IMAGE WHITE-BALANCE FOR MULTI-ILLUMINANT SCENES

ADITYA ARORA

A THESIS SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

GRADUATE PROGRAM IN COMPUTER SCIENCE
YORK UNIVERSITY
TORONTO, ONTARIO

AUGUST, 2024

© Aditya Arora, 2024

Abstract

Performing white-balance (WB) correction for scenes with multiple illuminants remains a challenging task in computer vision. Most previous methods estimate per-pixel scene illumination directly in the RAW sensor image space. Recent work explored an alternative fusion strategy, where a neural network fuses multiple white-balanced versions of the input image processed to sRGB using pre-defined white-balance settings. Inspired by this line of work, we present two contributions targeting fusion-based multi-illuminant WB correction. First, we introduce a large-scale multi-illumination dataset rendered from RAW images to support training fusion models and evaluation. The dataset comprises over 16,000 sRGB images with ground truth sRGB white-balance corrected images. Next, we introduce an attention-based architecture to fuse five white-balance settings. This architecture yields an improvement of up to 25% over prior work.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Waheguru for providing me with the strength, guidance, and wisdom to complete this journey. Your blessings have been my constant source of inspiration and resilience.

I would also like to thank my parents, whose unwavering support and encouragement have been the cornerstone of my achievements. Your sacrifices, love, and belief in me have been instrumental in reaching this milestone. I am eternally grateful for everything you have done.

To my partner, Akshita, thank you for your endless patience, understanding, and encouragement. Your support has been my anchor throughout this challenging process, and your love has been my motivation. I could not have done this without you by my side.

I am also grateful to my advisors Prof. Kosta Derpanis and Prof. Michael S. Brown who have provided invaluable guidance, feedback, and support during my research. Your contributions have been essential to the completion of this thesis. I want to thank Dr. Javier Vazquez-Corral and Dr. Mahmoud Afifi for their continuous feedback during this work.

I would also like to thank my friends for their constant encouragement, support, and companionship. Your belief in me has been a source of motivation, and I am truly grateful for your presence. Thank you all for being a part of this journey.

Contents

Abstract ii

Acknowledgements iii

Contents iv

List of Tables vi

List of Figures vii

1 Introduction 1

1.1 Motivation and problem 1

1.2 Thesis contribution 6

1.3 Summary 7

2 Related Work 8

2.1 Conventional white-balance correction 8

2.1.1 Statistical methods 8

2.1.2 Physics-based methods 9

2.1.3 Deep learning methods 10

2.2 Multiple-illuminant methods 11

2.3 White-balance manipulation in sRGB images 13

2.4 White-balance datasets 14

2.4.1 Single-illuminant datasets 15

2.4.2 Multi-illuminant datasets . . . 16
2.5 Self-attention in low-level vision . . . 18
2.6 Summary . . . 20

3 Multi-Illuminant sRGB Dataset **24**
3.1 Limitations . . . 28
3.2 Summary . . . 29

4 Method **31**
4.1 Baseline fusion-based methods . . . 32
4.2 Attention-based WB fusion . . . 34
4.3 Summary . . . 36

5 Experiments **37**
5.1 Datasets . . . 38
5.2 Metrics . . . 38
5.3 Training details . . . 41
5.4 Baselines . . . 43
5.5 Results . . . 44
5.6 Cross-camera generalization . . . 47
5.7 Results on a synthetic dataset . . . 49
5.8 Summary . . . 49

6 Conclusions **55**
6.1 Summary . . . 55
6.2 Societal Impacts . . . 56
6.3 Future Work . . . 59

Bibliography **62**

List of Tables

5.1 Quantitative evaluation for our model 41
5.2 Generalization of our methods 42
5.3 Results on the synthetic dataset 42

List of Figures

1.1 Motivation behind using multiple models for white-balance 2
1.2 Results by Afifi et al. and our proposed method 3

2.1 Examples from the Gehler dataset 21
2.2 Examples from the NUS dataset 22
2.3 Examples from the Afifi dataset 23

3.1 Dataset preparation 25
3.2 Examples of our dataset 30

4.1 Architecture of attention-based white-balance fusion method 34

5.1 Qualitative results for the Sony camera 51
5.2 Qualitative results for the Nikon camera 52
5.3 Qualitative results comparing our method versus Afifi et al. 53
5.4 Attention maps of our attention-based method 54

Chapter 1

Introduction

This chapter motivates the need for white-balance and discusses the limitations of current white-balance algorithms when a scene is illuminated by multiple light sources. The contributions of this thesis to addressing multi-illumination white-balance are described.

1.1 Motivation and problem

Proper white-balance is crucial for accurate color rendition in digital images. White-balance refers to removing color casts or tints caused by the environment's illumination from images so that objects that visually appear as white in the scene to a human observer are rendered as white in the image. Achieving high-quality white-balance is essential for many image-processing applications across various domains, including photography, computer vision, medical imaging, and remote sensing.

The need for white-balance arises because different light sources (e.g., daylight, incandescent, and fluorescent) emit light with varying spectral power distributions. For instance,



Figure 1.1: Motivation behind using multiple models for white-balance. For a scene with multiple illuminations. (Left) When the AWB algorithm uses outdoor illumination as a reference, it produces an orangeish output. (Mid) When the AWB algorithm uses indoor illumination as a reference, it produces a blueish output. (Right) Using a mix-illumination correction produces best results.

daylight creates a bluish cast in images, while incandescent bulbs produce a warmer, reddish color cast. The human visual system has an impressive ability to automatically compensate for such color shifts through chromatic adaptation, allowing us to perceive white objects as truly white under most illumination conditions [20]. However, digital imaging sensors lack this chromatic adaptation capability and thus require the application of white-balance adjustment to all images to produce accurate colors.

By addressing the color cast introduced by various light sources, from the warm glow of incandescent bulbs to the cool shade of natural daylight, white-balance correction techniques strive to neutralize these effects. This ensures that the resultant images convey a sense of realism and depth and maintain consistency in color representation, regardless of the lighting conditions under which they were captured.

Proper white-balance is not only necessary for producing visually appealing images, it is also indispensable for downstream computer vision tasks [6] that assume input im-

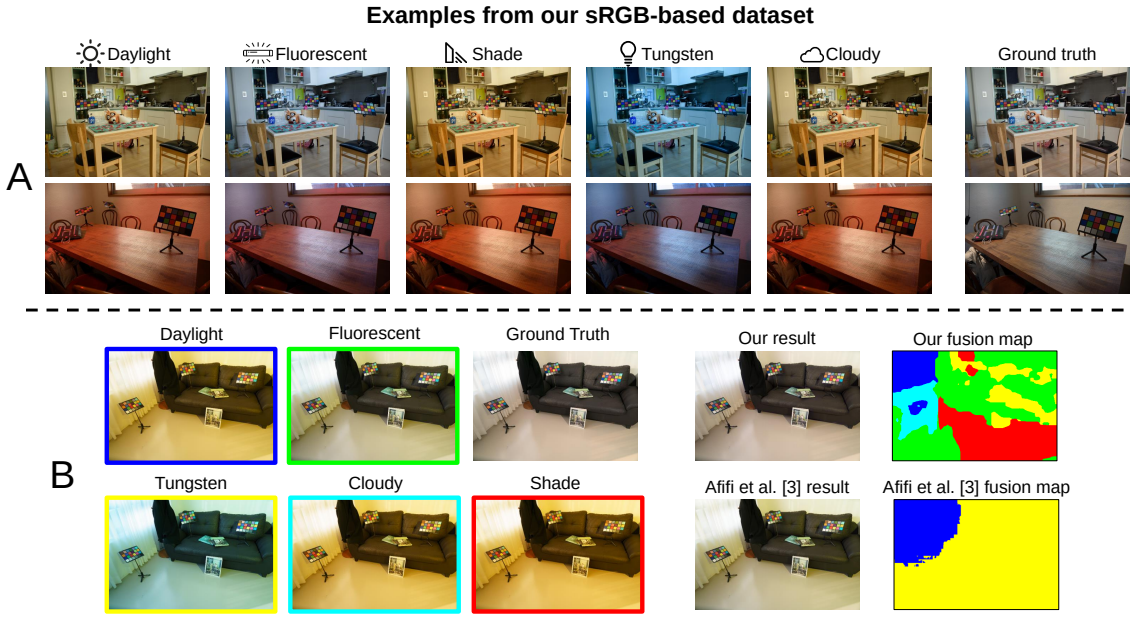


Figure 1.2: Results by Afifi et al. and our proposed method. (A) Examples from our sRGB dataset rendered from RAW images captured by Kim et al. [36]. The dataset provides the five preset white-balance sRGB outputs from the camera and ground truth sRGB image. (B) The dataset allows the training and testing of fusion-based methods for multi-illuminant white-balance correction. Shown are results by Afifi et al. [8] and our proposed method using attention-based fusion.

ages are standardized regarding their response to scene illumination. As computer vision systems rely heavily on accurate color information to analyze and understand images, any discrepancy in color representation can lead to errors in object recognition, scene understanding, and other automated processes. Therefore, by ensuring that the colors in an image are adjusted to reflect those seen by the human eye—making white areas appear truly white and accurately representing all other colors—white-balance correction is a foundational step in pre-processing images for computer vision applications.

Existing white-balance techniques can be broadly categorized into two main approaches:

computational color constancy algorithms and user-guided methods. Computational color constancy algorithms attempt to automatically estimate the scene's illuminant color and then apply the correct to the image based on this estimation. These algorithms typically exploit various statistical properties of images, such as exploiting the assumption that the average scene color should be achromatic (gray world assumption) or that the maximum RGB values in the image correspond to the illuminant color (max-RGB assumption). While computational methods are fully automatic, their performance can be limited in challenging scenarios with complex illumination conditions or violations of the underlying assumptions.

In contrast, user-guided white-balance methods rely on manual input or interaction from the user to specify reference white points or regions within the image. These methods leverage the user's visual perception and scene understanding to guide the white-balance process, often resulting in more accurate and reliable results than fully automatic techniques. However, user interaction can be time-consuming and not feasible or practical in certain applications requiring real-time or high-throughput processing.

A major assumption made by auto white-balance (AWB) methods is that the scene is illuminated by a single dominant light source. These methods, however, falter in the face of multi-illuminant environments, where varying light sources coexist, leading to color inaccuracies and casting challenges. As a result, conventional white-balancing techniques that offer satisfactory results in uniformly lit environments often fail to accurately correct images lit by multiple illuminants. Scenes with multiple light sources are commonplace in both outdoor and indoor settings, ranging from landscapes at twilight, where natural and artificial lights intermingle, to indoor spaces illuminated by a combination of daylight and

various artificial light sources.

As previously mentioned, standard AWB techniques (e.g., [15, 27, 55, 32, 10]) assume that a scene is illuminated by a single light source. AWB methods operate directly on the RAW sensor image to first estimate the color cast in the image in the form of a sensor-specific R, G, B vector. The illumination's color cast is removed by scaling each color channel of the RAW image such that the estimated illumination values become achromatic or "white" (i.e., R=G=B). The white-balanced RAW image is rendered through the camera's image signal processor (ISP) to render the final output image in a standard RGB (sRGB) color space.

Fig. 1.1 demonstrates the impact of applying conventional white-balance correction with its single illumination assumption to a scene with multiple light sources. The indoor scene has mixed illumination from indoor and outdoor lighting (via a window) often. If we correct the outdoor illumination, the indoor portion of the scene retains a reddish color cast. If we correct the indoor illumination, the outdoor portion of the scene retains a bluish color cast. The final image shows the result of performing multi-illuminant AWB to correct both illuminations.

There has been prior work that has examined ways to correct such multiple illuminations (e.g., [51, 57, 56]). These methods typically follow a two-step procedure that first tries to estimate the multiple illuminations in the scene, followed by spatially varying white-balance correction in the RAW sensor image based on the estimations. The need for both estimation and correction makes this a challenging problem. Another bottleneck for multi-illuminant methods is the difficulty in preparing per-pixel annotated RAW image datasets for training and testing.

Recent work by Afifi et al. [8] introduced a novel approach for multi-illuminant WB correction that avoided the need for explicit illumination estimation. Specifically, the method in Afifi et al. [8] exploited the availability of the camera's predefined white-balance values for common scene illuminations (i.e., the settings used by *manual WB mode* instead of AWB). The R,G,B sensor response to several common scene illuminations can be obtained directly through camera API calls or image metadata. This corrected an input image using a fixed selection of WB settings, namely: *tungsten*, *shade*, and *daylight*. These three images were then rendered to sRGB. A deep neural network was trained to predict per-pixel weights to blend the three post-WB images to produce a rendered sRGB output. These predefined illuminants define a solution-space gamut within which the blending is performed.

1.2 Thesis contribution

In this thesis, we revisit the fusion-based approach for multi-illumination WB correction with two contributions. First, we introduce a new multi-illuminant white-balance dataset in the sRGB domain. This dataset is constructed by repurposing the RAW multi-illuminant dataset by Kim et al. [36]. The dataset consists of 2714 scenes rendered to five white-balanced sRGB images and a corresponding ground truth white-balanced sRGB image. This dataset facilitates our blending-based fusion method's training, evaluation, and comparison. Fig. 1.2-(A) shows examples of the rendered sRGB images and our ground truth images. Based on this new dataset, we revisit the blending map estimation process by casting it as an attention-based architecture. We develop an innovative solution capable

of adeptly managing the complexities of multiple light sources in an image. This work introduces a novel approach that leverages deep learning techniques by critically analyzing the limitations inherent in conventional AWB methods, designed predominantly for single-illuminant scenarios. We show that the attention-based method produces significantly better results than that of Afifi et al. [8]. An example is shown in as shown in Fig. 1.2-(B).

1.3 Summary

This thesis tackles the challenge of achieving accurate white-balance in scenes with multiple light sources. Building on previous methods, we introduce a new approach that uses an attention-based fusion strategy to combine multiple white-balance settings effectively. We also present a large dataset of images with different lighting conditions and use advanced neural networks to improve color correction accuracy. The following chapters cover the related work, explain our method, show detailed experimental results, and discuss the broader impact and future possibilities of this research. Our goal is to significantly improve white-balance correction in complex lighting environments.

Chapter 2

Related Work

Related work is described in five areas: (i) conventional WB correction, (ii) multi-illuminant WB correction, (iii) methods focused on WB manipulation for processed sRGB images, (iv) white-balance datasets, and (v) self-attention in low-level vision.

2.1 Conventional white-balance correction

Traditional white-balance correction techniques attempt to estimate the sensor's response to illumination from a single RAW input image. These methods can be divided into three categories: statistics-based, physics-based, and deep-learning.

2.1.1 Statistical methods

Early statistical methods (e.g., [15, 39, 22, 52, 16, 16, 28, 17, 21, 47]) computed simple RAW image statistics as features for illumination estimation prediction. Finlayson et al.

[24] introduced color-by-correlation, a method for estimating scene illuminant color in digital imaging systems, significantly outperforming previous color constancy approaches through the use of correlation matrices and Bayesian statistics. Brainard et al. [14] introduced a Bayesian approach to color constancy, offering the novel maximum local mass estimator for superior illumination estimation accuracy. Gehler et al. [26] revisit Bayesian color constancy, contrasting with Brainard et al. [14] by employing a new, high-quality dataset of 568 images to refine illuminant and reflectance priors, significantly enhancing outdoor image illumination estimation over previous methods and grey-world algorithms. Sapiro et al. [48] presents a method for color and illuminant estimation in images by focusing on a voting mechanism that integrates multiple cues for more robust illuminant estimation. Vazquez et al. [54, 55] propose a new perceptual angular error measure for evaluating color constancy algorithms, emphasizing human preference for "naturalness" over mathematical accuracy, tested on a unique dataset with a calibrated camera and neutral grey sphere.

2.1.2 Physics-based methods

Physics-based methods operate under a physical image formation model, such as Shafer's dichromatic illumination model [49]. Funt et al. [25] developed a method to achieve color constancy by exploiting mutual reflection between surfaces, demonstrating that accurate surface spectral reflectance and ambient illumination can be recovered from color imaging data. Their approach, based on finite-dimensional linear models for illumination and surface reflectance, suggests that with sufficient sensor classes, it is possible to disentangle illumination and reflectance from the observed color signals. This method advances the

understanding of color perception under varied lighting conditions and has potential applications in improving computer vision systems, particularly in tasks requiring accurate color information independent of lighting. Klinker et al. [37] developed a comprehensive approach to color image understanding by integrating the Dichromatic Reflection Model [49] for analyzing color variations due to highlights and shading, enabling more accurate color image segmentation and interpretation by leveraging physical properties of light reflection.

2.1.3 Deep learning methods

Deep learning methods (e.g., [12, 30, 45, 46, 50, 61]) have now become the state of the art for illumination estimation. Because illumination estimation is performed on the sensor-specific RAW image format, these methods require large annotated data sets of images for each sensor. Afifi et al. [5] propose an innovative interactive white-balance correction method in camera-rendered images, enhancing user control by directly manipulating color cast vectors for more precise and efficient WB adjustments. Lo et al. [44] introduce a contrastive learning framework for color constancy (named CLCC), which improves accuracy and robustness through novel RAW-domain color augmentations and contrastive pair formulation [64].

The proposed method in this thesis aligns with this contemporary trajectory by harnessing the power of deep learning for white-balance correction in multi-illuminant scenarios. It builds upon the foundational principles of previous works while introducing a novel attention-based fusion strategy. This strategy enables the model to dynamically integrate information from multiple pre-defined white-balance settings, thereby enhancing its abil-

ity to handle complex illumination conditions without the need for explicit illumination estimation.

One limitation of traditional methods is their reliance on the assumption of a single dominant light source in the scene. When faced with scenes containing multiple light sources, such as indoor environments with artificial and natural lighting, traditional methods may fail to achieve accurate color reproduction.

In comparison to prior work, our method demonstrates a significant improvement in handling the intricacies of scenes with multiple illuminants. By leveraging an attention-based architecture, it surpasses traditional per-pixel illumination estimation techniques, which often struggle with the spatial variability of complex lighting environments. Our approach effectively mitigates these challenges, offering a more nuanced and precise white-balance correction that is essential for both visual fidelity and downstream computer vision tasks.

2.2 Multiple-illuminant methods

Various techniques have been specifically designed for multiple-illuminant white-balance correction to address the limitations of single-illuminant white-balance methods. Prior methods often relied on pixel-wise illumination estimation strategies [23, 29, 11]. Hsu et al. [31] developed a white-balance technique for images with mixed lighting, introducing a method to estimate light mixture per pixel and achieving enhanced white-balance and post-exposure relighting effects. Many of these methods require prior knowledge of the number of different lights present in the scene, which can be a limitation in real-world

scenarios.

Joze et al. [35] introduced a novel exemplar-based approach to color constancy, addressing the challenge of accurately estimating the illuminant color in images under varied lighting conditions. Their method, distinct from traditional techniques, utilizes texture features alongside weakly color constant values to identify and compare surfaces within an image to a pre-learned dataset, enabling more precise illuminant estimation. This approach outperforms existing algorithms across standard datasets and offers a robust solution to the often problematic scenario of images lit by multiple light sources, demonstrating its effectiveness through extensive evaluation and inter-dataset validation. Hussain et al. [33] proposed a color constancy algorithm for mixed-illuminant scenes that divides an image into segments, assesses color variability with Normalized Average Absolute Difference (a metric used to quantify the color variability within segments of an image), and combines segment-based weighting factors adjusted by pixel distances from segment centers.

Another hurdle for these methods is generating ground truth images for training and evaluation. Several works have captured datasets of real scenes with spatial illumination labels. For example, work by Bleier et al. [13] sprayed entire scenes with achromatic paint to provide per-pixel illumination labels but did so for only a handful of examples. Aghaei and Funt [9] placed a gray ball on a drone and flew it in a static scene, capturing approximately 100 spatial samples of the environment's illumination. Work by Kim et al. [36] prepared a large multi-illumination dataset by capturing several images of the same scene with a fixed camera with different light sources turned on or off. This last dataset represents the most comprehensive dataset for multi-illuminant research. We process this dataset further to be suitable for post-WB fusion methods targeted in this work.

In contrast to these traditional approaches, the proposed method in this thesis introduces a novel paradigm by leveraging an attention-based fusion strategy to address multi-illuminant white-balance correction. Rather than relying on explicit illumination estimation or requiring prior knowledge of the number of illuminants, our approach dynamically integrates multiple pre-defined white-balance settings through a neural network.

2.3 White-balance manipulation in sRGB images

The camera's ISP hardware applies white-balance as an early processing step. Several nonlinear processing steps are applied afterward to render the final sRGB image. This makes post-capture WB correction challenging. Recent work has investigated ways to perform white-balance manipulation to sRGB images [5]. Afifi et al. 2019 (a) [6] introduced a novel data-driven approach to correct improperly white-balanced images, leveraging a large dataset of over 65,000 image pairs to train a model that effectively reconstructs accurate color representation in images affected by incorrect white-balance settings. Afifi et al. 2019 (b) [7] introduced an imaging framework that allows accurate post-capture white-balance editing in sRGB images by rendering tiny versions of the original image with different white-balance settings to create color mapping functions, enabling adjustments to any color temperature with minimal data overhead. A notable advancement in this domain is the introduction of deep neural networks (DNNs) for correcting and editing the white-balance settings of camera-rendered sRGB images [4]. These methods have shown promising results in improving color accuracy, but they typically rely on post-capture correction techniques or user interaction, adding complexity and additional steps to the image

processing workflow.

As discussed in Sec. 1, Afifi et al. [8] proposed to address the multi-illuminant white-balance problem as a fusion problem in the sRGB domain. In particular, they pose the multi-illuminant white-balance problem by obtaining a correct weighting of three fixed illuminations. We show how to improve this approach by examining the selection of different predefined WB settings on the fusion algorithm.

While RAW-based methods offer detailed control over image adjustments, sRGB-based white-balance correction provides practical advantages in terms of workflow integration, computational efficiency, and leveraging existing image enhancements. The proposed method in this thesis, by operating in the sRGB domain, harnesses these benefits to deliver a robust and efficient solution for multi-illuminant white-balance correction.

2.4 White-balance datasets

The development and evaluation of white-balance correction algorithms necessitate robust datasets that encompass a wide variety of illumination conditions and scenes. These datasets are pivotal for training models, benchmarking their performance, and ensuring their generalizability across diverse real-world scenarios. Historically, white-balance datasets have evolved from containing single illuminant scenes to more complex multi-illuminant environments, reflecting the increasing sophistication and requirements of contemporary computational photography and computer vision applications. This section provides an overview of the key datasets utilized in the field, highlighting their unique contributions and limitations in supporting the advancement of white-balance correction methodologies.

By examining these datasets, we can better understand the challenges and opportunities they present for developing more accurate and efficient white-balance algorithms.

2.4.1 Single-illuminant datasets

Gehler Dataset The Gehler dataset Fig. 2.1, introduced by Gehler et al. [26], is a seminal dataset for evaluating color constancy algorithms. It consists of 568 high-quality linear images captured by two cameras (Canon 1D and Canon 5D) under diverse illumination conditions, including daylight, tungsten, and fluorescent lighting. The dataset provides ground truth illumination estimates, enabling quantitative evaluation of color constancy and white-balance algorithms. The images in the Gehler Dataset exhibit variations in orientation and size, as they were captured by two different camera models.

NUS Dataset The NUS dataset [17] provides an even larger collection of images Fig. 2.2 for benchmarking color constancy and white-balance algorithms. This dataset contains 1736 raw images, divided into eight subsets of approximately 210 images each, corresponding to different camera models. While the total number of images in the NUS dataset is larger than the Gehler dataset, it is important to note that each independent experiment or evaluation on the NUS dataset typically involves only about one-third of the number of images present in the reprocessed Gehler dataset. This is due to the division of images into distinct subsets based on camera models.

Rendered WB Dataset The Rendered WB dataset [6] is a large-scale dataset specifically designed for training and evaluating deep learning-based white-balance correction methods. This dataset addresses a critical need in the field by providing a diverse collection of image pairs, where each incorrectly white-balanced image is accompanied by its cor-

responding correctly white-balanced counterpart. The Rendered WB dataset consists of approximately 65,000 sRGB image pairs, rendered by different camera models and with varying white-balance settings. These images were captured by various DSLR and mobile phone cameras, providing a comprehensive representation of real-world scenarios encountered in computational photography and image processing applications. A notable aspect of this dataset is the inclusion of ground truth images rendered with accurate white-balance correction. For each incorrectly white-balanced image, there is a corresponding ground truth sRGB image rendered with the correct white-balance setting, effectively serving as the target for training and evaluation purposes. Furthermore, including images captured by different camera models and sensors allows for exploring generalization capabilities and developing robust white-balance correction techniques that can adapt to various imaging pipelines and device characteristics.

2.4.2 Multi-illuminant datasets

Mixed WB Dataset This dataset [8] provides a controlled and challenging testbed specifically designed to assess the performance of multi-illuminant white-balance methods. It consists of 30 synthetic images Fig. 2.3 rendered with five different illumination conditions, allowing for a comprehensive analysis of our method's capabilities in handling mixed-illumination scenarios. This dataset was carefully curated to provide a rigorous and realistic testing environment for multi-illuminant white-balance methods. It was generated using 3D scenes modeled in Autodesk 3Ds Max, with multiple light sources of varying color temperatures (e.g., 2850 K, 5500 K) added to each scene. The scenes were rendered using Vray rendering [2] to produce realistic photographs, resulting in a total of

150 images with mixed lighting conditions. For each rendered image, a corresponding ground-truth image was also generated by setting the color temperature of the virtual camera's white-balance and all light sources in the scene to 5500 K (daylight). This approach ensures that the ground-truth images have correctly white-balanced pixels, providing a reliable reference for evaluating the performance of multi-illuminant white-balance methods.

The dataset is the first to include images rendered with different white-balance settings, along with the corresponding ground-truth images rendered with correctly white-balanced pixels in the rendering pipeline. This dataset addresses a crucial limitation of existing multi-illuminant datasets, which often lack ground-truth references or fail to accurately capture the complexities of mixed-illumination scenes.

LSMI Dataset The LSMI dataset [36] contains RAW images of scenes captured with multiple cameras, where the illumination was carefully controlled. Specifically, the LSMI dataset was created by first capturing a scene under a single illumination source and then adding additional light sources step-by-step. This allowed for a per-pixel computation of the contribution of each light source in the RAW image data. The LSMI dataset was primarily intended for developing and evaluating RAW-based multi-illumination color constancy algorithms.

Our proposed method directly leverages the strengths of multi-illuminant datasets like the LSMI and Mixed WB datasets. By utilizing these datasets, we can train our attention-based fusion model on a diverse array of scenes with varying illumination conditions. This comprehensive training allows our model to effectively learn the complex interactions between different light sources and apply accurate white-balance corrections in multi-

illuminant environments. The integration of these datasets is critical to the success of our approach, as it provides the necessary variety and complexity needed to generalize well to real-world scenarios. Furthermore, the ground-truth images available in these datasets offer a robust benchmark for evaluating the performance of our proposed method, ensuring that it meets the high standards required for practical applications in computational photography and computer vision.

2.5 Self-attention in low-level vision

Computer vision has witnessed remarkable progress in recent years, largely driven by the success of deep learning models, particularly convolutional neural networks (CNNs). However, CNNs have an inherent limitation in capturing long-range dependencies due to their localized receptive fields. Self-attention mechanisms have been introduced to address this issue, allowing for modeling long-range interactions between different regions of an image.

One of the pioneering works that incorporated self-attention into computer vision tasks is the Non-Local Neural Networks [59]. This work introduced a non-local operation that captures long-range dependencies by computing the response at a position as a weighted sum of the responses from all positions. This approach improved performance on various tasks, such as object detection, instance segmentation, and video classification.

Building upon the success of self-attention in natural language processing (NLP), the Vision Transformer (ViT) [19] was proposed, which applies the transformer architecture [53] directly to image data. ViT splits an image into patches and treats them as tokens, en-

abling the self-attention mechanism to capture global interactions between these patches. Despite its simplicity, ViT achieved competitive or superior performance compared to state-of-the-art CNNs on various computer vision benchmarks.

Subsequent works have explored various extensions and improvements to the ViT architecture. For example, the Swin Transformer [43] introduces a hierarchical transformer that combines self-attention with shifted windowing to capture both global and local dependencies in an efficient manner. The Pyramid Vision Transformer (PVT) [58] employs a pyramidal structure to capture multi-scale features, while the Twins [18] model proposes a spatially separating self-attention mechanism to improve computational efficiency.

Previous works have successfully employed self-attention in various low-level vision tasks. For instance, SwinIR [42] used self-attention mechanisms for image super-resolution, demonstrating significant improvements in capturing fine details and texture information. Similarly, Uformer [60] applied self-attention to image denoising, showing that it helps in preserving important structures while removing noise. Moreover, Restormer [62] utilized self-attention for image restoration, allowing the model to better understand the context of the corrupted regions and generate more coherent results.

The self-attention has emerged as a powerful mechanism for capturing long-range dependencies in computer vision tasks, enabling significant performance improvements across various domains. As research continues, we can expect further advancements in self-attention-based models, leading to more efficient and interpretable computer vision systems.

2.6 Summary

In summary, the existing research on white-balance correction methods has evolved from basic statistical approaches to sophisticated deep learning techniques. Conventional methods, while foundational, often fall short in multi-illuminant scenarios. Recent advances leverage complex datasets and self-attention mechanisms to enhance accuracy and adaptability. Our proposed method builds on these advancements by incorporating an attention-based fusion approach, which dynamically integrates multiple white-balance settings. This strategy addresses the limitations of earlier methods, offering a more robust solution for scenes with diverse lighting conditions. Through comprehensive evaluation, we aim to demonstrate the effectiveness of this innovative approach in improving color accuracy and overall image quality in complex environments.



Figure 2.1: Examples from the Gehler dataset [26]. The images are linear RGB, are single illuminant, and are mostly indoor images hence lacking a variety.

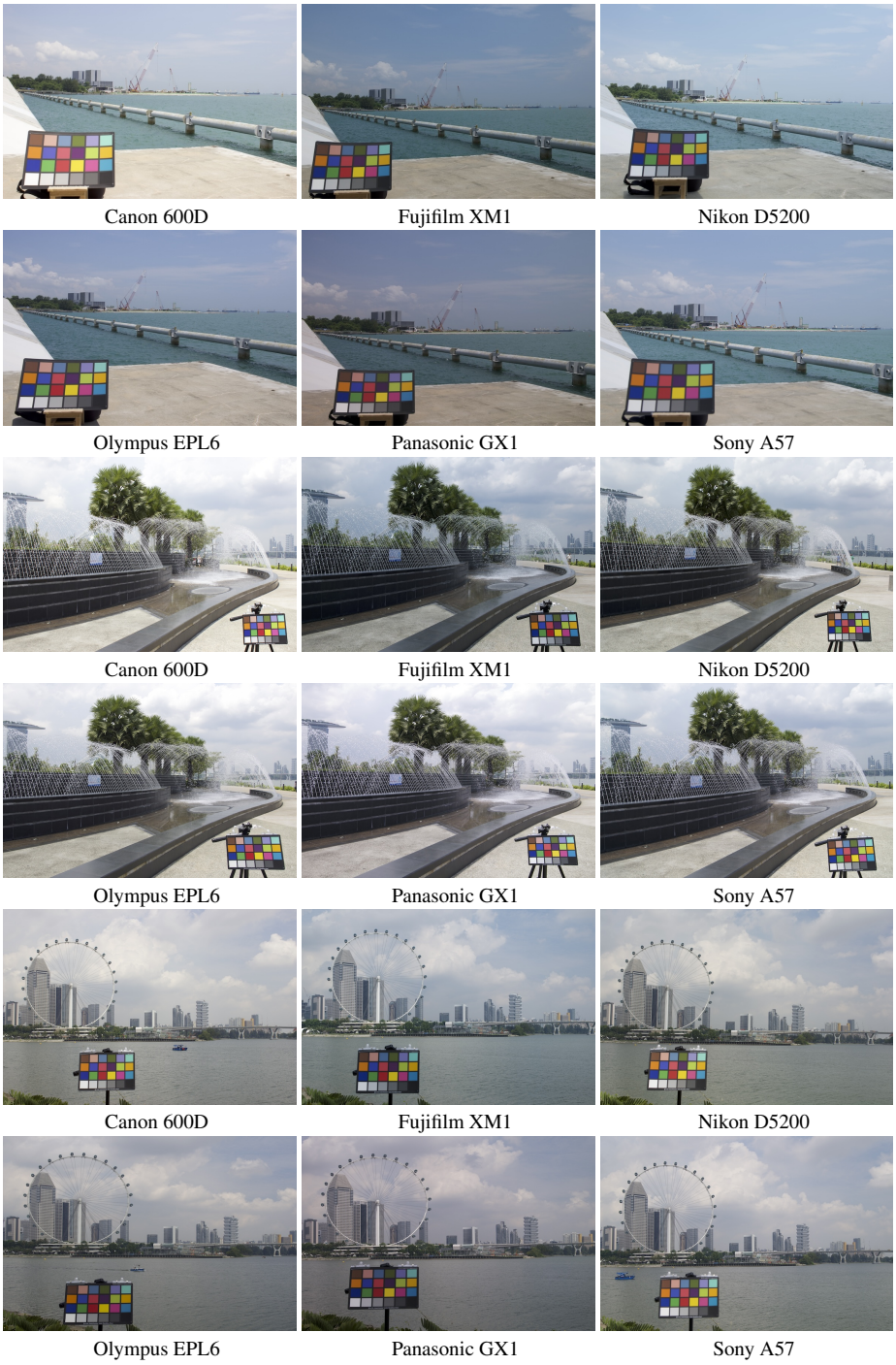


Figure 2.2: Examples from the NUS dataset [17]. The same scene is captured from different cameras.



Figure 2.3: Examples from the Afifi dataset [8]. There are 30 synthetically generated images of indoor scenes.

Chapter 3

Multi-Illuminant sRGB Dataset

We introduce a new dataset to address the task of fusion-based multi-illumination correction. The key objectives in creating this dataset are: (i) sRGB-rendered inputs for different camera white-balance settings: The dataset includes sRGB-rendered input images captured under various camera white-balance (WB) settings. This allows us to study the impact of different WB configurations on the final image quality and develop robust correction techniques. (ii) Corresponding white-balanced Ground Truth sRGB Images: For each input image with varying WB settings, the dataset also provides the corresponding ground truth sRGB image that has been properly white-balanced. This ground truth serves as the reference for evaluating the performance of fusion-based multi-illumination correction algorithms.

Our dataset builds upon the recently introduced Large-Scale Multi-Illuminant (LSMI) dataset by Kim et al. [36]. Our dataset leverages the LSMI dataset as a foundation, but extends it by focusing on the fusion-based multi-illumination correction task. Instead

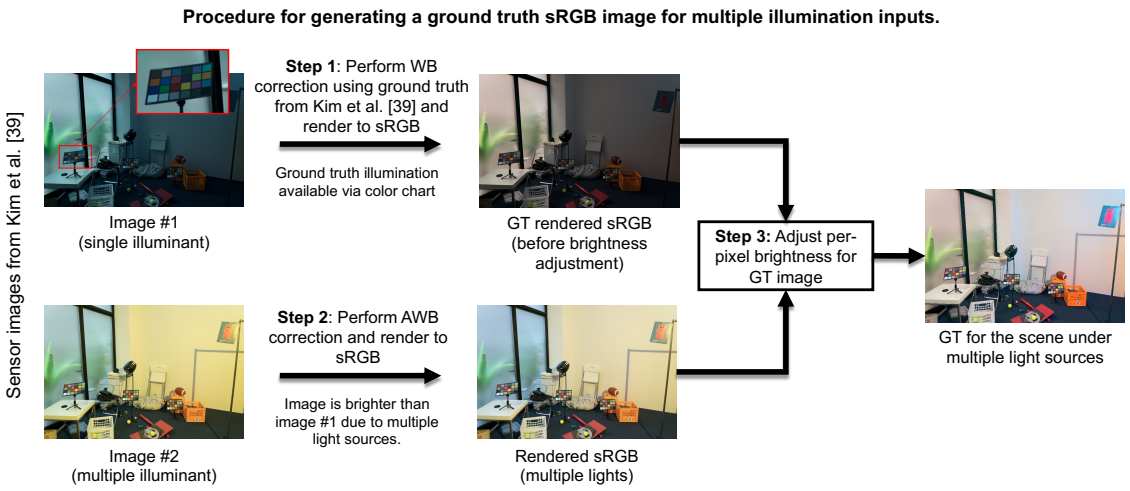


Figure 3.1: Dataset preparation. Our starting point is the dataset of Kim et al. [36]. Their approach starts by capturing a scene under a single illumination (typically daylight). Additional light sources are then added, one by one, to produce a multi-illuminant dataset. Note that their dataset focused on unprocessed RAW images. We use this initial single-illuminant image to produce our ground truth image, as shown at the top. This image is white-balanced using the ground truth illuminant value from the Macbeth color checker chart placed in the scene. However, this ground truth image is darker than images of the scene captured using multiple lights. To match the brightness level of other input images of the same scene but with more light sources, we adjust the per-pixel brightness of the ground truth image to match the multi-illuminant input (shown at the bottom).

of working with RAW images, we generate sRGB-rendered inputs with varying white-balance settings and their corresponding ground truth sRGB images. This allows us to explore more advanced fusion-based techniques that effectively combine information from multiple illumination sources to produce high-quality, color-corrected images. By building upon the LSMI dataset and expanding it to the fusion-based multi-illumination correction domain, we aim to provide a comprehensive resource for researchers and practitioners working in this field of computational photography.

We repurpose the Sony and Nikon camera data from the LSMI dataset in our work.

The LSMI dataset contains 955 scenes for the Nikon camera and 1317 scenes for the Sony camera, with each scene having two or three illuminants. To create our fusion-based multi-illumination correction dataset, we compute five sRGB-rendered images for each of the RAW DNG images in the LSMI dataset. These five images are generated using the five camera white-balance presets available in Adobe Camera RAW: daylight, cloudy, incandescent, shade, and fluorescent. It is important to note that we do not consider the Galaxy camera data from the LSMI dataset, as it cannot be handled by the Adobe Camera RAW [1] software used in our image rendering process.

One of the key aspects of our dataset generation process is the method we used to find the ground truth sRGB image for scenes with multiple light sources, as shown in Fig. 3.1. We leveraged the unique properties of the original LSMI dataset to derive these ground truth references. The LSMI dataset was carefully curated to include, for each scene, an image captured under a single illuminant. This single illuminant image is a crucial reference point in our ground truth generation process. Since all the images in the LSMI dataset were captured using a fixed exposure, the single illuminant image is inherently darker than those with multiple illuminations. This is because the single illuminant provides less overall light in the scene compared to the multiple illuminant scenarios.

Our method generates the ground truth as follows: First, identify the single illuminant image in each scene in the LSMI dataset. By taking advantage of the single illuminant images in the LSMI dataset, we are able to generate high-quality ground-truth sRGB references for scenes with multiple light sources. This is a crucial step in creating our fusion-based multi-illumination correction dataset, as it provides the necessary ground truth for evaluating the performance of our algorithms.

However, the initial ground truth image derived from the single illuminant reference is typically darker than the corresponding image of the same scene with multiple light sources. To address this brightness discrepancy, we employ a two-step process: (i) sRGB Rendering of the Multi-Illuminant Image: We render the multi-illuminant image to sRGB using a standard automatic white-balance (AWB) procedure. This results in an incorrectly white-balanced image due to the multiple illuminations. (ii) Per-Pixel Brightness Normalization: We then use the incorrectly white-balanced sRGB image from step (i) to perform per-pixel brightness normalization on the ground truth image. This allows us to produce a properly white-balanced sRGB version of the scene that matches the brightness of the multi-illuminant sRGB input.

By applying this brightness normalization step, we ensure that the ground truth images in our dataset accurately represent the scene under multiple illuminations, both in terms of color and brightness. The final dataset for the Nikon and Sony camera sets contains a total of 16,284 sRGB images, comprising the sRGB-rendered inputs with varying white-balance settings and their corresponding ground truth sRGB images.

We have included example images from our dataset in Fig. 3.2 to illustrate the diversity and characteristics of the data. The top five images correspond to the Nikon dataset, while the five bottom ones are from the Sony dataset. These examples showcase the different white-balance presets applied to the sRGB-rendered inputs, including Tungsten, Fluorescent, Cloudy, and Shade. As can be observed, the Tungsten and Fluorescent presets result in images with a bluish tone, as they correct the image to compensate for the reddish illuminant in the scene. Conversely, the Cloudy and Shade presets produce images with a more yellowish tone, as they correct the image towards a bluish illuminant.

3.1 Limitations

While our dataset provides a valuable resource for studying white-balance and illumination estimation, there are certain limitations to consider:

1. **Controlled Environment**: The dataset was captured in a controlled studio environment with carefully set lighting conditions. This may not fully represent the complexity and variability of illumination encountered in real-world, uncontrolled imaging scenarios.

2. **Limited Illuminant Diversity**: Although the dataset includes scenes with multiple illuminants, the number of distinct illuminant types is relatively small (daylight, cloudy, incandescent, shade, fluorescent). Many atypical illuminants found in specialized domains (e.g., stage lighting, underwater, etc.) are not represented.

3. **Fixed Exposures**: All images in the dataset are captured using a fixed exposure setting. This does not account for the varying exposure conditions that photographers often face, which can impact white-balance estimation.

4. **Camera Limitations**: The dataset only includes images from a limited set of camera models (Nikon, Sony). While our method demonstrates cross-camera generalization, the dataset may not fully capture the diversity of imaging pipelines and sensor characteristics across all camera manufacturers.

5. **Synthetic Ground Truth**: The ground truth images are generated synthetically by rendering RAW images with single and multi-illuminant. While care was taken to

ensure accuracy, this process may introduce subtle unwanted artifacts or deviations from true real-world ground truth data.

6. **Dataset Size**: While the LSMI dataset is large-scale, it may still be limited in its coverage of all possible scene compositions, materials, and illumination combinations that exist in the physical world.

3.2 Summary

In summary, our newly introduced multi-illuminant sRGB dataset provides a valuable resource for developing and testing white-balance correction methods. By repurposing the LSMI dataset and generating sRGB-rendered inputs with various white-balance settings, we offer a comprehensive and diverse set of training and evaluation images. This dataset allows for more effective training of fusion-based models, enabling better handling of complex lighting scenarios. The examples included demonstrate the range and quality of the data, highlighting its potential to advance the field of multi-illuminant white-balance correction. Through this dataset, we aim to support the development of more accurate and reliable color correction techniques in real-world applications.

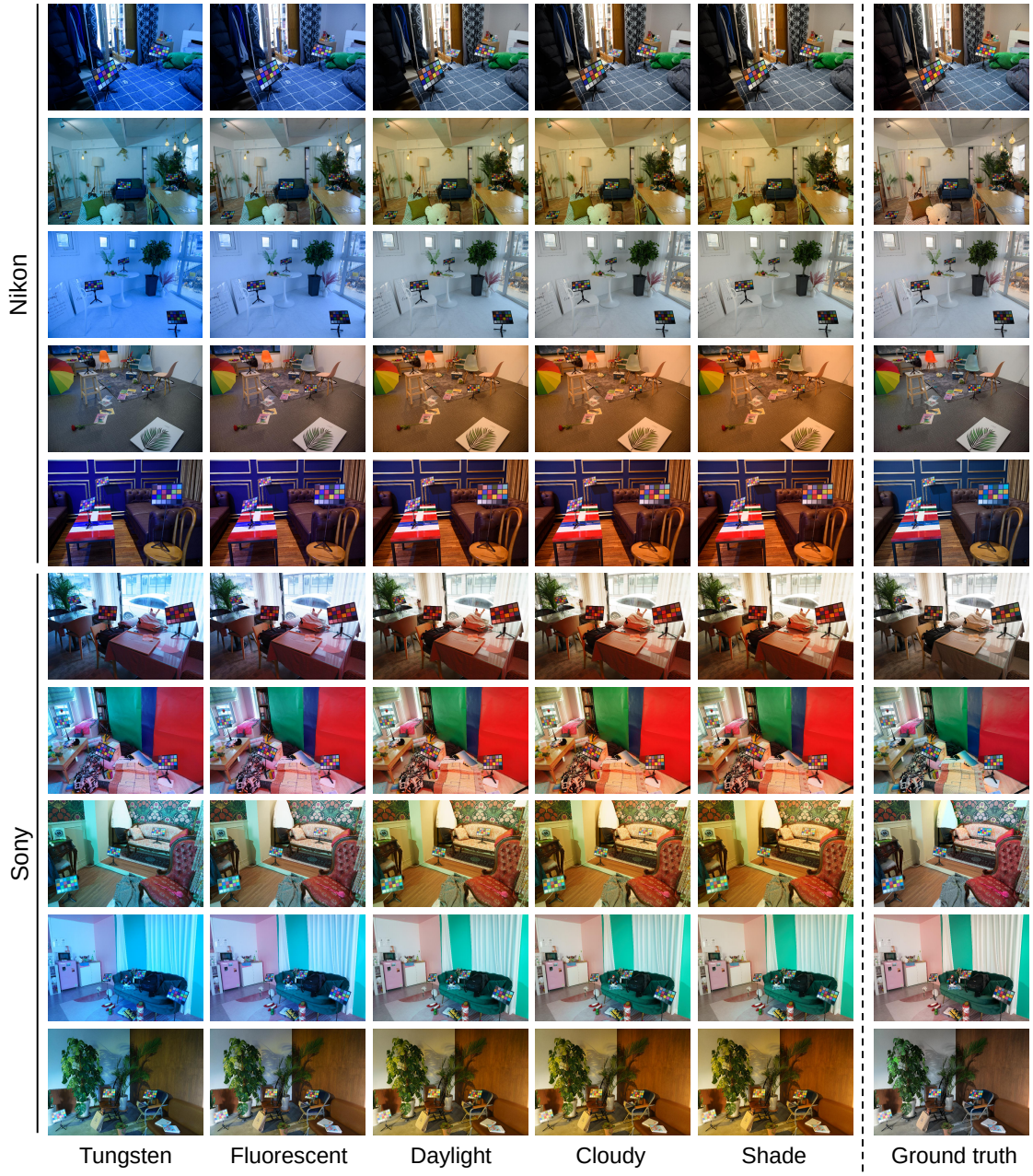


Figure 3.2: Examples of our dataset using images from the Nikon and Sony cameras. Tungsten and Fluorescent preset suppose the light in the room is reddish, and therefore, they correct the image to a more bluish tone. Conversely, cloudy and shaded, suppose the illuminant in the scene is bluish, correcting the image to a more yellowish tone. The last column presents the image of the ground truth.

Chapter 4

Method

In this chapter, we present the methodology developed to address the complex task of multi-illuminant white-balance correction. Building on the foundational work of Afifi et al. [8], we introduce a novel attention-based fusion approach that leverages recent advancements in deep learning to improve the accuracy and robustness of white-balance correction in scenes illuminated by multiple light sources. We begin by discussing the foundational steps, including the rendering of RAW images to various white-balance settings and the essential calibration processes. Following this, we introduce our neural network-based fusion mechanism, which dynamically blends these settings to achieve optimal white-balance. Additionally, we outline the dataset employed for training our model and compare our method against established baselines. Through detailed analysis and evaluation, this chapter showcases the significant advancements our method brings to the field, underscoring its effectiveness and applicability in real-world scenarios of computational photography and computer vision.

4.1 Baseline fusion-based methods

We begin by reviewing the method proposed by Afifi et al. [8]. This approach leverages a fusion-based technique to address the problem of multi-illumination correction. The key steps in Afifi et al.'s method are as follows:

- **RAW Sensor Image Rendering**: A RAW sensor image is rendered to three predefined white-balance (WB) settings: daylight (D), shade (S), and tungsten (T) (also known as incandescent). This step involves using the camera's internal settings to create multiple versions of the same image, each adjusted for different lighting conditions. Rendering the RAW sensor image with these settings ensures that the intrinsic color information is preserved, allowing for more accurate white-balance correction later in the process.

- **WB Setting Calibration**: The sensor responses to these three illumination settings are calibrated in the factory and can be found by setting the camera in manual mode, capturing an image, and examining the image's metadata. This calibration process is crucial because it provides a reference point for the camera's interpretation of color under different lighting conditions. By understanding how the camera responds to various light sources, the algorithm can more accurately predict and correct for the colors in the final image.

- **sRGB Rendering**: The three images corresponding to the D, S, and T WB settings are then rendered to sRGB using the respective white-balance settings. Converting the images to sRGB, the standard color space for digital images, allows for consistent and comparable results across different devices and applications. This step

involves applying a series of color transformations and corrections to ensure that the rendered images closely match what the human eye would perceive under the given lighting conditions.

- **Neural Network-based Fusion**: A neural network is used to receive the three sRGB-rendered images and output a per-pixel weight map to blend the images. The weight map represents the relative contribution of each white-balance setting (daylight, shade, and tungsten) at every pixel in the image. The neural network is trained to understand the subtle variations in color and lighting across the different WB settings and to combine them in a way that produces a uniformly white-balanced image. This fusion process is dynamic and context-aware, leveraging the strengths of each rendered image to correct for the color casts introduced by multiple light sources.

For training the neural network, Afifi et al. relied on the Rendered-WB dataset [6], which contains approximately 65,000 sRGB images rendered with various color temperatures. From this dataset, they selected a subset of 9,200 training images that were rendered using the "camera standard" photofinishing technique. These 9,200 training images encompass a range of color temperatures, including 2850 Kelvin (K), 3800 K, 5500 K, 6500 K, and 7500 K, which correspond to the five white-balance presets: tungsten (T), fluorescent (F), daylight (D), cloudy (C), and shade (S). While Afifi et al. initially tried to fuse all five images, they empirically found that their network performed best when the fusion was applied to only three images: the D, S, and T presets.

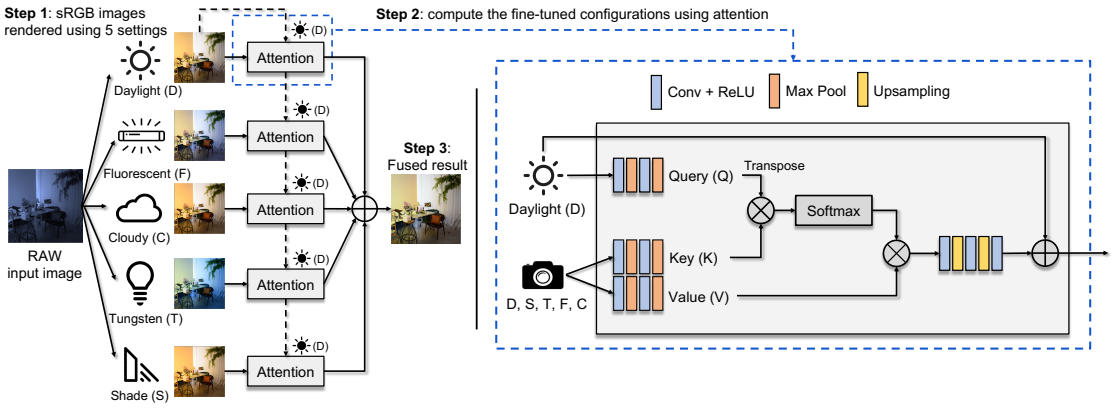


Figure 4.1: Architecture of attention-based white-balance fusion method. Input RAW image is rendered using the five predefined WB settings. Each sRGB image is fed in an attention-based model. The sRGB images provide the *key* and *value*, while the *daylight* image provides us with the *query*. We fuse the outputs of the five attention modules.

4.2 Attention-based WB fusion

As previously mentioned, Afifi et al. [8] found that their model produced the best results when the fusion was limited to only three input images. We revisit the fusion problem, but in contrast, we explore the fusion of all five pre-defined white-balance inputs. Inspired by existing attention-based approaches [40, 19, 63], we leverage an attention-based architecture to compute the weighting among the camera presets adaptively.

We employ an attention module for every distinct rendered sRGB image. Each sRGB image provides us with both the 'key' and 'value' inputs, while the 'daylight' image serves as the 'query' to create a globally adjusted white-balance setting. The final result is then achieved by fusing these fine-tuned configurations. Figure 4.1 provides an overview of this approach.

Specifically, the 'daylight' image and the other four white-balance rendered images

are fed into convolutional and max-pooling layers to transform them into a reduced feature space. We employ two convolutional layers to obtain 16 convolutional feature maps and two max-pooling layers to enlarge the receptive field and reduce the computational burden. In particular, we obtain $Q \in \mathbb{R}^{h/16 \times w/16 \times 16}$ from the 'daylight' image and $K \in \mathbb{R}^{h/16 \times w/16 \times 16}$ and $V \in \mathbb{R}^{h/16 \times w/16 \times 16}$ from the input rendered white-balance setting.

We use a single-head attention module to compute the attention map and model their spatial dependencies. The attention map captures the relationships between the 'query' (daylight image) and the 'key' (rendered white-balance setting), enabling the model to selectively attend to relevant regions and features for effective fusion.

Next, the output of the attention stage is processed by a sequence of convolutional layers with interspersed upsampling layers to recover the original spatial resolution. The upsampled feature maps are then additively combined with the initial daylight input through a residual connection. This residual connection facilitates the propagation of low-level details and helps preserve important image structures during the fusion process.

Finally, the white-balanced output image is obtained via a pixel-wise average among the outputs of the previous attention modules. This averaging operation effectively fuses the information from the multiple attention modules, each of which has learned to attend to different aspects of the input images and their illumination characteristics.

By leveraging an attention-based architecture, our approach can adaptively weight and combine the information from the different white-balance rendered inputs, allowing it to selectively attend to the most relevant regions and features for accurate white-balance correction. The attention mechanism enables the model to capture long-range dependencies and spatial relationships, enhancing its ability to fuse the complementary information

present in the various input images.

Moreover, our approach considers all five pre-defined white-balance inputs, in contrast to the approach of Afifi et al. [8], which was limited to three inputs. By utilizing a more comprehensive set of input images with diverse illumination conditions, our model can potentially capture a wider range of illumination cues and produce more accurate white-balance estimates, particularly in complex mixed-illumination scenarios.

The combination of the attention-based fusion strategy and the utilization of all available white-balance rendered inputs enables our approach to effectively leverage the complementary information present in the different input images, leading to improved white-balance correction performance, as demonstrated by our quantitative results.

4.3 Summary

In summary, this chapter describes a novel attention-based fusion approach for multi-illuminant white-balance correction, addressing the limitations of previous methods by leveraging the power of self-attention mechanisms and exploiting the complementary information present in a comprehensive set of white-balance rendered input images.

Chapter 5

Experiments

In this chapter, we present a comprehensive evaluation of our proposed attention-based fusion method for white-balance correction. We begin by detailing our experimental setup, including the dataset configurations and evaluation metrics employed. Following this, we introduce our neural network-based fusion mechanism and describe the training process. The chapter then presents a thorough comparison of our method against several state-of-the-art baselines, showcasing quantitative results that demonstrate the superior performance of our approach across different camera models and metrics. Additionally, we provide qualitative results with visual comparisons and an analysis of the model's attention maps. The chapter further explores the cross-camera generalization capabilities of our method and its performance on a synthetic multi-illuminant dataset. Through this detailed analysis and evaluation, we underscore the significant advancements our method brings to the field, highlighting its effectiveness and applicability in real-world scenarios of computational photography and computer vision.

5.1 Datasets

We evaluate the performance of our proposed attention-based fusion method on our introduced dataset described in Sec. 3. We conduct extensive quantitative experiments on three different dataset configurations: Sony, Nikon, and a combination of both Sony and Nikon datasets. These datasets comprise images captured by Sony and Nikon cameras under various illumination conditions, providing a diverse and challenging testbed for our method. The quantitative results for these three dataset configurations are presented in Table 5.1.

We use the same splits as in the original LSMI dataset. Therefore, the images for the Sony and Nikon cameras are organized into training (70%), test (10%), and validation (20%) splits. Specifically, the Sony split consists of 1175 training, 335 validation, and 171 test images, while the Nikon split consists of 722 training, 204 validation, and 107 test images.

5.2 Metrics

We quantitatively evaluate the accuracy of our proposed white-balance correction method using three well-established evaluation metrics: mean squared error (MSE), mean angular error (MAE), and ΔE2000. These metrics are widely used in the field of color imaging and provide a comprehensive assessment of the performance of our method.

The mean squared error (MSE) is a widely used metric that measures the cumulative squared error between the estimated and the ground truth values. It is computed as the squared difference between the estimated and ground truth values, averaged over all the pixels in the image:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{5.1}$$

where: n is the number of observations, y_i is the actual value, $\hat{y}_i$ is the predicted value. A lower MSE value indicates a better performance of the white-balance correction method.

The mean angular error (MAE) is a metric specifically designed for evaluating color constancy algorithms. It measures the angular difference between the estimated and ground truth illumination vectors, averaged over all the pixels in the image:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\cos^{-1}\left(\frac{y_i \cdot \hat{y}_i}{\|y_i\|\|\hat{y}_i\|}\right) \tag{5.2}$$

where: n is the number of observations, y_i and $\hat{y}_i$ are the actual and predicted vectors, $\cdot$ represents the dot product between vectors, $\|y_i\|$ and $\|\hat{y}_i\|$ are the magnitudes of the vectors, $\cos^{-1}$ is the inverse cosine function, which computes the angle between the vectors.

The MAE is particularly useful in assessing the accuracy of the estimated illumination color, which is crucial for effective white-balance correction.

The ΔE2000 metric is a perceptually uniform color difference metric that takes into account the non-linear response of the human visual system to color differences. It provides a more accurate measure of the perceived color difference between the corrected and ground truth images. It is defined as:

$$\Delta E_{00} = \sqrt{\left(\frac{\Delta L'}{k_L S_L}\right)^2 + \left(\frac{\Delta C'}{k_C S_C}\right)^2 + \left(\frac{\Delta H'}{k_H S_H}\right)^2 + R_T\left(\frac{\Delta C'}{k_C S_C}\right)\left(\frac{\Delta H'}{k_H S_H}\right)}, \tag{5.3}$$

where: $\Delta L'$, $\Delta C'$, and $\Delta H'$ are the differences in lightness, chroma, and hue, respectively, S_L, S_C, and S_H are the weighting functions, and k_L, k_C, and k_H are the parametric factors (usually set to 1). R_T is a rotation term.

A lower ΔE2000 value indicates a better preservation of the original image colors after white-balance correction.

For each of these metrics, we compute the mean, median, and trimean values. The mean provides an overall measure of the central tendency, while the median and trimean are robust measures that are less influenced by outliers. By reporting all three central tendency measures, we provide a comprehensive picture of the performance distribution across the dataset. The trimean combines elements of both the median and the mean, and is calculated as follows:

$$\text{Trimean} = \frac{Q_1 + 2 \cdot Q_2 + Q_3}{4} \tag{5.4}$$

where: Q_1 is the first quartile (25th percentile), Q_2 is the median (50th percentile), Q_3 is the third quartile (75th percentile).

To ensure a fair and rigorous evaluation, we do not utilize the color charts for training, validation, and testing. The color charts are masked to prevent any potential biases or leakage of information to our model. This approach follows best practices similar to previous white-balance methods and ensures that our results are reliable and generalizable.

Table 5.1: Quantitative evaluation for our model. We compare our approaches against a well-known single image white-balance method [32], a recent multi-illuminant white-balance method, SWBNet [41], and the results of the Afifi et al. [8] model. Best results are marked in **bold** for each metric.

		ΔE2000 ↓			MSE ↓			MAE ↓		
	Method	Mean	Median	Trimean	Mean	Median	Trimean	Mean	Median	Trimean
Sony	FC4 [32]	12.07	10.93	11.23	501.12	272.77	319.01	9.88º	8.15º	8.71º
	SWBNet [41]	10.86	10.75	10.67	474.23	268.56	307.11	8.37º	7.71º	7.84º
	DeepWB [4]	8.53	8.14	8.18	217.11	160.53	174.26	6.80º	6.41º	6.47º
	Afifi et al. (DST) [8]	9.71	9.08	9.11	253.67	202.98	209.83	7.20º	6.78º	6.83º
	Afifi et al. (DSTFC) [8]	8.90	8.21	8.27	211.88	153.97	165.03	6.21º	5.74º	5.82º
	Afifi et al. (best: DFC) [8]	9.18	8.80	8.86	227.58	165.76	182.29	6.52º	5.77º	5.96º
	Afifi et al. "Oracle" [8]	8.56	8.42	8.49	208.83	150.67	178.90	6.06º	5.72º	5.81º
	Ours (DST)	8.32	8.12	8.12	143.90	101.15	107.60	6.02º	5.91º	6.03º
	Ours (DSTFC)	**7.24**	**6.90**	**7.01**	**124.85**	**88.27**	**90.56**	**5.09º**	**4.87º**	**4.81º**
Nikon	FC4 [32]	11.51	10.36	10.61	449.86	240.18	264.76	10.57º	8.02º	8.80º
	SWBNet [41]	10.57	9.95	10.18	507.84	228.66	298.26	7.52º	6.89º	7.01º
	DeepWB [4]	8.76	8.20	8.23	211.70	145.54	164.39	6.20º	5.81º	5.92º
	Afifi et al. (DST) [8]	10.10	9.09	9.35	268.02	169.37	191.43	7.06º	5.99º	6.22º
	Afifi et al. (DSTFC) [8]	9.85	8.63	9.16	257.65	153.91	176.01	7.02º	5.86º	6.16º
	Afifi et al. (best: DST) [8]	10.10	9.09	9.35	268.02	169.37	191.43	7.06º	5.99º	6.22º
	Afifi et al. "Oracle" [8]	9.76	9.43	9.54	249.52	150.12	168.33	7.00º	5.77º	6.08º
	Ours (DST)	8.43	7.80	7.96	148.71	104.57	109.08	6.55º	5.97º	6.02º
	Ours (DSTFC)	**6.71**	**6.44**	**6.45**	**78.12**	**53.52**	**61.52**	**4.32º**	**3.90º**	**3.92º**
Sony + Nikon	FC4 [32]	11.85	10.68	11.04	481.61	253.23	292.32	10.14º	8.07º	8.73º
	SWBNet [41]	10.75	10.45	10.51	487.02	253.18	303.91	8.04º	7.30º	7.45º
	DeepWB [4]	8.56	8.04	8.12	215.02	154.76	170.46	6.57º	6.18º	6.26º
	Afifi et al. (DST) [8]	9.66	9.11	9.15	231.51	177.71	187.40	6.79º	6.05º	6.22º
	Afifi et al. (DSTFC) [8]	9.31	8.45	8.65	219.51	148.71	162.40	6.36º	5.60º	5.72º
	Afifi et al. (best: DSF) [8]	9.53	8.70	8.87	249.51	166.71	181.40	6.79º	5.80	6.00
	Afifi et al. "Oracle" [8]	8.93	8.34	8.51	202.51	143.71	160.40	6.16º	5.89	5.97
	Ours (DST)	8.09	7.64	7.66	140.02	111.52	115.99	7.77º	5.49º	5.56º
	Ours (DSTFC)	**7.62**	**7.39**	**7.45**	**121.53**	**90.65**	**92.52**	**5.18º**	**4.78º**	**4.36º**

5.3 Training details

For our baseline comparisons with the work of Afifi et al. [8], we employ the same deep neural network (DNN) architecture and train all models using the same loss function, which is a combination of mean squared error (MSE) loss and Sobel loss. We follow the same training procedure as Afifi et al.[8].

Table 5.2: Generalization of our methods. We present the results for all the metrics when training our models in one camera and testing it on the other camera. Best results are in bold.

			ΔE2000 ↓			MSE ↓			MAE ↓		
Train	Test	Model	Mean	Median	Trimean	Mean	Median	Trimean	Mean	Median	Trimean
Nikon	Sony	Afifi et al. (DST) [8]	9.46	8.86	8.91	241.71	199.76	200.39	6.72º	6.23º	6.32º
		Afifi et al. (DSTFC) [8]	9.17	8.60	8.53	246.42	207.80	203.91	6.89º	6.27º	6.45º
		Afifi et al. "Oracle" [8]	9.03	8.57	8.67	221.67	180.45	184.67	6.61º	6.11º	6.23º
		Ours (DSTFC)	**8.48**	**8.10**	**8.20**	**198.15**	**133.93**	**143.88**	**5.85º**	**5.31º**	**5.43º**
Sony	Nikon	Afifi et al. (DST) [8]	10.23	9.54	9.67	289.91	234.54	254.0	7.22º	6.31º	6.42º
		Afifi et al. (DSTFC) [8]	9.96	9.70	9.76	289.91	234.54	254.0	7.74º	7.34º	7.54º
		Afifi et al. "Oracle" [8]	9.84	9.59	9.62	261.69	154.21	169.17	7.16º	5.91º	5.98º
		Ours (DSTFC)	**8.31**	**7.69**	**7.96**	**129.96**	**92.31**	**100.50**	**5.65º**	**5.05º**	**5.05º**

Table 5.3: Results on the Synthetic Multi-Illuminant Dataset from Afifi et al. [8]. Best results are in bold.

	ΔE2000 ↓			MSE ↓			MAE ↓		
Method	Mean	Median	Trimean	Mean	Median	Trimean	Mean	Median	Trimean
DeepWB [4]	10.93	9.82	10.04	1130.60	886.32	917.09	5.53º	5.55º	5.19º
Afifi et al. (DST) [8]	10.40	10.07	10.29	438.47	408.76	411.63	5.13º	5.02º	5.06º
Afifi et al. (DSTFC) [8]	10.43	10.35	10.48	386.39	357.92	363.57	**4.75º**	4.56º	4.69º
Afifi et al. "Oracle" [8]	10.26	9.89	9.97	336.57	301.94	313.82	4.96º	4.81º	4.88º
Ours (DSTFC)	**9.17**	**8.27**	**8.71**	**188.08**	**169.41**	**168.06**	5.10º	**4.17º**	**4.44º**

We conduct experiments with two predefined sets of white-balance (WB) settings: (i) 5500 K, 7500 K, 2850 K, representing daylight, shade, and tungsten illumination (DST), respectively, and (ii) 5500 K, 7500 K, 2850 K, 3800 K, 6500 K, representing daylight, shade, tungsten, fluorescent, and cloudy illumination (DSTFC), respectively.

For our attention-based fusion model, we train the model using an Adam optimizer and a cosine annealing learning rate scheduler. The initial learning rate is set to $1e$-3, and the final learning rate is $1e$-5. We select the model checkpoint that achieves the best ΔE2000 metric on the validation set.

5.4 Baselines

To comprehensively evaluate the performance of our attention-based fusion model, we compare it against several well-established and state-of-the-art methods for white-balance correction. By including these diverse set of methods, we aim to provide a comprehensive comparison that covers both single-illumination and multi-illuminant approaches, as well as traditional learning-based methods and the most recent state-of-the-art techniques. This rigorous comparative analysis will allow us to thoroughly assess the strengths and limitations of our proposed attention-based fusion model and contextualize its performance within the broader landscape of white-balance correction methods. Specifically, we include the following methods in our comparative analysis:

FC4 [32]: The FC4 method employed deeper network architectures based on AlexNet [38] and SqueezeNet [34] pre-trained models, with the SqueezeNet variant having a model size of 5.89 MB. It was trained on relatively small real-world datasets, specifically the reprocessed Color Checker Dataset [26] (568 images) and the NUS 8-Camera Dataset [17] (1736 images).

SWBNet [41]: The SWBNet approach trained exclusively on 12,000 synthetic images from the Rendered WB dataset.

Deep WB [4]: The Deep WB method was trained similarly to MixedWB, using 12,000 image patches from two partitions of the Rendered WB dataset.

Afifi et al. [8]: Afifi et al MixedWB method trained on 9,200 images from the Rendered WB dataset, with the images rendered across various color temperatures. With a model size in the range of 5.1-5.9 MB, MixedWB was designed to be efficiently deployable on devices with limited computational resources.

Variants on Afifi et al. [8] fusion models. We extend Afifi et al. [8] to work with all illuminant combinations that contain the *daylight* WB setting. These are six models in total, denoted as: DST, DSF, DSC, DTF, DTC, DFC. We did not include the cases where *daylight* is not included as we found that it is important that this setting always is present as one of the inputs to produce good results, given it has the median temperature (5500 K) among the presets. Given all these combinations, we report the best-fixed combination of illuminants providing the best results in each specific scenario and the results for an "Oracle" version of the method. This "Oracle" selects the illuminant combination providing the best results for each specific image, therefore setting the best possible result that can be obtained using Afifi et al.'s [8] post-WB approach.

5.5 Results

Our method was trained on the Rendered WB dataset in a similar fashion to Afifi et al., but achieves significantly better performance on the LSMI dataset. Specifically, using its top-performing DSTFC (daylight, shade, tungsten, fluorescent, and cloudy illumination) configuration, it obtained a mean ΔE2000 of 7.24 on the Sony subset, with a mean MSE of 124.85 and mean MAE of 5.09° as shown in Table 5.1. On the Nikon subset, the corresponding numbers were 6.71, 78.12, and 4.32°, respectively. These are by far the lowest errors achieved by any method on this challenging real-world dataset. The outstanding performance of the proposed approach, despite being trained only on the synthetic Rendered WB data, signifies its excellent ability to generalize learnt features to distributions divergent from the training data. The strength of the DSTFC configuration suggests that

leveraging different model components and combining their strengths is highly beneficial. However, the relatively high errors even for this top-performing method underscore the need for more diverse, real-world training data to further elevate generalization capabilities.

In comparison, the baseline methods showed the following performance: FC4[32] achieved a mean ΔE2000 error of 12.07 on the Sony subset and 11.51 on the Nikon subset. The mean MSE was 501.12 and 449.86, while the mean MAE was 9.88° and 10.57° on the Sony and Nikon subsets respectively. SWBNet[41] achieved a mean ΔE2000 of 10.86, mean MSE of 474.23, and mean MAE of 8.37° on the Sony subset. On the Nikon subset, the corresponding numbers were 10.57, 507.84, and 7.52°. Deep WB[4] achieved a mean ΔE2000 of 8.53, mean MSE of 217.11, and mean MAE of 6.80° on the Sony subset. On the Nikon subset, the corresponding numbers were 8.76, 211.70, and 6.20°. Afifi et al. [8] best-performing DFC variant achieved a mean ΔE2000 of 9.18 on the Sony subset, with a mean MSE of 227.58 and mean MAE of 6.52°. On the Nikon subset, the DFC variant's performance was a mean ΔE2000 of 10.10, mean MSE of 268.02, and mean MAE of 7.06°. Our approach outperforms all the other methods included in the comparison for all three dataset configurations and across all evaluation metrics. The superior performance is consistent across the mean, median, and trimean values, indicating the robustness and stability of our method.

Our approach outperforms all the other methods included in the comparison for all three dataset configurations and across all evaluation metrics. The superior performance is consistent across the mean, median, and trimean values, indicating the robustness and stability of our method.

Notably, when considering the combined dataset consisting of both Sony and Nikon images, our method maintains its advantage over the other approaches. This result demonstrates the generalizability of our attention-based fusion model, as it can effectively handle images from different camera sources and characteristics, without requiring any specific adaptation or tuning.

A particularly remarkable observation is that our method outperforms even the 'Oracle' version of Afifi et al.'s method [8], which represents an upper bound on their performance. This superior performance highlights the effectiveness of our attention-based fusion strategy in blending multiple input images for white-balance correction. By learning to selectively attend to and fuse the most relevant information from the input images, our model can better capture the underlying illumination characteristics and produce more accurate white-balanced outputs.

These quantitative results provide strong empirical evidence for the efficacy of our proposed approach, showcasing its ability to outperform state-of-the-art methods for white-balance correction across diverse datasets and camera configurations. The consistent and significant performance gains demonstrate the potential of our attention-based fusion model as a powerful solution for the challenging problem of multi-illuminant white-balance correction.

We show qualitative results in Fig. 5.1 and 5.2 for the Sony and Nikon camera, respectively. From left to right we show SWBNet [41], FC4 [32], DeepWB [4], Afifi et al. [8], and our method. The last column shows the ground truth image. Each image shows the ΔE2000 error in its bottom-right corner. In these figures both Afifi et al. and our method present results for the DSTFC case, as it provided better results in Table 5.1.

When focusing on the Sony images Fig. 5.1, we can see how our method outperforms all the others. This is more evident when examining the floor (1st and 2nd images), the white and gray stripes (4th image), the green wall (6th image), and the yellow and pink mattress (8th image). For the Nikon images in Fig. 5.2, our method also produces the best subjective results. This is evident looking at the wall at the back (1st image), the green wall (2nd image), the floor (3rd image), and the windows (6th, 7th, and 8th images).

Fig. 5.3 presents a comparison of Afifi et al. [8] method and our approach when using the input configurations, DST and DSTFC. We see that the attention-based approach produces the best results. In particular, improvements can be discerned on the floor (1st image), the yellow mattress (2nd image), the white wall (3rd image), and the green wall (4th image).

Finally, Fig. 5.4 shows the attention maps of three scenes of our re-purposed dataset. The diagonal of the attention maps reshaped to the size of the *Query*, *Key*, and *Value* is displayed. For visualization purposes, we clip the maps to $A \in [0, 0.01]$. These maps help reveal which part of the different images are contributing to the final output. It potentially offers a way to interpret and understand our model's decision-making process in correcting white-balance in complex, multi-illuminant scenes.

5.6 Cross-camera generalization

To assess the generalization capabilities of our attention-based fusion model and the baseline methods, we conduct an experiment that evaluates their performance on images captured by unseen camera models. Specifically, we test models trained on the Sony cam-

era dataset on the Nikon image set, and vice versa. This cross-camera evaluation provides valuable insights into the robustness and adaptability of the different approaches to camera-specific characteristics and imaging pipelines.

The results of this generalization experiment are presented in Table 5.2. We report the mean, median, and trimean values for the standard evaluation metrics: mean squared error (MSE), mean angular error (MAE), and ΔE2000. These metrics capture different aspects of the white-balance correction performance, providing a comprehensive assessment of the models' ability to generalize to unseen camera data.

Across all evaluation metrics and visual outputs, our attention-based fusion model consistently outperforms the different versions of the method by Afifi et al. [8], including their 'Oracle' approach, which represents an upper bound on their method's performance. The superior generalization ability of our model is particularly noteworthy, as it demonstrates its robustness to camera-specific factors and its ability to effectively handle image data from diverse imaging pipelines.

The strong generalization performance of our attention-based fusion model can be attributed to its flexible and adaptive nature. By leveraging self-attention mechanisms, our model can selectively attend to the most relevant features and illumination cues present in the input images, regardless of the specific camera characteristics. This attention-based fusion strategy allows our model to capture the underlying illumination properties and effectively combine the information from multiple input images rendered under different illumination conditions, even when presented with data from an unseen camera source.

In contrast, the baseline methods by Afifi et al. [8] appear to be more susceptible to camera-specific biases and variations, as evidenced by their relatively lower performance

in this cross-camera evaluation. Their approaches, while effective on the datasets they were trained on, appear to struggle to generalize effectively to unseen camera data due to potential limitations in their fusion strategies or their reliance on specific camera characteristics learned during training.

5.7 Results on a synthetic dataset

Finally, we evaluate our attention-based fusion approach on the Synthetic Multi-Illuminant Dataset introduced in the prior work of Afifi et al. [8]. The results for our method on the Synthetic Multi-Illuminant Dataset are presented in Table 5.3. We report the mean, median, and trimean values for the same set of evaluation metrics used in our previous experiments: mean squared error (MSE), mean angular error (MAE), and ΔE2000. Across these metrics, our attention-based approach achieves the best performance in all but one measure, demonstrating its effectiveness in handling synthetic mixed-illumination scenes.

By achieving state-of-the-art performance on this challenging dataset, our attention-based fusion approach demonstrates its ability to handle complex mixed-illumination scenarios, even in controlled and synthetically generated environments. These results further validate the efficacy of our method and its potential for practical applications in real-world scenarios involving multi-illuminant white-balance correction.

5.8 Summary

In summary, while all the methods evaluated on the LSMI dataset are bottlenecked by training only on the limited older WB datasets, their performance correlates with the model

capacity and architectural complexity employed. Our method can extract maximum value from this dataset through its high-capacity multi-component design. Methods like FC4 are trained on relatively smaller real-world datasets fare worse compared to those trained on larger synthetic data. Nonetheless, the necessity of incorporating real-world images during training to improve generalization to practical deployments cannot be overstated.

The results of the generalization experiment in Sec. 5.6 highlight the versatility and robustness of our attention-based fusion model, demonstrating its ability to adapt to diverse imaging conditions and camera pipelines. This strong generalization capability is particularly valuable in practical applications, where the deployed system may encounter image data from various camera models or imaging sources. By outperforming state-of-the-art methods, including their 'Oracle' versions, our approach exhibits promising potential for real-world deployment and applicability across a wide range of camera setups and imaging scenarios.



Figure 5.1: Qualitative results for the Sony camera. From left to right, the results of SWBNet [41], FC4 [32], DeepWB [4], Afifi et al. [8] (DSTFC), our attention approach (DSTFC), and the ground truth. The ΔE2000 value for each image is shown in the bottom-right corner.



Figure 5.2: Qualitative results for the Nikon camera. From left to right, the results of SWBNet [41], FC4 [32], DeepWB [4], Afifi et al. [8] (DSTFC), our attention approach (DSTFC), and the ground truth. The ΔE2000 value for each image is shown in the bottom-right corner.



Figure 5.3: Qualitative results comparing our method versus Afifi et al. [8] for both the three (DST) and the 5 (DSTFC) illuminant settings. In this figure we can see the same trend as in Table 5.1. The five illuminant versions work better than their three illuminant counterparts. Also, any version of our method is better than both versions from Afifi et al.

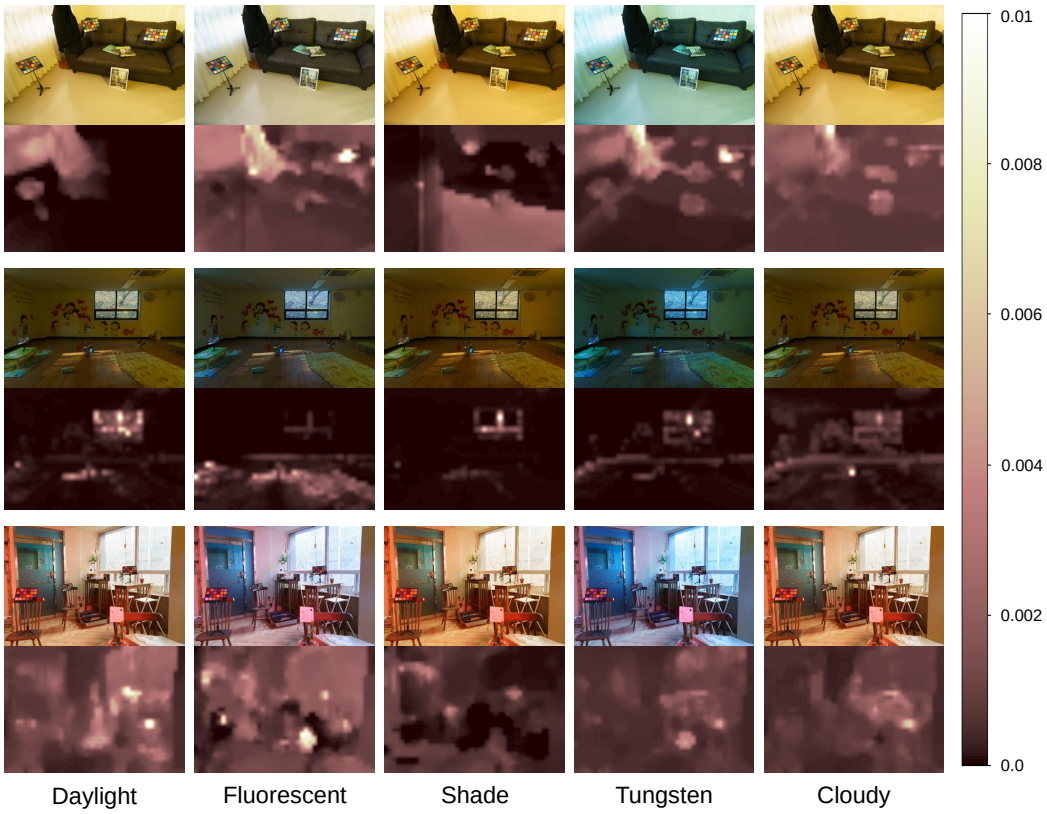


Figure 5.4: Attention maps of our attention-based method for all the presets. First, third and fifth rows: The five renders with pre-defined color temperatures that correlate to common lighting conditions. Second, fourth and sixth rows: Attention maps of the attention module for each preset.

Chapter 6

Conclusions

In this final chapter, we summarize the main findings and contributions of this thesis, focusing on the progress made in correcting white-balance in scenes with multiple light sources. We review how our proposed attention-based fusion method outperforms existing techniques. The chapter also explores the broader significance of this work for fields like photography and computer vision, highlighting practical uses and benefits. Additionally, we consider the societal impacts of our research, including both positive advancements and ethical concerns. Finally, we suggest possible future directions for research, offering ideas for further improvements that could make white-balance correction methods even more effective in handling complex lighting situations.

6.1 Summary

We introduced a new sRGB image dataset by re-purposing an existing multi-illuminant dataset [36] aimed at multi-illumination white-balance correction in the RAW domain.

Our dataset provides ground-truth sRGB images along with five white-balance presets. The purpose of this dataset is to support the development of fusion-based multi-illuminant methods that offer an interesting alternative strategy to white-balance correction over traditional per-pixel illumination estimation. To achieve this, we introduced a new fusion-based method utilizing attention-based fusion. We demonstrate that this attention-based approach surpasses several variants on the previous state-of-the-art models [8] by over 25% and exceeds the performance of an "oracle" method that selects the best per-image result using the previous method. Moreover, our approach demonstrates strong cross-camera generalization capabilities. We showed that by working in the sRGB domain, we achieve good performance across different camera models. This is a significant advantage over RAW-based methods, which often struggle with cross-camera generalization due to the sensor-specific nature of RAW data. Our sRGB-based approach allows the model to learn more generalizable features, making it more robust and adaptable to various camera sensors and imaging pipelines. We believe this new sRGB-based dataset will be useful in assisting future development in fusion-based multi-illuminant white-balance correction.

6.2 Societal Impacts

The ability to perform accurate white-balance correction, especially in complex multi-illuminant environments, has significant societal implications across various domains. White-balance is a crucial preprocessing step in many computer vision and image processing pipelines, and its proper handling can have far-reaching effects on the performance and reliability of downstream applications.

In the realm of computational photography and image editing, effective multi-illuminant white-balance correction can greatly enhance the visual quality and realism of digital images. Photographers and content creators often capture scenes with mixed illumination sources, such as indoor and outdoor lighting, or a combination of natural and artificial light. Accurate white-balance correction in these scenarios can help preserve the natural color rendition and faithfully represent the intended visual appearance, leading to more aesthetically pleasing and visually appealing images.

Moreover, multi-illuminant white-balance correction is crucial for numerous computer vision applications that rely on accurate color information. For instance, in the field of autonomous vehicles and advanced driver-assistance systems (ADAS), robust color constancy is essential for reliable object detection, traffic sign recognition, and scene understanding under varying illumination conditions. Failure to properly account for multi-illuminant scenarios could lead to missed detections or false positives, potentially compromising the safety and decision-making capabilities of these systems.

Similarly, in medical imaging and computational pathology, accurate color representation is critical for reliable diagnosis and analysis. Many medical imaging modalities, such as endoscopy, dermatoscopy, and microscopy, involve capturing images under different illumination sources or mixed lighting conditions. Effective multi-illuminant white-balance correction can help preserve the true color information of tissues, lesions, and cellular structures, enabling more accurate diagnoses and better-informed treatment decisions.

Beyond these specific domains, multi-illuminant white-balance correction can also contribute to the development of more robust and generalizable computer vision models. By accurately handling complex lighting conditions, these models can better adapt to

real-world scenarios and exhibit improved performance in a wide range of applications, such as object recognition, image retrieval, and scene understanding. Accurate white-balance not only improves the visual quality of images but also enhances the performance of downstream tasks such as object detection and image segmentation. Properly calibrated colors reduce the likelihood of deep learning models being misled by artifacts introduced by poor illumination, as highlighted in the work by Afifi et al. 2019 (c) [3], where the robustness of deep learning models was shown to degrade under specific manipulations of image quality

Additionally, the development and deployment of multi-illuminant white-balance algorithms should be done in an inclusive and responsible manner, ensuring that the algorithms are trained on diverse datasets and do not perpetuate or amplify existing biases or disparities related to race, ethnicity, or other protected characteristics.

Overall, the advancements in multi-illuminant white-balance correction have the potential to positively impact various sectors and enable more reliable and accurate computer vision applications.

However, it is important to acknowledge the potential negative societal impacts that such technologies might entail.

One of the primary concerns is the potential for these technologies to be misused in ways that manipulate visual content, leading to misinformation or deceptive practices. For instance, in journalism, advertising, or social media, enhanced white-balance correction could be used to alter the perceived reality of an image, thereby influencing public perception or consumer behavior in a misleading manner. This raises ethical questions about the authenticity of visual media and the responsibility of content creators and distributors to

maintain truthful representation.

Another issue relates to the cultural implications of color accuracy in images. Different cultures may interpret and value color differently, and the use of white-balance correction algorithms that standardize color representation could inadvertently impose a particular cultural bias. This could result in the erasure of cultural nuances in images, particularly in global media or cross-cultural communication, where color can carry significant cultural meaning.

It is crucial to consider the ethical implications and societal impacts of these technologies, and to develop and deploy them in a responsible and inclusive manner, prioritizing fairness, privacy, and the well-being of individuals and communities.

6.3 Future Work

While our attention-based fusion approach for multi-illuminant white-balance correction has demonstrated promising results, several avenues for future research can be explored to further advance the field.

One promising direction is the investigation of end-to-end trainable architectures for multi-illuminant white-balance. Current approaches, including ours, rely on pre-rendered input images under different illumination conditions. However, an end-to-end trainable model that can directly process raw sensor data or low-level image representations could potentially learn more effective representations and fusion strategies tailored for the multi-illuminant problem. Such an approach may alleviate the need for pre-rendering and could enable more efficient and streamlined inference pipelines.

An interesting direction for future research would be to investigate whether using a white-balance setting other than Daylight as the query to the attention mechanism could improve performance. Certain lighting conditions may respond better to alternative white-balance presets, potentially leading to more accurate and context-aware corrections. This exploration could help refine the adaptability of the attention-based fusion method, allowing it to select the most suitable reference white balance dynamically based on the scene.

Another area of interest is the exploration of self-supervised or unsupervised learning techniques for multi-illuminant white-balance. While our current approach relies on supervised training with ground truth white-balanced images, self-supervised or unsupervised methods could potentially leverage the inherent structure and statistics of multi-illuminant scenes to learn effective white-balance correction without the need for laborious data annotation. This could facilitate the development of more scalable and adaptable solutions, especially in scenarios where ground truth data is scarce or difficult to obtain.

Furthermore, extending the multi-illuminant white-balance framework to handle more complex and dynamic illumination conditions, such as those encountered in outdoor scenes with varying weather conditions or in indoor environments with multiple localized light sources, could be a valuable research direction. Developing models that can robustly handle such scenarios could enable more reliable and practical applications in fields like autonomous vehicles, augmented reality, and computational photography.

Additionally, exploring the applicability of multi-illuminant white-balance techniques to other modalities beyond traditional RGB images, such as multispectral or hyperspectral imaging, could open up new avenues for research and applications. These modalities often involve complex spectral and illumination characteristics, and effective white-balance

correction could unlock their potential in domains like remote sensing, medical imaging, and material analysis.

In summary, the field of multi-illuminant white-balance correction offers numerous opportunities for future exploration, ranging from architectural innovations and novel learning paradigms to domain expansions and ethical considerations. Addressing these challenges will not only advance the technical capabilities of white-balance correction but also contribute to the development of more robust, reliable, and responsible computer vision systems for a wide range of real-world applications.

Bibliography

[1] Color and camera raw. https://helpx.adobe.com/ca/photoshop-elements/using/color-camera-raw.html. Accessed: 2021-07-20. 26

[2] V-ray for 3ds max. https://www.chaosgroup.com/3d-rendering-software. Accessed: 2021-07-20. 16

[3] Mahmoud Afifi and Michael S Brown. What else can fool deep learning? Addressing color constancy errors on deep neural network performance. In *ICCV*, 2019. 58

[4] Mahmoud Afifi and Michael S Brown. Deep white-balance editing. In *CVPR*, 2020. 13, 41, 42, 43, 45, 46, 51, 52

[5] Mahmoud Afifi and Michael S Brown. Interactive White Balancing for Camera-Rendered Images. In *CIC*, 2020. 10, 13

[6] Mahmoud Afifi, Brian Price, Scott Cohen, and Michael S Brown. When Color Constancy Goes Wrong: Correcting Improperly White-Balanced Images. In *CVPR*, 2019. 2, 13, 15, 33

[7] Mahmoud Afifi, Abhijith Punnappurath, Abdelrahman Abdelhamed, Hakki Can Karaimer, Abdullah Abuolaim, and Michael S Brown. Color temperature tuning: Allowing accurate post-capture white-balance editing. In *CIC*, 2019. 13

[8] Mahmoud Afifi, Marcus A Brubaker, and Michael S Brown. Auto White-Balance Correction for Mixed-Illuminant Scenes. In *WACV*, 2022. 3, 6, 7, 14, 16, 23, 31, 32, 34, 36, 41, 42, 43, 44, 45, 46, 47, 48, 49, 51, 52, 53, 56

[9] Hoda Aghaei and Brian Funt. A flying gray ball multi-illuminant image dataset for color research. In *CIC*, 2020. 12

[10] Jonathan T. Barron and Yun-Ta Tsai. Fast Fourier Color Constancy. In *CVPR*, 2017. 5

[11] Shida Beigpour, Christian Riess, Joost Van De Weijer, and Elli Angelopoulou. Multi-illuminant estimation with conditional random fields. *TIP*, 2013. 11

[12] Simone Bianco and Claudio Cusano. Quasi-Unsupervised Color Constancy. In *CVPR*, 2019. 10

[13] Michael Bleier, Christian Riess, Shida Beigpour, Eva Eibenberger, Elli Angelopoulou, Tobias Tröger, and André Kaup. Color constancy and non-uniform illumination: Can existing algorithms work? In *ICCV Workshops*, 2011. 12

[14] David H Brainard and William T Freeman. Bayesian color constancy. *JOSA A*, 1997. 9

[15] George Buchsbaum. A Spatial Processor Model for Object Colour Perception. *Journal of the Franklin Institute*, 1980. 5, 8

[16] Ayan Chakrabarti, Keigo Hirakawa, and Todd Zickler. Color Constancy Beyond Bags of Pixels. In *CVPR*, 2008. 8

[17] Dongliang Cheng, Dilip K Prasad, and Michael S Brown. Illuminant Estimation for Color Constancy: Why spatial domain methods work and the role of the color distribution. *JOSA A*, 2014. 8, 15, 22, 43

[18] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *NeurIPS*, 2021. 19

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 18, 34

[20] Mark D Fairchild. *Color appearance models*. John Wiley & Sons, 2013. 2

[21] Graham D Finlayson. Corrected-Moment Illuminant Estimation. In *ICCV*, 2013. 8

[22] Graham D Finlayson and Elisabetta Trezzi. Shades of Gray and Colour Constancy. In *Color and Imaging Conference*, 2004. 8

[23] Graham D Finlayson, Brian V Funt, and Kobus Barnard. Color constancy under varying illumination. In *ICCV*, 1995. 11

[24] Graham D Finlayson, Paul M Hubel, and Steven Hordley. Color by Correlation. In *CIC*, 1997. 9

[25] Brian V Funt, Mark S Drew, and Jian Ho. Color Constancy from Mutual Reflection. *IJCV*, 1991. 9

[26] Peter Vincent Gehler, Carsten Rother, Andrew Blake, Tom Minka, and Toby Sharp. Bayesian Color Constancy Revisited. In *CVPR*, 2008. 9, 15, 21, 43

[27] Arjan Gijsenij, Theo Gevers, and Joost van de Weijer. Computational Color Constancy: Survey and Experiments. *TIP*, 2011. 5

[28] Arjan Gijsenij, Theo Gevers, and Joost Van De Weijer. Improving Color Constancy by Photometric Edge Weighting. *TPAMI*, 2011. 8

[29] Arjan Gijsenij, Rui Lu, and Theo Gevers. Color constancy for multiple light sources. *TIP*, 2011. 11

[30] Daniel Hernandez-Juarez, Sarah Parisot, Benjamin Busam, Ales Leonardis, Gregory Slabaugh, and Steven McDonagh. A Multi-Hypothesis Approach to Color Constancy. In *CVPR*, 2020. 10

[31] Eugene Hsu, Tom Mertens, Sylvain Paris, Shai Avidan, and Frédo Durand. Light mixture estimation for spatially varying white balance. In *ACM SIGGRAPH*, 2008. 11

[32] Yuanming Hu, Baoyuan Wang, and Stephen Lin. FC4: Fully Convolutional Color Constancy with Confidence-weighted Pooling. In *CVPR*, 2017. 5, 41, 43, 45, 46, 51, 52

[33] Md Akmol Hussain and Akbar Sheikh Akbari. Color constancy algorithm for mixed-illuminant scene images. *IEEE Access*, 2018. 12

[34] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 43

[35] Hamid Reza Vaezi Joze and Mark S Drew. Exemplar-based color constancy and multiple illumination. *TPAMI*, 2013. 12

[36] Dongyoung Kim, Jinwoo Kim, Seonghyeon Nam, Dongwoo Lee, Yeonkyung Lee, Nahyup Kang, Hyong-Euk Lee, ByungIn Yoo, Jae-Joon Han, and Seon Joo Kim. Large Scale Multi-Illuminant (LSMI) Dataset for Developing White Balance Algorithm under Mixed Illumination. In *ICCV*, 2021. 3, 6, 12, 17, 24, 25, 55

[37] Gudrun J Klinker, Steven A Shafer, and Takeo Kanade. A Physical Approach to Color Image Understanding. *IJCV*, 1990. 10

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 2012. 43

[39] Edwin H Land and John J McCann. Lightness and Retinex Theory. *JOSA A*, 1971. 8

[40] Chongyi Li, Chunle Guo, Shangchen Zhou, Qiming Ai, Ruicheng Feng, and Chen Change Loy. FlexiCurve: Flexible Piecewise Curves Estimation for Photo Retouching. In *CVPR*, 2023. 34

[41] Chunxiao Li, Xuejing Kang, Zhifeng Zhang, and Anlong Ming. SWBNet: A Stable White Balance Network for sRGB Images. In *AAAI*, 2023. 41, 43, 45, 46, 51, 52

[42] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV Workshops*, 2021. 19

[43] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 19

[44] Yi-Chen Lo, Chia-Che Chang, Hsuan-Chao Chiu, Yu-Hao Huang, Chia-Ping Chen, Yu-Lin Chang, and Kevin Jou. CLCC: Contrastive Learning for Color Constancy. In *CVPR*, 2021. 10

[45] Lou, Zhongyu and Gevers, Theo and Hu, Ninghang and Lucassen, Marcel P and others. Color Constancy by Deep Learning. In *BMVC*, 2015. 10

[46] Seoung Wug Oh and Seon Joo Kim. Approaching the Computational Color Constancy as a Classification Problem through Deep Learning. *PR*, 2017. 10

[47] Yanlin Qian, Joni-Kristian Kamarainen, Jarno Nikkanen, and Jiri Matas. On Finding Gray Pixels. In *CVPR*, 2019. 8

[48] Guillermo Sapiro. Color and Illuminant Voting. *TPAMI*, 1999. 9

[49] Steven A Shafer. Using Color to Separate Reflection Components. *Color Research & Application*, 1985. 9, 10

[50] Wu Shi, Chen Change Loy, and Xiaoou Tang. Deep Specialized Network for Illuminant Estimation. In *ECCV*, 2016. 10

[51] Oleksii Sidorov. Conditional GANs for Multi-Illuminant Color Constancy: Revolution or Yet Another Approach? In *CVPR Workshops*, 2019. 5

[52] Joost Van De Weijer, Theo Gevers, and Arjan Gijsenij. Edge-Based Color Constancy. *TIP*, 2007. 8

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 18

[54] Javier Vazquez-Corral, C Alejandro Párraga, Maria Vanrell, and Ramon Baldrich. Color Constancy Algorithms: Psychophysical Evaluation on a New Dataset. *JIST*, 2009. 9

[55] Javier Vazquez-Corral, Maria Vanrell, Ramon Baldrich, and Francesc Tous. Color Constancy by Category Correlation. *TIP*, 2012. 5, 9

[56] Donik Vršnak, Ilija Domislović, Marko Subašić, and Sven Lončarić. Autoencoder-based training for multi-illuminant color constancy. *JOSA A*, 2022. 5

[57] Donik Vršnak, Ilija Domislović, Marko Subašić, and Sven Lončarić. Illuminant segmentation for multi-illuminant scenes using latent illumination encoding. *Signal Processing: Image Communication*, 2022. 5

[58] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021. 19

[59] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 18

[60] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, 2022. 19

[61] Bolei Xu, Jingxin Liu, Xianxu Hou, Bozhi Liu, and Guoping Qiu. End-to-End Illuminant Estimation based on Deep Metric Learning. In *CVPR*, 2020. 10

[62] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 19

[63] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 34

[64] Simone Zini, Marco Buzzelli, Simone Bianco, and Raimondo Schettini. COCOA: Combining Color Constancy Algorithms for Images and Videos. *TCI*, 2022. 10