \tag{2.16}$$

where γ is the correction factor, which is typically set to 2.2 or 2.4 for *sRGB* displays.

2.4.8 Tone Mapping

Digital camera sensors can capture a wide range of brightness values in a scene, often exceeding the display capabilities of common devices. High Dynamic Range (HDR) images encode this wide range of luminance, but most consumer displays operate within a limited dynamic range due to physical constraints. Tone mapping is a key process in the ISP pipeline, particularly when handling images captured in HDR. The primary goal of tone mapping is to compress the wide dynamic range of luminance values present in HDR images into a range that can be displayed on conventional Low Dynamic Range (LDR)

Figure 2.8: *Illustration of tone mapping effects on an HDR image.*

(a) Under-exposed scene (b) Corrected scene (c) Over-exposed scene

devices (*i.e.*, standard monitors or mobile screens) while maintaining important visual details across both bright and dark regions of the scene. Without this operation, attempting to display an HDR image directly on these devices would lead to significant clipping, where bright areas become overexposed and details in dark areas may be lost, as illustrated in Figure 2.8.

In general, tone mapping can be viewed as a transformation applied to the luminance values of the image, formalized as

$$I_{\text{tone}} = TMO(I_{\text{HDR}}) \tag{2.17}$$

where I_{tone} represents the tone-mapped image, I_{HDR} represents the original high dynamic range image, and TMO refers to the tone mapping operator. Examples of tone mapping operators include Drago [65], Reinhard [66], Mantiuk [67], Flash [68], and Dawn [69].

2.4.9 Post-Processing Operations

After key stages like demosaicing, WB correction, and tone mapping, the final steps in the ISP pipeline involve a series of post-processing operations that aim to improve the overall quality and aesthetics of the image. These operations, which are often optional and can be included or excluded based on the specific purpose of the imaging, include denoising and sharpening. The former reduces unwanted noise caused by low-light conditions or sensor limitations while the latter aims to enhance edges and textures

to improve clarity. Additionally, memory color enhancement [70] can be applied to boost colors that are commonly perceived in certain objects, such as the blue of the sky or the green of the grass. This operation leads to aligning the image with the way viewers expect certain objects to appear. Further enhancements, such as contrast adjustments, vignette correction, and color grading, are often applied to fine-tune brightness, shadows, or saturation, to create a visually appealing final output. Examples of these adjustments include increasing contrast for better depth or applying vibrance to intensify muted colors without oversaturating the overall image. These post-processing steps are optional and collectively ensure that the image is polished and ready for display across various devices, depending on the imaging goals.

3. LITERATURE REVIEW

WB correction is a fundamental process in digital imaging aimed at removing color casts caused by scene illumination. Various approaches have been proposed over the years, ranging from traditional, physically motivated algorithms to more recent deep learning-based methods. This section reviews the key methods in illuminant estimation and WB correction, highlighting their contributions, methodologies, and limitations. Table 3.1 summarizes the comparative analysis of various methods, highlighting their adaptability, computational cost, robustness to multi-illuminant conditions, accuracy, real-time capability, and additional comments.

3.1 Traditional Methods

One of the earliest methods, introduced by Buchsbaum [13], estimates the illumination of the scene by calculating the spatial average reflectance under the assumption that the illuminant is *uniform* across the entire scene (*i.e.*, the Gray-world assumption). Although simple, this approach forms the basis for subsequent advancements. Building upon this, Brainard and Wandell [14] proposed an analysis on the Retinex Theory [71] (*i.e.*, the maximum response in the *RGB*-channels is caused by a perfect reflectance), which estimates illumination by computing *lightness* values invariant to lighting conditions, thus offering robustness in more complex scenes. Gershon *et al.* [72] introduced an improved approach to this assumption by *segmenting* the image and computing the average color for each segment, rather than for all pixels. This method mitigates the sensitivity to large uniformly colored areas injected in earlier algorithms, which often violates the assumption of a balanced color scene.

In the evolution of the Gray-world assumption, Barnard *et al.* [73] introduced

the General Gray-World hypothesis, which extends the basic concept by improving *reflectance estimates* across varying surfaces and lighting environments. This method corrected some of the limitations of earlier techniques by offering more robust illuminant estimations under real-world conditions. Finlayson and Trezzi [15] later demonstrated that these methods (*i.e.*, Gray-world and Retinex) could be viewed as *extremes* of the Minkowski family norm (*i.e.*, $\mathcal{L}_1$ and $\mathcal{L}_\infty$), which suggests that intermediate values can provide more accurate illumination estimates. Xiong *et al.* [74] proposed an extension to the Gray-World assumption, which first identifies colors likely to originate from *real gray surfaces*, then averages only those colors.

A surface with perfect reflectance properties reflects the full spectrum of the captured light, which means its color is exactly that of the light source [75]. The max-*RGB* algorithm alleviates the assumption of perfect reflectance by estimating the illuminant through the maximum response in each color channel separately. Related approaches, such as those of Gijsenij and Gevers [76] and Ebner [77], apply averaging (*i.e.*, sort of *smoothing*) before the illuminant estimation, which reduces the impact of noisy pixels and thus improves the accuracy of the white-patch algorithm. Funt and Shi [78, 79] further analyze the max-*RGB* algorithm, which demonstrates that both *dynamic range* and *pre-processing* strategies significantly affect the overall performance.

The Gray-Edge hypothesis by Van de Weijer *et al.* [80] introduced the idea that the largest variation in *color derivatives* corresponds to the direction of the light source, which can be estimated using the Minkowski norm of these derivatives. Chakrabarti *et al.* [81] presented an approach that explicitly models *spatial* dependencies between pixels, where it offers a more efficient way to capture these relationships compared to the Gray-Edge hypothesis. This approach allows the model to learn pixel dependencies more effectively for improved color constancy. Gijsenij *et al.* [82] further improved the edge-based color constancy by computing a weighted average of various *edge* types, which improves the

overall accuracy in edge-dominated scenes. Joze *et al.* [83] demonstrated that selecting the *brightest* 20% of the pixels yields superior results across existing methods (*i.e.*, Bright pixels). Cheng *et al.* [12] took another *spatial* approach, where the study focused on analyzing the relationship between spatial and color information for better illumination estimation.

3.2 Gamut Mapping Solutions

Gamut-based solutions are based on the principle that only a limited range of colors (*i.e.*, gamut) can be observed under a specific illuminant. Forsyth [84] originally introduced the gamut mapping algorithm, which assumes that, for a given illuminant, one observes only a subset of all possible colors, and any deviation from the canonical gamut indicates a shift in the light source. Later works, such as Finlayson *et al.* [85] and Finlayson *et al.* [86], introduced extensions that relax the assumption operating under a *diagonal model* to improve robustness in cases where it fails, such as by augmenting the canonical gamut or incorporating nonlinear transformations.

Following early work on gamut mapping solutions, Finlayson and Hordley [87] introduced a *gamut-based* constraints solution to estimate illumination by mapping diagonal matrices from unknown lighting conditions to reference colors. Gijsenij *et al.* [88] extended this by incorporating *linear filter* output, instead of the diagonal model, which improves the robustness of the solution under varying lighting conditions. Lastly, Mosny *et al.* [89] proposed a simplified version of the gamut mapping algorithm by using a simple cube representation of pixel values, rather than the more complex convex hull. This approach reduces computational complexity while still providing an effective illuminant estimate.

3.3 Low-level Statistical Methods

Color by correlation, introduced by Finlayson *et al.* [90], replaces the canonical gamut with a *correlation matrix*, which splits the chromaticity space into cells and calculates the probability of occurrence under each illuminant, which is then matched to the input image to estimate the most likely light source. Rosenberg *et al.* [91] extended this method by incorporating *Kullback–Leibler divergence* to select the scene illuminant based on the divergence between the correlation matrix of the input image and those of possible light sources.

Bayesian methods have also played a significant role in WB correction. Brainard and Freeman [92] developed a Bayesian model that captures the relationship among illuminants, surfaces, and photo-sensor responses where the prior distributions are used to describe the likelihood of specific illuminants and surfaces in the scene. Sapiro [93] proposed a framework for estimating illuminant and reflectance in natural images using the generalized probabilistic Hough transform. Each pixel in the image casts a *vote* for potential illuminants, and the final estimation is determined by aggregating these votes, providing a robust solution for illumination estimation. Moreover, another Bayesian approach, proposed by Tsin *et al.* [94] to classify outdoor scenes, which incorporates a likelihood model that accounts for the physics of image formation, sensor noise distribution, and prior distributions over geometry, material types, and illuminant spectrum parameters. Rosenberg *et al.* [95] presents the Bayesian approach that relaxes the assumption of Gaussian-distributed reflectance factors by employing a non-Gaussian probabilistic model for the image formation process. Lastly, Gehler *et al.* [96] analyzed that precise priors for illumination and reflectance can achieve competitive results compared to state-of-the-art methods.

Basic machine learning techniques such as *support vector regression* have been used by Xiong and Funt [97], and Wang *et al.* [98], which utilize statistical models to

estimate the illuminant based on the input data. Similarly, Agarwal *et al.* [99] have explored linear regression techniques such as *ridge regression* and *kernel regression*, which offer efficient solutions to estimate the illumination of the scene. An alternative approach by Xiong *et al.* [100] employs *thin-plate spline interpolation* for illuminant estimation, which interpolates the color of the light source over a non-uniformly sampled input space (*e.g.*, a collection of training images).

In addition, studies have shown that traditional methods can be enhanced by incorporating various statistical techniques for improvement. Gao *et al.* [101] proposed the Locally Normalized Reflectance Estimation (LNRE) method, inspired by *retinal feedback mechanisms*. By normalizing local patches and computing the ratio of global intensity summations, they estimate the illuminant with minimal computational cost and only one free parameter. Afifi *et al.* [102] introduced a projective transformation approach for post-estimate *bias correction*, which adapts to the input *RGB* vector to improve illumination estimation accuracy. Banić *et al.* [103] demonstrated fine-tuning of illumination estimation parameters using only *non-calibrated images*, (*i.e.*, the green stability assumption), which can allow us to eliminate the need for time-consuming sensor-specific calibration without any significant performance loss compared to training on calibrated images. Qian *et al.* [104] proposed the mean-shifted gray pixel method, which statistically approximates pixels assumed to be *gray* under neutral illumination. Extending this idea, Qian *et al.* [105] developed the Grayness Index (GI) using Shafer's Dichromatic Reflection Model (DRM) [106], which mainly allows for the *ranking* of pixels by their grayness and thus improves multi-illuminant estimation.

3.4 Methods Using Scene Semantics

Gijsenij and Gevers [107] proposed a dynamic selection approach for WB correction algorithms, which chooses the appropriate method based on the known scene semantics of the image. Building on this idea, Bianco *et al.* [108] developed a method to classify scenes into three categories—indoor, outdoor, and unsure—and learn the optimal correction algorithm for each. Lu *et al.* [109] introduced a method that uses *3D geometry* models to classify images into stages, segment them into different regions with hard and soft segmentation, and select the optimal color constancy algorithm for each geometrical segment, enabling light source estimation to be adapted to the geometry of the entire scene. Van De Weijer *et al.* [110] proposed a method that improves illuminant estimation by applying several approaches to compute the possible set of illuminants, then selecting the one that results in the most semantically likely image—based on *prior knowledge of the world*—by modeling the image as a mixture of semantic classes (*e.g.*, sky, grass, road) using probabilistic latent semantic analysis. Rahtu *et al.* [111] extended this approach by introducing the concept of memory color, which refers to colors that are specifically associated with certain object categories, further improving the accuracy of illuminant estimation.

3.5 Neural Networks

In recent years, neural networks have become increasingly prominent in illuminant estimation and WB correction, which offers data-driven approaches that use the ability of deep learning to capture complex relationships between scene chromaticity and lighting conditions. Cardei *et al.* [112] introduced one of the first approaches to illuminant estimation using neural networks, where the input to the network is a binarized *chromaticity histogram* of the input image, and the output consists of two chromaticity values representing the estimated illuminant. Following this, Stanikunas *et al.* [113] proposed

an approach where the neural network calculates *color differences* between foreground and background factors, using a color vector as the output signal. By being trained with the backpropagation algorithm, the network was designed to identify the color of Munsell samples [114] under varying illuminants to adapt the network to diverse lighting conditions.

Lou *et al.* [115] reformulated WB correction as a Deep Neural Network (DNN)-based *regression* task, which estimates the color of the light source directly. Addressing limitations in previous methods, it notes that traditional approaches rely on specific assumptions that prevent them from serving as universal predictors. Bianco *et al.* [116] introduced a CNN-based architecture for estimating scene illumination directly from image patches in the *spatial domain*, which aims to eliminate the reliance on hand-crafted features used in previous work. This network contains a convolutional layer with max pooling, a fully connected layer, and three output nodes, and also combines feature learning and regression within a unified optimization. In parallel with [116], Barron [16] reformulated WB correction as a 2-dimensional spatial localization task in *log-chrominance space*, which goes beyond traditional statistical modeling of natural object colors and illuminates.

Shi *et al.* [117] proposed a more advanced network architecture, namely DS-Net, to address estimation ambiguities in color constancy. Their model includes two interacting sub-networks: *HypNet*, a two-branch network that generates multiple illumination hypotheses to capture various illuminant modes, and *SelNet*, which adaptively selects the best estimate among these hypotheses. On the other hand, Bianco *et al.* [118] presented a three-stage method for illuminant estimation in *RAW* images, which combines CNN-generated local estimates with a *support vector regressor* for refinement, which adapts to single and multiple illuminant scenes. This approach advances the overall performance in illuminant estimation by effectively integrating local and global estimations through

non-linear aggregation.

Oh *et al.* [119] proposed a deep learning approach to estimate scene illumination by considering the color constancy problem as an *illumination classification* task where CNN-based model is designed to directly compute the illuminant color under uniform lighting, which demonstrates superior feature extraction for illumination estimation. Hu *et al.* [17] addressed estimation ambiguity in patch-based CNNs for color constancy by introducing a *fully convolutional* network that applies confidence weights to patches, which enhances overall performance. Their contribution is related to adding a custom pooling layer that merges local estimates into a global solution, and thus allowing the network to automatically learn what to learn and how to pool without requiring additional supervision. Afifi *et al.* [6] leveraged the *k-nearest neighbor* strategy for correcting improperly white-balanced images by identifying similar examples within a large dataset. From these examples, the method effectively removes color casts by constructing a non-linear color correction transform to be applied. Afifi *et al.* [120] investigated how strong color casts from improperly applied WB affect neural network performance in *downstream vision tasks* (*i.e.*, image classification and segmentation).

Recent advances in WB correction have introduced various optimization and learning strategies, which aim to improve the overall performance under complex and diverse lighting conditions. Banić *et al.* [1] introduced an unsupervised learning method that estimates parameters without calibrated ground truth data, which effectively allows *inter-camera adaptation* and eliminates the need for sensor-specific calibration. Hernandez-Juarez *et al.* [121] proposed a Bayesian multi-hypothesis framework that applies multiple candidate illuminants to a scene and then learns an achromatic likelihood model via a camera-agnostic CNN. This enables effective multi-camera training and improves sensor generalization. Bianco *et al.* [19] presented a *quasi-unsupervised* learning strategy where

a deep CNN is trained to detect achromatic pixels in grayscale images. This enables effective illuminant estimation without requiring specific illuminant information while achieving competitive results in both unsupervised and supervised settings. Xu *et al.* [122] developed a *deep metric learning* approach, which utilizes triplet networks [123]. Their architecture produces a discriminative but robust feature space by grouping images based on similar illuminant conditions rather than the content, thus achieving robust illuminant estimation across varying scenarios. Finally, Lo *et al.* [23] presented *CLCC*, a contrastive learning framework, that leverages *illuminant-dependent features* through the augmentation of color in the raw domain, and achieves good performance with fewer parameters and enhanced robustness in data-sparse regions. Li *et al.* [124] proposed *SWBNet*, a specialized network for WB correction, which stabilizes color correction in varying color temperatures by learning temperature-insensitive features, employing a contrastive loss and a color temperature-oriented transformer architecture.

Recent advances by Afifi *et al.* [125, 20, 126] have introduced novel frameworks to address the challenge of modifying WB in *sRGB* images post-capture. Traditional in-camera imaging pipelines apply WB early, followed by non-linear color adjustments, which challenge post-capture WB correction. To solve this, Afifi *et al.* [125] proposed a method that takes advantage of lower-resolution versions of an image with varying color temperatures, which enables learning *color mapping functions*, to adjust the full-size *sRGB* image to different color temperatures with minimal data overhead. Following this, Afifi *et al.* [20] designed a specialized DNN architecture that maps an *sRGB* image to *various WB settings*, which achieves greater accuracy and flexibility compared to the leading methods of its time. Another extension by Afifi *et al.* [126] enables *interactive* WB editing on camera-rendered images by linking non-linear color-mapping functions directly to user-selected colors. This allows efficient, user-driven WB adjustments even on camera-rendered images with memory and run-time efficiency improvements. Ulucan *et*

al. [127] conducted a *comprehensive analysis* of how strong color casts affect traditional, learning-based, and data-driven WB correction algorithms, particularly for illuminants at the edges and beyond the color temperature curve.

Furthermore, Buzzelli *et al.* [128] introduced a different deep learning approach to illuminant estimation that bypasses the need for ground-truth illuminants by leveraging an object recognition loss function as an *auxiliary task*. Afifi *et al.* [129] presented CNN-based cross-camera WB correction architecture, namely *C5*, which is a learning-based method that adapts dynamically to the spectral properties of unseen cameras for illuminant estimation. Distinct from earlier models, *C5* employs transductive inference by utilizing additional unlabeled images during testing, enabling *real-time adaptation* to new camera sensors without requiring calibration. Ulucan *et al.* [130] introduced a computational color correction method inspired by biological principles, which aims to emulate the hierarchical color perception of *the human visual system*. This model integrates aspects such as focal and peripheral vision, the retinotopic structure, double-opponent cell responses, and the visual cortex's saliency map, thereby mimicking how humans achieve color constancy and respond to color assimilation illusions. Based on findings from human visual perception, Ulucan *et al.* [131] leveraged the luminance of the brightest patches and the space-average color as indicators for illuminant estimation, inspired by the innate ability of the human visual system to *discount illumination* in perceiving object colors.

With the introduction of datasets containing multi-illuminant scenarios [132, 133, 134, 135], especially like the LSMI dataset [135], research has increasingly focused on handling complex, non-uniform illumination scenarios. This prompts recent studies to develop approaches specifically aimed at addressing the challenges posed by multiple light sources in a scene. In earlier studies focusing on WB correction for non-uniform illuminated scenes, Gijsenij *et al.* [133] introduced a methodology to extend traditional algorithms by applying them to *individual patches* within an image, rather than globally,

and then combining these estimates for a more accurate correction in scenes with multiple light sources. Bleier *et al.* [132] adapted existing WB correction algorithms to estimate illumination locally by segmenting images into *super-pixels*, each with its own illuminant estimate, and then combining these estimates to better approximate the model for complex lighting environments. Joze *et al.* [136] proposed an exemplar-based learning framework for WB correction that addresses the challenges of multi-illuminant scenes by estimating illumination based on *local surface statistics*, rather than assuming a uniform scene illuminant. Through unsupervised learning, this approach builds models for each surface in training scenes and then leverages nearest-neighbor surfaces to estimate illumination in test scenes. Beigpour *et al.* [134] proposed an energy minimization framework within a Conditional Random Field (CRF) to estimate both the colors and *spatial distribution* of multiple illuminants.

The following studies build upon the need to correct color balance in real-world settings where multiple light sources often co-exist. Afifi *et al.* [2] proposed a WB correction method relying on learning *the weighting maps* of multiple pre-defined WB settings (*i.e.*, color temperatures) to obtain a corrected image by blending them to effectively handle mixed lighting conditions. Akazawa *et al.* [137] introduced an exceptional method for WB correction, namely *N-white balancing*, which adjusts WB by aligning *multiple source white points* (*i.e.*, N) rather than relying on the number of light sources. Under multi-illuminant conditions, this approach effectively matches each source point to its ground truth value, which reduces lighting effects even if N exceeds the actual illuminants. Li *et al.* [138] introduce a multi-task learning framework with auxiliary tasks, such as achromatic-pixel detection and surface-color similarity prediction, to improve local lighting and surface color estimation under complex illuminant conditions. Domislović *et al.* [139] designed a CNN-based model that operates *patch-wise*, assuming each patch contains a single illuminant, but leverages image-wide features to enhance local

illuminant accuracy under variable light sources. Entok *et al.* [140] presented a pixel-wise multi-illuminant model incorporating total variation loss and bilateral filtering. With the help of these additional factors for optimization, the model can maintain smooth illumination transitions across spatial dependencies. Finally, Kim *et al.* [9] proposed a Transformer-based WB architecture with the addition of a novel slot attention strategy to separately represent individual illuminants, which then fuse into a comprehensive illumination map. This strategy allows for scene illumination editing and achieving leading performance across multi-illuminant benchmarks. Collectively, these methods push the boundaries of WB correction in complex lighting scenarios, which offer promising results for non-uniform illumination correction.

Despite advancements in WB correction, existing methods still face challenges in multi-illuminant environments, generalization across diverse scenes, and computational efficiency. Traditional statistical approaches, such as Gray-World [13] and Gamut Mapping [84, 85, 86], rely on global assumptions that fail under spatially varying lighting conditions. Scene semantics-based methods incorporate high-level priors but lack adaptability to unseen illuminants, while deep learning-based approaches improve accuracy, but often fail to model illumination's statistical impact on image features. Additionally, many methods focus on pixel-wise corrections, overlooking higher-order feature relationships crucial for perceptual color consistency.

To overcome these limitations, this dissertation introduces a feature distribution matching-based WB correction framework, treating lighting as a style factor for robust adaptation under diverse illumination conditions. By leveraging EFDM and a well-designed optimization process, the proposed method aims to exactly align the image color distributions with an ideal white-balanced reference during training, minimizing color distortions at a feature level rather than solely at the pixel level.

Table 3.1: *Comparison of White Balance Correction Methods.*

Method	Adaptability	Computational Cost	Multi-Illuminant Robustness	Accuracy	Real-Time Capability	Comments
Traditional	Low	Low	Poor	Moderate	Yes	Effective for simple scenes but struggles in complex lighting conditions.
Deep Learning-Based	High	High	Moderate	High	No	Delivers moderate or high accuracy but not advanced on multiple illuminant scenarios.
Semantics-Based	Moderate	Moderate	Moderate	Moderate	No	Balances adaptability and robustness, though limited in real-time applications.
Lower-order Statistics [1]	High	Moderate	Moderate	High	Yes	Efficient and robust in less diverse lighting conditions.
Distribution Matching [2]	High	Moderate	High	Very High	Yes	Offers superior accuracy in diverse lighting conditions.

[1] Style WB
[2] FDM WB & FDM Loss

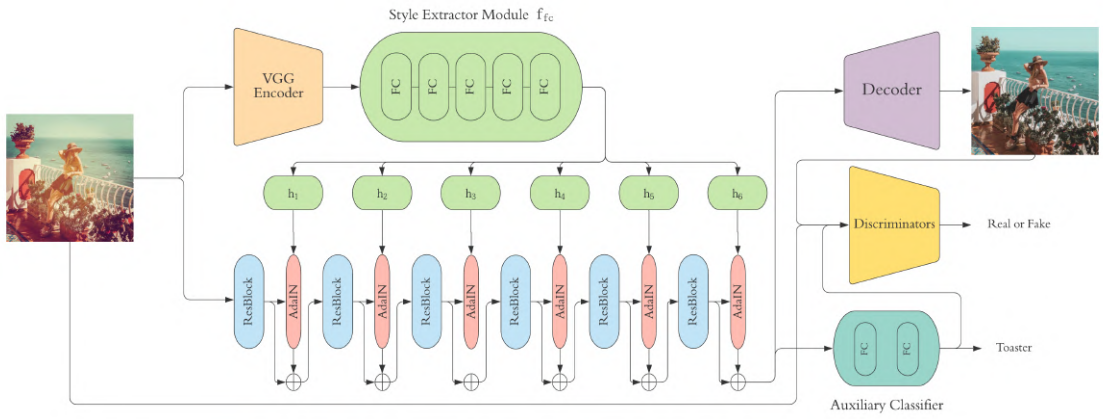
4. METHODOLOGY

This chapter outlines the methodological framework that underpins the proposed approach to WB correction, which is developed through a series of progressive studies that conceptualize lighting as a style factor within scenes. The foundation for this approach was initially established in a study addressing Instagram filter removal in fashionable images [41], where filters were considered as style factors injected into the images. This work demonstrates how style factors could be isolated and removed within an encoder architecture. This idea is further extended through patch-wise contrastive style learning [141] to enhance filter removal. Although initially unrelated to WB correction, these studies introduce a critical insight: disruptive visual elements, *including lighting*, could be considered as style factors and effectively swept away with the help of a deep learning framework. Building upon this insight, subsequent work [24] focuses directly on WB correction, which aims to model lighting variations as a style factor to enhance WB correction performance. This shift led to the development of the use of statistics from feature maps to construct style representations in scenes, which culminated in the present study. Here, feature distribution matching is utilized both as a fundamental component of the architecture [28] and as an objective function [29] to represent illumination as a style to achieve robust WB correction.

4.1 Foundational Study

This section summarizes the foundational study that forms the basis for modeling disruptive elements within a scene as style factors, the main approach proposed in this thesis.

Figure 4.1: *Overall architecture of Instagram Filter Removal Network (IFRNet).*



4.1.1 Instagram Filter Removal on Fashionable Images

The process of WB correction fundamentally aims to mitigate unwanted color shifts caused by varying illumination conditions. Although WB correction is traditionally addressed within an ISP pipeline, previous research has shown that certain artificial modifications, such as social media filters, can introduce stylistic alterations that affect the perceived color of the scene. From a computational perspective, these modifications share similarities with illumination-induced color distortions, as they impose systematic transformations on the color characteristics of an image. This study initially explored the removal of artificial stylistic alterations to gain a deeper understanding of the representation of the style factor, which ultimately laid the foundation for modeling lighting as a style factor in WB correction.

Model Architecture and Objective: The Instagram Filter Removal Network (IFRNet) is designed as an encoder-decoder framework with an adaptive feature normalization mechanism that effectively removes external stylistic influences while preserving the content structure. The central objective of IFRNet is to restore images altered by social media

filters by treating these filters as extraneous style factors that can be isolated and removed.

The architecture incorporates AdaIN, which aligns feature distributions between the filtered and original images in multiple layers of the encoder. The model style extractor module, a fully connected five-layer network, maps high-level features extracted from a pretrained VGG network into a latent style representation. This representation is subsequently used to predict affine transformation parameters that drive feature normalization. The overall architecture of IFRNet is shown in Figure 4.1.

Mathematically, given a feature representation $\mathbf{z}_{vgg}$ extracted from a pretrained VGG network, the predicted affine transformation parameters y_i for each normalization layer are computed as

$$y_i = h_i(f_{fc}(\mathbf{z}_{vgg})), \tag{4.1}$$

where $f_{fc}(\cdot)$ denotes the fully connected style extractor, and $h_i(\cdot)$ represents the layer-specific transformation function.

The AdaIN operation aligns the mean and variance of the input feature maps x with those of the extracted style representation y:

$$\text{AdaIN}(x, y) = \sigma(y) \left(\frac{x - \mu(x)}{\sigma(x)} \right) + \mu(y), \tag{4.2}$$

where $\mu(x)$ and $\sigma(x)$ represent the mean and standard deviation of the input feature maps, while $\mu(y)$ and $\sigma(y)$ correspond to those of the extracted style representation. This adaptive normalization process ensures that the stylistic artifacts introduced by filters are systematically suppressed through the heads $h_i(\cdot)$, which acts as a reverse style transfer, allowing the model to restore images to their original unaltered color distributions.

To maintain content integrity while removing the style factor injected by filters, the encoder is constructed with six residual blocks, each incorporating an AdaIN layer to normalize feature maps using the affine parameters learned by the style extractor. Skip connections are introduced to preserve essential content information, ensuring that the model

selectively removes filter-induced distortions without compromising semantic structure. The effect of these skip connections can be expressed as

$$o_i = r_i(v_i, y_i) + v_i, \tag{4.3}$$

where $r_i(\cdot)$ represents the residual block, which receives feature maps v_i and affine parameters y_i as input, producing output o_i. Note that when y_i is set to zero, AdaIN at that layer is effectively neutralized, which means that normalization does not alter feature maps. This implies that no external style information (*e.g.*, visual artifacts injected by filter) is being applied or retained in the feature maps at that level. By setting y_i to zero, the model preserves only the features related to the pure style of the image, which allows the encoder to sweep away any additional style information and retain the original content of the scene with its pure style. This approach provides fine-grained control over the extent of style removal at each layer in the encoder.

Objective Function: The IFRNet is trained using a multi-component loss objective designed to ensure texture fidelity, semantic consistency, and structural coherence in the restored images. The objective function integrates a patch-wise texture loss, a semantic consistency loss, and an adversarial loss, each contributing to the overall stability and accuracy of filter removal.

To preserve fine-grained details, patch-based texture loss is employed using the ID-MRF formulation [142], which enforces local texture similarity between the restored image $\mathbf{I}_{out}$ and the reference image $\mathbf{I}_{gt}$. This loss is defined as follows.

$$\mathcal{L}_{tex} = \sum_{p} \min_{q \in N(p)} \|\phi_p(\mathbf{I}_{out}) - \phi_q(\mathbf{I}_{gt})\|_2^2, \tag{4.4}$$

where $\phi_p(\cdot)$ represents feature extraction at location p, and $N(p)$ denotes the nearest-neighbor set of patches.

To ensure feature-level consistency, a semantic consistency loss is introduced, which aligns the intermediate feature representations of the restored and ground truth images. This loss is computed as

$$\mathcal{L}_{sem} = \sum_{p=0}^{P-1} \frac{1}{C_p H_p W_p} \left\| \Phi_p(\mathbf{I}_{out}) - \Phi_p(\mathbf{I}_{gt}) \right\|_2^2, \tag{4.5}$$

where $\Phi_p(\cdot)$ denotes the feature map of the p^{th} layer extracted from a pretrained VGG16 network [48].

To enhance perceptual realism and ensure that restored images maintain structural integrity, an adversarial loss is applied using a PatchGAN discriminator [143]. The adversarial objective is formulated as

$$\mathcal{L}_{adv} = \mathbb{E}[\log D(\mathbf{I}_{gt})] + \mathbb{E}[\log(1 - D(\mathbf{I}_{out}))], \tag{4.6}$$

where $D(\cdot)$ represents the discriminator network.

To stabilize training, a gradient penalty term is incorporated

$$\mathcal{L}_{gp} = \lambda_{gp} \mathbb{E}\left[(\|\nabla_{\hat{\mathbf{I}}} D(\hat{\mathbf{I}})\|_2 - 1)^2 \right], \tag{4.7}$$

where $\hat{\mathbf{I}}$ is an interpolated image between $\mathbf{I}_{out}$ and $\mathbf{I}_{gt}$.

In addition, an auxiliary classification loss is introduced to improve robustness, ensuring that the network correctly predicts the removed filter type,

$$\mathcal{L}_{cls} = \lambda_{cls} \sum_{i=1}^{N} y_i \log(\hat{y}_i), \tag{4.8}$$

where y_i and $\hat{y}_i$ denote the true and predicted filter labels, respectively.

The final loss function integrates all components as follows.

$$\mathcal{L} = \lambda_{tex}\mathcal{L}_{tex} + \lambda_{sem}\mathcal{L}_{sem} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{gp}\mathcal{L}_{gp} + \lambda_{cls}\mathcal{L}_{cls}. \tag{4.9}$$

With its encoder-decoder architecture and the integration of AdaIN layers, IFRNet effectively distills and removes filter-induced style information from the feature maps.

This established a basis for considering lighting as a similar extrinsic factor in scenes for WB correction. This methodology of modeling external style factors in encoder layers provides a transferable approach, which allows us to model lighting as a style factor, which aligns with the broader scope of this research on style factor-based WB correction.

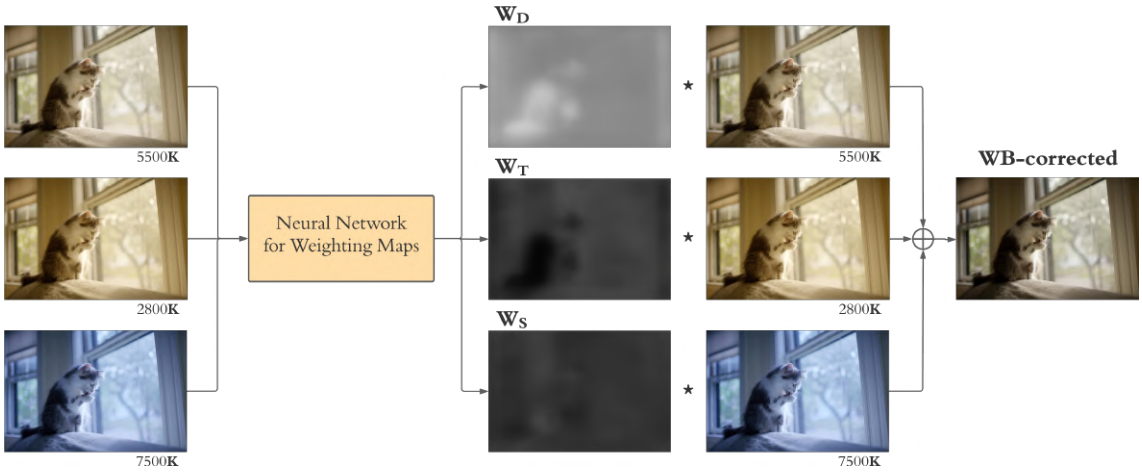
Significance for WB Correction: Although IFRNet was initially developed for filter removal, its underlying style-based modeling framework provided a crucial insight: *illumination conditions can be treated as a style factor that alters the color distribution of an image*. The methodology employed in IFRNet, where stylistic distortions are systematically removed by aligning the feature distribution, directly influenced the development of proposed approaches for WB correction.

4.2 Learning Style Factors for White Balance Correction

In machine learning-based computer vision, *style* often denotes a broad spectrum of abstract image attributes (*i.e.*, artistic elements, personal features or textures), which are shaped by the learned feature space of the machine learning model. For example, a model's feature space can capture aspects such as the artistic style of a painting, the hairstyle of a person, the texture of clothing, or even the color characteristics of an animal. Previous studies, including [41], have shown that disruptive image modifications, such as social media filters that corrupt the attributes of the original image, can be effectively modeled as style factors. Based on this principle, we propose that lighting in scenes, whether from a single or multiple illuminants, can similarly be treated as an injected style factor.

However, this approach diverges from conventional style transfer methods. Instead of transferring the stylistic qualities of one image to another, our goal is to normalize or eliminate injected *style* information, where illumination serves as the primary

Figure 4.2: *Example of predictions for the weighting maps and White Balance correction results by blending these maps.*



style factor. In this context, the model learns to adaptively adjust varying lighting conditions to achieve consistent white balance. Our initial approach is designed to address both uniform and mixed illumination settings by integrating style removal through adaptive feature normalization. The following iteration advances this by employing EFDM [30], instead of naive feature alignment, and leverages the capabilities of a Transformer-based architecture (*i.e.*, Uformer [5]). Finally, the most recent version replaces the style removal module with a novel color distribution matching loss term, which enhances precision in modeling the lighting as style.

4.2.1 Illumination as Style with Adaptive Feature Normalization

Similarly to the foundational work, our initial approach introduces a novel design for WB correction that treats the lighting in the scenes with multiple illuminants as a style injected into the scene by different light sources.

4.2.1.1 Modified Camera ISP for WB Correction

Following the prior work on WB correction [20, 2], we design a method to produce the high-resolution image with fixed WB settings (*i.e.*, daylight) and additional small images rendered with a set of predefined WB settings, which are {t, f, d, c, s} and {t, d, s}. {t, f, d, c, s} refers to tungsten (2800**K**), fluorescent (3800**K**), daylight (5500**K**), cloudy (6500**K**), and shade (7500**K**), respectively. The formula for rendering the small images can be described as follows

$$\hat{I}_{c_i} = M_{c_i}\phi(I_{init}) \tag{4.10}$$

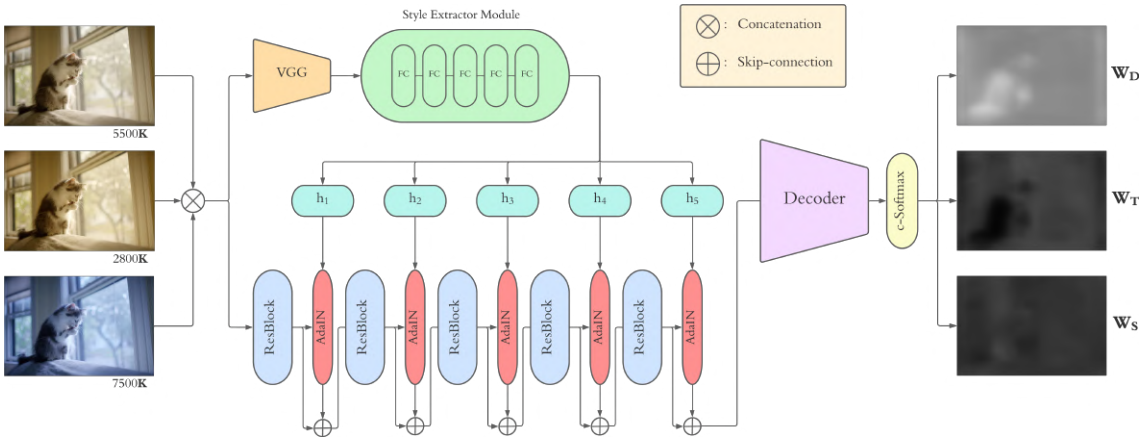
where I_{init} is the initial high-resolution image rendered with a fixed WB setting (*i.e.* daylight), $\hat{I}_{c_i}$ represents the output image mapped to the target WB setting, M_{c_i} is the matrix that maps the colors of the initial image represented in a higher-dimensional space, and $\phi(\cdot)$ is a polynomial kernel function projecting the colors of the initial image into the higher-dimensional space. $\phi(\cdot)$ is optimized by minimizing the sum-squared error between the colors of the target and the source images, as in [2]. As distinct from [2], we consider this part as a preprocessing for training, and save the target images before training, instead of computing them on the fly.

After extracting the small images, following the method in [2], we employ a learning mechanism for the weighting maps of different scenes with a predefined set of WB settings. We use these learned weighting maps for generating the final *sRGB* output image by linearly combining them with the small images, as shown in the following equation

$$\tilde{I}_{corr} = \sum_i W_i \odot \tilde{I}_{c_i} \tag{4.11}$$

where $\tilde{I}_{corr}$ is the corrected small *sRGB* image, $\odot$ is Hadamard product, W_i represents the weighting map for i^{th} WB setting (*i.e.*, c_i), and $\tilde{I}_{c_i}$ denotes the small image rendered with c_i. This operation for WB correction is illustrated in Figure 4.2. Apart from prior

Figure 4.3: *Overall design of proposed learning mechanism for the weighting maps of different White Balance settings.*



WB correction methods [20, 2], we employ a learning-based style factor approach to effectively learn weighting maps, allowing the model to adapt to diverse illumination conditions in scenes.

4.2.1.2 Learning Mechanism

Given a set of small images $\tilde{I}_{c_i}$, our proposed learning mechanism learns to estimate $\{W_i\}$. Our initial work adapts a style removal network proposed in [41] as the learning mechanism of the weighting maps. The architecture consists of an encoder-decoder structure that employs an adaptive feature normalization strategy to all layers of the encoder part. With the help of this strategy, illumination that comes from different light sources can be modeled as an external style, which needs to be discarded or adjusted in another style. The main component to achieve this is AdaIN [44] for each encoder layer, which transfers the feature statistics computed across spatial locations. AdaIN aligns the channel-wise mean μ and variance σ of the feature maps of the content image x with the style input statistics y, as formulated in Equation 4.2.

To extract the style input for the images, we use a multi-head mapping module

that maps the feature representations encoded by a pretrained VGG network [48] to the style latent space. The style latent code $\mathbf{w}$ is fed into different heads for different encoder levels, and each head h_i is attached to a projection layer p_i (*i.e.*, fully-connected), which adapts the affine parameters y_i of each normalization layer in the encoder.

$$\begin{aligned} \mathbf{w} &= M(\mathbf{z}), \\ y_i &= p_i(h_i(\mathbf{w})) \end{aligned} \tag{4.12}$$

where $\mathbf{z}$ is the feature representation of the input image x extracted by VGG, and M denotes the style extractor module mapping the input latent space to the style latent space.

In our design, the style extractor module is made up of a five-layer MLP block. The encoder contains five residual blocks, each of which has a specific AdaIN layer to normalize the feature maps with the affine parameters projected by the corresponding head. The network takes the concatenated feature representations of the small images rendered with different WB settings as input, and learns to produce the weighting maps for these WB settings. As suggested in [41], we use skip connections between encoder layers to preserve the related information (*i.e.*, the content assumed under pure white-light illumination) while distilling the style (*e.g.*, additional illumination led to color cast). The overall design of the proposed learning mechanism for the weighting maps of different WB settings is shown in Figure 4.3.

Through an encoder-decoder architecture with aligning feature statistics at each encoder layer, our method captures and corrects unwanted color casts from varying WB settings by treating the affine parameters as latent style factors. These affine parameters are dynamically learned via a style extractor module that maps features from a pretrained VGG network into a style latent space. This style-informed approach effectively normalizes lighting variations across diverse WB settings, which mitigates the impact of mixed or inconsistent illuminants.

To ensure accurate WB correction and maintain color consistency, the model is

optimized using a combination of reconstruction and smoothing loss terms, following prior work [2]. The reconstruction loss minimizes the discrepancy between the WB-corrected image patches and their corresponding ground truth patches. Given an input patch P_{c_i} rendered under a specific WB setting c_i, and the ground truth patch P_{gt}, the reconstruction loss is formulated as

$$\mathcal{L}_r = \left\| P_{gt} - \sum_i \hat{W}_i \odot P_{c_i} \right\|_F^2 \tag{4.13}$$

where $\hat{W}_i$ represents the weighting map for each WB setting c_i, and $\odot$ denotes the Hadamard product. This objective ensures that the corrected image is aligned with the true color-balanced reference by enforcing a direct per-pixel fidelity constraint.

To enforce spatial consistency and prevent artifacts in the corrected images, a smoothing loss is introduced, which regularizes the weighting maps by penalizing abrupt variations across spatial dimensions. This is implemented using horizontal and vertical Sobel filters with kernel size 3×3, denoted as ∇_x and ∇_y, respectively.

$$\mathcal{L}_s = \sum_i \left\| \hat{W}_i * \nabla_x \right\|_F^2 + \left\| \hat{W}_i * \nabla_y \right\|_F^2 \tag{4.14}$$

where $*$ represents the convolution operation. By enforcing smooth transitions in the weighting maps, this loss mitigates discontinuities in the corrected image, ensuring perceptually coherent WB adaptation across different regions.

The final optimization objective combines both loss components as follows.

$$\mathcal{L} = \mathcal{L}_r + \lambda \mathcal{L}_s \tag{4.15}$$

where λ is the regularization coefficient, set to 100 in our experiments. This formulation balances accurate WB correction with spatial consistency across the weighting maps.

4.2.1.3 Post-processing

Two post-processing steps are used to further refine the learned weighting maps and enhancing the final *sRGB* image quality. First, *multi-scale ensembling* generates

multi-scale weighting maps, which are then bilinearly upsampled to a high resolution and averaged to achieve smooth and accurate weighting. Next, *edge-aware smoothing* is applied to the weighting maps using a fast bilateral solver [144], guided by the high-resolution input image to preserve edges and fine details in the final corrected image. It is worth noting that these operations are not proposed in this work, but are applied to ensure a fair comparison with prior methods. Together, these enhancements improve the consistency and realism of the output, which ensures that the corrected image closely resembles a naturally white-balanced photograph.

4.2.2 Illumination as Distribution Statistics

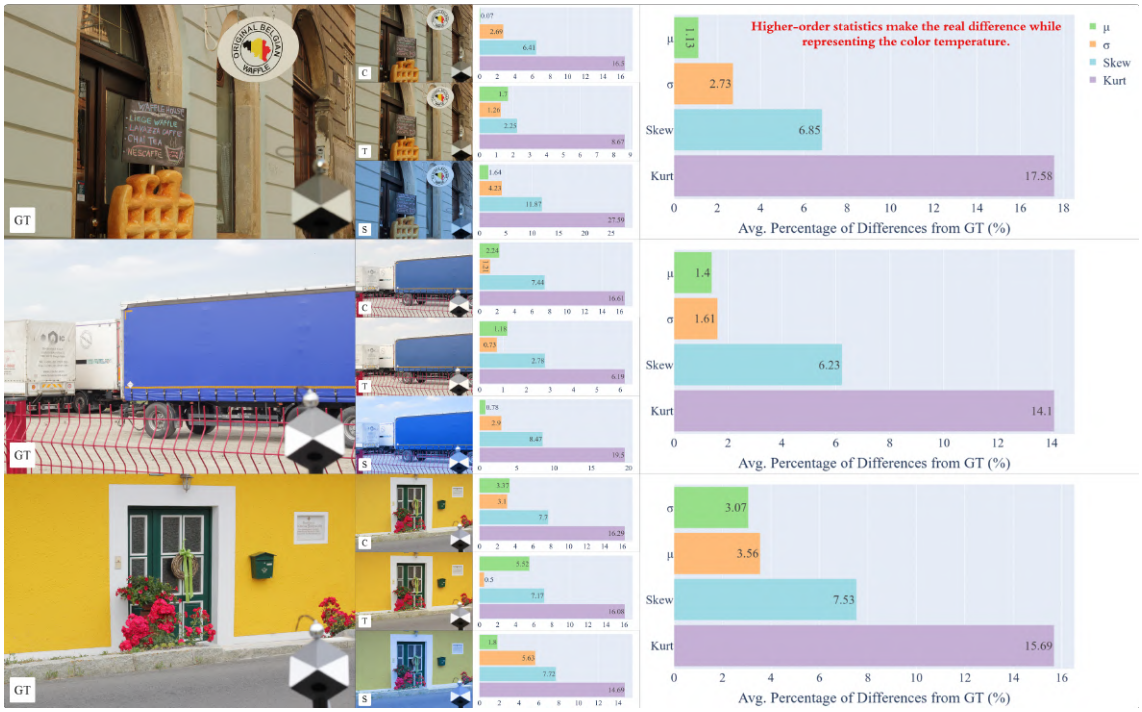
Following this, we conducted several analyses to explore how different illumination settings, particularly those with multiple light sources, affect chromaticity channels and feature representations in images corrected for white balance. These analyses, presented in Figures 4.5, 4.6, and 4.7, focus on two key aspects: chromaticity channel distributions and feature distribution statistics. By examining chromaticity distributions in the U and V channels and [CLS] token feature statistics (*i.e.*, mean, standard deviation, skewness and kurtosis) under varying lighting conditions, we aim to reveal the limitations of our first WB correction approach that relies solely on lower-order statistics representing the style factor.

4.2.2.1 Feature Distribution Discrepancies Under Varying WB Settings

To examine the impact of WB settings on feature representations, we analyze statistical variations in VGG-derived features under different illumination conditions. This analysis highlights the limitations of traditional WB correction approaches that align only lower-order statistics and motivates the use of exact distribution matching.

Figure 4.4 illustrates the statistical differences in the VGG features under three WB settings- Cloudy, Shade, and Tungsten- using three representative images from the

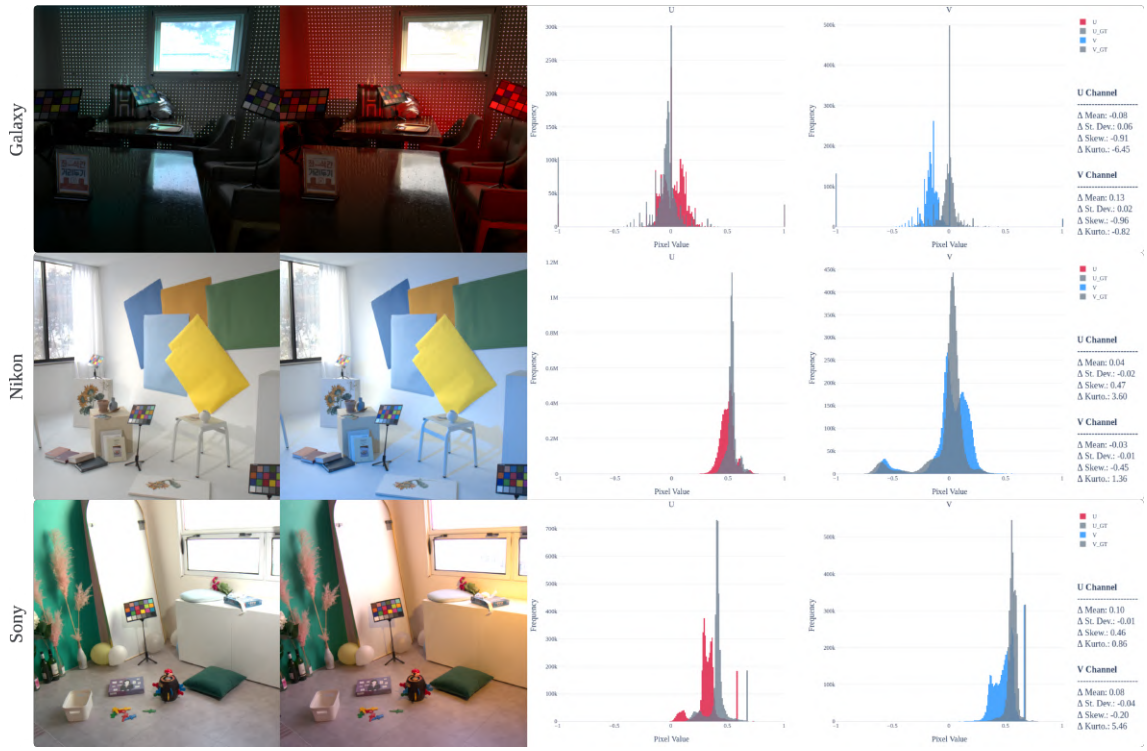
Figure 4.4: *Chromaticity channel distributions under different lighting conditions.*



Cube+ dataset [26]. The left column displays ground truth images along with their altered versions, while the right column presents the percentage differences in statistical measures (*i.e.*, mean, variance, skewness and kurtosis) relative to ground truth images. The results demonstrate that, while mean and variance show moderate variations, higher-order statistics such as skewness and kurtosis exhibit significant deviations, particularly under strong color temperature shifts.

These findings indicate that aligning only lower-order statistics may be insufficient for robust WB correction, as it fails to capture the full distributional changes induced by different lighting conditions. The observed discrepancies motivate the use of EFDM, which explicitly aligns entire feature distributions rather than focusing solely on first- and second-order moments. By addressing lower- and higher-order differences, EFDM enhances the robustness of color correction, particularly representing different WB settings.

Figure 4.5: *Chromaticity channel distributions under different lighting conditions.*



4.2.2.2 Chromaticity Channel Distributions under Multiple Illuminants

Figure 4.5 illustrates the variations in the chromaticity channel distributions under different lighting conditions, using samples from three different camera models (*i.e.*, Galaxy, Nikon, and Sony). Each row presents, from left to right: a ground truth white balanced image, an image with multiple illuminants, and the corresponding histograms for the U and V chromaticity channels. This analysis provides insight into how different lighting conditions, specifically multi-illuminant scenes, impact the distribution of color information within the U and V channels.

As observed, the U and V channel histograms for the white balanced images show relatively narrow distributions with lower skewness and kurtosis values, which reflects a more uniform color spread. This narrower range suggests that in the absence of complex

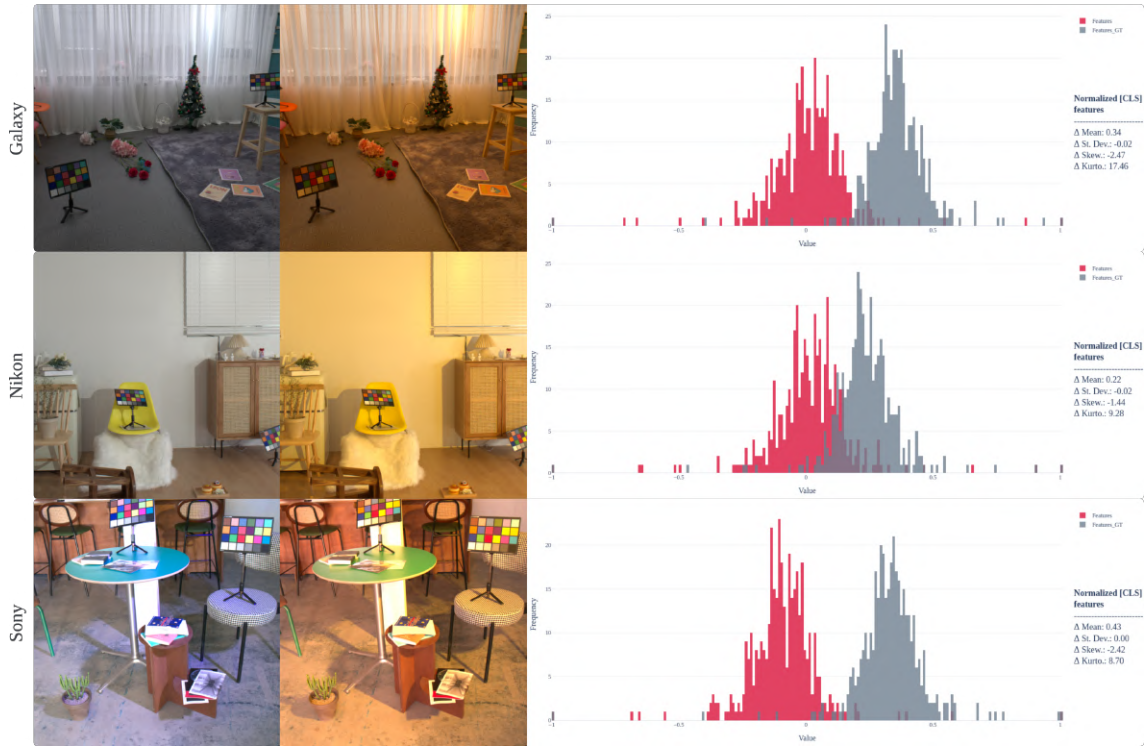
lighting variations, the color information remains consistent across the image, suggesting a balanced chromaticity. Such uniform distributions imply that the color channels maintain a predictable pattern, with minimal deviation from the mean, as expected in a controlled, single-illuminant setting.

However, in images with mixed illumination, the U and V channel distributions become significantly wider and more skewed. This widening of the distributions is indicative of a broader range of color values being introduced by the varied lighting sources, with each source contributing its own unique spectral characteristics to the scene. This shift towards wider, skewed distributions underscores the complexity introduced by multi-illuminant scenarios, where each light source interacts with scene elements in distinct ways, which results in unpredictable chromatic shifts. The elevated skewness reflects an asymmetry in the color distribution, where certain hues may become more dominant due to specific illuminants, while kurtosis points to the presence of extreme chromatic variations or outliers within the distribution, which leads to a *tailed* spread of color values. Such variations are indicative of non-uniform color casts imposed by the overlapping effects of multiple illuminants, which lead to higher-order distortions in the chromaticity distributions. These distortions not only affect the global color balance but also introduce localized shifts in chromatic values. This further complicates the WB correction process.

4.2.2.3 Multi-Illuminant Effects on Feature Distributions across Cameras

In this analysis, we present three samples from different cameras (*i.e.*, Galaxy, Nikon, and Sony, ordered by row) that illustrate the ground truth white balanced image, an illuminated scene with at least two light sources, and the corresponding [CLS] token feature distributions. The histograms on the right depict the distribution of normalized [CLS] token features under these lighting conditions, which allows us to observe the impact of illumination on feature statistics.

Figure 4.6: *CLS token distribution statistics for samples from different cameras.*



The variations observed in feature distributions underscore the significant impact of multiple illuminants on feature representation. This reveals that methods aligning only basic statistics, as in our prior approach, can fall short. Although our earlier method effectively aligned mean and variance to handle simpler, more uniform lighting scenarios, it faces limitations in more complex settings with multiple light sources. Such environments often involve interactions between various illuminants with different spectral properties, leading to color shifts that are neither predictable nor uniformly distributed across the image. This complexity suggests a need for a more nuanced approach that extends beyond lower-order statistical alignment to fully capture the chromatic shifts introduced by multi-illuminant lighting.

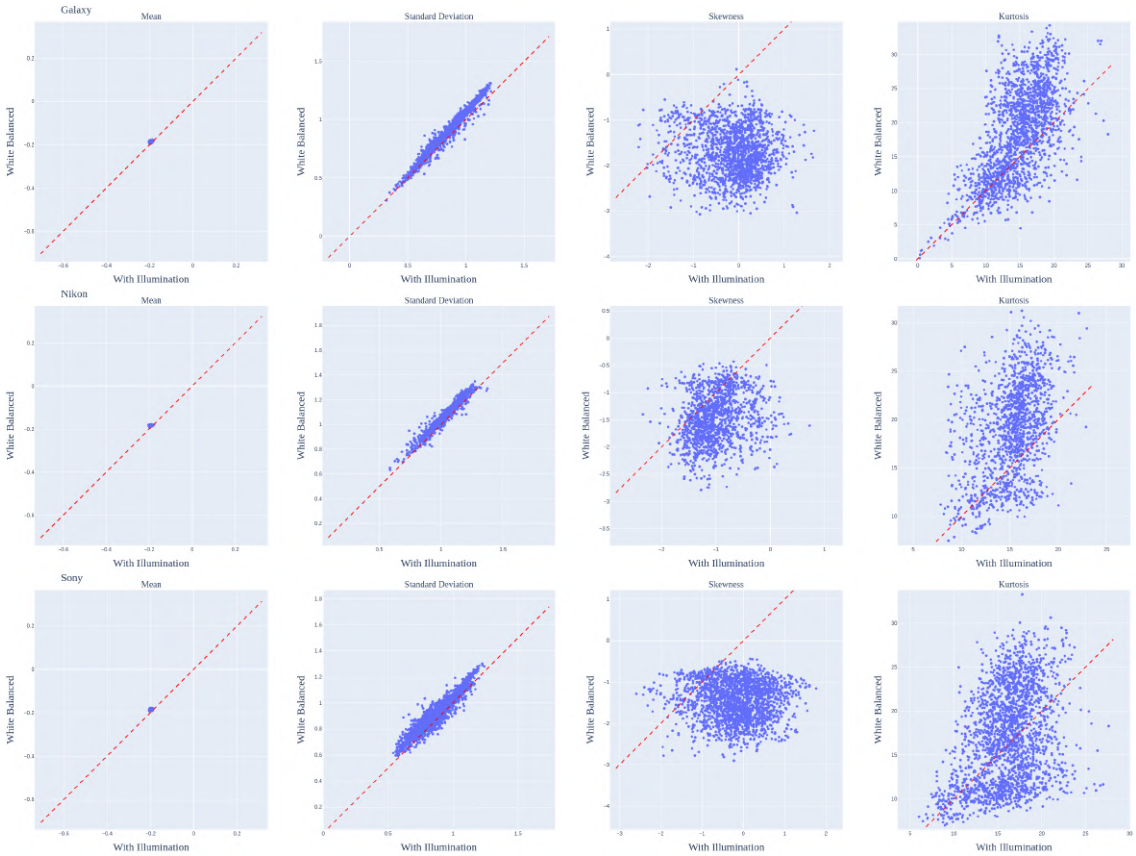
Our analysis shows that higher-order statistics, such as skewness and kurtosis,

display pronounced variations between images corrected for white balance and those captured under multi-illuminant settings. This indicates that multi-illuminant conditions introduce non-Gaussian characteristics into the feature space, which results in distributions that are both asymmetrical (*skewed*) and have heavier tails (*high kurtosis*). These deviations capture the intricate chromatic interactions introduced by each illuminant, reflecting the influence of multiple light sources on color representation. The limitations of aligning only mean and variance in such cases become apparent, as this approach may inadequately represent the full scope of chromatic distortions, which leads to suboptimal WB correction under diverse lighting conditions.

4.2.2.4 Analysis of Feature Distribution Statistics across Illumination Settings

To evaluate the impact of illumination on feature representations in WB corrected images, we analyzed the distribution statistics of the [CLS] token extracted from ViT [4] across three different cameras (*i.e.*, Galaxy, Nikon, and Sony) in the LSMI dataset [8]. Specifically, we compared the distribution characteristics (*i.e.*, mean, standard deviation, skewness, and kurtosis) of the [CLS] token for images under different illumination settings (*i.e.*, single and multiple illuminants) and their corresponding WB corrected versions. As discussed in Section 2.3.2.1, the [CLS] token, which serves as a global representation in ViT, encapsulates style information effectively. Its aggregation of feature-level details provides a comprehensive view of image characteristics, including illumination. To substantiate this, we perform an analysis on the capacity of the [CLS] token to capture and represent style-related information, particularly in terms of illumination variations across different scenes.

Figure 4.7: *CLS token distribution statistics for illuminated vs. white balanced images across three cameras in the LSMI dataset.*



The results, displayed in Figure 4.7, indicate that higher-order statistics (*i.e.*, skewness and kurtosis) capture more substantial deviations between illuminated and WB corrected images than lower-order statistics (*i.e.*, mean and standard deviation). This is particularly evident across all three cameras, where the mean and standard deviation exhibit relatively minor differences between illumination conditions, largely aligning along the diagonal line that represents equality between illuminated and WB settings. In contrast, skewness and kurtosis show a significant spread, which reflects the non-Gaussian characteristics of scene lighting in real-world conditions.

These findings highlight that while lower-order statistics, such as mean and standard deviation, provide some alignment between different illumination settings, they fall short in capturing the complex variations introduced by non-uniform or multi-source lighting. Higher-order statistics (skewness and kurtosis) offer a more nuanced representation of these variations, which reveal intricate illumination effects on feature distributions. This underscores the need for full distribution matching over simple alignment of lower-order moments. By incorporating higher-order statistics through EFDM, our framework can provide a more comprehensive representation of illumination, thus enhancing WB correction accuracy, particularly in challenging multi-illuminant scenarios. EFDM thus enables a more robust illumination modeling by preserving both lower- and higher-order statistical moments, which facilitates realistic WB corrected outputs across diverse lighting conditions.

4.2.3 Leveraging Feature Distribution Matching for WB Correction

Building upon our initial approach of learning style factors for WB correction, our next approach further refines the style representation by moving from a simple alignment of feature statistics to exact matching them (via Exact Feature Distribution Matching (EFDM) [30]). While alignment-based strategies like AdaIN only align mean and variance, they often fail to capture the intricate and complex lighting variations found in real-world scenes. Relying solely on lower-order statistics limits the model's ability to generalize across scenes under illumination based on non-Gaussian lighting distributions, which leads to suboptimal WB correction, especially in multi-illuminant scenarios.

To address the limitations of alignment-based strategies, we propose a novel deep learning architecture, namely FDM WB, which utilizes feature distribution matching to capture the full distribution of features across varying illuminants. Inspired by the approach in StyleGAN [45] to constructing style spaces from random noise, we employ a

pretrained VGG network [48] to generate a style feature space by applying projection matrices to features extracted under different WB conditions. Unlike the previous approach, which aligns only basic statistics such as mean and variance, this model directly matches the cumulative distribution functions (CDFs) of feature statistics while constructing a style space of illumination in scenes. This preserves higher-order moments and thus provides a more accurate and global representation of lighting as a style factor. The proposed method (*i.e.*, WB correction via feature distribution matching) establishes a foundation for generating corrected images by adjusting WB settings in a way that better captures the nuanced effects of different lighting conditions.

The implementation leverages a U-shaped Transformer-based architecture [5], which is well suited to handle spatial dependencies over large areas, capturing subtle lighting variations across different parts of the scene. With a similar practice in our first approach, but integrating EFDM within new architecture instead of AdaIN, our network learns to effectively capture and match the distributional shifts of features under various WB conditions. This alignment of empirical distributions, rather than only mean and variance, facilitates a more robust and adaptable color correction process, which achieves greater fidelity in WB correction for both single- and mixed-illuminant environments.

4.2.3.1 From Alignment to Exact Matching with EFDM

As mentioned in Section 2.2.3, EFDM distinguishes itself from conventional statistical alignment methods, such as AdaIN or GAN-based approaches, by focusing on aligning the entire feature distribution rather than solely lower-order statistics, such as mean and variance. While traditional methods ensure alignment of basic statistical properties, they fail to capture higher-order moments, including skewness and kurtosis, which are critical for representing the complex and non-Gaussian characteristics of illumination distributions in real-world scenes.

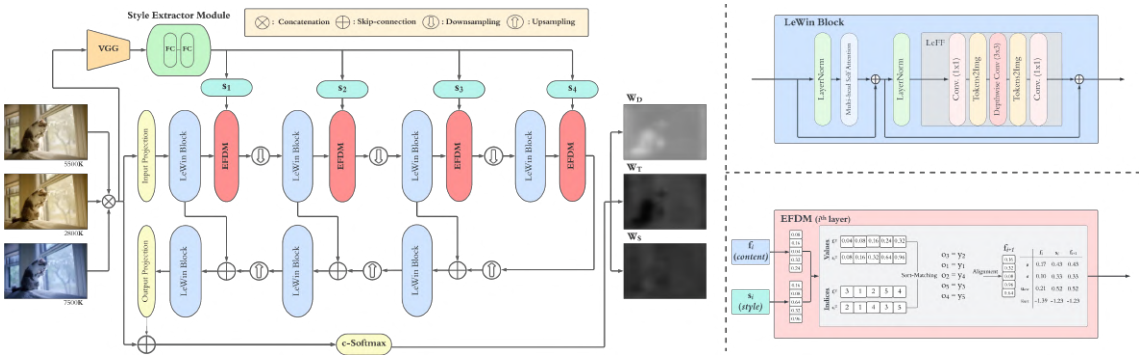
To conceptualize this difference, standard alignment methods can be likened to matching the central tendencies and overall spread of two distributions, analogous to aligning the height and width of hills. However, this approach neglects finer structural details, such as asymmetries or additional peaks within the distributions, which can significantly influence the performance of WB correction, especially under challenging multi-illuminant conditions. EFDM addresses this limitation by aligning the cumulative distribution functions (CDFs) of feature statistics, thereby ensuring a more complete representation of the illumination distribution.

By aligning the entire distribution, EFDM enables the model to accurately capture and correct both subtle and complex lighting variations. This capability is particularly advantageous in multi-illuminant scenarios, where lighting variations are inherently intricate and non-uniform. Consequently, EFDM provides a robust foundation for WB correction, which can achieve improved global consistency and local color fidelity compared to our prior methods that rely on simpler statistical alignment techniques. This makes EFDM uniquely suited to address the challenges of multi-illuminant WB correction.

4.2.3.2 Style Extraction Mechanism

Our proposed architecture comprises a multi-head *Style Extractor* module that projects features, encoded by a pretrained VGG network, into a latent style vector $\mathbf{w}$. Each WB setting (*e.g.*, Tungsten, Daylight, Shade) is processed through VGG, with concatenated features forming the input for the Style Extractor module. This module generates style feature vectors $\mathbf{s}_i$, with each vector linked to a specific encoder level. The multi-head structure of this module applies affine transformations at each level, which adapts feature

Figure 4.8: *Our proposed architecture (FDM WB) for White Balance correction.*



statistics to accommodate illumination complexity. This process is formalized as follows

$$\begin{aligned} \mathbf{z} &= \text{concat}\left(VGG(\mathbf{I_T}), VGG(\mathbf{I_D}), VGG(\mathbf{I_S})\right), \\ \mathbf{w} &= SE(\mathbf{z}), \\ \mathbf{s}_i &= s_i(\mathbf{w}), \end{aligned} \tag{4.16}$$

where $\mathbf{I_T}$, $\mathbf{I_D}$, and $\mathbf{I_S}$ denote input images under different WB conditions, and SE represents the Style Extractor module.

4.2.3.3 Uformer with Feature Distribution Matching

The proposed model adopts a U-shaped Transformer architecture, which utilizes Locally-enhanced Window Transformer (LeWin) blocks, with EFDM layers for robust WB correction. LeWin blocks can capture long-range dependencies within localized windows while controlling computational overhead. Within each block, injected style information, color temperature or illumination in our scenario, is processed. Through EFDM, the model shifts feature distributions to align with target style spaces generated by the Style Extractor, which learns to represent a pure white-balanced style. This is achieved by optimizing an objective function that guides the generation of accurate weighting maps for each WB setting. Analogously to StyleGAN's mapping network that transforms noise

into a style space for facial attributes, the Style Extractor in our method generates a distinct style space tailored for white balance, enabling precise adjustment under various illumination conditions. This process is formalized as

$$\begin{aligned} \mathbf{f_i} &= LeWin(\mathbf{c_i}), \\ \tilde{\mathbf{f}}_\mathbf{i} &= EFDM(\mathbf{f_i}, \mathbf{s_i}), \\ \mathbf{f_{i+1}} &= DS(\tilde{\mathbf{f}}_\mathbf{i}) \end{aligned} \tag{4.17}$$

where $\mathbf{c_i}$ is the input at the $\mathbf{i}^{th}$ encoder level, $\mathbf{s_i}$ denotes the target style feature vector, $\tilde{\mathbf{f}}_\mathbf{i}$ is the EFDM-aligned feature map, and $\mathbf{f_{i+1}}$ is the downsampled output. The decoder mirrors the encoder's structure, using skip connections to ensure spatial detail preservation. As aforementioned before, the final *sRGB* outputs are produced by linearly blending the input images and the corresponding weighting maps, as illustrated in Figure 4.2 and formulated in Equation 4.11. The overall architecture of Uformer with Feature Distribution Matching for WB correction (FDM WB) is illustrated in Figure 4.8.

To facilitate a fair comparison with prior studies, our optimization combines reconstruction and smoothing losses with the same hyperparameters. The reconstruction loss $\mathcal{L}_r$ minimizes the discrepancy between corrected and ground truth images as

$$\mathcal{L}_r = ||\mathbf{I_{gt}} - \mathbf{I_c}||_F^2 \tag{4.18}$$

where $\mathbf{I_{gt}}$ represents the ground truth image, and $||\cdot||_F^2$ denotes the Frobenius norm. The smoothing loss $\mathcal{L}_s$ regularizes weighting map edges using Sobel filters, addressing artifacts along horizontal and vertical axes as

$$\mathcal{L}_s = \sum\nolimits_{t \in WB} ||W_t * \nabla_x||_F^2 + ||W_t * \nabla_y||_F^2 \tag{4.19}$$

where ∇_x and ∇_y are Sobel filters with 3×3 kernels. Our final objective function is as

$$\mathcal{L} = \mathcal{L}_r + \lambda \mathcal{L}_s \tag{4.20}$$

where λ is set to 100, aligning with previous studies to ensure comparability.

4.2.4 Feature Distribution Statistics as a Loss Objective

Building upon our foundational approaches of treating illumination as a style factor, this final approach leverages EFDM applied specifically to the [CLS] token in Vision Transformers (ViT) [4]. Unlike earlier methods that introduced specialized architectural modules for adaptive feature normalization or feature alignment through exact matching within the Uformer architecture, this approach simplifies the architecture while targeting the global style representation captured by the [CLS] token. By incorporating a novel EFDM-based loss function, we enable precise alignment of the distributional characteristics of illuminated and ground truth white balanced scenes, focusing particularly on the higher-order statistics, which are critical to capturing complex illumination variations in scenes, as presented in analysis in Section 4.2.2.

In Vision Transformers (ViTs), the [CLS] token is commonly used as a global representation of the entire image, as it aggregates feature information across all patches in the input image [55]. This aggregation property makes the [CLS] token particularly well-suited for capturing scene-wide style characteristics, which includes the variations in illumination. Unlike lower-level features, which primarily represent localized spatial details, the [CLS] token encapsulates global attributes. These attributes include color balance, illumination sources, and chromatic consistency, thus serving as an ideal proxy for style information within images.

Previous analyses, as discussed in Section 4.2.2, highlight that lighting in multi-illuminant scenes introduces non-Gaussian characteristics in feature distributions, which are manifested as skewness and kurtosis variations. These higher-order moments are crucial for capturing the full extent of lighting variations in complex scenes, as simple alignment of mean and variance fails to address these intricacies. The [CLS] token, because of its global representation, inherently includes these higher-order statistical nuances. Leveraging EFDM on the [CLS] token allows the model to match not just the basic statistics

but the entire feature distribution, which encompasses both the lower- and higher-order moments that define illumination-based style variations.

4.2.4.1 Proposed objective function

The core of the proposed approach is an EFDM-based loss function that aims to exactly match the [CLS] token distributions between the predicted and ground truth white balanced images. This matching, as an optimization objective, ensures that the model captures and corrects both simple and complex lighting variations across scenes, which enables more accurate WB correction in multi-illuminant scenarios. Our proposed loss function is formulated as a feature distribution alignment process, where the [CLS] token distributions of the predicted image are matched with those of the ground truth image by minimizing their distributional discrepancy.

Using a pretrained ViT as a feature extractor, we obtain the [CLS] token from both the predicted and ground truth images. This [CLS] token, denoted as f_{pred} for the predicted image and f_{gt} for the ground truth, acts as a compact and comprehensive representation of the entire image. By design, the [CLS] token aggregates information across all input patches during the attention mechanism of the ViT, which makes it a powerful summary feature for global image characteristics.

The [CLS] token exhibits the unique capability of encapsulating the style of each image, particularly capturing illumination characteristics such as intensity, color temperature, and the interplay of multiple light sources. This characteristic makes it highly suitable for tasks that require modeling complex illumination conditions. Unlike patch-level features, which retain localized information, the [CLS] token inherently captures the global illumination context by considering inter-dependencies among all patches in the image. This global perspective is critical for accurate WB correction, as lighting variations often manifest in both local and global contexts within a scene.

Furthermore, by focusing on the [CLS] token, we significantly reduce the complexity of feature matching. Instead of matching high-dimensional feature maps or patch-level representations, the model considers a single vector per image. This reduction in dimensionality simplifies the computational process, while ensuring that the critical global illumination information is retained. The compact representation provided by the [CLS] token enables efficient learning of matching feature distributions, particularly in scenarios involving complex, non-uniform lighting conditions. This approach also ensures scalability, allowing the model to generalize effectively across diverse datasets and lighting scenarios while maintaining computational efficiency.

EFDM ensures that the [CLS] token distributions between the predicted and ground truth images are matched across their entire cumulative distribution functions (CDFs), capturing all moments of the distribution, including skewness and kurtosis. This approach effectively accounts for the non-Gaussian characteristics introduced by multi-illuminant conditions, so that the model can manage complex lighting variations with high accuracy. Using the [CLS] token as the basis for feature distribution matching, the approach balances computational simplicity with representational richness. The [CLS] token, as a compact yet holistic representation of global image characteristics, effectively encapsulates illumination-related style information. This synergy between EFDM and distribution matching of the [CLS] token creates a robust foundation for precise color and illumination adjustments, making it particularly effective for WB correction in complex, multi-illuminant scenarios.

Our proposed loss function is defined as

$$L_{\text{EFDM}}(f_{\text{pred}}, f_{\text{gt}}) = \frac{1}{n}\sum_{i=1}^{n}\left[f_{\text{pred}}[i] - f_{\text{gt}}[\text{rank}(f_{\text{pred}}[i])]\right]^2 \tag{4.21}$$

where f_{pred} and f_{gt} are the [CLS] token vectors of the predicted and ground truth images, respectively. $\text{rank}(f_{\text{pred}}[i])$ returns the index corresponding to the rank of the i-th value of f_{pred} in the sorted vector f_{gt}. n is the number of elements in the [CLS] token feature vector

(*i.e.*, the dimension). This loss function effectively matches the sorted distributions of f_{gt} to the rank-sorted distribution f_{pred}, which ensures the precise matching of the higher-order statistics.

The practical implementation of EFDM involves computing a rank-based transformation of the feature values in the predicted and target feature maps, followed by a point-wise comparison. Instead of relying on absolute intensity values, EFDM operates by sorting the feature values in both the predicted and reference feature representations and applying a transformation based on their relative ranks. This approach makes the loss function invariant to scale and intensity variations, thus ensuring that feature alignments remain consistent under different conditions.

The rank transformation, which is applied independently per feature channel, ensures that corresponding features are matched in a way that reflects their statistical relationships rather than their absolute magnitudes. This process enables EFDM to effectively capture distributional discrepancies between feature representations and enforce feature-level alignment without being affected by variations in scale, intensity, or structural distortions. Unlike conventional loss functions that rely on mean and variance normalization, EFDM aligns the entire feature distribution, allowing for a more precise optimization target in the context of WB correction.

A simplified pseudocode representation of this implementation is provided in Algorithm 1, detailing the step-by-step process of computing our proposed loss function. The rank transformation within the algorithm is implemented using the `argsort` function, which determines the sorted order of the feature values, followed by `gather`, which maps these sorted indices back to their original positions to produce rank indices. This formulation ensures that the ranking process is numerically stable and is consistently applied across feature channels.

Our latest proposed approach marks a paradigm shift in optimization by enforcing

EFDM on the [CLS] token as the objective of optimization. This shifts from enhancing neural network modules, yet depending basic objective functions to developing a sophisticated loss function that explicitly focuses on full distribution alignment. This shift enables the model to better capture both lower- and higher-order moments of the feature distribution, and helps to address complex illumination dynamics in multi-illuminant scenes. Unlike previous methods that focus on basic statistics such as mean and variance, EFDM effectively aligns the entire distribution to the target, including asymmetries and tail characteristics. This refined optimization perspective ensures the preservation of intricate lighting details while maintaining global illumination consistency and local color fidelity, which makes it a more suitable solution for challenging WB correction tasks.

Instead of introducing a new architecture or retaining the *Style Modulator* module injected into Uformer in our previous approach, we focus solely on applying the Uformer [5] and UNet [59] architectures, both with and without the proposed loss function, to address the challenging multi-illuminant scenarios of the LSMI dataset [8]. The integration of EFDM as the objective demonstrates significant improvements in handling the complex lighting variations inherent to such scenes, which showcases the adaptability of the proposed loss function and its critical role in enhancing the performance of these architectures under diverse illumination conditions.

5. EXPERIMENTS

This chapter presents the experimental framework designed to evaluate the three proposed approaches for WB correction under single- and multi-illuminant conditions. The progression of these approaches—from learning style factors for WB correction to leveraging exact feature matching within architectures, and finally simplifying the model to use exact feature matching as a standalone objective function—provides a comprehensive exploration of the trade-offs between model complexity and performance.

The experiments aim to address the following key objectives:

- **Quantitative validation:** Assessing the effectiveness of each proposed approach in correcting WB across diverse illumination scenarios, particularly mixed-illuminant conditions (*i.e.*, single and multi).
- **Impact of exact distribution matching via EFDM:** Evaluating the contribution of EFDM, both as a module integrated into the architecture and as a standalone objective function, in achieving superior alignment of feature distributions.
- **Model simplification and performance trade-offs:** Investigating how the removal of architecture-specific modules and the sole reliance on EFDM as a loss function affect performance, generalizability, and computational efficiency.

For the evaluation, the RenderedWB dataset [20] is used for single-illuminant benchmarks, while their synthetic Multi-Illuminant Evaluation Set provided a controlled multi-illuminant testing environment for the earlier methods. For the final approach, the LSMI dataset [8] is used, as it offers a well-curated benchmark for challenging real-world multi-illuminant scenarios. By leveraging these datasets, the experiments transition from single-illuminant benchmarks to multi-illuminant challenges, which highlights the strengths and limitations of each proposed approach. This progression underscores

the practical relevance of the proposed methodology in handling increasingly complex lighting scenarios.

5.1 Experimental Setup

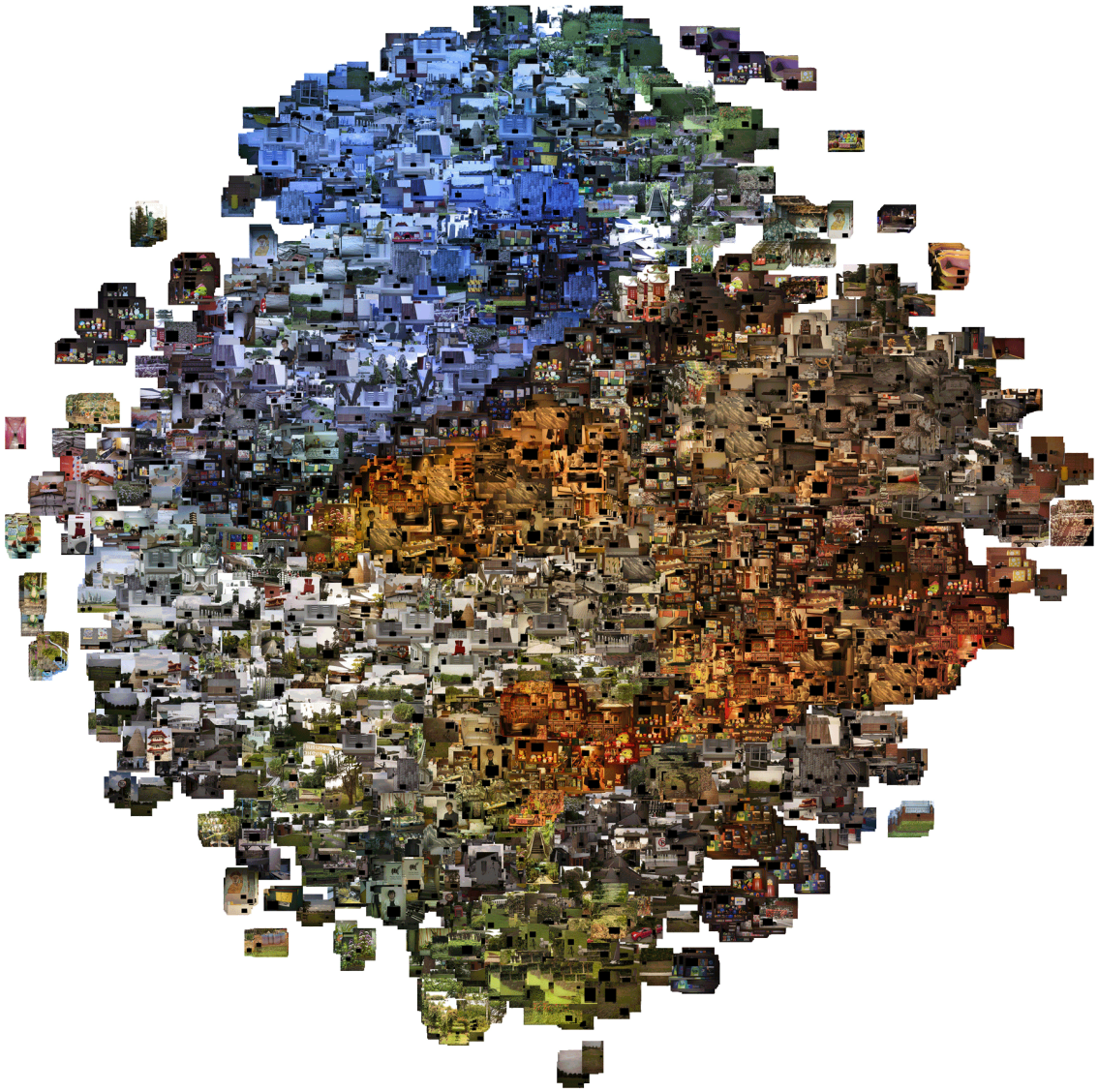
This section outlines the datasets and training details that are used to evaluate the proposed approaches for WB correction. It includes details on the training datasets, evaluation protocols, metrics, and computational environments used in all experiments. All experimental details for the proposed approaches are presented in Table 5.1.

5.1.1 Datasets and Evaluation Protocols

The first two proposed approaches were trained and evaluated on the RenderedWB dataset [6], which contains 65,000 *sRGB* images captured by various cameras with specific pre-defined WB settings. These WB settings include two configurations: {`t, f, d, c, s`} (*i.e.*, Tungsten at 2800**K**, Fluorescent at 3800**K**, Daylight at 5500**K**, Cloudy at 6500**K**, and Shade at 7500**K**) and {`t, d, s`} (*i.e.*, a subset of the former). Each image in the dataset has a corresponding ground truth image that is accurately white-balanced. Figure 5.1 presents the t-SNE visualization of the training images of the RenderedWB dataset, derived from their corresponding Principal Component Analysis (PCA) feature vectors. For quantitative evaluation, we utilized Cube+ [1] and the synthetic multi-illuminant evaluation set proposed by Afifi et al. [2]. Cube+ comprises 1,707 single-illumination, color-calibrated images captured using a Canon EOS 550D camera across different environments. The synthetic multi-illuminant evaluation set consists of 150 rendered images, which are created using Autodesk 3Ds Max, featuring multiple illumination scenarios. Furthermore, qualitative evaluations were performed on the MIT-Adobe FiveK dataset [7], which includes 5,000 images captured by DSLR cameras and retouched by professional photographers.

The Large Scale Multi-Illuminant (LSMI) dataset, introduced by Kim *et al.* [8],

Figure 5.1: *t-SNE visualization of the training images of the RenderedWB dataset, based on their corresponding PCA feature vectors. Obtained from [6].*



is used for training and evaluation in the final proposed approach, due to its unique suitability for multi-illuminant scenarios. The LSMI dataset comprises 7,486 meticulously annotated images captured in more than 2,700 diverse indoor and outdoor scenes, utilizing three different camera models: Samsung Galaxy Note 20 Ultra, Sony α9, and Nikon

Figure 5.2: *Example images from the LSMI dataset (first row) alongside their corresponding illuminant coefficient maps (second row).*

D810. Each image in the dataset provides pixel-wise ground truth information about illuminant chromaticity, per-pixel illumination levels, and the mixture ratios of multiple illuminants within the scene, as illustrated in Figure 5.2. This rich annotation enables an unprecedented level of detail in modeling and correcting complex illumination conditions, which makes the dataset a benchmark choice for advancing WB correction methods in multi-illuminant scenarios. Furthermore, the dataset includes images with varying color temperatures, scene compositions, and lighting complexities, which reflect real-world diversity and challenges in illumination scenarios. Its extensive annotations and challenging conditions offer an ideal testbed for evaluating the efficacy of our last proposed approach, which aims to capture higher-order statistical characteristics of illumination distributions.

5.1.2 Training Details

The training procedures for the first two approaches adhere to a consistent setup. RenderedWB images were randomly cropped at resolutions of 64×64 and 128×128, and no data augmentation techniques were applied. For the first approach, Adam Optimizer [145] is used with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, a learning rate of 1×10^{-4}, and a batch size of 32. The second approach utilized the AdamW optimizer [146] with the same hyperparameters, but the batch size was increased to 40, composed of 4 patches from 10 distinct samples per batch. Both approaches were trained for 200 epochs without learning rate scheduling.

Post-processing operations were applied to enhance the weighting maps before generating the final WB corrected outputs. These included multi-scale inference of the weighting maps (*ms*) and edge-aware smoothing (*eas*) using a fast bilateral solver [144], as mentioned in Section 4.2.1.3. For inference, the models processed low-resolution input images (*i.e.*, 384×384) rendered under pre-defined WB settings, which produces weighting maps resized to the input resolution for final blending.

For the last approach, the training is performed using both U-Net [59] and Uformer [5] architectures, with and without the proposed objective function. The architectures were adapted to have three input channels and two output channels for predicting *UV* chromaticity channels for WB correction. The input resolution is fixed at 256×256 pixels to balance computational efficiency and model performance. The training process spans 2,400 epochs, with a batch size of 32, using the Adam optimizer with hyperparameters set to $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The initial learning rate is set to 2×10^{-4}, with a linear decay starting after 800 epochs to ensure stable convergence. To simulate real-world variations, data augmentation techniques are incorporated for illumination sources, including random cropping, illumination augmentation, and adjustments to *Saturation*, and *Value* ranges. The saturation is varied between 0.2 and 0.8, while the value ranges are restricted

Table 5.1: *Experimental details for the proposed approaches*

	Style WB	**FDM WB**	**FDM Loss**
Dataset	RenderedWB [6]		LSMI [8]
Architecture	IFRNet [41] Style Modulator AdaIN [44]	Uformer [5] Style Modulator EFDM [30]	U-Net [59] & Uformer [5] FDM Loss EFDM [30]
Input Size	64×64 128×128		256×256
Batch Size	32	40	32
Optimizer	Adam [145]	AdamW [146]	Adam [145]
Learning Rate	1×10^{-4}	1×10^{-4}	2×10^{-4}
Epochs	200	200	2400
Augmentation	None		Illumination Color Random Cropping
Post-Processing	Multi-scale inference Edge-aware smoothing		None
Evaluation Metrics	Pixel Accuracy Angular Error Color Difference		Angular Error
GPU Setup	$2\times$ RTX 2080Ti	$2\times$ RTX 2080Ti	$2\times$ A100
Framework	PyTorch [147]		

between 0.01 and 0.99 to avoid extreme lighting distortions. Random cropping ensures that the model learns from varying spatial contexts within the dataset, and the augmentation of illumination helps to emulate real-world lighting inconsistencies. This training framework is designed to fully exploit the detailed annotations of the LSMI dataset and to evaluate the robustness and efficacy of the proposed method in correcting multi-illuminant WB scenarios.

5.1.3 Metrics for Evaluation

To assess the performance of the proposed approaches, different evaluation metrics are employed based on the approach and its specific goals.

- **Mean Squared Error (Mean Squared Error (MSE)):** Measuring the pixel-wise difference between the ground truth and corrected images. This metric is used

for the first two approaches to evaluate pixel-level accuracy in single- and multi-illuminant scenarios.

- **Mean Angular Error (MAE):** Evaluating the angular difference between the vector forms of the ground truth and the predicted images in *RGB* color space. For the last approach, both mean and median MAE are specifically reported to balance the evaluation of average correction accuracy and robustness to outliers in the challenging multi-illuminant scenarios of the LSMI dataset.

- **Color Difference ($\triangle$E 2000):** Quantifying the perceptual color differences in the *L*C*H** color space between the ground truth and the predicted images. This metric is applied only for the first two approaches for single- and multi-illuminant benchmarks.

To provide a comprehensive analysis for the first two approaches, we report the mean values and quantile averages (*i.e.*, first, second, and third quantiles) of the metrics for each evaluation set. For the last approach, the focus is on the mean and median values of MAE to emphasize its robustness in multi-illuminant scenarios.

5.1.4 Computational Environment

The experiments for the first two approaches are conducted on a computational setup equipped with $2\times$ NVIDIA RTX 2080Ti GPUs. For the last approach, the experiments utilize a computational setup with $2\times$ NVIDIA A100 GPUs, which provides the necessary resources to manage the large dataset and high computational demands. All implementations are carried out using the PyTorch framework [147], building upon prior works in WB correction [20, 2, 8].

5.2 Results and Discussion

This section presents the evaluation of the proposed approaches under various experimental setups. The results are analyzed comprehensively to highlight the performance of the methods in both single- and multi-illuminant scenarios. Benchmark datasets, including the Cube+ [1] dataset, the LSMI [8] dataset, and the synthetic mixed-illuminant evaluation set [2], are used to provide quantitative and qualitative evaluations. In addition, ablation studies are conducted to validate design choices and practical applications, such as night photography, are explored to demonstrate the robustness of the proposed approaches.

5.2.1 Experimental Results for Style WB

Our first proposed approach, namely *Style WB,* models lighting in both single- and mixed-illuminant scenarios as a style factor to enhance WB correction. This strategy extends WB correction method proposed by Afifi *et al.* [2] as aiming to refine performance by leveraging detailed weighting maps and avoiding explicit illuminant estimation. Below, the results of the method are discussed in single-illuminant and mixed-illuminant contexts, followed by insights from ablation studies and its application in night photography.

5.2.1.1 Benchmark on Single-Illuminant Scenarios

To evaluate the performance in single illumination scenarios, we conduct experiments using the Cube+ dataset [1]. The results, summarized in Table 5.2, compare the proposed method with state-of-the-art approaches [17, 19, 6, 126, 2]. Metrics such as Mean Squared Error (MSE), Mean Angular Error (MAE), and Color Difference (ΔE 2000) are reported for patch sizes $p = 64$ and $p = 128$, along with different WB settings ($\{\texttt{t,d,s}\}$ and $\{\texttt{t,f,d,c,s}\}$).

Table 5.2: *Benchmark on single-illuminant Cube+ dataset [1]. The top results are indicated with colored cells as, the best: green, the second: yellow, the third: red.*

Method	MSE				MAE				ΔE 2000				Size
	Mean	Q1	Q2	Q3	Mean	Q1	Q2	Q3	Mean	Q1	Q2	Q3	
FC4 [17]	371.90	79.15	213.41	467.33	6.49°	3.34°	5.59°	8.59°	10.38	6.60	9.76	13.26	5.89 MB
Quasi-U CC [19]	292.18	15.57	55.41	261.58	6.12°	1.95°	3.88°	8.83°	7.25	2.89	5.21	10.37	622 MB
KNN WB [6]	194.98	27.43	57.08	118.21	4.12°	1.96°	3.17°	5.04°	5.68	3.22	4.61	6.70	21.8 MB
Interactive WB [126]	159.88	21.94	54.76	125.02	4.64°	2.12°	3.64°	5.98°	6.20	3.28	5.17	7.45	**38 KB**
Deep WB [20]	**80.46**	15.43	33.88	74.42	3.45°	1.87°	2.82°	4.26°	4.59	2.68	3.81	5.53	16.7 MB
Mixed WB [2]													
$p = 64$, WB=$\{\texttt{t},\texttt{d},\texttt{s}\}$	168.38	8.97	19.87	105.22	4.20°	1.39°	2.18°	5.54°	5.03	2.07	3.12	7.19	5.09 MB
$p = 64$, WB=$\{\texttt{t},\texttt{f},\texttt{d},\texttt{c},\texttt{s}\}$	161.80	9.01	19.33	90.81	4.05°	1.40°	2.12°	4.88°	4.89	2.16	3.10	6.78	5.10 MB
$p = 128$, WB=$\{\texttt{t},\texttt{f},\texttt{d},\texttt{c},\texttt{s}\}$	176.38	16.96	35.91	115.50	4.71°	2.10°	3.09°	5.92°	5.77	3.01	4.27	7.71	5.10 MB
Style WB (ours)													
$p = 64$, WB=$\{\texttt{t},\texttt{d},\texttt{s}\}$	92.65	**6.52**	**14.23**	**35.01**	**2.47°**	**0.82°**	**1.44°**	**2.49°**	**2.99**	**1.36**	**2.04**	**3.32**	61.0 MB
$p = 64$, WB=$\{\texttt{t},\texttt{f},\texttt{d},\texttt{c},\texttt{s}\}$	151.38	29.49	56.35	125.33	4.18°	2.13°	3.03°	4.81°	5.42	3.11	4.42	6.76	61.1 MB
$p = 128$, WB=$\{\texttt{t},\texttt{d},\texttt{s}\}$	88.03	7.92	17.73	45.01	2.61°	0.93°	1.58°	2.85°	3.24	1.50	2.30	3.95	61.2 MB
$p = 128$, WB=$\{\texttt{t},\texttt{f},\texttt{d},\texttt{c},\texttt{s}\}$	100.24	10.77	37.74	70.18	3.09°	1.15°	2.61°	3.87°	3.96	1.59	3.55	5.51	61.3 MB

Our method achieves the best performance in most cases, particularly with smaller patch sizes (*i.e.*, $p = 64$) and fewer WB settings (*i.e.*, $\{\texttt{t},\texttt{d},\texttt{s}\}$). These configurations enable for more precise modeling of the illuminants by reducing the complexity of blending multiple WB settings. Although increasing the number of WB settings introduces additional channels during training, it also increases the complexity of modeling correlations, which can hinder performance for specific metrics.

In addition to quantitative results, we evaluate the qualitative performance of our method using the MIT-Adobe FiveK dataset [7]. Figure 5.3 presents examples of predicted weighting maps and WB correction results. The images demonstrate that our method represents the illuminants in a detailed and interpretable manner, accurately differentiating between multiple illuminants within the same scene. This capability enables our approach to outperform prior methods, as it produces more precise weighting maps and achieves competitive per-pixel performance on WB correction in the *sRGB* space. Moreover, in Figure 5.4, we introduce the comparison of the qualitative results of our initial WB correction approach and recent methods [20, 2], along with the default AWB versions, on the selected samples from the same dataset. It shows that our proposed method achieves competitive per-pixel performance on WB correction in the *sRGB* space compared to the recent methods.

Figure 5.3: *Example of predictions for the weighting maps and WB correction results by blending these maps. We render the linear raw DNG files for the images in MIT-Adobe FiveK dataset [7] (id:* 323, 2808*) in different WB settings.*

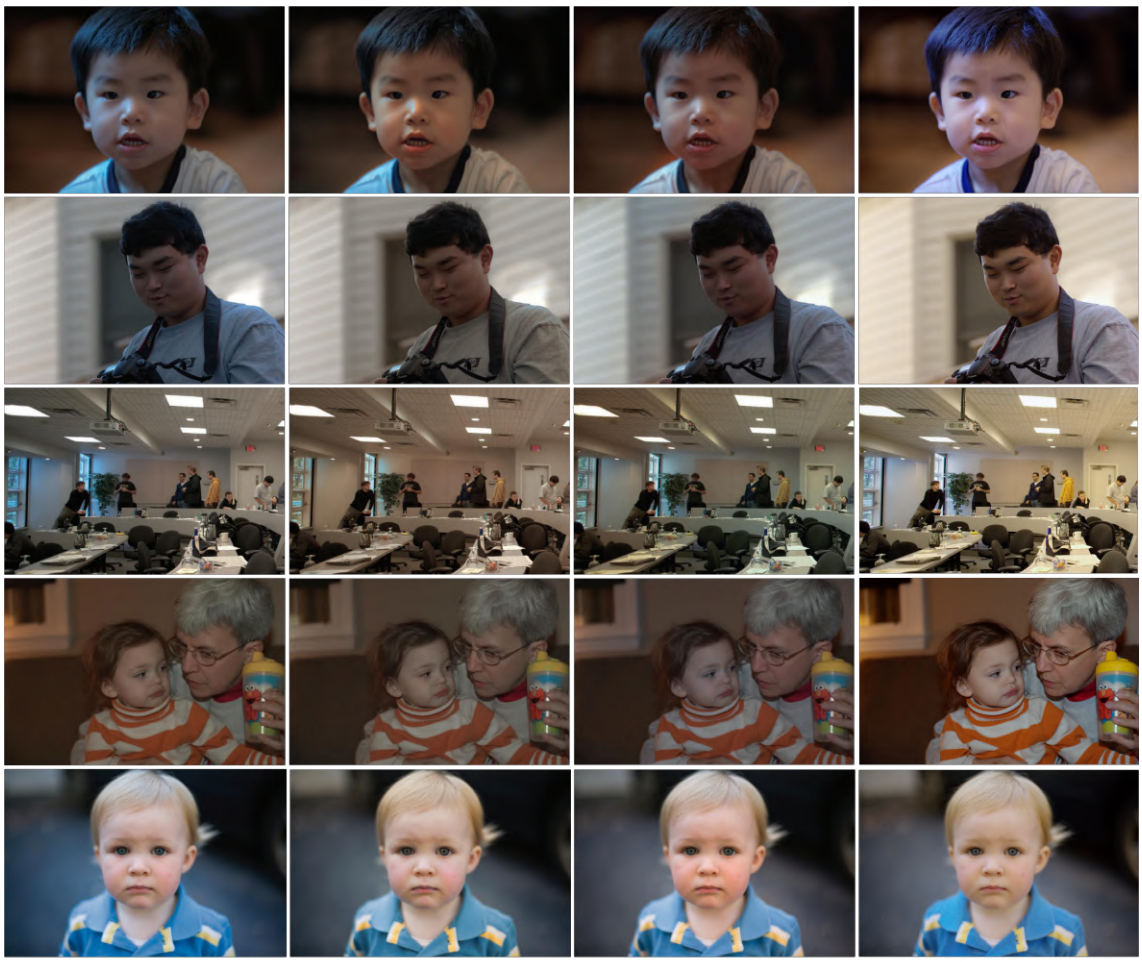
Table 5.3: *Benchmark on mixed-illuminant evaluation set [2]. The top results are indicated with colored cells as, the best: green, the second: yellow, the third: red.*

Method	MSE				MAE				Δ E 2000			
	Mean	Q1	Q2	Q3	Mean	Q1	Q2	Q3	Mean	Q1	Q2	Q3
Gray Pixel [104]	4959.20	3252.14	4209.12	5858.69	19.67°	11.92°	17.21°	27.05°	25.13	19.07	22.62	27.46
Grayness index [105]	1345.47	727.90	1055.83	1494.81	6.39°	4.72°	5.65°	7.06°	12.84	9.57	12.49	14.60
KNN WB [6]	1226.57	680.65	1062.64	1573.89	5.81°	4.29°	5.76°	6.85°	12.00	9.37	11.56	13.61
Interactive WB [126]	1059.88	616.24	896.90	1265.62	5.86°	4.56°	5.62°	6.62°	11.41	8.92	10.99	12.84
Deep WB [20]	1130.60	621.00	886.32	1274.72	**4.53°**	**3.55°**	4.19°	**5.21°**	10.93	**8.59**	**9.82**	11.96
Mixed WB [2]												
$p = 64$, WB=$\{$t,d,s$\}$	**819.47**	655.88	845.79	1000.82	5.43°	4.27°	4.89°	6.23°	**10.61**	9.42	10.72	**11.81**
$p = 64$, WB=$\{$t,f,d,c,s$\}$	938.02	757.49	961.55	1161.52	4.67°	3.71°	**4.14°**	5.35°	12.26	10.80	11.58	12.76
$p = 128$, WB=$\{$t,d,s$\}$	830.20	584.77	853.01	**992.56**	5.03°	3.93°	4.78°	5.90°	11.41	9.76	11.39	12.53
$p = 128$, WB=$\{$t,f,d,c,s$\}$	1089.69	846.21	1125.59	1279.39	5.64°	4.15°	5.09°	6.50°	13.75	11.45	12.58	15.59
Style WB (ours)												
$p = 64$, WB=$\{$t,d,s$\}$	868.01	649.36	889.00	1026.98	5.73°	4.48°	5.42°	6.34°	12.11	10.42	12.12	13.36
$p = 64$, WB=$\{$t,f,d,c,s$\}$	1051.07	760.86	1024.00	1332.50	6.30°	4.43°	6.01°	7.69°	14.43	11.90	13.11	16.15
$p = 128$, WB=$\{$t,d,s$\}$	822.77	**576.52**	**840.67**	1025.26	5.11°	3.93°	4.85°	5.51°	11.65	10.63	11.86	13.02
$p = 128$, WB=$\{$t,f,d,c,s$\}$	834.28	629.95	842.71	1005.59	5.71°	4.57°	5.54°	6.19°	11.79	9.84	12.19	13.00

5.2.1.2 Benchmark on Multi-Illuminant Scenarios

We evaluate the performance of our first proposed approach on the synthetic mixed-illuminant evaluation set [2] to further analyze its robustness under challenging multi-illuminant scenarios. The quantitative results are summarized in Table 5.3. These

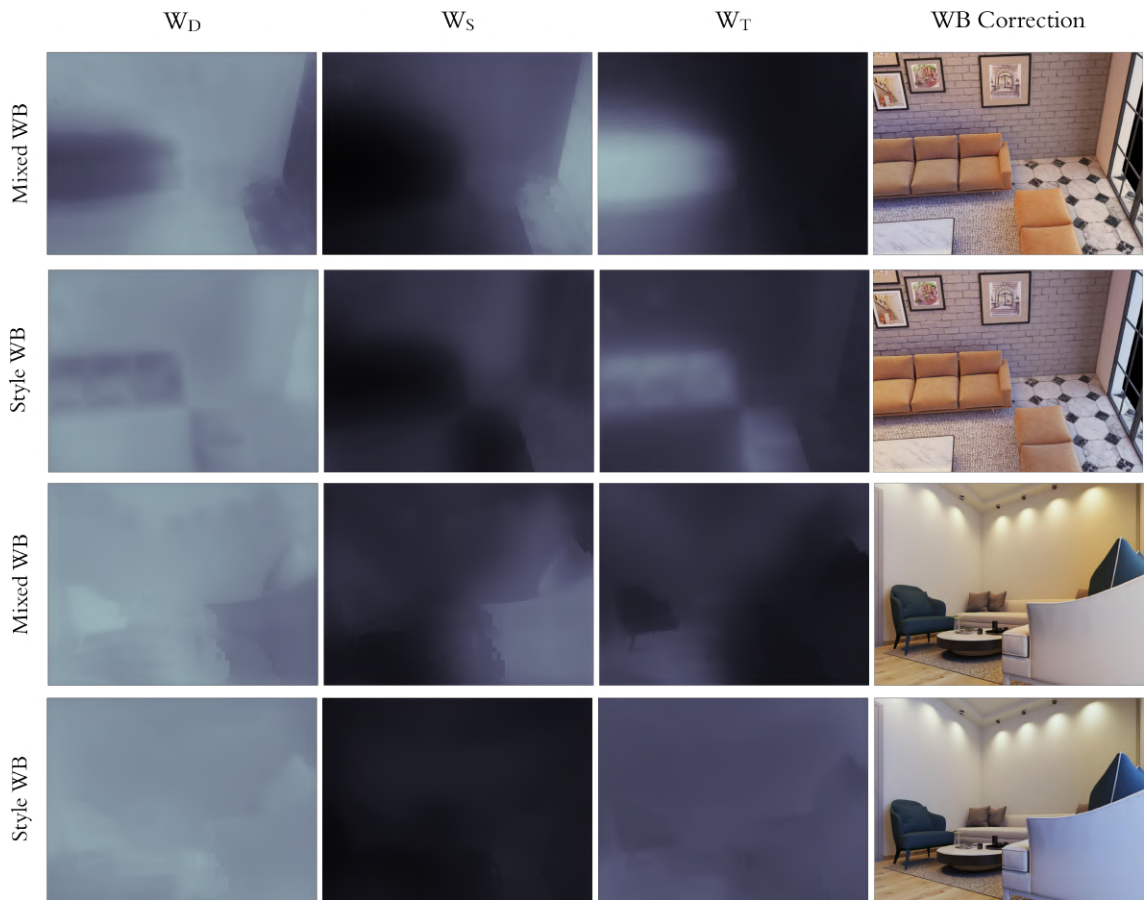
Figure 5.4: *Comparison of the qualitative results of our WB correction method and the other methods on the selected samples from MIT-Adobe FiveK dataset [7]. Image indices from top to bottom:* 2882, 606, 659, 2431, 2550.



results indicate that no single method consistently outperforms others in all metrics, which simply highlight the complexity of multi-illuminant scenes. However, our method achieves superior performance in terms of MSE, which demonstrates its strength in minimizing pixel-wise intensity differences. For MAE and ΔE 2000 metrics, our method produces competitive results compared to the state-of-the-art approaches.

The qualitative comparisons presented in Figure 5.5 further validate the efficacy

Figure 5.5: *Comparison of the performance of the prior work [4] and our method on mixed-illuminant dataset.*



of our approach. The weighting maps generated by our network show significant improvements in detail-oriented representation of the lighting, especially in object regions affected by multiple illumination sources. These maps accurately differentiate between various illuminants that impact the same object, which enhances the blended WB correction results. This capability is achieved through modeling the lighting as a style factor, which captures more nuanced illumination characteristics compared to prior methods.

Despite these strengths, the results on the synthetic dataset reveal certain limitations. Synthetic data often contains sharper edges and transitions compared to real-world images, and our blending strategy, while highly detailed, does not include such synthetic

Table 5.4: *The ablation study on using multi-scale (ms) weighting maps and applying edge-aware smoothing (eas) to weighting maps.*

Models	**MSE**	**MAE**	**ΔE 2000**
Single-illuminant dataset, WB = {t, d, s}, $p = 64$			
$ms = 0$, $eas = 0$	98.55	2.71°	3.32
$ms = 1$, $eas = 0$	93.78	2.59°	3.15
$ms = 0$, $eas = 1$	97.20	2.66°	3.28
$ms = 1$, $eas = 1$	**92.65**	**2.47°**	**2.99**
Mixed-illuminant dataset, WB = {t, d, s}, $p = 128$			
$ms = 0$, $eas = 0$	878.58	5.05°	12.12
$ms = 1$, $eas = 0$	843.50	**5.04°**	11.70
$ms = 0$, $eas = 1$	843.64	**5.04°**	11.98
$ms = 1$, $eas = 1$	**822.77**	5.11°	**11.65**

samples during training. This exclusion may lead to color discrepancies at the edge of the object in the final output, which negatively affects quantitative metrics such as ΔE 2000 and MAE. Nevertheless, the overall results demonstrate the potential of our method to handle multi-illuminant scenarios with high accuracy and interpretability.

5.2.1.3 Ablation Study on Post-processing

To evaluate the contribution of multi-scale weighting maps (*ms*) and edge-aware smoothing (*eas*) used as post-processing, ablation studies are carried out on single and mixed illumination datasets. The results, summarized in Table 5.4, demonstrate that the application of these operations achieves the best overall performance in most metrics.

On the single-illuminant dataset, with experimental settings of WB={t, d, s} and $p = 64$ where p is the patch size, incorporating both techniques improves the MSE from 98.55 to 92.65, the MAE from $2.71°$ to $2.47°$, and the ΔE 2000 from 3.32 to 2.99. Similarly, for the mixed-illuminant evaluation set, with experimental settings of WB={t, d, s} and $p = 128$ where p is the patch size, applying both operations shows the best MSE

(822.77) and ΔE 2000 (11.65), while achieving comparable MAE (5.11°) to other configurations. These results highlight the ability of *ms* to better capture illuminant variations and of *eas* to refine blending at edges, which leads to better overall WB correction.

Although *ms* alone provides significant improvements, the addition of *eas* further enhances the results, especially in reducing color differences. These findings validate the effectiveness of both techniques in improving WB correction accuracy. Although alternative approaches, such as total variation regularization, could be explored, their integration would require significant modifications and is beyond the scope of this work.

5.2.2 Experimental Results for FDM WB

Our second proposed approach, referred as *FDM WB*, leverages feature distribution matching to address the challenges of WB correction under diverse lighting conditions. This method builds upon the principles of modeling illumination as a style factor, which introduces a distribution-based learning mechanism to ensure robustness across varying scenarios. By aligning feature distributions between input and reference domains, FDM WB achieves improved color constancy without relying on explicit illuminant estimation. The following sections provide a detailed discussion of the results, which covers both single- and mixed-illuminant datasets, and are supported by qualitative comparisons and ablation studies that highlight the efficacy of the proposed method.

5.2.2.1 Benchmark on Single-Illuminant Scenarios

Our second proposed approach, FDM WB, demonstrates exceptional performance in single-illuminant scenarios. This evaluation operates on the Cube+ dataset [1], where our method achieves the lowest values in all quantitative metrics, including MSE, MAE, and color difference (ΔE 2000). These results highlight the framework's ability to deliver superior WB correction, which outperforms both recent and traditional methods. Quantitative comparisons, presented in Table 5.5, reveal that FDM WB achieves remarkable

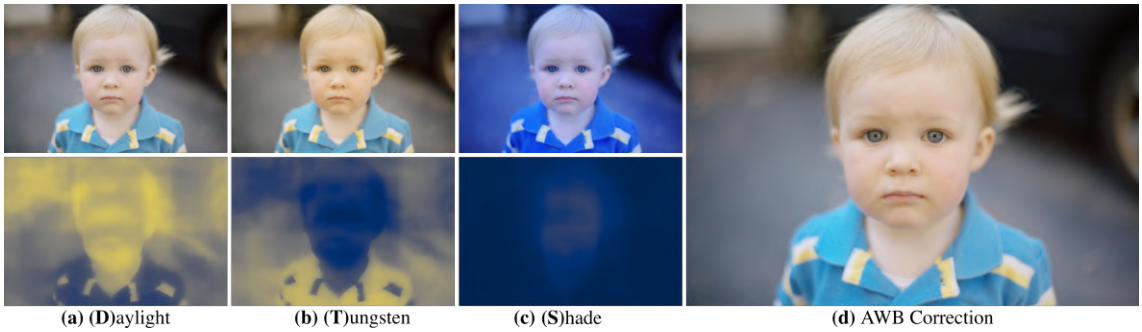
Table 5.5: *Benchmark on single-illuminant Cube+ dataset [1]. ↓ denotes that lower is better.*

	MSE ↓				MAE ↓				ΔE 2000 ↓			
Methods	Mean	Q1	Q2	Q3	Mean	Q1	Q2	Q3	Mean	Q1	Q2	Q3
FC4 [17]	371.90	79.15	213.41	467.33	6.49°	3.34°	5.59°	8.59°	10.38	6.60	9.76	13.26
Quasi-U CC [19]	292.18	15.57	55.41	261.58	6.12°	1.95°	3.88°	8.83°	7.25	2.89	5.21	10.37
KNN WB [6]	194.98	27.43	57.08	118.21	4.12°	1.96°	3.17°	5.04°	5.68	3.22	4.61	6.70
Interactive WB [126]	159.88	21.94	54.76	125.02	4.64°	2.12°	3.64°	5.98°	6.20	3.28	5.17	7.45
Deep WB [20]	80.46	15.43	33.88	74.42	3.45°	1.87°	2.82°	4.26°	4.59	2.68	3.81	5.53
MIMT [138]	-	-	-	-	2.52°	0.98°	1.38°	2.96°	2.88	1.94	2.42	2.87
Mixed WB [2]	161.80	9.01	19.33	90.81	4.05°	1.40°	2.12°	4.88°	4.89	2.16	3.10	6.78
Style WB [24]	88.03	7.92	17.73	45.01	2.61°	0.93°	1.58°	2.85°	3.24	1.50	2.30	3.95
DeNIM + Mixed WB [148]	99.70	13.89	24.71	43.88	2.49°	1.07°	1.62°	2.41°	3.44	1.95	2.74	3.78
DeNIM + Style WB [148]	83.41	13.23	21.46	37.44	1.93°	0.77°	1.09°	1.70°	2.73	1.62	2.03	2.71
FDM WB (**ours**)	**79.35**	**6.46**	**16.84**	**35.76**	**1.35°**	**0.56°**	**1.01°**	**1.66°**	**1.40**	**0.98**	**1.41**	**2.55**

improvements, with an MSE of 79.35, an MAE of $1.35°$, and a ΔE 2000 score of 1.40. These results set a new benchmark for single-illuminant correction, which showcases approximately 42% improvement in ΔE 2000 and 48% in MAE compared to our first proposed method. These significant improvements highlight the effectiveness of exact distribution matching in modeling lighting as a style factor, leading to improved spatial coherence and perceptual quality.

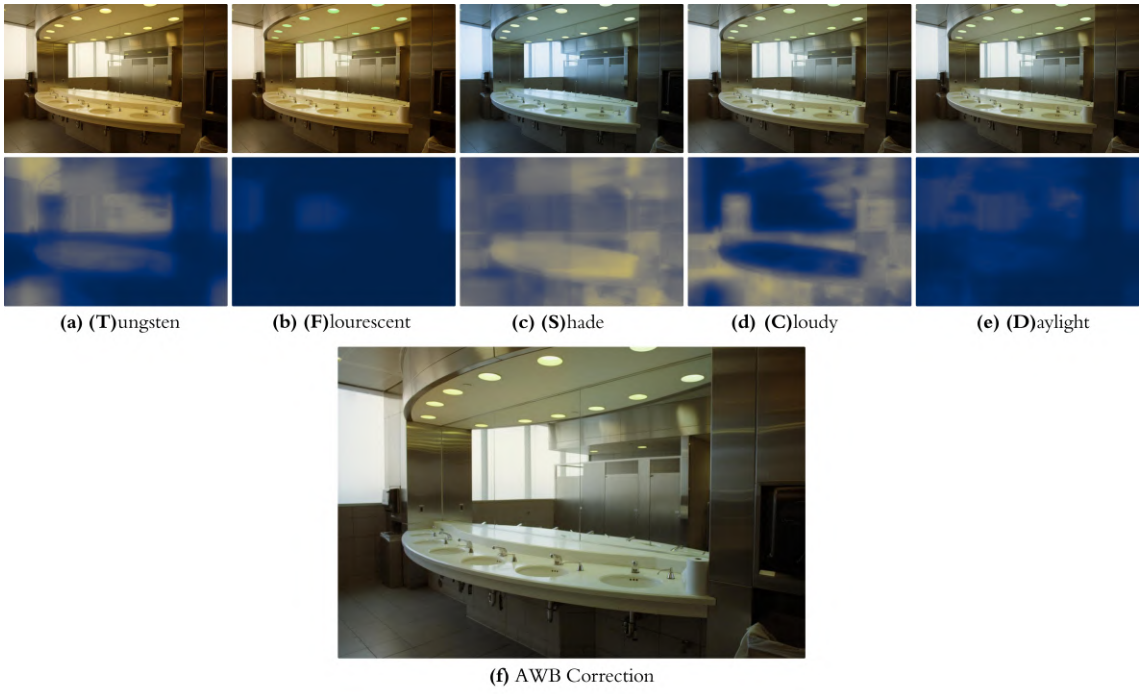
The qualitative results for single-illuminant scenarios, depicted in Figure 5.6 and Figure 5.7, illustrate the performance of FDM WB under 3 and 5 WB settings. The former highlights the corrected outputs for 3 WB settings (*i.e.*, **D**, **S**, **T**), where the weighting maps demonstrate the model's ability to learn spatially consistent blending of WB settings for effective correction. Specifically, our model generates perceptually accurate images by preserving natural tones and details while achieving minimal color discrepancies across the entire scene. For example, natural skin tones are accurately reproduced without introducing undesirable color shifts, which emphasizes spatial consistency. The latter represents the results under the more complex 5 WB setting (*i.e.*, **D**, **S**, **T**, **F**, **C**). This figure further validates the robustness of our proposed approach. As shown in Figure 5.7(**f**), the corrected image aligns closely with human perception of white balance, while the weighting maps reflect enhanced adaptability to varying regions in the scene.

Figure 5.6: *Illustration of WB correction result of FDM WB by learning the weighting maps for 3 WB settings. Sample* 2550 *in MIT-Adobe FiveK dataset.*



(a) (D)aylight **(b)** (T)ungsten **(c)** (S)hade **(d)** AWB Correction

Figure 5.7: *Illustration of WB correction result of FDM WB by learning the weighting maps for 5 WB settings. Sample* 892 *in MIT-Adobe FiveK dataset.*



(a) (T)ungsten **(b)** (F)lourescent **(c)** (S)hade **(d)** (C)loudy **(e)** (D)aylight

(f) AWB Correction

For instance, the sink and mirror areas in the bathroom scene, which are affected by differing illuminants, are corrected uniformly, and this showcases the ability of proposed approach to blend contributions from all five settings and generate a globally consistent result.

Figure 5.8: *Qualitative comparison of the visual results of FDM WB with the prior works on the selected samples from MIT-Adobe FiveK dataset [7]. Image indices from top to bottom:* 323, 659, 2053, 2431.

(a) Deep WB **(b)** Mixed WB **(c)** Style WB **(d)** FDM WB **(ours)**

These visual results emphasize the benefits of integrating EFDM into the Uformer backbone to model illumination as a style factor. Through exact matching of feature statistics, the proposed approach effectively captures contextual information across diverse image regions, and achieves improved perceptual quality and spatial consistency, as demonstrated by the smooth weighting maps and lack of artifacts.

Moreover, Figure 5.8 offers a comparative analysis of our results against state-of-the-art methods such as Deep WB [20], Mixed WB [2], and our first proposed approach [24], namely Style WB. Our approach consistently delivers perceptually superior results, effectively mitigating color distortions and preserving fine details. This highlights the

Table 5.6: *Benchmark on mixed-illuminant evaluation set [2]. ↓ denotes that lower is better.*

	MSE ↓				MAE ↓				ΔE 2000 ↓			
Methods	Mean	Q1	Q2	Q3	Mean	Q1	Q2	Q3	Mean	Q1	Q2	Q3
Gray Pixel [104]	4959.2	3252.1	4209.1	5858.7	19.67°	11.92°	17.21°	27.05°	25.13	19.07	22.62	27.46
Grayness In. [105]	1345.5	727.9	1055.8	1494.8	6.39°	4.72°	5.65°	7.06°	12.84	9.57	12.49	14.60
KNN WB [6]	1226.6	680.7	1062.6	2573.9	5.81°	4.29°	5.76°	6.85°	12.00	9.37	11.56	13.61
Interact. WB [126]	1059.9	616.2	896.9	1265.6	5.86°	4.56°	5.62°	6.62°	11.41	8.92	10.99	12.84
Deep WB [20]	1130.6	621.0	886.3	1274.7	**4.53°**	**3.55°**	**4.19°**	**5.21°**	10.93	**8.59**	9.82	11.96
Mixed WB [2]	819.5	655.9	845.8	1000.8	5.43°	4.27°	4.89°	6.23°	10.61	9.42	10.72	11.81
Style WB [24]	822.8	576.5	840.7	1025.3	5.11°	3.93°	4.85°	5.51°	11.65	10.63	11.86	13.02
FDM WB (**ours**)	**761.9**	**513.9**	**818.4**	**969.3**	5.95°	4.64°	5.88°	6.90°	**10.16**	8.75	**9.81**	**11.69**

advantage of combining exact feature statistics matching with a Transformer-based architecture, which enhances spatial consistency while preserving a high level of detail.

5.2.2.2 Benchmark on Multi-Illuminant Scenarios

The results of the mixed-illuminant benchmark, presented in Table 5.6, underscore the effectiveness of FDM WB compared to the existing state-of-the-art approaches. In particular, our method achieves the lowest MSE score of 761.9, significantly improving over prior works, including Mixed WB [2] (819.5) and Style WB [24] (822.8). This reduction in MSE highlights the model capability of FDM WB to minimize luminance errors even under complex mixed-illuminant scenarios. In terms of color difference ($\Delta E2000$), FDM WB shows better performance with a mean score of 10.16, which outperforms both Mixed WB (10.61) and Style WB (11.65). The improvement is particularly noticeable in the third quartile (**Q3**), where FDM WB achieves 11.69, compared to 13.02 for our first proposed approach. This highlights the performance advantage of exactly matching feature statistics over solely aligning them, particularly in a wide range of challenging cases.

Although MAE for FDM WB ($5.95°$) is slightly higher than Mixed WB ($5.43°$) and Style WB ($5.11°$), it remains competitive. This trade-off can be attributed to the method's focus on achieving better perceptual color correction and spatial consistency,

Figure 5.9: *Qualitative comparison of the visual results of FDM WB with the prior works on the selected samples from the mixed-illuminant evaluation set [2]. Image indices from top to bottom:* 5, 16, 20, 24.

(a) Deep WB **(b)** Mixed WB **(c)** Style WB **(d)** FDM WB **(ours)** **(e)** GT

which can be indicated as evidence for improvements on $\Delta E2000$ scores. In general, these results validate the power of combining feature distribution matching with a Transformer architecture. By leveraging exact statistical alignment of features from the scenes with mixed-illuminant regions, FDM WB consistently achieves robust and accurate WB correction, which establishes itself as a strong candidate for spatially varying illumination scenarios.

The qualitative evaluation results, presented in Figure 5.9, visually compare the performance of FDM WB with the state-of-the-art approaches, namely Mixed WB, Style WB, and Deep WB, in various mixed illumination scenarios. It is important to note that, due to the synthetic nature of the dataset, all methods, including the proposed approach, could not exhibit visually promising results. None of the methods achieve consistently

high performance from a qualitative perspective, as the inherent differences in detail and complexity between synthetic data and real-world scenarios prevent the former from fully replicating the latter's challenges.

Despite these limitations, FDM WB demonstrates notable strengths in specific aspects of perceptual quality. For example, in the first row, the intricate interplay of shadows and highlights in the dark-toned living room is preserved without introducing artifacts, which can demonstrate the ability of FDM WB to maintain detail under challenging lighting conditions. Similarly, in the second row, the focus shifts to the wall area, where FDM WB effectively balances the cool and warm tones, which ensures accurate color correction without compromising the integrity of the brick texture. Unlike prior methods, which either over-saturate or leave residual color casts, FDM WB achieves a neutral tone and enhances the realism of the scene while preserving the fine details of the bricks. These results highlight the potential of FDM WB to handle mixed illumination scenarios more effectively than its predecessors, even if qualitative differences remain subtle in synthetic environments.

5.2.2.3 Ablation Study

To gain a deeper understanding, we perform an ablation study to systematically evaluate the individual contributions of each component in the proposed method and examine the effects of various parameters on the training process.

Impact of Style Extractor and EFDM: Initially, we investigate the impact of building a style feature space using the *Style Extractor* module and the feature distribution matching learning strategy in the proposed method. Table 5.7 presents the quantitative results that compare the Uformer-only architecture with the proposed method in both evaluation datasets. The results clearly demonstrate that, while the Uformer-only architecture offers a notable performance improvement, particularly when compared to prior works, most

Table 5.7: *Ablation study on the impact of employing the Style Extractor module and EFDM on Cube+ dataset [1] and mixed-illuminant evaluation set [2].*

Method	MSE ↓	MAE ↓	ΔE 2000 ↓
Cube+ dataset			
$p = 64$, Uformer [5]	107.38	2.80°	3.46
$p = 64$, FDM WB	91.34	2.38°	2.88
$p = 128$, Uformer [5]	105.68	2.77°	3.39
$p = 128$, FDM WB	**79.35**	**1.35**°	**1.40**
Mixed-illuminant evaluation set			
$p = 64$, Uformer [5]	939.52	4.98°	12.97
$p = 64$, FDM WB	780.74	4.85°	10.84
$p = 128$, Uformer [5]	1067.37	5.99°	14.43
$p = 128$, FDM WB	**761.95**	**5.95**°	**10.16**

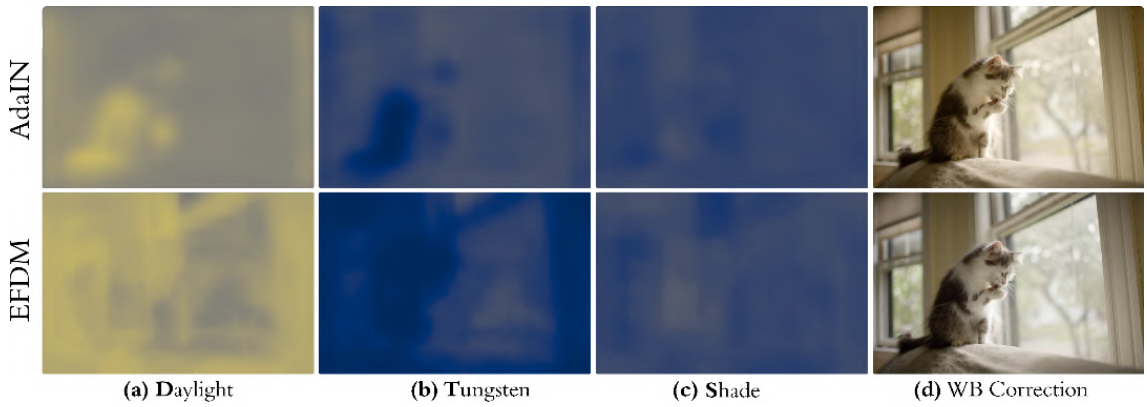
Table 5.8: *Ablation study on style factor learning strategy on Cube+ dataset [1] and mixed-illuminant evaluation set [2].*

Method	MSE ↓	MAE ↓	ΔE 2000 ↓
Cube+ dataset			
AdaIN [44]	92.47	1.78°	1.94
EFDM	**79.35**	**1.35**°	**1.40**
Mixed-illuminant evaluation set			
AdaIN [44]	818.99	**5.41**°	11.01
EFDM	**761.95**	5.95°	**10.16**

of the observed performance gains are attributable to the feature distribution matching learning strategy. For example, on the Cube+ dataset, integrating this strategy with the Uformer architecture leads to significant reductions in all metrics. This effect is even more pronounced with larger patch sizes and is also evident in the mixed-illuminant evaluation set. Furthermore, it is important to note that the performance of the Uformer-only architecture lags behind previous methods, including Mixed WB [2] and Style WB [24].

EFDM vs. AdaIN: This ablation study investigates the impact of using EFDM for feature statistics alignment compared to employing AdaIN, which focuses on aligning low-level statistics without achieving exact matching. Table 5.8 summarizes the results

Figure 5.10: *Analyzing the impact of aligning and matching feature distributions on the weighting maps generated by our proposed model on the selected sample from MIT-Adobe FiveK dataset [7]. Image index:* 2808.



for the datasets used in our quantitative evaluation, comparing the effectiveness of learning style factors via EFDM and AdaIN. In single-illuminant scenarios, employing EFDM demonstrates significant improvements in all metrics, including a 15% reduction in MSE, a 24% reduction in MAE, and a 27% reduction in ΔE 2000, compared to AdaIN. Similarly, while both strategies show promising results on the mixed-illuminant evaluation set, EFDM outperforms AdaIN, particularly in MSE and color difference metrics.

The qualitative results of the MIT-Adobe FiveK dataset [7] further validate the superior ability of EFDM to model lighting as a style factor, and surpassing AdaIN in both perceptual quality and spatial consistency. As illustrated in Figure 5.10, the exact matching of feature statistics yields more precise weighting coefficients in the output maps compared to AdaIN's alignment-based strategy, which relies solely on low order statistics. This precision significantly improves the overall quality of WB correction.

Patch Size and WB Settings: The effect of varying patch sizes and input WB settings is detailed in Table 5.9. Larger patch sizes (*i.e.*, $p = 128$) consistently yield better performance in both datasets. For example, on the Cube+ dataset, increasing the patch size from

Table 5.9: *Ablation study on changing patch size and using different WB settings on Cube+ dataset [1] and mixed-illuminant evaluation set [2].*

Method	MSE ↓	MAE ↓	ΔE 2000 ↓
Cube+ dataset			
$p = 64$, {t, d, s}	91.34	2.38°	2.88
$p = 64$, {t, f, d, c, s}	118.51	3.65°	4.56
$p = 128$, {t, d, s}	79.35	**1.35**°	**1.40**
$p = 128$,{t, f, d, c, s}	**78.76**	1.54°	1.69
Mixed-illuminant evaluation set			
$p = 64$, {t, d, s}	780.74	4.85°	10.84
$p = 64$, {t, f, d, c, s}	815.24	4.82°	11.36
$p = 128$, {t, d, s}	**761.95**	5.95°	**10.16**
$p = 128$,{t, f, d, c, s}	822.12	**4.73**°	11.08

64 to 128 reduces MSE from 91.34 to 79.35 and MAE from 2.38° to 1.35°. Similarly, on the mixed-illuminant evaluation set, the larger patch size achieves superior MSE and ΔE 2000 values, which underscores the role of greater contextual awareness in accurate WB correction.

When considering the simplified WB setting spectrum (*i.e.*, {t, d, s}), the impact is significant for both datasets. For $p = 64$, our approach achieves an MSE of 91.34 and an MAE of 2.38° on the Cube+ dataset, while it achieves an MSE of 780.74 and a ΔE 2000 of 10.84 on the mixed-illuminant evaluation set. In contrast, using the full spectrum (*i.e.*, {t, f, d, c, s}) results in higher MSE and MAE values for both datasets. Specifically, the results presented report 118.51 for MSE and 3.65° for MAE on the Cube+ dataset, while it shows 815.24 for MSE and 11.36 for ΔE 2000 on the mixed-illuminant evaluation set. For $p = 128$, this trend continues, with the simplified spectrum consistently outperforming the full spectrum. On the mixed-illuminant evaluation set, the simplified spectrum achieves an MSE of 761.95, compared to 822.12 for the full spectrum. These findings underscore that reducing the number of WB settings simplifies the non-linear interpolation process during the generation of weighting maps. This reduction improves

Table 5.10: *Ablation study on the effect of post-processing operation on the performance of our proposed model on Cube+ dataset [1].*

Method	MSE ↓	MAE ↓	ΔE 2000 ↓	Time (s)
ms ✗, eas ✗	85.20	1.33°	1.35	0.292
ms ✗, eas ✓	80.11	**1.29**°	**1.32**	11.051
ms ✓, eas ✗	80.72	1.37°	1.41	0.337
ms ✓, eas ✓	**79.35**	1.35°	1.40	11.228

spatial consistency and overall performance. We can conclude that striking an appropriate balance between patch size and WB setting complexity allows the model to more effectively adapt to diverse illumination scenarios.

Post-Processing: Building on the foundation of our first approach, as discussed in Section 5.2.1.3, we extend our analysis to the second approach by evaluating the effects of two post-processing operations applied during inference. These experiments systematically assess the contributions of these operations to the overall performance of our proposed method. Table 5.10 presents the quantitative results in the selected metrics. Although post-processing operations, particularly edge-aware smoothing (eas), offer marginal improvements in performance, their advantages are diminished when considering the additional processing time they require. Importantly, even without any post-processing operations, FDM WB outperforms other techniques that rely on such operations, which confirms its robustness and efficiency.

Complexity Analysis: To evaluate the impact of adopting a Transformer-based architecture on computational cost, we conduct a comprehensive analysis comparing the complexities of our proposed method with those of previous methods, both with and without post-processing operations. Table 5.11 summarizes the average running time in seconds, the total number of parameters, and the number of floating-point operations for each model architecture. Specifically, Mixed WB uses the GridNet architecture [149],

Table 5.11: *Comparison of the complexity of FDM WB and the prior methods with their post-processing tricks.*

Method	Time (s)	# of Params (M)	FLOPs (G)
Mixed WB [2] + ms + eas	10.390	**1.32**	**9.78**
Mixed WB [2] + eas	10.279		
Mixed WB [2] + ms	0.228		
Mixed WB [2]	**0.212**		
Style WB [24] + ms + eas	10.342	15.31	126.60
Style WB [24] + eas	10.307		
Style WB [24] + ms	0.232		
Style WB [24]	0.217		
FDM WB (ours) + ms + eas	11.228	20.53	61.92
FDM WB (ours) + eas	11.041		
FDM WB (ours) + ms	0.337		
FDM WB (ours)	0.292		

Style WB utilizes the IFRNet architecture [41], while our proposed method is built on the Uformer architecture.

Despite incorporating a Transformer-based backbone, our analysis shows that the proposed method maintains FLOPs and model parameter count comparable to or even lower than those of IFRNet-based approaches used in Style WB. This efficiency is achieved through architectural optimizations, including a reduced number of layers in both the encoder-decoder structure and the projector layers of the *Style Extractor* module. These modifications ensure a balanced trade-off between computational efficiency and performance while keeping the computational demands within a manageable range.

Although global self-attention mechanisms in Transformer-based architectures inherently increase memory usage and inference time, the Uformer backbone demonstrates its ability to model global dependencies without excessive computational overhead. In contrast, IFRNet-based methods rely on localized convolutions, which, despite their limited capacity to capture global features, contribute significantly to higher FLOPs. This

Figure 5.11: *Qualitative comparison of the visual results of FDM WB with the prior works under challenging lighting conditions. Image indices from top to bottom:* 596, 619, 581 *in MIT-Adobe FiveK dataset [7].*

(a) Deep WB **(b)** Mixed WB **(c)** Style WB **(d)** FDM WB **(ours)**

highlights the advantage of Uformer in spatial WB correction, where global feature interactions play a critical role.

The proposed method achieves substantial improvements in spatial WB correction performance, justifying the modest increase in parameter count and runtime. Furthermore, its runtime efficiency ensures near-real-time applicability, which makes it a viable candidate for further optimization and practical deployment in real-world imaging tasks.

Challenging Scenarios: To further demonstrate the effectiveness of our proposed method in challenging lighting conditions, we present Figure 5.11, which features examples from

scenarios such as nighttime photography [150, 151], mixed illumination with strong highlights and complex multi-color lighting. For this comparison, we use images with indices 596, 619, and 581. Notably, existing methods often struggle to fully preserve color fidelity, leading to issues such as oversaturation or undercorrection in these difficult environments. In contrast, our approach effectively adapts to these conditions, delivering visually balanced outputs with accurate illumination correction while preserving the natural appearance of the scene.

5.2.3 Experimental Results for FDM Loss

Tables 5.12, 5.13, and 5.14 present the benchmark results on the LSMI dataset [8], which showcases the performance of various methods including our third proposed approach, Uformer with FDM Loss, in single-, multi-, and mixed-illuminant scenarios for *Galaxy*, *Nikon*, and *Sony* cameras, respectively. The benchmark results on the LSMI dataset in three camera setups demonstrate the effectiveness of our third proposed approach, *Uformer with FDM Loss*, which integrates the EFDM-based loss function into the optimization process of the Uformer architecture for WB correction. By leveraging a Transformer-based architecture, our method ensures robust performance under challenging multi-illuminant conditions while maintaining accuracy in single-illuminant scenarios.

5.2.3.1 A Novel Metric for Generalization: The Multi-to-Single Ratio

To further evaluate the robustness of WB correction methods, we introduce the Multi-to-Single Ratio (MSR), defined as the ratio of the mean MAE in multi-illuminant scenarios to the mean MAE in single-illuminant scenarios. Assuming a sufficient number of samples are used to evaluate both scenarios, this metric quantifies how effectively a model generalizes to conditions under multiple illumination without overfitting to single-illuminant scenarios. A lower MSR indicates better adaptability, as it reflects minimal

Table 5.12: *Benchmark results on the LSMI dataset for the Galaxy camera. The Multi-to-Single Ratio reflects the robustness of the models in multi-illuminant scenarios.*

Model	Single		Multi		Mixed		MSR
	Mean	**Median**	**Mean**	**Median**	**Mean**	**Median**	
Pix2Pix [143]	6.53	2.17	4.28	2.63	5.66	2.44	**0.66**
Gijsenij *et al.* [82]	7.49	6.04	12.38	9.57	10.09	7.43	1.65
Bianco *et al.* [118]	4.15	3.30	5.56	4.33	4.89	3.83	1.34
HDRNet [152] r. [8]	2.85	2.20	3.13	2.70	3.06	2.54	1.10
HDRNet [152] r. [9]	-	-	-	-	3.06	2.54	-
UNet [59] r. [8]	2.95	1.86	2.35	2.00	2.63	1.91	0.80
UNet [59] r. [9]	2.85	-	2.55	-	2.68	2.17	0.90
One-Net [139]	**1.19**	0.75	2.16	1.53	**1.57**	0.93	*1.82*
AID [9]	**1.19**	-	2.03	-	1.66	1.41	*1.71*
Uformer + FDM (**ours**)	1.78	1.48	**1.87**	1.69	1.83	1.62	***1.05***

performance degradation between single- and multi-illuminant conditions.

We propose this ratio to provide a deeper understanding of the robustness of the model. Overfitting to single-illuminant scenarios often leads to a sharp decline in performance under multi-illuminant conditions, which are more representative of real-world settings. By balancing performance across these scenarios, our third proposed approach demonstrates its ability to address practical challenges in WB correction.

5.2.3.2 Benchmark on the LSMI dataset

In multi-illuminant scenarios, the Uformer with FDM Loss achieves state-of-the-art performance across all three camera setups, as demonstrated by both the mean and median MAE values. For the *Galaxy* camera, Uformer with FDM Loss records a mean MAE of 1.87 and a median of 1.69, outperforming key competitors such as One-Net (2.16, 1.53). Similarly, for the *Nikon* camera, the proposed method achieves a mean MAE of 1.54 and a median of 1.12, surpassing One-Net (1.99, 1.43) and other notable methods. On the *Sony* camera, Uformer with FDM Loss achieves a mean MAE of 1.67 and a median of 1.57, which further solidifies its superiority under multi-illuminant conditions. The

Table 5.13: *Benchmark results on the LSMI dataset for the Nikon camera. The Multi-to-Single Ratio reflects the robustness of the models in multi-illuminant scenarios.*

Model	Single		Multi		Mixed		MSR
	Mean	**Median**	**Mean**	**Median**	**Mean**	**Median**	
Pix2Pix [143]	6.1	2.27	4.18	2.76	5.41	2.49	**0.77**
Bianco *et al.* [118]	3.18	2.61	4.65	4.19	3.93	3.48	1.18
HDRNet [152] r. [8]	2.76	2.43	3.2	3.01	2.99	2.61	1.07
HDRNet [152] r. [9]	-	-	-	-	2.99	2.61	-
UNet [59] r. [8]	1.51	1.14	2.36	1.84	1.95	1.45	1.21
UNet [59] r. [9]	1.49	-	2.30	-	1.92	1.54	1.20
One-Net [139]	1.27	0.67	1.99	1.43	1.53	0.85	*1.30*
AID [9]	**1.11**	-	2.26	-	1.71	1.34	*1.32*
Uformer + FDM (**ours**)	1.26	0.97	**1.54**	1.13	**1.48**	1.05	***1.22***

Table 5.14: *Benchmark results on the LSMI dataset for the Sony camera. The Multi-to-Single Ratio reflects the robustness of the models in multi-illuminant scenarios.*

Model	Single		Multi		Mixed		MSR
	Mean	**Median**	**Mean**	**Median**	**Mean**	**Median**	
Pix2Pix [143]	4.08	1.72	4.37	3.26	4.20	2.20	**1.07**
Bianco *et al.* [118]	3.25	2.62	4.38	3.93	3.86	3.19	1.35
HDRNet [152] r. [8]	-	-	-	-	3.21	2.89	-
HDRNet [152] r. [9]	2.76	2.43	3.2	3.01	2.99	2.61	**1.07**
UNet [59] r. [8]	2.83	2.44	3.04	2.78	2.94	2.66	**1.07**
UNet [59] r. [9]	1.92	-	2.34	-	2.15	1.74	1.22
One-Net [139]	1.45	0.60	2.23	1.65	1.76	0.93	*1.54*
AID [9]	**1.01**	-	2.16	-	1.66	1.35	*2.14*
Uformer + FDM (**ours**)	1.52	1.39	**1.67**	1.57	**1.61**	1.47	***1.10***

consistent performance across all three camera setups highlights the robustness of the proposed method in handling challenging lighting conditions. By using EFDM as the loss function to ensure the exact match of feature distributions, the method effectively models the interaction of multiple light sources, resulting in superior perceptual quality and accurate color correction.

Several competing methods demonstrate strong performance in single-illuminant

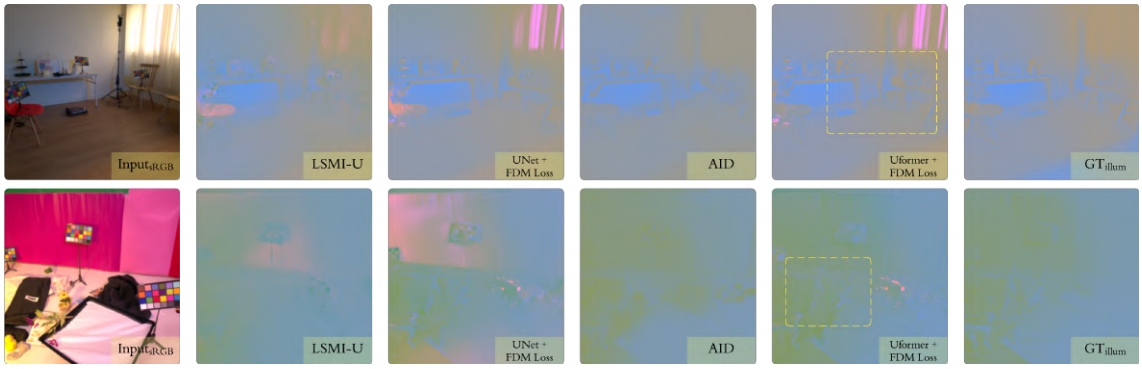
scenarios but exhibit significant performance degradation under multi-illuminant conditions, which suggests an overfitting tendency. For instance, on the *Galaxy* camera, OneNet achieves a mean MAE of 1.19 in single-illuminant scenarios but increases to 2.16 in multi-illuminant scenarios, which results in a high MSR value of 1.82. Similarly, AID on the *Sony* camera achieves a MSR value of 2.14, which reflects its overreliance on single-illuminant conditions.

In contrast, the Uformer with FDM Loss demonstrates a much lower MSR value across all setups, and this indicates its ability to generalize effectively. For the *Galaxy*, *Nikon*, and *Sony* cameras, the ratios are 1.05, 1.22 and 1.10, respectively. These values underscore the ability of the method to maintain robust performance in multi-illuminant conditions without sacrificing accuracy in single-illuminant scenarios.

Traditional methods, such as HDRNet and UNet, while competitive in single-illuminant scenarios, fail to maintain similar levels of accuracy under multi-illuminant conditions. For example, HDRNet, reported for the *Galaxy* camera, achieves a mean MAE of 2.85 in single-illuminant scenarios but only improves marginally to 3.13 in multi-illuminant settings. Similarly, UNet achieves a mean MAE of 2.83 on the *Sony* camera under single-illuminant conditions, but its performance degrades to 3.04 in multi-illuminant settings.

The results emphasize the importance of balancing single- and multi-illuminant performance for practical real-world applications. Our proposed approach demonstrates strong generalization across diverse lighting conditions without overfitting, making it a highly reliable solution for WB correction. Its superior performance in multi-illuminant environments, reflected in an MSR value closer to 1, highlights its robustness and adaptability. This resilience stems from the feature distribution matching mechanism, which aligns feature representations directly rather than relying on pixel-wise intensity comparisons or statistical normalization techniques.

Figure 5.12: *Qualitative comparison of illumination estimation results among LSMI-U [8], AID [9], and our proposed method. Image indices:* 525 *(Galaxy),* 757 *(Sony).*



Compared methods, such as HDRNet [152] and One-Net [139], optimize illumination by minimizing direct pixel-level errors. However, this often results in overfitting to single-illuminant settings and instability when handling complex illumination variations. In contrast, EFDM maintains the structural integrity of the feature distributions, ensuring that the learned representations remain consistent in different lighting conditions and preventing excessive adaptation to a specific illuminant type. This stability is particularly critical in multi-illuminant scenarios, where conventional pixel-intensity-based loss functions struggle with the non-linear interactions of multiple light sources. The consistently low MSR values achieved further validate that distribution matching mitigates performance degradation under complex illumination settings, making it a more reliable objective function for real-world applications.

5.2.3.3 Comparison with State-of-the-art

To further assess the effectiveness of our proposed approach, we provide a qualitative comparison with LSMI-U [8] and AID [9], as shown in Figure 5.12. Due to the lack of publicly available models for state-of-the-art methods, direct reproduction of their results on the LSMI test set was not feasible. Instead, qualitative results for AID were

extracted from their original paper to make a fair comparison possible. However, OneNet [139] does not include test set comparisons in its qualitative evaluations, making its inclusion impractical.

The results reveal distinct differences in illumination estimation among the methods. LSMI-U, as a baseline dataset-derived approach, performs reasonably well in WB correction, but struggles with residual color shifts in complex multi-illuminant scenarios. AID, designed for illumination decomposition rather than perceptual correction, estimates chromatic illumination components, which can lead to noticeable deviations in certain regions. In contrast, both variants of our method generate more consistent illumination maps with fewer artifacts, demonstrating the effectiveness of EFDM as a training objective to enforce feature distribution alignment and achieve robust WB correction.

Notably, in scenes with strong color casts, the proposed approach, namely *Uformer with FDM Loss*, exhibits superior adaptation to multi-illuminant conditions, particularly in the highlighted regions. The method effectively reduces unwanted tints and maintains global color consistency. However, minor deviations remain in illumination maps, especially in highly saturated regions, which may suggest potential areas for future improvements in illuminant adaptation.

5.2.3.4 Ablation Study

Ablation studies systematically evaluate the contributions of the proposed FDM loss function and its adaptability in different architectures. By isolating the effect of the loss function and comparing its performance within different architectures, we provide compelling evidence for the superiority of the proposed methodology in addressing the complexities of WB correction under diverse illumination scenarios.

The results presented in Table 5.15 illustrate the significant advantages of employing the FDM loss function over the traditional pixel-wise mean squared error loss. The

Table 5.15: *Ablation study on the proposed loss function using the Uformer architecture.*

Camera	Loss	Single		Multi		Mixed		MSR
	Function	Mean	Median	Mean	Median	Mean	Median	
Galaxy	MSE	2.20	1.65	2.03	1.73	2.05	1.64	0.88
	FDM	**1.78**	1.48	**1.87**	1.69	**1.83**	1.62	1.05
Nikon	MSE	1.39	1.01	1.72	1.15	1.56	1.10	1.10
	FDM	**1.31**	0.98	**1.54**	1.12	**1.43**	1.05	1.08
Sony	MSE	2.15	1.54	2.03	1.73	2.08	1.68	0.94
	FDM	**1.52**	1.39	**1.67**	1.57	**1.61**	1.47	1.10

key strength of the FDM loss lies in its ability to perform exact feature distribution matching, which ensures that higher-order statistics, such as skewness and kurtosis, are effectively aligned between the predicted and ground truth images. This capability enables the model to better handle non-Gaussian and highly variable illumination conditions, which are common in multi- and mixed-illuminant scenarios.

For example, applying the FDM loss to the Uformer architecture significantly improves performance in multi-illuminant scenarios. For the *Galaxy* camera, the mean MAE improves from 2.03 with MSE loss to 1.87 with FDM loss. This demonstrates the robustness of our proposed loss function under complex lighting conditions. Similarly, for the *Nikon* camera, the mean MAE is reduced from 1.56 with MSE loss to 1.43 with FDM loss, which reflects improved mixed-illuminant performance. These gains result from the holistic approach of FDM loss to matching feature distributions, effectively capturing both global and local illumination characteristics.

Unlike MSE loss, which focuses on minimizing pixel-wise intensity differences, FDM loss captures the intricate relationships among image regions, maintaining global context while ensuring spatial consistency. This is especially critical in scenarios with uneven lighting variations throughout the scene. By leveraging higher-order statistical insights derived from feature information, the proposed loss function enhances the model's ability to generalize across diverse scenes without overfitting. This capability is reflected

Table 5.16: *Ablation study on the proposed loss function using the UNet architecture.*

Camera	Loss	Single		Multi		Mixed		MSR
	Function	Mean	Median	Mean	Median	Mean	Median	
Galaxy	MSE	2.95	1.86	2.35	2.00	2.63	1.91	0.80
	FDM	**2.42**	1.81	**2.14**	1.74	**2.27**	1.79	0.88
Nikon	MSE	1.51	1.14	2.36	1.84	1.95	1.45	1.21
	FDM	**1.40**	1.17	**1.89**	1.33	**1.66**	1.25	1.14
Sony	MSE	2.83	2.44	3.04	2.78	2.94	2.66	1.07
	FDM	**1.96**	1.63	**2.10**	1.74	**2.04**	1.67	1.07

in the balanced MSR observed across all three camera setups.

The results in Table 5.16 demonstrate the versatility of the proposed FDM loss through its significant performance improvements within the UNet architecture. For the *Sony* camera, the multi-illuminant mean MAE decreases from 3.04 with MSE loss to 2.10 with FDM loss. Similarly, the mixed-illuminant mean MAE improves from 2.94 to 2.04. These results highlight the adaptability of FDM loss, effectively enhancing the performance of a convolutional network like UNet, despite its architectural limitations compared to Uformer.

The seamless integration of FDM loss with both Uformer and UNet architectures underscores its robustness and generalizability as an optimization objective. While Uformer leverages its Transformer-based design to inherently capture both local and global image features, the proposed loss function ensures that even UNet, with its more localized receptive field, achieves substantial performance gains. This adaptability demonstrates the effectiveness of FDM loss in meeting various architectural requirements while consistently providing superior performance in various illumination scenarios.

The qualitative comparisons in Figure 5.13, Figure 5.14, and Figure 5.15 show the efficacy of our last proposed approach, Uformer with FDM loss, in addressing the challenges posed by multi-illuminant scenarios in the LSMI dataset. Across the three

camera setups (*i.e.*, *Galaxy*, *Nikon*, *Sony*), our proposed approach consistently outperforms its counterparts, including those utilizing conventional loss functions such as MSE and traditional architectures such as UNet.

The introduction of FDM-based loss represents a significant advancement in performance by enabling exact matching of feature distributions for the [CLS] token, a critical representation of global image characteristics, as previously detailed in Section 2.3.2.1. By harnessing the [CLS] token's representational power in the global context, the FDM loss not only ensures spatial consistency but also preserves intricate lighting details and color fidelity, particularly in complex multi-illuminant scenarios. Furthermore, as analyzed in Section 4.2.2, the qualitative results substantiate that this advanced optimization strategy enhances the model's ability to seamlessly integrate local and global corrections. This leads to improved robustness and accuracy in WB correction, as evidenced by the reduced MAE and enhanced uniformity of illumination across all scenes evaluated.

The Uformer architecture further amplifies these gains due to its attention mechanisms, which enhance contextual awareness and spatial consistency. While the UNet architecture struggles with global illumination adjustments, which often introduces residual color casts, the Uformer architecture excels in preserving fine details and achieving uniform WB correction. These benefits are particularly noticeable in scenes with high illumination complexity, such as those in Figure 5.13 and Figure 5.15.

The superiority of Uformer with FDM loss is evident in challenging scenarios where mixed or multi-illuminant conditions dominate. The results on *Galaxy* camera (*i.e.*, Figure 5.13) highlight the ability of our proposed approach to neutralize strong color casts while preserving texture and detail. Similarly, the results on *Nikon* and *Sony* cameras (*i.e.*, Figure 5.14 and Figure 5.15) emphasize the model's ability to achieve spatially consistent WB corrections without sacrificing color fidelity.

Figure 5.13: *Visual comparison of WB correction outputs on the Galaxy camera from the LSMI dataset. Image indices:* 312, 323, 896.

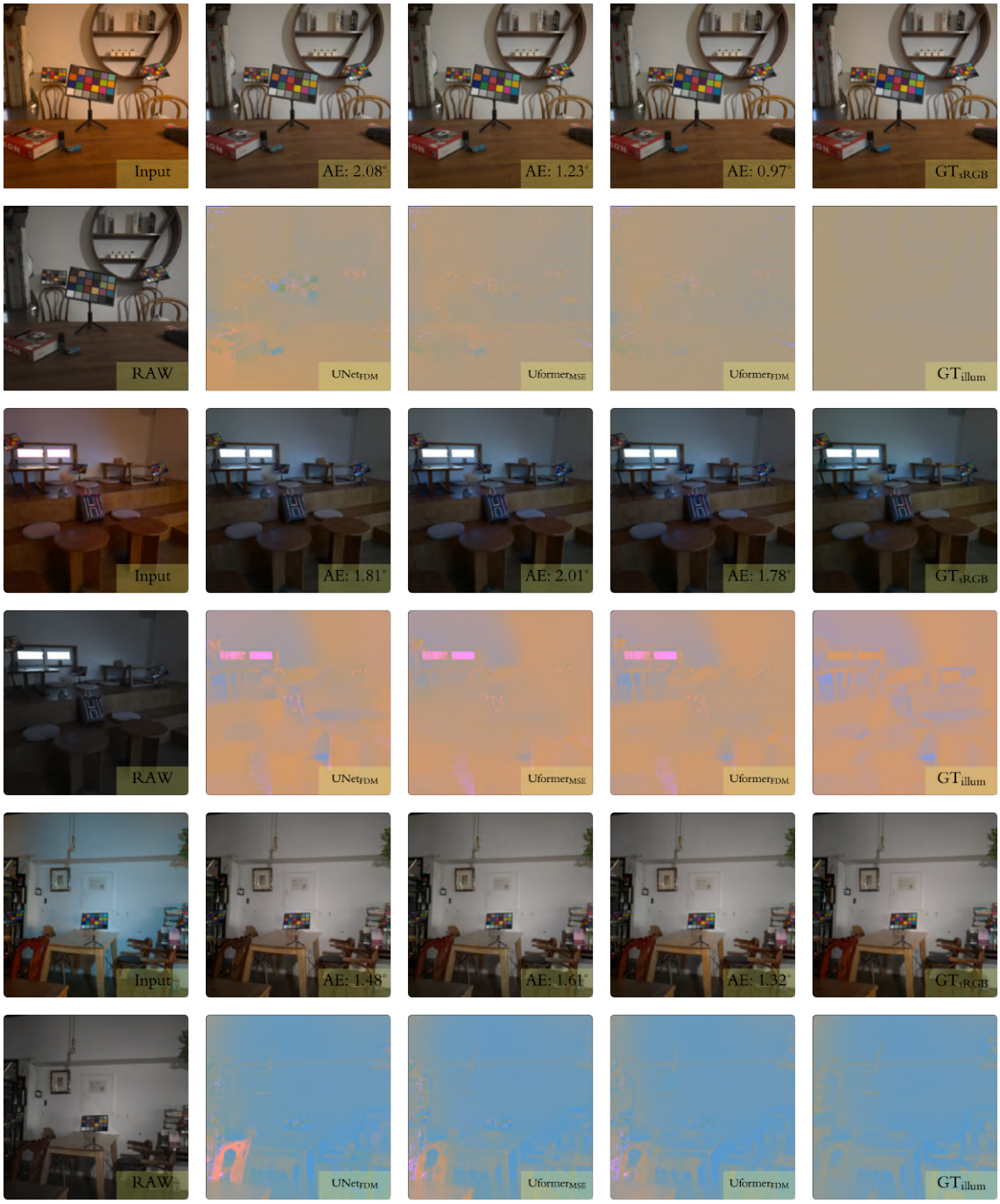


Figure 5.14: *Visual comparison of WB correction outputs on the Nikon camera from the LSMI dataset. Image indices:* 63, 221, 934.

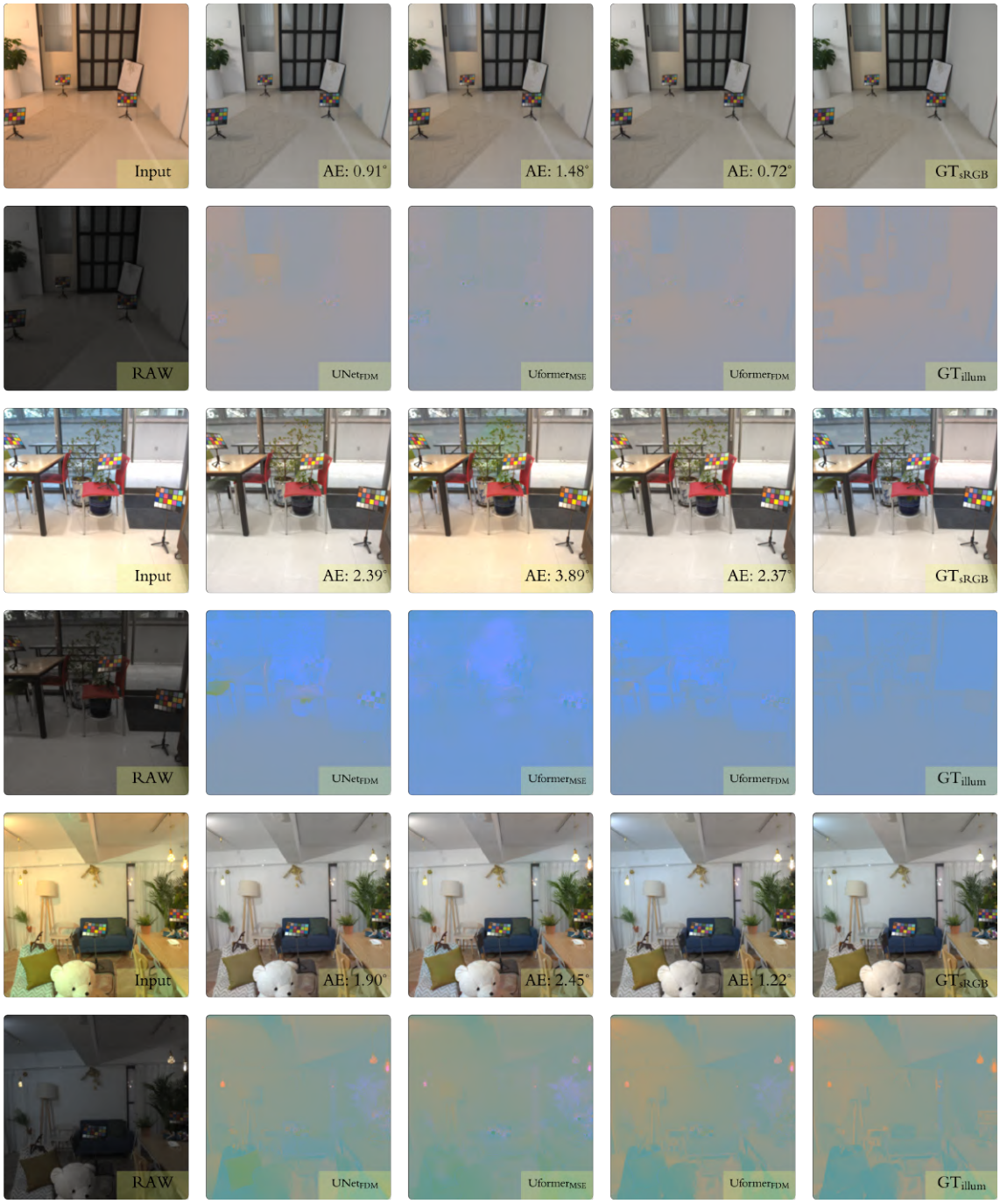


Figure 5.15: *Visual comparison of WB correction outputs on the Sony camera from the LSMI dataset. Image indices:* 790, 1202, 1314.

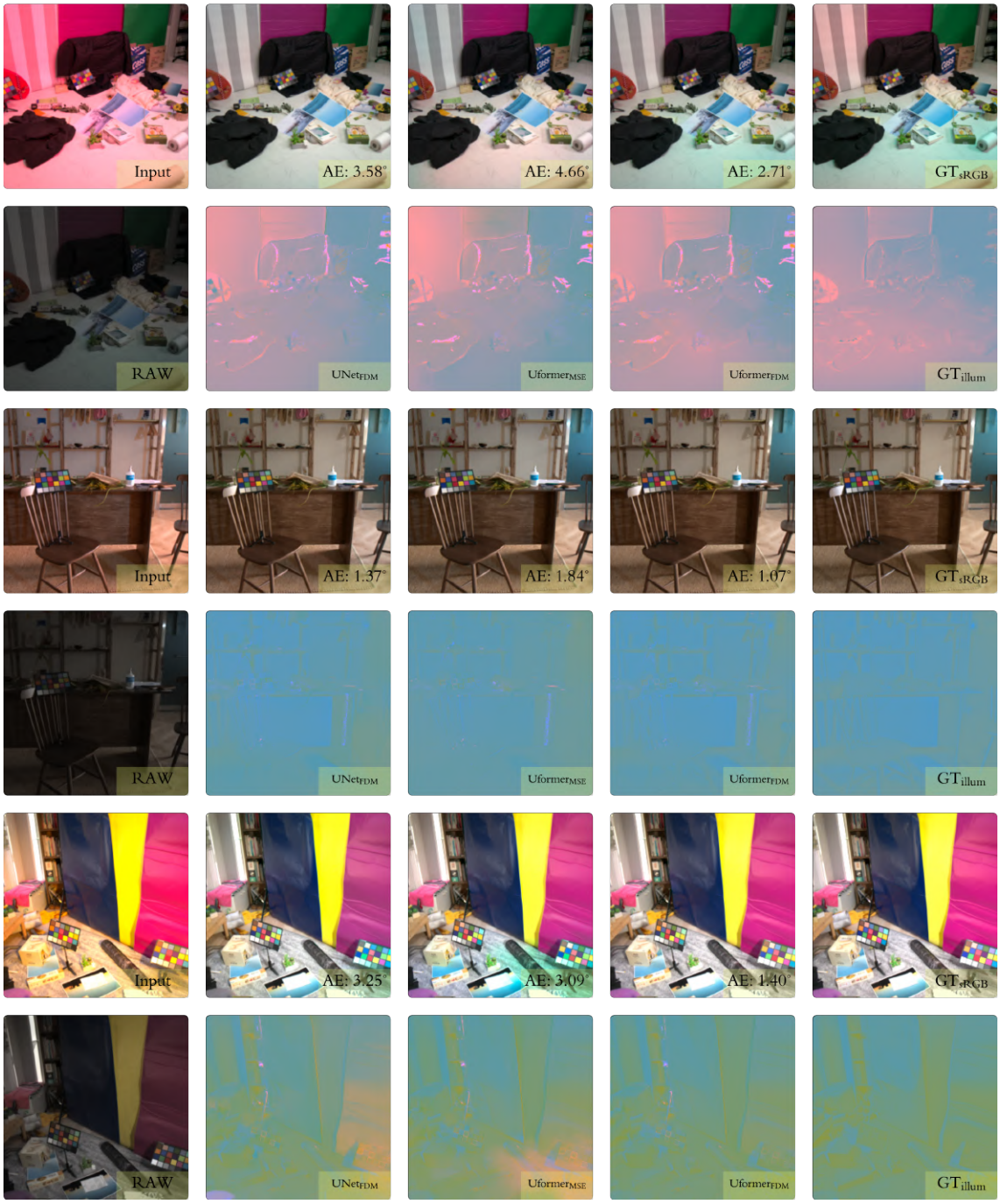
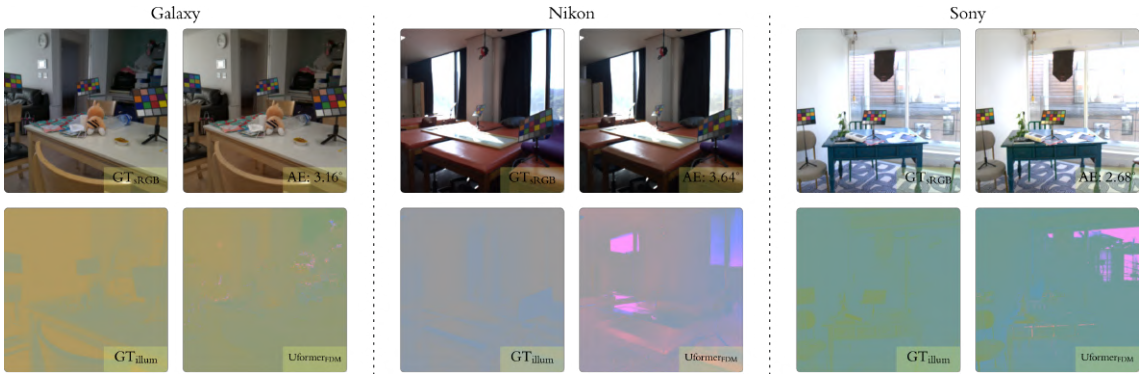


Figure 5.16: *Illustration of failure cases observed in our proposed method. Image indices:* 144, 20, 242.



Failure Case Analysis: Although the proposed method demonstrates strong WB correction performance in various scenarios, certain failure cases highlight its limitations under complex illumination conditions. Figure 5.16 showcases instances where the method struggles with challenging lighting environments, resulting in deviations from the ground truth. In these cases, the predicted illumination maps diverge from the reference, particularly in regions with mixed or highly non-uniform lighting distributions. As a result, the corrected images exhibit noticeable color shifts, with angular error values remaining higher than in other cases, and this indicates the impact of these challenges on overall performance.

A key difficulty arises in handling complex multi-illuminant regions, particularly when multiple light sources interact, such as natural daylight combined with artificial indoor lighting. Although the proposed method effectively infers the illumination correctly in most cases, it can introduce artifacts in these scenarios, where dominant illuminants vary spatially. This suggests that EFDM may require further refinement to ensure consistent alignment when multiple illuminants influence different regions of the scene. Consequently, corrected images may retain residual color shifts in highly complex lighting environments.

Additionally, in cases where strong directional illumination creates pronounced shadowed and brightly lit areas, minor chromatic inconsistencies can appear in the predicted illumination maps. This does not necessarily indicate a failure in generalization, but instead highlights the challenges posed by extreme lighting variations, where global feature alignment may not fully capture local illumination nuances. While pixel-wise approaches rely on intensity adjustments to handle such variations, EFDM focuses on the alignment of the feature distribution, which enhances robustness against global illumination shifts, but does not explicitly enforce spatial coherence in extreme cases.

Moreover, highly reflective surfaces or strongly saturated regions can introduce slight deviations in the predicted illumination maps, particularly where illumination discontinuities are sharply localized. Despite these limitations, the proposed method consistently delivers strong performance in diverse lighting conditions. These observations suggest potential areas for further refinement, such as incorporating spatial priors to enhance stability in extreme illumination scenarios.

Complexity Analysis: Since EFDM is integrated as a training objective rather than an architectural modification, it does not increase computational complexity during inference or deployment. While inference time and model size are determined by the backbone architecture, a balanced evaluation is conducted by applying EFDM as a loss function to both CNN-based (*i.e.*, UNet) and Transformer-based (*i.e.*, Uformer) architectures. Among the compared methods, AID [9] uses an attention-based approach, making its computational characteristics comparable to Uformer, whereas One-Net [139] and HDRNet [152] rely on CNN-based feature extraction, similar to UNet. This ensures that the experimental setup encompasses a diverse range of architectures with similar computational properties, which demonstrates that integrating EFDM does not introduce additional inference overhead beyond what is typically expected in existing approaches.

In conclusion, our final approach effectively delivers robust and reliable performance, despite certain limitations, in diverse camera setups and challenging illumination conditions. These results highlight the practical applicability and versatility of our proposed method in real-world imaging scenarios.

6. APPLICATIONS AND EXTENSIONS

This chapter discusses the applications and extensions of the proposed methods by presenting three key works published during this research: Deterministic Neural Illumination Mapping (Deterministic Neural Illumination Mapping (DeNIM)) for resource-constrained environments and two challenges focusing on night photography rendering, organized by one of the leading venues in the field of image restoration, which showcases real-world implementations of the approaches proposed in this thesis. These works demonstrate how the main contributions of this thesis can be adapted and extended to address specific challenges in image restoration and enhancement.

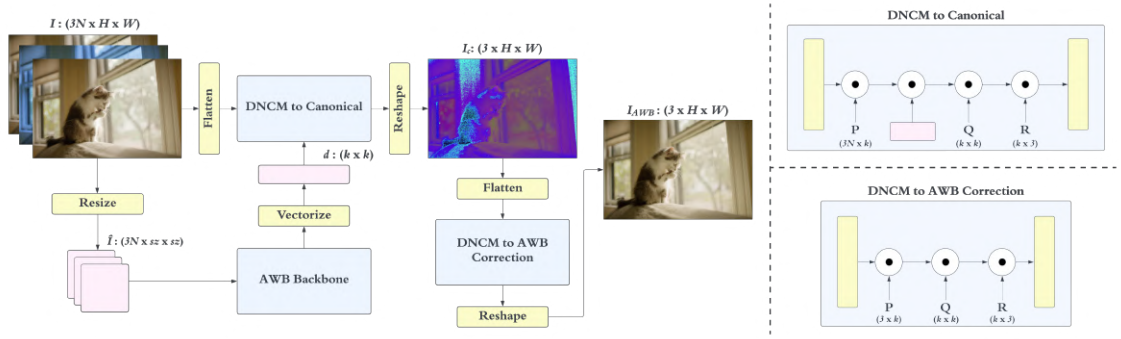
6.1 Deterministic Neural Illumination Mapping (DeNIM)

This extension [148] introduces an efficient WB correction framework for high-resolution images. It uses deterministic color mapping to align the colors of pixels in various illumination conditions using learnable projection matrices. DeNIM is designed to be both resolution-agnostic and model-agnostic, enabling seamless integration with pre-trained WB correction networks, such as Mixed WB and Style WB.

6.1.1 Architecture

Given a set of high-resolution images with different WB settings I, this proposed strategy focuses on achieving deterministic illumination color mapping for efficient WB correction. Although previous work [2, 24] has shown success in learning weighting maps in low-resolution space and rendering high-resolution WB corrected images through multi-scale inference, these methods often require additional post-processing steps, such as smoothing after resizing weighting maps, which limit their practicality. Inspired by deterministic color style transfer [153], we introduce an illumination mapping strategy that eliminates the need for such post-processing by directly aligning color distributions

Figure 6.1: *Overall design of Deterministic Neural Illumination Mapping (DeNIM), proposed illumination mapping strategy for high-resolution images.*



in high-resolution space, as depicted in Figure 6.1.

First, we downsample the input images I to a resolution compatible with the architectures of prior works (*i.e.*, 256×256). Using only the encoder portion of these architectures, low-resolution images $\hat{I}$ are processed to extract rich feature information across different WB settings. A 1×1 convolutional layer, followed by Gaussian Error Linear Unit (GeLU) activation [154], is applied to the extracted feature maps to obtain latent representations. These representations are then used to compute an image-adaptive color mapping matrix d, which is fed into the Deterministic Neural Color Mapping (DNCM) module [153] to generate the canonical form.

$$d^{(k \times k)} = V(E(\hat{\mathbf{I}})) \tag{6.1}$$

where E denotes the AWB encoder (*i.e.*, as proposed in [2] or [24]), V stands for the vectorization operation performed using a 1×1 convolutional layer followed by an activation function. It is important to note that DeNIM utilizes pre-trained weights for E and keeps these weights frozen during training.

For the *DNCM to canonical* module, the process begins by unfolding the high-resolution image I into a 2-dimensional matrix with dimensions $(HW \times 3N)$, where

N represents the number of WB settings, and H and W denote height and width, respectively. Each pixel in I is then transformed into a k-dimensional vector through a projection matrix P of size $(3N \times k)$. The parameter k can be adjusted based on computational resources, but in our design it is set to 32. The extracted image-adaptive color mapping matrix d is then applied to this k-dimensional vector, enriching the projected space with contextual information. Following this, two learnable projection matrices, Q of size $(k \times k)$ and R of size $(k \times 3)$, are utilized to produce the canonical form. This module, termed *DNCMc*, is mathematically represented as follows

$$DNCMc(\mathbf{I}, d) = \mathbf{I}^{(HW \times 3)} \cdot \mathbf{P}^{(3 \times k)} \cdot d^{(k \times k)} \cdot \mathbf{Q}^{(k \times k)} \cdot \mathbf{R}^{(k \times 3)} \tag{6.2}$$

where $\cdot$ denotes matrix multiplication.

The canonical form is then processed by the *DNCM to AWB correction* module (*DNCMa*). Unlike *DNCMc*, this module does not include fusion capabilities but focuses on directly mapping pixel values from the canonical form to their corrected versions for white-balanced output. Each pixel in the canonical form I_c is first projected onto a k-dimensional vector using a projection matrix P of size $(3 \times k)$. Following a design similar to *DNCMc*, two additional learnable matrices, Q of size $(k \times k)$ and R of size $(k \times 3)$, are applied to transform the embedded k-dimensional vector back into the *RGB* color space. This process results in the final WB corrected output I_{AWB}. The formal definition of *DNCMa* is presented in Equation 6.3.

$$DNCMa(\mathbf{I_c}) = \mathbf{I_c}^{(HW \times 3)} \cdot \mathbf{P}^{(3 \times k)} \cdot \mathbf{Q}^{(k \times k)} \cdot \mathbf{R}^{(k \times 3)} \tag{6.3}$$

Apart from the self-supervised learning mechanism for DNCM adopted in [153], our learning objective focuses on minimizing the reconstruction error between the ground truth and the WB corrected output, as defined in Equation 6.4.

$$\mathcal{L} = ||\mathbf{I_{GT}} - \mathbf{I_{AWB}}||_F^2 \tag{6.4}$$

where $\mathbf{I_{GT}}$ and $\mathbf{I_{AWB}}$ represent the ground truth image and the WB corrected output, respectively. To maintain simplicity and tractability in the training process, we do not incorporate smoothing loss [2] or perceptual loss [155] into the final objective function.

This design eliminates the decoder component responsible for generating weighting maps in prior works and instead directly computes the illumination color mapping through two distinct DNCM modules for the canonical form and the WB corrected output. This approach removes the necessity for post-processing weighting maps, thereby reducing time complexity without compromising performance. Furthermore, the pixel-wise mapping capability, enabled by matrix multiplications, allows the correction model to operate independently of resolution. Furthermore, the flexibility of our design allows for the seamless integration of any WB correction method to extract rich information from low-resolution inputs across different WB settings, which makes it inherently model-agnostic.

6.1.2 Experiments

The experimental setup follows the methodology employed in Style WB, where it utilizes the same dataset and evaluation metrics. Any data augmentation techniques were not applied during training. DNCM modules were trained from scratch, while the AWB backbone weights was kept frozen. The resolution of the input images was set to 256 pixels, the training process employing the AdamW optimizer [146], a batch size of 16, and a learning rate of $1e^{-4}$. No post-processing operations were applied during inference.

The benchmark results on the single-illuminant Cube+ dataset [1], presented in Table 6.1, follow the experimental setup outlined in previous works [2, 24]. Two patch sizes (*i.e.*, 64 and 128) were employed for the backbone network, and input images were designed with two sets of WB settings: a default configuration including **T**ungsten, **D**aylight, and **S**hade, and an extended version incorporating **F**luorescent and **C**loudy color

Table 6.1: *Benchmark of DeNIM on single-illuminant Cube+ dataset [1]. The top results are indicated with colored cells as, the best: green, the second: yellow, the third: red.*

Method	MSE				MAE				ΔE 2000				Size
	Mean	Q1	Q2	Q3	Mean	Q1	Q2	Q3	Mean	Q1	Q2	Q3	
FC4 [17]	371.90	79.15	213.41	467.33	6.49°	3.34°	5.59°	8.59°	10.38	6.60	9.76	13.26	5.89 MB
Quasi-U CC [19]	292.18	15.57	55.41	261.58	6.12°	1.95°	3.88°	8.83°	7.25	2.89	5.21	10.37	622 MB
KNN WB [6]	194.98	27.43	57.08	118.21	4.12°	1.96°	3.17°	5.04°	5.68	3.22	4.61	6.70	21.8 MB
Interactive WB [126]	159.88	21.94	54.76	125.02	4.64°	2.12°	3.64°	5.98°	6.20	3.28	5.17	7.45	**38 KB**
Deep WB [20]	80.46	15.43	33.88	74.42	3.45°	1.87°	2.82°	4.26°	4.59	2.68	3.81	5.53	16.7 MB
MIMT [138]	-	-	-	-	2.52°	0.98°	1.38°	2.96°	2.88	1.94	2.42	2.87	-
Mixed WB [2]													
$p = 64$, WB=$\{$t,d,s$\}$	168.38	8.97	19.87	105.22	4.20°	1.39°	2.18°	5.54°	5.03	2.07	3.12	7.19	5.09 MB
$p = 64$, WB=$\{$t,f,d,c,s$\}$	161.80	9.01	19.33	90.81	4.05°	1.40°	2.12°	4.88°	4.89	2.16	3.10	6.78	5.10 MB
$p = 128$, WB=$\{$t,f,d,c,s$\}$	176.38	16.96	35.91	115.50	4.71°	2.10°	3.09°	5.92°	5.77	3.01	4.27	7.71	5.10 MB
Style WB [24]													
$p = 64$, WB=$\{$t,d,s$\}$	92.65	**6.52**	**14.23**	35.01	2.47°	0.82°	1.44°	2.49°	2.99	**1.36**	2.04	3.32	61.0 MB
$p = 64$, WB=$\{$t,f,d,c,s$\}$	151.38	29.49	56.35	125.33	4.18°	2.13°	3.03°	4.81°	5.42	3.11	4.42	6.76	61.1 MB
$p = 128$, WB=$\{$t,d,s$\}$	88.03	7.92	17.73	45.01	2.61°	0.93°	1.58°	2.85°	3.24	1.50	2.30	3.95	61.2 MB
$p = 128$, WB=$\{$t,f,d,c,s$\}$	100.24	10.77	37.74	70.18	3.09°	1.15°	2.61°	3.87°	3.96	1.59	3.55	5.51	61.3 MB
DeNIM + Mixed WB [2]													
$p = 64$, WB=$\{$t,d,s$\}$	120.14	36.39	77.40	152.96	2.57°	1.53°	2.17°	3.19°	5.26	3.38	4.71	6.64	28.7 MB
$p = 64$, WB=$\{$t,f,d,c,s$\}$	129.01	14.39	27.69	57.90	2.67°	0.99°	1.45°	2.29°	3.96	2.10	2.85	4.24	28.7 MB
$p = 128$, WB=$\{$t,d,s$\}$	158.58	60.14	115.66	198.59	4.20°	2.38°	3.77°	5.63°	5.69	3.91	5.41	7.10	28.8 MB
$p = 128$, WB=$\{$t,f,d,c,s$\}$	99.70	13.89	24.71	43.88	2.49°	1.07°	1.62°	2.41°	3.44	1.95	2.74	3.78	28.8 MB
DeNIM + Style WB [24]													
$p = 64$, WB=$\{$t,d,s$\}$	**65.80**	10.06	16.98	**28.82**	2.03°	0.88°	1.23°	1.93°	2.95	1.79	2.33	3.18	196.3 MB
$p = 64$, WB=$\{$t,f,d,c,s$\}$	83.41	13.23	21.46	37.44	**1.93°**	**0.77°**	**1.09°**	**1.70°**	**2.73**	1.62	**2.03**	**2.71**	196.3 MB
$p = 128$, WB=$\{$t,d,s$\}$	80.53	17.59	27.80	44.35	2.16°	0.88°	1.34°	2.16°	3.08	1.86	2.37	3.30	196.4 MB
$p = 128$, WB=$\{$t,f,d,c,s$\}$	89.10	11.27	19.34	43.01	2.49°	1.24°	1.64°	2.92°	3.16	1.87	2.53	3.35	196.4 MB

temperatures to enhance the versatility of the method. The quantitative results demonstrate increased efficiency and improved performance in all patch sizes and WB settings, as evidenced by the evaluation metrics. In particular, the configuration with a patch size of 64 and all WB settings outperform other configurations by achieving superior performance. The results indicate a significant improvement in the third quantiles of all evaluation metrics, which reflects the robustness of the strategy for challenging samples. Furthermore, smaller patch sizes continue to show better illuminant modeling and, in this case, improved learning of color mappings, consistent with previous observations. However, MSE displays inconsistencies compared to the other metrics, suggesting that it may not adequately capture the quality of the color correction and may not be the most suitable metric for assessing WB correction performance.

The results in Table 6.2 highlight the efficiency of the proposed strategy compared to previous works in various metrics. Efficiency is assessed based on processing time (*i.e.*, Time (s)), model complexity in terms of parameter count (*i.e.*, Param (M)), and

Table 6.2: *Comparison of the complexity of DeNIM and the prior methods with their post-processing tricks. ms: multi-scale weighting maps, eas: edge-aware smoothing.*

Model Architecture	Time (s)	Param (M)	FLOPS (G)
Mixed WB [2] + ms + eas	10.390	**1.32**	82.68
Mixed WB [2] + ms	0.228		
Mixed WB [2] + eas	10.279		
Mixed WB [2]	0.212		
Style WB [24] + ms + eas	10.342	15.31	76.80
Style WB [24] + ms	0.232		
Style WB [24] + eas	10.307		
Style WB [24]	0.217		
DeNIM + Mixed WB [2]	**0.006**	1.67	**2.14**
DeNIM + Style WB [24]	**0.010**	16.19	26.89

computational load measured in Floating Point Operations Per Second (*i.e.*, *FLOPS* (G)). In terms of processing time, DeNIM demonstrates a significant reduction in the time required for WB correction. This improvement is achieved by eliminating post-processing operations, such as multi-scale inference and edge-aware smoothing, and replacing the decoder with simple learnable projection matrices. DeNIM exhibits a remarkable speed advantage, being at least 35 times faster than earlier models and up to 1.700 times faster when post-processing is considered.

The complexity of the model, measured by parameter count, shows a slight increase in DeNIM compared to previous work, despite discarding the decoder in the baseline models. This increase is attributed to the use of fully-connected layers as projection matrices, instead of convolutional layers. Fully-connected layers inherently require more parameters due to their dense connections between input and output neurons. Although this design choice slightly increases the complexity of the model, it does not significantly affect the processing time. Lastly, the computational load, represented by *FLOPS*, reflects another key advantage of DeNIM. When trained with the Mixed WB backbone, DeNIM achieves an approximately 97% reduction in *FLOPS*. Similarly, training with the Style WB backbone reduces *FLOPS* by about 65%. These substantial reductions underscore the exceptional efficiency of the proposed strategy compared to previous methods.

6.2 Night Photography Rendering

The NTIRE 2023 Challenge on Night Photography Rendering [151] presents a unique opportunity to address the complexities of nighttime image rendering, including multi-illuminant scenarios and accumulated noise from consecutive processing steps. The *VGL OzU* team, representing the application of Style WB within an ISP pipeline, develops a comprehensive solution tailored to the requirements of the challenge.

The cornerstone of our approach is the integration of Style WB as a WB correction module. This method excels in handling mixed illuminants by modeling lighting conditions as a style factor and reversing their effects to achieve white balance. As post-processing, the images are processed using Restormer [156], a transformer-based model optimized for efficient, high-resolution image restoration. Furthermore, an adaptive auto-contrast strategy was employed, dynamically adjusting minimum cut-off values based on histogram outliers to improve the quality of darker regions.

The ISP pipeline begins with essential preprocessing steps, including black-level normalization and hot/bad pixel correction. For demosaicing, directional filtering (*i.e.*, Menon's algorithm) [61] replaces traditional CFA interpolation, which ensures high-quality raw data processing. Preliminary illumination estimation is performed in the *raw-RGB* domain using a random subsampling-based White Patch algorithm [157]. The transformation from *raw-RGB* to *sRGB* space incorporates Color Component Transfer Function (CCTF) encoding, followed by wavelet-based denoising with adaptive noise thresholding [158] to enhance image quality.

Once the data is in the *sRGB* domain, Style WB plays a pivotal role in mitigating the effects of different illuminants frequently observed in night scenes. To further enhance natural color representation, a memory color enhancement algorithm neutralizes specific colors such as sky and grass. The images were subsequently resized, oriented according to the metadata specifications, and subjected to final denoising using Restormer.

Figure 6.2: *Overall pipeline of proposed ISP for night photography rendering challenge.*

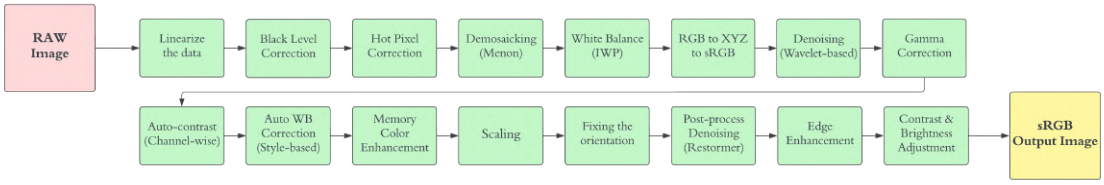


Table 6.3: *People's choice ranking results of night photography rendering challenge.*

Rank	Team	Mean Score
1	IVLTeam	0.670
2	DH_ImageAlgo	0.645
3	MiAlgo	0.626
4	BSSC	0.606
5	DH-AISP	0.583
6	Manual image enhancement	0.491
7	OzUVGL (**ours**)	0.453
8	The Majestic Mavericks	0.444
9	JMUCVLAB	0.439
10	NTU607	0.376
11	Baseline ISP	0.345

Due to Restormer's computational complexity at high resolutions, the images were processed in smaller grids to balance efficiency with quality. The pipeline is finished with unsharp masking to enhance edge clarity and fine-tuned contrast and brightness adjustments for optimal visual output. Figure 6.2 illustrates the proposed ISP pipeline for night photography rendering in this challenge.

The competition results, presented in Table 6.3, highlight the robustness and accuracy of our proposed pipeline. The integration of Style WB effectively tackles the challenges of multi-illuminant scenarios, while advanced denoising and contrast adjustment strategies can deliver visually appealing, high-quality nighttime photographs. Although the team did not secure top-tier rankings, the solution showcased significant potential for practical applications in night photography rendering, which provides a solid foundation

for further refinement and exploration in real-world scenarios.

As illustrated in Figure 6.3, the comparison between the methods highlights that Mixed WB, while effective in simpler cases, struggles to resolve localized inconsistencies in complex mixed lighting scenarios, whereas DeNIM with Style WB backbone achieves superior results by balancing accurate white balancing with realistic rendering, which produces visually coherent and natural outputs across all scenes tested. The comparison reveals significant differences in their ability to handle mixed illuminant scenarios. Mixed WB provides a straightforward approach to balancing color casts from multiple light sources, which can effectively harmonize global illumination in simpler scenes. However, its limitations become apparent in complex scenarios, where it struggles to fully mitigate residual tints and inconsistencies between localized lighting regions. For example, in the scene with the snow-covered sculpture, Mixed WB addresses the competing cold ambient light and warm artificial sources, but fails to achieve seamless integration, leaving visible color mismatches disrupting the scene's natural coherence.

In contrast, DeNIM with the Style WB backbone demonstrates a distinct advantage, enabling it to preserve the natural vibrancy of scenes while effectively resolving challenges that arise in mixed-illumination scenarios. In the snow-covered sculpture scene, it not only balances competing light sources but also maintains the visual integrity of the snow's cold tones and the warm glow of the streetlights. Similarly, in the illuminated building with decorative lights, this approach achieves an ideal balance by mitigating excessive yellow casts while retaining the vibrancy of the scene, which results in a natural yet aesthetically pleasing outcome. These results underscore that DeNIM with the Style WB backbone excels in addressing complex real-world scenarios by achieving a nuanced equilibrium between technical accuracy and stylistic coherence, solidifying its position as the most robust method among the evaluated approaches.

Figure 6.3: *Comparison of the night photography rendering results of our WB correction strategies with Mixed WB [2] on the selected samples from Night Photography Rendering Challenge 23' evaluation set. Image indices:* 8678, 8210, 8817, 8894, 8941.

7. CONCLUSIONS

This dissertation presents a comprehensive exploration of WB correction by modeling lighting conditions as a style factor. Through the development of exact matching of the feature distribution (*i.e.*, EFDM) as a novel loss function and its integration into state-of-the-art deep learning architectures, this research addresses key challenges in handling complex illumination scenarios. The approaches proposed in this study demonstrate significant advancements over traditional approaches, as it improves spatial consistency, perceptual accuracy, and robustness in multi-illuminant scenarios, thereby accomplishing the research objectives.

This research introduces EFDM as an innovative optimization strategy that enables the exact matching of higher-order feature statistics through the [CLS] token representation for WB correction. This compact yet powerful approach improves the global contextual understanding of lighting conditions, which results in superior performance in WB correction. The integration of EFDM with architectures such as Uformer and UNet validates its adaptability, which produces consistent improvements across single-, mixed- and multi-illuminant scenarios. The single-to-multi ratio (MSR) metric, introduced in this study, provides additional insights into robustness under varying illumination conditions, highlighting the practical relevance of the proposed methods.

Although feature distribution matching-based approaches prove highly effective, their implementation within the scope of this study is limited to Uformer and UNet architectures. This limitation presents an opportunity to explore their integration with more complex and advanced architectures, potentially further enhancing performance. Furthermore, the integration of DeNIM-based mechanisms into FDM WB remains unexplored. Incorporating the deterministic pixel mapping technique of DeNIM for illumination mapping within the FDM WB could enhance its ability to handle fine-grained illumination variations and adapt to device-specific constraints while maintaining the adaptability and

precision of WB correction. Moreover, while EFDM demonstrates notable generalization and scalability under various illumination conditions, future studies could extend its application to real-world datasets with even more challenging and diverse lighting scenarios. Such efforts would ensure that these methods remain robust and generalizable in uncontrolled environments.

Future research aims to extend the findings of this work by exploring diffusion models as a means to model complex data distributions for WB correction. Flow matching and neural ordinary differential equations (ODEs) also emerge as promising frameworks for distribution-based optimization, which can potentially enhance the scalability and precision of EFDM in high-dimensional feature spaces. Expanding evaluations to real-world datasets and integrating the proposed methods into practical imaging systems further solidifies their applicability, particularly for mobile photography and professional imaging tools.

In conclusion, this dissertation offers a significant contribution to WB correction by advancing the understanding and application of distribution-based modules and optimization strategies. The proposed methods provide a robust foundation for addressing complex illumination dynamics, paving the way for further innovation in image restoration and enhancement under diverse lighting conditions.

REFERENCES

[1] N. Banić, K. Koščević, and S. Lončarić, "Unsupervised learning for color constancy," *arXiv preprint arXiv:1712.00436*, 2017.

[2] M. Afifi, M. A. Brubaker, and M. S. Brown, "Auto white-balance correction for mixed-illuminant scenes," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1210–1219, 2022.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[5] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17683–17693, 2022.

[6] M. Afifi, B. Price, S. Cohen, and M. S. Brown, "When color constancy goes wrong: Correcting improperly white-balanced images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1535–1544, 2019.

[7] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input/output image pairs," in *CVPR 2011*, pp. 97–104, IEEE, 2011.

[8] D. Kim, J. Kim, S. Nam, D. Lee, Y. Lee, N. Kang, H.-E. Lee, B. Yoo, J.-J. Han, and S. J. Kim, "Large scale multi-illuminant (lsmi) dataset for developing white balance algorithm under mixed illumination," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2410–2419, 2021.

[9] D. Kim, J. Kim, J. Yu, and S. J. Kim, " Attentive Illumination Decomposition Model for Multi-Illuminant White Balancing ," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 25512–25521, IEEE Computer Society, June 2024.

[10] M. Ebnar, "Color reproduction," in *Color Constancy*, ch. 4, pp. 67–85, John Wiley & Sons, Ltd, 2006.

[11] A. Gilchrist, *Seeing black and white*. Oxford University Press, 2006.

[12] D. Cheng, D. K. Prasad, and M. S. Brown, "Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution," *JOSA A*, vol. 31, no. 5, pp. 1049–1058, 2014.

[13] G. Buchsbaum, "A spatial processor model for object colour perception," *Journal of the Franklin Institute*, vol. 310, no. 1, pp. 1–26, 1980.

[14] D. H. Brainard and B. A. Wandell, "Analysis of the retinex theory of color vision," *JOSA A*, vol. 3, no. 10, pp. 1651–1661, 1986.

[15] G. D. Finlayson and E. Trezzi, "Shades of gray and colour constancy," in *Color and Imaging Conference*, vol. 12, pp. 37–41, Society of Imaging Science and Technology, 2004.

[16] J. T. Barron, "Convolutional color constancy," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 379–387, 2015.

[17] Y. Hu, B. Wang, and S. Lin, "Fc4: Fully convolutional color constancy with confidence-weighted pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4085–4094, 2017.

[18] J. T. Barron and Y.-T. Tsai, "Fast fourier color constancy," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 886–894, 2017.

[19] S. Bianco and C. Cusano, "Quasi-unsupervised color constancy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12212–12221, 2019.

[20] M. Afifi and M. S. Brown, "Deep white-balance editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1397–1406, 2020.

[21] B. Xu, J. Liu, X. Hou, B. Liu, and G. Qiu, "End-to-end illuminant estimation based on deep metric learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3616–3625, 2020.

[22] M. Afifi, J. T. Barron, C. LeGendre, Y.-T. Tsai, and F. Bleibel, "Cross-camera convolutional color constancy," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1981–1990, 2021.

[23] Y.-C. Lo, C.-C. Chang, H.-C. Chiu, Y.-H. Huang, C.-P. Chen, Y.-L. Chang, and K. Jou, "Clcc: Contrastive learning for color constancy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8053–8063, 2021.

[24] F. Kınlı, D. Yılmaz, B. Özcan, and F. Kıraç, "Modeling the lighting in scenes as style for auto white-balance correction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4903–4913, 2023.

[25] L. Shi, "Re-processed version of the gehler color constancy dataset of 568 images," *http://www. cs. sfu. ca/˜ color/data/*, 2000.

[26] E. Ershov, A. Savchik, I. Semenkov, N. Banić, A. Belokopytov, D. Senshina, K. Koščević, M. Subašić, and S. Lončarić, "The cube++ illumination estimation dataset," *IEEE access*, vol. 8, pp. 227511–227527, 2020.

[27] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[28] F. Kınlı, B. Özcan, and F. Kıraç, "Advancing white balance correction through deep feature statistics and feature distribution matching," *Journal of Visual Communication and Image Representation*, vol. 108, p. 104412, 2025.

[29] F. Kınlı and F. Kıraç, "Feature distribution statistics as a loss objective for robust white balance correction," *Machine Vision and Applications*, vol. 36, p. 58, Mar 2025.

[30] Y. Zhang, M. Li, R. Li, K. Jia, and L. Zhang, "Exact feature distribution matching for arbitrary style transfer and domain generalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8035–8045, 2022.

[31] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Computation*, vol. 12, no. 6, pp. 1247–1283, 2000.

[32] G. E. Hinton and R. Zemel, "Autoencoders, minimum description length and helmholtz free energy," *Advances in neural information processing systems*, vol. 6, 1993.

[33] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel, "The helmholtz machine," *Neural Computation*, vol. 7, no. 5, pp. 889–904, 1995.

[34] G. E. Hinton and Z. Ghahramani, "Generative models for discovering sparse distributed representations," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 352, no. 1358, pp. 1177–1190, 1997.

[35] D. J. Heeger and J. R. Bergen, "Pyramid-based texture analysis/synthesis," in *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, pp. 229–238, 1995.

[36] M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3828–3836, 2015.

[37] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," *Advances in neural information processing systems*, vol. 28, 2015.

[38] T. Park, J.-Y. Zhu, O. Wang, J. Lu, E. Shechtman, A. Efros, and R. Zhang, "Swapping autoencoder for deep image manipulation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7198–7211, 2020.

[39] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller, "Recognizing image style," *arXiv preprint arXiv:1311.3715*, 2013.

[40] Z. Wu, Z. Wu, B. Singh, and L. Davis, "Recognizing instagram filtered images with feature de-stylization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 12418–12425, 2020.

[41] F. Kinli, B. Ozcan, and F. Kirac, "Instagram filter removal on fashionable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 736–745, 2021.

[42] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," 2015.

[43] G. Ghiasi, H. Lee, M. Kudlur, V. Dumoulin, and J. Shlens, "Exploring the structure of a real-time, arbitrary neural artistic stylization network," *arXiv preprint arXiv:1705.06830*, 2017.

[44] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1501–1510, 2017.

[45] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019.

[46] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," *Advances in neural information processing systems*, vol. 33, pp. 12104–12114, 2020.

[47] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *Advances in neural information processing systems*, vol. 34, pp. 852–863, 2021.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[49] S. Ioffe, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[50] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[51] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2017.

[52] A. W. Van der Vaart, *Asymptotic statistics*, vol. 3. Cambridge university press, 2000.

[53] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 855–868, 2008.

[54] A. Graves, *Long Short-Term Memory*, pp. 37–45. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[55] N. Tumanyan, O. Bar-Tal, S. Bagon, and T. Dekel, "Splicing vit features for semantic appearance transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10748–10757, 2022.

[56] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biology*, vol. 52, pp. 99–115, 1990.

[57] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, IEEE, 2009.

[58] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[59] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of Medical image computing and computer-assisted intervention (MICCAI 2015), part III 18*, pp. 234–241, Springer, 2015.

[60] H. S. Malvar, L.-w. He, and R. Cutler, "High-quality linear interpolation for demosaicing of bayer-patterned color images," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, pp. 485–488, IEEE, 2004.

[61] D. Menon, S. Andriani, and G. Calvagno, "Demosaicing with directional filtering and a posteriori decision," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 132–141, 2006.

[62] T. Smith and J. Guild, "The cie colorimetric standards and their use," *Transactions of the optical society*, vol. 33, no. 3, p. 73, 1931.

[63] S. Süsstrunk, R. Buckley, and S. Swen, "Standard rgb color spaces," in *Color and Imaging Conference*, vol. 7, pp. 127–134, Society of Imaging Science and Technology, 1999.

[64] C. Poynton, *Digital Video and HD: Algorithms and Interfaces*. Morgan Kaufmann Publishers, 2012.

[65] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, "Adaptive logarithmic mapping for displaying high contrast scenes," in *Computer Graphics Forum*, vol. 22, pp. 419–426, John Wiley & Sons, Incorporated, 2003.

[66] E. Reinhard and K. Devlin, "Dynamic range reduction inspired by photoreceptor physiology," *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, no. 1, pp. 13–24, 2005.

[67] R. Mantiuk, S. Daly, and L. Kerofsky, "Display adaptive tone mapping," in *ACM SIGGRAPH 2008 papers*, pp. 1–10, ACM, 2008.

[68] N. Banic and S. Loncaric, "Flash and storm: Fast and highly practical tone mapping based on naka-rushton equation.," in *VISIGRAPP (4: VISAPP)*, pp. 47–53, 2018.

[69] F. Kınlı, B. Özcan, and F. Kıraç, "Dawn: A robust tone mapping operator for multi-illuminant and low-light scenarios," in *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 3: VISAPP*, pp. 62–68, INSTICC, SciTePress, 2025.

[70] S. Xue, M. Tan, A. McNamara, J. Dorsey, and H. Rushmeier, "Exploring the use of memory colors for image enhancement," in *Human Vision and Electronic Imaging XIX*, vol. 9014, pp. 266–275, SPIE, 2014.

[71] E. H. Land, "The retinex theory of color vision," *Scientific american*, vol. 237, no. 6, pp. 108–129, 1977.

[72] R. Gershon, A. D. Jepson, and J. K. Tsotsos, "From [r, g, b] to surface reflectance: Computing color constant descriptors in images.," in *IJCAI*, pp. 755–758, 1987.

[73] K. Barnard, V. Cardei, and B. Funt, "A comparison of computational color constancy algorithms. i: Methodology and experiments with synthesized data," *IEEE Transactions on Image Processing*, vol. 11, no. 9, pp. 972–984, 2002.

[74] W. Xiong, B. Funt, L. Shi, S.-S. Kim, B.-H. Kang, S.-D. Lee, and C.-Y. Kim, "Automatic white balancing via gray surface identification," in *Color and Imaging Conference*, vol. 15, pp. 143–146, Society of Imaging Science and Technology, 2007.

[75] A. Gijsenij, T. Gevers, and J. Van De Weijer, "Computational color constancy: Survey and experiments," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2475–2489, 2011.

[76] A. Gijsenij and T. Gevers, "Color constancy by local averaging," in *14th International Conference of Image Analysis and Processing-Workshops (ICIAPW 2007)*, pp. 171–174, IEEE, 2007.

[77] M. Ebner, "Color constancy based on local space average color," *Machine Vision and Applications*, vol. 20, no. 5, pp. 283–301, 2009.

[78] B. Funt and L. Shi, "The rehabilitation of maxrgb," in *Proceedings of the Eighteenth Color Imaging Conference*, Simon Fraser University, 2010.

[79] B. Funt and L. Shi, "The effect of exposure on maxrgb color constancy," in *Human Vision and Electronic Imaging XV*, vol. 7527, pp. 282–288, SPIE, 2010.

[80] J. Van De Weijer, T. Gevers, and A. Gijsenij, "Edge-based color constancy," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2207–2214, 2007.

[81] A. Chakrabarti, K. Hirakawa, and T. Zickler, "Color constancy beyond bags of pixels," in *2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–6, IEEE, 2008.

[82] A. Gijsenij, T. Gevers, and J. Van De Weijer, "Improving color constancy by photometric edge weighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 918–929, 2011.

[83] H. R. V. Joze, M. S. Drew, G. D. Finlayson, and P. A. T. Rey, "The role of bright pixels in illumination estimation," in *Color and Imaging Conference*, vol. 2012, pp. 41–46, Citeseer, 2012.

[84] D. A. Forsyth, "A novel algorithm for color constancy," *International Journal of Computer Vision*, vol. 5, no. 1, pp. 5–35, 1990.

[85] G. D. Finlayson, "Color in perspective," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 10, pp. 1034–1038, 1996.

[86] G. Finlayson and S. Hordley, "Improving gamut mapping color constancy," *IEEE Transactions on Image Processing*, vol. 9, no. 10, pp. 1774–1783, 2000.

[87] G. D. Finlayson, S. D. Hordley, and I. Tastl, "Gamut constrained illuminant estimation," *International Journal of Computer Vision*, vol. 67, pp. 93–109, 2006.

[88] A. Gijsenij, T. Gevers, and J. Van De Weijer, "Generalized gamut mapping using image derivative structures for color constancy," *International Journal of Computer Vision*, vol. 86, no. 2-3, p. 127, 2010.

[89] M. Mosny and B. Funt, "Cubical gamut mapping colour constancy," in *Conference on Colour in Graphics, Imaging, and Vision*, vol. 5, pp. 466–470, Society of Imaging Science and Technology, 2010.

[90] G. D. Finlayson, S. D. Hordley, and P. M. Hubel, "Color by correlation: A simple, unifying framework for color constancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1209–1221, 2001.

[91] C. Rosenberg, M. Hebert, and S. Thrun, "Color constancy using kl-divergence," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 1, pp. 239–246, IEEE, 2001.

[92] D. H. Brainard and W. T. Freeman, "Bayesian color constancy," *JOSA A*, vol. 14, no. 7, pp. 1393–1411, 1997.

[93] G. Sapiro, "Color and illuminant voting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 11, pp. 1210–1215, 1999.

[94] Y. Tsin, R. T. Collins, V. Ramesh, and T. Kanade, "Bayesian color constancy for outdoor object recognition," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, pp. I–I, IEEE, 2001.

[95] C. Rosenberg, A. Ladsariya, and T. Minka, "Bayesian color constancy with non-gaussian models," *Advances in neural information processing systems*, vol. 16, 2003.

[96] P. V. Gehler, C. Rother, A. Blake, T. Minka, and T. Sharp, "Bayesian color constancy revisited," in *2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, IEEE, 2008.

[97] W. Xiong and B. Funt, "Estimating illumination chromaticity via support vector regression," *Journal of Imaging Science and Technology*, vol. 50, no. 4, pp. 341–348, 2006.

[98] N. Wang, D. Xu, and B. Li, "Edge-based color constancy via support vector regression," *IEICE Transactions on Information and Systems*, vol. 92, no. 11, pp. 2279–2282, 2009.

[99] V. Agarwal, A. Gribok, A. Koschan, B. Abidi, and M. Abidi, "Illumination chromaticity estimation using linear learning methods," *Journal of Pattern Recognition Research*, vol. 4, no. 1, pp. 92–109, 2009.

[100] W. Xiong, L. Shi, B. Funt, S.-S. Kim, B.-H. Kang, S.-D. Lee, and C.-Y. Kim, "Illumination estimation via thin-plate spline interpolation," in *Color and Imaging Conference*, vol. 15, pp. 25–29, Society of Imaging Science and Technology, 2007.

[101] S. Gao, W. Han, K. Yang, C. Li, and Y. Li, "Efficient color constancy with local surface reflectance statistics," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pp. 158–173, Springer, 2014.

[102] M. Afifi, A. Punnappurath, G. Finlayson, and M. S. Brown, "As-projective-as-possible bias correction for illumination estimation algorithms," *JOSA A*, vol. 36, no. 1, pp. 71–78, 2018.

[103] N. Banić and S. Lončarić, "Green stability assumption: Unsupervised learning for statistics-based illumination estimation," *Journal of Imaging*, vol. 4, no. 11, p. 127, 2018.

[104] Y. Qian, S. Pertuz, J. Nikkanen, J.-K. Kämäräinen, and J. Matas, "Revisiting gray pixel for statistical illumination estimation," *arXiv preprint arXiv:1803.08326*, 2018.

[105] Y. Qian, J.-K. Kamarainen, J. Nikkanen, and J. Matas, "On finding gray pixels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8062–8070, 2019.

[106] S. A. Shafer, "Using color to separate reflection components," *Color Research & Application*, vol. 10, no. 4, pp. 210–218, 1985.

[107] A. Gijsenij and T. Gevers, "Color constancy using natural image statistics and scene semantics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 687–698, 2010.

[108] S. Bianco, G. Ciocca, C. Cusano, and R. Schettini, "Improving color constancy using indoor–outdoor image classification," *IEEE Transactions on Image Processing*, vol. 17, no. 12, pp. 2381–2392, 2008.

[109] R. Lu, A. Gijsenij, T. Gevers, V. Nedović, D. Xu, and J.-M. Geusebroek, "Color constancy using 3d scene geometry," in *2009 IEEE 12th International Conference on Computer Vision (ICCV)*, pp. 1749–1756, IEEE, 2009.

[110] J. Van De Weijer, C. Schmid, and J. Verbeek, "Using high-level visual information for color constancy," in *2007 IEEE 11th International Conference on Computer Vision (ICCV)*, pp. 1–8, IEEE, 2007.

[111] E. Rahtu, J. Nikkanen, J. Kannala, L. Lepistö, and J. Heikkilä, "Applying visual object categorization and memory colors for automatic color constancy," in *Image Analysis and Processing–ICIAP 2009: 15th International Conference Vietri sul Mare, Italy, September 8-11, 2009 Proceedings 15*, pp. 873–882, Springer, 2009.

[112] V. C. Cardei, B. Funt, and K. Barnard, "Estimating the scene illumination chromaticity by using a neural network," *JOSA A*, vol. 19, no. 12, pp. 2374–2386, 2002.

[113] R. Stanikunas, H. Vaitkevicius, and J. J. Kulikowski, "Investigation of color constancy with a neural network," *Neural Networks*, vol. 17, no. 3, pp. 327–337, 2004.

[114] J. Kulikowski and H. Vaitkevicius, "Colour constancy as a function of hue," *Acta Psychologica*, vol. 97, no. 1, pp. 25–35, 1997.

[115] Z. Lou, T. Gevers, N. Hu, M. P. Lucassen, *et al.*, "Color constancy by deep learning.," in *BMVC*, pp. 76–1, 2015.

[116] S. Bianco, C. Cusano, and R. Schettini, "Color constancy using cnns," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 81–89, 2015.

[117] W. Shi, C. C. Loy, and X. Tang, "Deep specialized network for illuminant estimation," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 371–387, Springer, 2016.

[118] S. Bianco, C. Cusano, and R. Schettini, "Single and multiple illuminant estimation using convolutional neural networks," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4347–4362, 2017.

[119] S. W. Oh and S. J. Kim, "Approaching the computational color constancy as a classification problem through deep learning," *Pattern Recognition*, vol. 61, pp. 405–416, 2017.

[120] M. Afifi and M. S. Brown, "What else can fool deep learning? addressing color constancy errors on deep neural network performance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 243–252, 2019.

[121] D. Hernandez-Juarez, S. Parisot, B. Busam, A. Leonardis, G. Slabaugh, and S. McDonagh, "A multi-hypothesis approach to color constancy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[122] B. Xu, J. Liu, X. Hou, B. Liu, and G. Qiu, "End-to-end illuminant estimation based on deep metric learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[123] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Similarity-based pattern recognition: third international workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*, pp. 84–92, Springer, 2015.

[124] C. Li, X. Kang, Z. Zhang, and A. Ming, "Swbnet: A stable white balance network for srgb images," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 1278–1286, Jun. 2023.

[125] M. Afifi, A. Punnappurath, A. Abdelhamed, H. C. Karaimer, A. Abuolaim, and M. S. Brown, "Color temperature tuning: Allowing accurate post-capture white-balance editing," in *Color and Imaging Conference*, vol. 27, pp. 1–6, Society for Imaging Science and Technology, 2019.

[126] M. Afifi and M. S. Brown, "Interactive white balancing for camera-rendered images," *arXiv preprint arXiv:2009.12632*, 2020.

[127] O. Ulucan, D. Ulucan, and M. Ebner, "Color constancy beyond standard illuminants," in *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 2826–2830, IEEE, 2022.

[128] M. Buzzelli, J. van de Weijer, and R. Schettini, "Learning illuminant estimation from object recognition," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 3234–3238, 2018.

[129] M. Afifi, J. T. Barron, C. LeGendre, Y.-T. Tsai, and F. Bleibel, "Cross-camera convolutional color constancy," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1981–1990, October 2021.

[130] O. Ulucan, D. Ulucan, and M. Ebner, "Bio-cc: Biologically inspired color constancy.," in *33rd British Machine Vision Conference 2022 (BMVC)*, 2022.

[131] O. Ulucan, D. Ulucan, and M. Ebner, "Multi-scale color constancy based on salient varying local spatial statistics," *The Visual Computer*, vol. 40, no. 9, pp. 5979–5995, 2024.

[132] M. Bleier, C. Riess, S. Beigpour, E. Eibenberger, E. Angelopoulou, T. Tröger, and A. Kaup, "Color constancy and non-uniform illumination: Can existing algorithms work?," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 774–781, 2011.

[133] A. Gijsenij, R. Lu, and T. Gevers, "Color constancy for multiple light sources," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 697–707, 2012.

[134] S. Beigpour, C. Riess, J. van de Weijer, and E. Angelopoulou, "Multi-illuminant estimation with conditional random fields," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 83–96, 2014.

[135] D. Kim, J. Kim, S. Nam, D. Lee, Y. Lee, N. Kang, H.-E. Lee, B. Yoo, J.-J. Han, and S. J. Kim, "Large scale multi-illuminant (lsmi) dataset for developing white balance algorithm under mixed illumination," in *Proceedings of the IEEE/CVF*

International Conference on Computer Vision (ICCV), pp. 2410–2419, October 2021.

[136] H. R. V. Joze and M. S. Drew, "Exemplar-based color constancy and multiple illumination," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 860–873, 2014.

[137] T. Akazawa, Y. Kinoshita, S. Shiota, and H. Kiya, "N-white balancing: White balancing for multiple illuminants including non-uniform illumination," *IEEE Access*, vol. 10, pp. 89051–89062, 2022.

[138] S. Li, J. Wang, M. S. Brown, and R. T. Tan, "Mimt: Multi-illuminant color constancy via multi-task local surface and light color learning," *arXiv preprint arXiv:2211.08772*, 2022.

[139] I. Domislović, D. Vršnjak, M. Subašić, and S. Lončarić, "Color constancy for non-uniform illumination estimation with variable number of illuminants," *Neural Comput. Appl.*, vol. 35, p. 14825–14835, Mar. 2023.

[140] U. C. Entok, F. Laakom, F. Pakdaman, and M. Gabbouj, "Pixel-wise color constancy via smoothness techniques in multi-illuminant scenes," 2024.

[141] F. Kınlı, B. Özcan, and F. Kıraç, "Patch-wise contrastive style learning for instagram filter removal," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 578–588, 2022.

[142] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 768–783, 2018.

[143] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[144] B. Poole and J. Barron, "The fast bilateral solver," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 617–632, Springer, 2016.

[145] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[146] I. Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[147] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[148] F. Kınlı, D. Yılmaz, B. Özcan, and F. Kıraç, "Deterministic neural illumination mapping for efficient auto-white balance correction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1139–1147, 2023.

[149] D. Fourure, R. Emonet, E. Fromont, D. Muselet, A. Tremeau, and C. Wolf, "Residual conv-deconv grid network for semantic segmentation," *arXiv preprint arXiv:1707.07958*, 2017.

[150] E. Ershov, A. Savchik, D. Shepelev, N. Banić, M. S. Brown, R. Timofte, K. Koščević, M. Freeman, V. Tesalin, D. Bocharov, *et al.*, "Ntire 2022 challenge on night photography rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1287–1300, 2022.

[151] A. Shutova, E. Ershov, G. Perevozchikov, I. Ermakov, N. Banić, R. Timofte, R. Collins, M. Efimova, A. Terekhin, S. Zini, *et al.*, "Ntire 2023 challenge on night photography rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1982–1993, 2023.

[152] J. Li and P. Fang, "Hdrnet: Single-image-based hdr reconstruction using channel attention cnn," in *ICMSSP '19*, (New York, NY, USA), p. 119–124, Association for Computing Machinery, 2019.

[153] Z. Ke, Y. Liu, L. Zhu, N. Zhao, and R. W. Lau, "Neural preset for color style transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14173–14182, June 2023.

[154] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[155] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 694–711, Springer, 2016.

[156] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5728–5739, 2022.

[157] N. Banić and S. Lončarić, "Improving the white patch method by subsampling," in *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 605–609, IEEE, 2014.

[158] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Transactions on Image Processing*, vol. 9, no. 9, pp. 1532–1546, 2000.

VITA

Osman Furkan Kınlı

Education

Ph.D. in Computer Science @ Özyeğin University (2019 – 2025)

M.Sc. in Computer Science @ Özyeğin University (2018 – 2019)

B.Sc. in Computer Science in Engineering @ Özyeğin University (2012 – 2018)

Work Experience

Co-Founder & AI Director: T-Fashion (Canada & Türkiye) (2019 – *Present*)

Research Engineer (CV & DL): Fishency Innovation (Norway - Remote) (2020 – 2024)

Engineer (ML): eBay & GittiGidiyor (Türkiye) (2017 – 2018)

Intern Engineer (BI & Data): Türk Telekom (Türkiye) (06.2017 – 08.2017)

Awards and Scholarships

Recipient: ML Reproducibility Challenge 2022

Nomination for Outstanding Reviewer: ML Reproducibility Challenge 2022

4th Place: NTIRE 2022 Challenge on Night Photography Rendering

7th Place: NTIRE 2023 Challenge on Night Photography Rendering

Publications

– F. Kınlı, F. Kıraç. ”Feature distribution statistics as a loss objective for robust white balance correction.” Machine Vision and Applications, 36(3), 1-20, 2025.

– F. Kınlı, B. Özcan, F. Kıraç. ”Advancing white balance correction through deep feature statistics and feature distribution matching.” Journal of Visual Communication and Image Representation, 108, 104412, 2025.

– N. Banić, E. Ershov, A. Panshin, ..., F. Kınlı, ..., R. Timofte. ”NTIRE 2024 challenge on night photography rendering.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2024.

– B. Özcan, F. Kınlı, F. Kıraç. ”Generalization to unseen viewpoint images of objects via alleviated pose attentive capsule agreement.” Neural Computing and Applications, 2023, 35(4), 3521-3536.

– A. Shutova, E. Ershov, G. Perevozchikov, ..., F. Kınlı, ..., R. Timofte. ”NTIRE 2023 challenge on night photography rendering.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2023, pp. 1982–1993.

– F. Kınlı, B. Özcan, F. Kıraç. ”[Re] Exact Feature Distribution Matching for Arbitrary Style Transfer and Domain Generalization.” ML Reproducibility Challenge, 2023.

– F. Kınlı, D. Yılmaz, B. Özcan, F. Kıraç. ”Deterministic Neural Illumination Mapping for Efficient Auto-White Balance Correction.” Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 1139–1147.

– F. Kınlı, D. Yılmaz, B. Özcan, F. Kıraç. ”Modeling the Lighting in Scenes as Style for Auto White-Balance Correction.” Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 4903–4913.

– F. Kınlı, B. Özcan, F. Kıraç. ”Reversing image signal processors by reverse style transferring.” Computer Vision–ECCV 2022 Workshops, 2023, pp. 688–698.

– F. Kınlı, S. Menteş, B. Özcan, F. Kıraç, R. Timofte, Y. Zuo, Z. Wang, X. Zhang, ... ”Aim 2022 challenge on instagram filter removal: Methods and results.” Computer Vision–ECCV 2022 Workshops, 2023, pp. 27–43.

– F. Kınlı, B. Özcan, F. Kıraç. ”Patch-wise contrastive style learning for instagram filter removal.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 578–588.

– M.V. Conde, R. Timofte, Y. Huang, ..., F. Kınlı, ..., R. Timofte. ”Reversed image signal processing and RAW reconstruction. AIM 2022 challenge report.” European Conference on Computer Vision, 2022, pp. 3–26.

– E. Ershov, A. Savchik, D. Shepelev, N. Banić, M.S. Brown, R. Timofte, ..., F. Kınlı, ... ”NTIRE 2022 challenge on night photography rendering.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2022, pp. 1287–1300.

– F. Kınlı, B. Özcan, F. Kıraç. ”Instagram filter removal on fashionable images.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 736–745.

– B. Özcan, F. Kınlı, F. Kıraç. ”Quaternion capsule networks.” 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 6858–6865.

– F. Kınlı, B. Özcan, F. Kıraç. ”A benchmark for inpainting of clothing images with irregular holes.” Computer Vision–ECCV 2020 Workshops, 2020, pp. 182–199.

– S. Menteş, F. Kınlı, B. Özcan, F. Kıraç. ”[Re] Spatial-Adaptive Network for Single Image Denoising.” ML Reproducibility Challenge, 2020.

– F. Kınlı, B. Özcan, F. Kıraç. ”Description-aware fashion image inpainting with convolutional neural networks in coarse-to-fine manner.” Proceedings of the 2020 6th International Conference on Computer and Technology Applications (ICCTA), 2020, pp. 74–79.

– F. Kınlı, F. Kıraç. ”FashionCapsNet: Clothing classification with capsule networks.” Bilişim Teknolojileri Dergisi, 13(1), 87-96, 2020.

– F. Kınlı, B. Özcan, F. Kıraç. ”Fashion image retrieval with capsule networks.” Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 36.

– F. Kınlı. Clothing image retrieval with triplet capsule networks. Ozyegin University, 2019.