Machine Vision and Applications (2025) 36:58
https://doi.org/10.1007/s00138-025-01680-1

RESEARCH

# Feature distribution statistics as a loss objective for robust white balance correction

**Furkan Kınlı**[1,2] · **Furkan Kıraç**[1]

Received: 10 December 2024 / Revised: 1 March 2025 / Accepted: 7 March 2025 / Published online: 19 March 2025
© The Author(s) 2025

**Abstract**
White balance (WB) correction is critical for accurate color reproduction in digital images, especially under complex, multi-illuminant lighting conditions. Traditional methods, such as the Gray-World assumption, rely on global statistics and struggle in real-world, non-uniform lighting scenarios. Modern deep learning approaches, including convolutional and attention-based architectures, have significantly advanced WB correction but often fail to explicitly account for higher-order feature distribution statistics, which may limit their robustness in challenging environments. This study introduces a novel framework that leverages Exact Feature Distribution Matching (EFDM) as a loss objective to align feature distributions across multiple moments, including mean, variance, skewness, and kurtosis. By modeling lighting as a style factor, the method explicitly addresses distributional shifts caused by complex illumination, offering a robust solution for WB correction. The framework integrates EFDM with a Vision Transformer architecture, enabling precise handling of global and local lighting variations. Extensive experiments on the large-scale multi-illuminant (LSMI) dataset demonstrate the superiority of the proposed approach over state-of-the-art methods and commonly used loss functions when applied to the same architecture. Qualitative and quantitative evaluations highlight its effectiveness in achieving perceptually accurate WB correction, particularly in multi-illuminant environments. By bridging statistical modeling with modern deep learning, this work establishes the critical role of feature distribution alignment in advancing WB correction and sets a new benchmark for robustness and generalization in complex lighting scenarios.

**Keywords** White balance correction · Feature distribution alignment · Multi-illuminant environments · EFDM · Vision transformers

## 1 Introduction

Accurate reproduction of colors under varying lighting conditions remains a fundamental challenge in digital image processing, particularly in scenarios characterized by multi-illuminant environments or non-uniform lighting distributions. White Balance (WB) correction, a critical component of the Image Signal Processor (ISP) pipeline, aims to ensure that colors in digital images are perceived correctly under diverse illuminants by neutralizing unwanted color casts. Traditional approaches to WB correction often rely on simplistic assumptions, such as the Gray-World hypothesis [1], which assume uniform illumination throughout the scene. Although computationally efficient, these methods struggle in complex real-world scenarios, where the interplay of multiple light sources can result in significant color distortions that such techniques cannot adequately address [2].

The complexity of these scenarios is further illustrated in Fig. 1, which demonstrates how different spectral signatures of light sources can alter the overall appearance of a scene. Such variations highlight the challenges posed by multi-illuminant conditions, where simple assumptions about illumination often fail to account for the intricate interactions between light sources and scene elements.

Recent advances in deep learning have brought notable improvements to WB correction by leveraging data-driven approaches to model complex lighting conditions.

✉ Furkan Kınlı
furkan.kinli@ozyegin.edu.tr

Furkan Kıraç
furkan.kirac@ozyegin.edu.tr

[1] Department of Computer Science, Özyeğin University, Istanbul 34782, Turkey

[2] T-Fashion Inc., Toronto, ON, Canada




**Fig. 1** Different spectral signatures of the light sources in a sample scene. The columns represent various illumination setups: 'Illum-1' corresponds to the scene illuminated by a single light source, 'Illum-1-2' and 'Illum-1-3' represent the combined illumination from two different sources, and 'Illum-1-2-3' represents all three sources turned on simultaneously

Convolutional Neural Networks (CNNs) [3–11] and, more recently, attention-based architectures [12, 13] have demonstrated the ability to learn mappings between input images and their desired white-balanced outputs. These models excel in capturing both global and local lighting variations, which makes them well suited to handle intricate illumination scenarios. However, most existing deep learning approaches focus on implicit feature extraction and do not explicitly consider the role of feature distributions in the learning process. As a result, they often fail to generalize to challenging multi-illuminant environments, where higher-order statistical shifts in feature distributions play a critical role in the perceived quality of the correction.

This work builds upon the foundation of treating lighting as a style factor [14, 15], a perspective that views illumination as a variable that influences the statistical properties of an image's features. By aligning the distributions of these features across various lighting conditions, it becomes possible to neutralize the stylistic effects of lighting while preserving the content and structure of the scene. Although prior work [15] has utilized statistical measures such as mean and variance to address such variations, these approaches remain limited in their ability to capture higher-order distributional characteristics, which are often critical for achieving perceptually accurate color corrections in complex lighting scenarios.

The approach presented in this study introduces a novel optimization framework that utilizes feature distribution statistics as a loss objective for WB correction. This framework employs Exact Feature Distribution Matching (EFDM) [16] to align not only first- and second-order moments, but also higher-order statistics, such as skewness and kurtosis, of the feature distributions between input and target images. By treating these distributional shifts as a *style factor* to be corrected, the proposed method extends beyond the capabilities of existing techniques, which offers a more robust solution for multi-illuminant environments and non-uniform lighting conditions. The integration of EFDM with a Vision Transformer-based feature extraction [17] and a UNet-like Transformer-based architecture [18] further enhances the model's ability to adapt to varying scenarios, thus ensuring consistent and reliable performance across diverse datasets.

This study provides a comprehensive analysis of feature distribution statistics under multi-illuminant conditions and demonstrates the efficacy of the proposed framework in addressing the limitations of both traditional and modern WB correction methods. By aligning the statistical properties of feature distributions, the method achieves a perceptual consistency that is both quantitative and qualitative in nature, which establishes a new benchmark for robustness in challenging lighting environments. Our key contributions are as follows:

- EFDM is used as a training objective to enforce feature-level alignment, ensuring the predicted feature distribution aligns with the ground truth illumination representation, instead of relying on pixel-wise intensity differences.
- Through experimental validation, we show that higher-order statistics play a crucial role in capturing illumination shifts, and this makes EFDM a superior alternative to traditional loss functions used in WB correction.
- EFDM is designed as an architecture-independent training objective and can be integrated into both CNN-based (e.g., UNet) and Transformer-based (e.g., Uformer) architectures without additional inference complexity.

- EFDM ensures robust adaptation to diverse lighting conditions, particularly with improved generalization in multi-illuminant environments.

## 2 Related work

The problem of WB correction has been extensively studied, with research evolving from simple statistical models to advanced deep learning techniques. The early approaches [1, 19–28] are based heavily on global statistical assumptions, such as the Gray-World hypothesis, which assumes that the average color of a scene should be gray under neutral lighting conditions [1]. These methods focus on global illumination estimates to correct the entire image uniformly. Similarly, solutions based on gamut mapping [29–32] also offer computationally efficient techniques by restricting possible illuminant colors to a feasible gamut, ensuring that corrections remained within realistic limits. Both approaches are computationally efficient and straightforward in design, which makes them suitable for early WB correction tasks. However, these assumptions fail in complex environments where lighting conditions are not uniform, which results in overcorrection or undercorrection of color casts, particularly in scenes with mixed or multi-illuminant lighting [2, 11].

In addition to these methods, low-level statistical methods [28, 33–42] can be considered as an alternative, where they aim to refine illuminant estimation using statistics. These methods incorporate specialized statistical approaches to dynamically adjust illumination throughout the image, which improve the robustness of the corrections in some cases. However, their reliance on hand-crafted features and simplistic assumptions about scene structure often constrain their adaptability and effectiveness in diverse real-world environments. Scene semantics-based approaches [43–47] further extend these ideas by integrating a high-level understanding of objects and their contexts, which allows models to use scene information to infer lighting conditions. Although it is promising, these methods relies heavily on predefined models and prior knowledge of the scenes, which makes them less suitable for uncontrolled or dynamic settings where new or unexpected lighting conditions could occur.

The introduction of deep learning marked a significant shift in WB correction, where CNNs were used to directly learn the mappings between raw images and their white-balanced counterparts [3–11]. These approaches demonstrate superior performance compared to traditional methods by capturing both global and local lighting variations through their strength in representing related features. With their ability to learn from large-scale datasets, CNN-based models bring greater flexibility and adaptability to WB correction, particularly in scenarios involving complex lighting [11]. More recently, attention-based architectures [12, 13] have introduced new capabilities by using self-attention mechanisms, which allow modeling long-range dependencies and global context in images. Despite these advancements, many deep learning methods still fail to explicitly model the distributional properties of image features, which this study demonstrates to be critical for handling multi-illuminant environments. By relying only on implicit feature extraction, they often lack the robustness needed to generalize to scenarios where higher-order statistics of feature distributions, such as skewness and kurtosis, play a critical role in achieving perceptually accurate corrections.

The idea of treating feature distributions as style factors originates from research in style transfer and domain adaptation [14, 15], where techniques such as Adaptive Instance Normalization (AdaIN) [48] shows that aligning feature statistics such as mean and variance could allow effective style manipulation without altering content. This approach demonstrates the importance of explicitly aligning statistical properties between source and target distributions, which inspires its application to other tasks such as image synthesis [49, 50] and restoration [14]. However, while these methods effectively capture first- and second-order statistics, they fail to address higher-order statistical characteristics that often play a crucial role in modeling complex variations introduced by style factors. To overcome these limitations, EFDM [16] is proposed, which aligns all the moments of the feature distributions utilizing the empirical cumulative distribution function (eCDF), thus providing a more comprehensive and precise mechanism for feature alignment. Although EFDM has shown its potential in tasks such as style transfer and domain adaptation, as evidenced by the reproducibility study [51], its application to WB correction remains largely underexplored. This study builds upon the foundational concept of modeling lighting as a style factor and extends it by integrating EFDM into a novel optimization framework. By explicitly aligning feature distributions across multiple moments, the proposed approach addresses the limitations of both traditional methods and existing deep learning-based techniques, offering a robust and flexible solution for WB correction. This framework not only incorporates higher-order statistics to enhance generalization, but also bridges the gap between statistical modeling and modern deep learning architectures, which enables effective handling of multi-illuminant environments where existing approaches often fall short.

## 3 Methodology

This section presents a WB correction framework with the proposed loss objective that emphasizes the role of feature distribution statistics in modeling and addressing complex illumination scenarios. Figure 2 provides an overview of the proposed framework for White Balance correction, highlighting its integration with FDM Loss. We first present the definition of the proposed objective function and then follow by an exploration of the critical role that distributional shifts play in driving color changes within a scene, particularly as they manifest in chromaticity channels and feature representations. The analysis delves into the global style information captured by Vision Transformers (ViT) [17], with a focus on the [CLS] token and its ability to encapsulate illumination-related variations. Finally, the statistical behavior of the chromaticity channel and feature distributions is examined under various illumination conditions. This underscores the limitations of traditional methods and demonstrates the robustness of the proposed approach to address these challenges.

### 3.1 Feature distribution statistics as a loss objective

The proposed framework employs EFDM [16] as a loss objective, applied specifically to the [CLS] token in ViT [17]. EFDM was introduced to address the limitations of Gaussian-based assumptions made by earlier feature alignment methods, such as those used in [48]. By aligning the empirical cumulative distribution functions (eCDFs) of feature maps, EFDM ensures that not only first- and second-order statistics (i.e., mean and variance) but also higher-order moments such as skewness and kurtosis are aligned. This provides a more comprehensive method for handling complex and non-linear distortions introduced by style factors.

The theoretical foundation of EFDM is grounded in the Glivenko-Cantelli theorem [52], which guarantees the uniform convergence of the eCDF to the true CDF as the sample size increases:

$$\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0 \text{ as } n \to \infty, \tag{1}$$

where $\hat{F}_n(x)$ is the empirical CDF, $F(x)$ is the true CDF, and the convergence is uniform. Using this principle, EFDM matches feature distributions with precision by aligning their eCDFs. This allows the proposed method to handle intricate chromatic variations and complex lighting distortions more effectively than traditional Gaussian-based methods.

The practical implementation of EFDM involves computing a rank-based transformation of the feature values in the predicted and target feature maps, followed by a pointwise comparison. Instead of relying on absolute intensity values, EFDM operates by sorting the feature values in both the predicted and reference feature representations and applying a transformation based on their relative ranks. This approach makes the loss function invariant to scale and

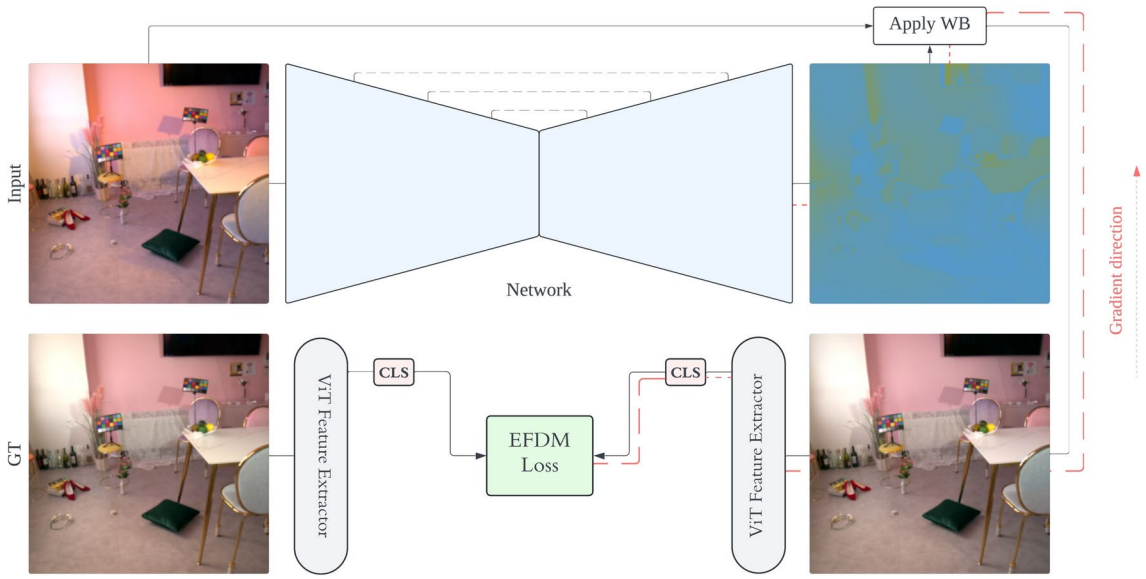


**Fig. 2** Overview of the proposed framework for White Balance correction leveraging FDM Loss. The network (i.e., Uformer, UNet) processes the input image (top left) to generate an illumination map in the *UV* color space (top right), which is subsequently used to reconstruct the predicted white-balanced *RGB* image (bottom right). ViT feature extractor is employed to extract the [CLS] tokens from both the predicted and ground truth images, which are then input into the EFDM-based loss function to align their feature distributions effectively

intensity variations, thus ensuring that feature alignments remain consistent under different conditions.

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

**X**: input vector, **Y**: target vector
_, IndexX = torch.sort(**X**)
SortedY, _ = torch.sort(**Y**)
InverseIndex = IndexX.argsort(−1)
**return** **X**+ SortedY.gather(−1, InverseIndex) −**X**.detach()

**Algorithm 1** PyTorch-like pseudo-code for EFDM

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

In ViTs, the [CLS] token, 384-dimensional feature vector, serves as a global representation of the image by aggregating information from all input patches [53]. This property makes the [CLS] token particularly effective in capturing global style characteristics, including illumination and chromatic consistency, which are critical for accurate WB correction. Using this global representation, the EFDM precisely aligns the predicted and ground-truth [CLS] token distributions to ensure robust illumination modeling across diverse scenes.

Our proposed FDM loss function is mathematically defined as:

$$\mathcal{L}_{\mathrm{EFDM}}(f_{\mathrm{pred}}, f_{\mathrm{gt}}) = \frac{1}{n}\sum_{i=1}^{n}\left[f_{\mathrm{pred}}[i] - f_{\mathrm{gt}}[\mathrm{rank}(f_{\mathrm{pred}}[i])]\right]^2, \tag{2}$$

where $f_{\mathrm{pred}}[i]$ and $f_{\mathrm{gt}}[i]$ are elements of the predicted and ground-truth [CLS] token vectors, respectively. The function $\mathrm{rank}(\cdot)$ maps each element of $f_{\mathrm{pred}}$ to its corresponding position in the sorted vector $f_{\mathrm{gt}}$. By aligning the ranks, EFDM ensures precise matching of the empirical distributions of the feature maps, effectively addressing variations introduced by multi-illuminant environments.

By incorporating EFDM as a loss objective, named FDM Loss, the proposed framework achieves a robust WB correction. The ability to match exact feature distributions, beyond simple mean and variance alignment, allows the method to handle the intricate variations caused by complex lighting conditions. This ensures perceptual consistency and accurate color correction across diverse and challenging scenarios, solidifying EFDM's role as a cornerstone in the proposed methodology.

## 3.2 Style information in vision transformers

The [CLS] token in ViTs has demonstrated versatility in representing global appearance characteristics across different scenes and images. This token aggregates key visual elements while remaining invariant to spatial configurations, and this capability makes it effective for tasks that require capturing style information. For example, it has been successfully applied in semantic appearance transfer tasks, where the global appearance of one image is applied to another while maintaining structural integrity [53].

This flexibility extends to modeling lighting conditions, which can be considered as a style factor. The [CLS] token's ability to encapsulate global appearance properties, including illumination variations, makes it particularly suitable for tasks involving multi-illuminant settings. Unlike patch-level features, which focus on localized details, the [CLS] token inherently captures global illumination characteristics, such as intensity, color temperature, and chromatic balance across the scene. This makes it ideal for EFDM-based alignment, which targets global style consistency in WB correction.

## 3.3 Analysis of illumination as distribution statistics

To validate the proposed framework, an extensive analysis is conducted to understand the impact of varying illumination conditions, particularly in multi-illuminant scenarios, on chromaticity channels and feature distributions.

### 3.3.1 Chromaticity channel distributions under multiple illuminants

Figure 3 illustrates how multi-illuminant conditions affect the chromaticity distributions in the $U$ and $V$ channels. The left columns show the *RGB*-rendered white-balanced and illuminated images that highlight the visual differences introduced by multi-illuminant lighting. The right columns present histograms for the $U$ and $V$ channels, comparing their distributions in the illuminated scenes (i.e., red and blue

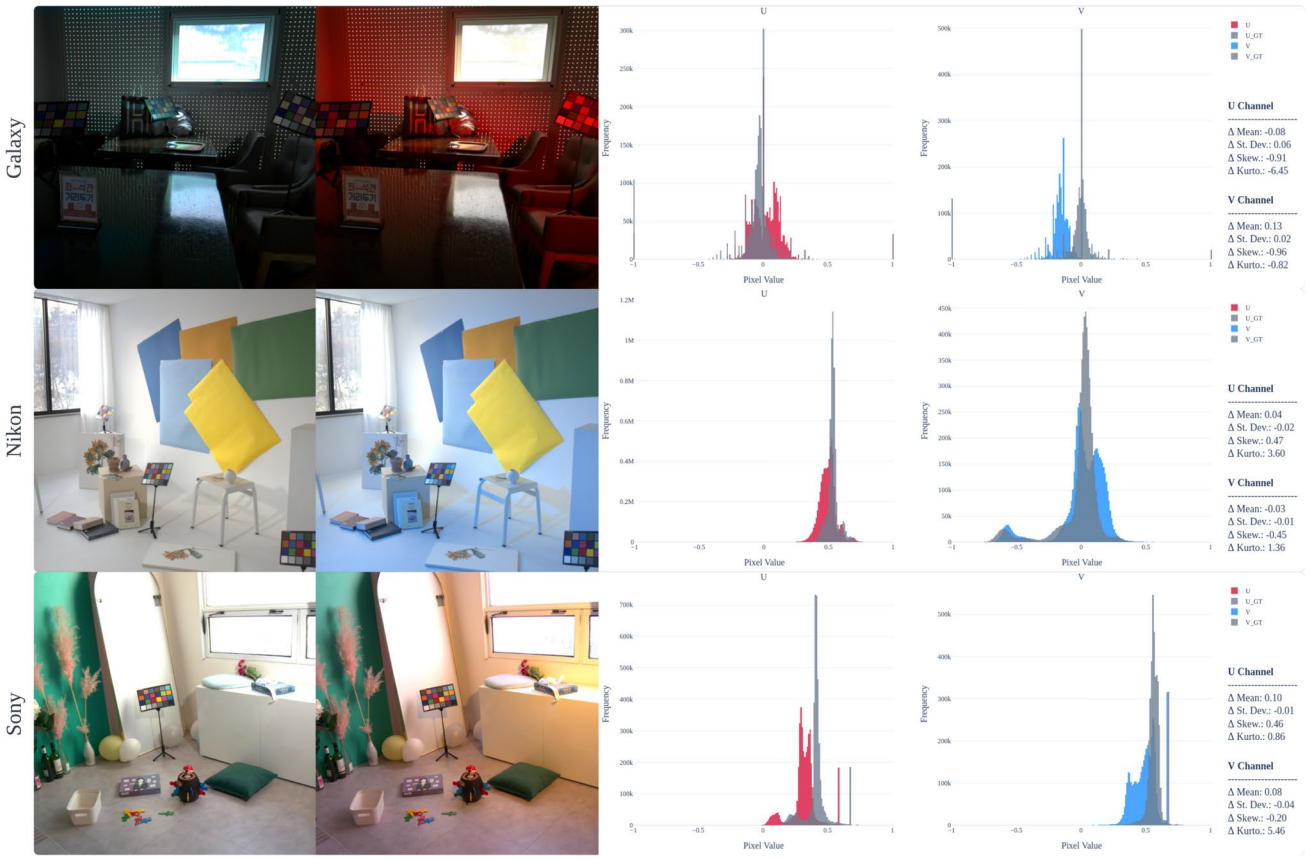


**Fig. 3** Chromaticity channel distributions under different lighting conditions

histograms) with those of the ground truth white-balanced images (i.e., gray histograms). Across three camera models (i.e., Galaxy, Nikon, Sony), the histograms for white-balanced images exhibit narrow distributions with low skewness and kurtosis, which indicates uniform chromaticity. In contrast, images captured with multiple illuminants show wider and more skewed distributions, which reflects the complex spectral interactions introduced by multiple light sources.

The widening and asymmetry in these distributions reveal the higher-order distortions caused by multi-illuminant lighting. Elevated kurtosis values indicate the presence of extreme chromatic variations or outliers, while skewness reflects asymmetry in the color distribution. These distortions challenge previous methods that align only mean and variance, which underscores the need to align higher-order moments to ensure a robust WB correction.

### 3.3.2 Multi-illuminant effects on feature distributions

Figure 4 highlights the impact of multi-illuminant lighting on the feature distributions of the [CLS] token. Each sample, from different camera models (i.e., Galaxy, Nikon, and Sony), illustrates a ground truth white-balanced image, an illuminated scene with at least two light sources, and the corresponding [CLS] token feature distributions. The histograms on the right depict the distribution of normalized [CLS] token features under these varying lighting conditions to offer insights into the impact of illumination on feature statistics.

The observed variations in feature distributions underscore the significant challenges posed by multiple illuminants in feature representation. These challenges reveal the limitations of methods that align only basic statistics. Although these methods can handle simpler lighting scenarios, in other words, more uniform lighting scenarios, they struggle in complex settings involving multiple light sources. Multi-illuminant environments often result in interactions between illuminants with distinct spectral properties. This can introduce color shifts that are neither predictable nor uniformly distributed across the image. The findings emphasize the importance of adopting a more nuanced approach that surpasses lower-order statistical alignment to effectively capture the chromatic shifts caused by complex illumination conditions.

Our analysis reveals that higher-order statistics, including skewness and kurtosis, exhibit pronounced variations between white-balanced images and those captured under

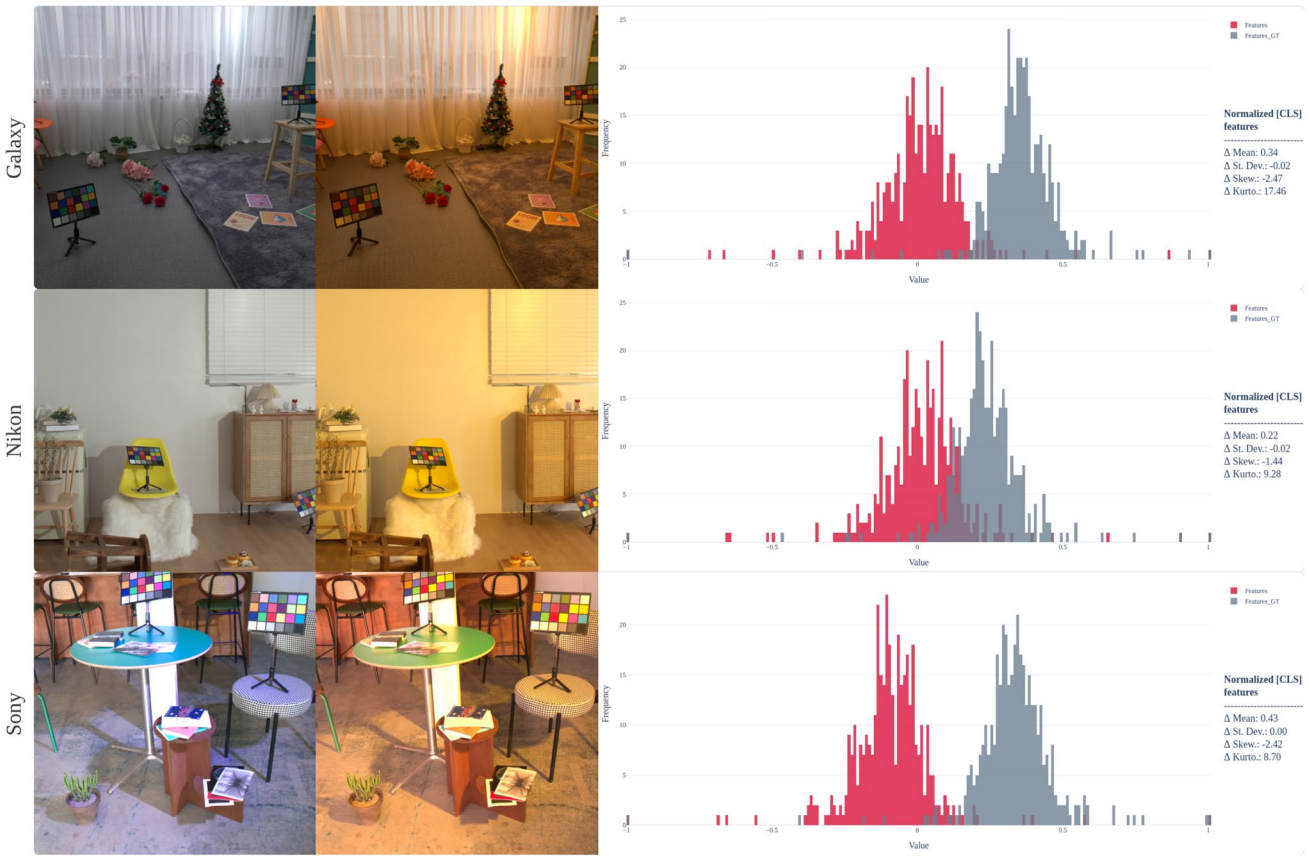


**Fig. 4** [CLS] token distribution statistics for samples from different cameras

multi-illuminant conditions. Specifically, these higher-order deviations reflect the non-Gaussian characteristics induced by the interplay of multiple light sources. Skewness captures the asymmetry of color distributions, which indicates a dominance of certain hues due to specific illuminants, while kurtosis reflects the presence of extreme chromatic variations, or *heavy tails* in the distribution. These characteristics illustrate how multi-illuminant environments can amplify complex chromatic interactions, which may further complicate WB correction.

The analysis demonstrates that aligning only mean and variance inadequately represents the full scope of chromatic distortions in such scenarios. For instance, while lower-order statistics might capture general trends in illumination, they fail to address the intricate variations caused by multiple light sources, resulting in a suboptimal WB correction. By incorporating higher-order statistics, such as those modeled through EFDM, these limitations can be overcome.

### 3.3.3 Feature distribution statistics across illumination settings

Figure 5 presents the mean, standard deviation, skewness, and kurtosis of the [CLS] token distributions across all samples in the LSMI dataset under different illumination settings. Lower-order statistics, such as mean and standard deviation, exhibit minor variations between illuminated and white-balanced images, largely aligning along the diagonal. In contrast, higher-order statistics show significant deviations, which reflects the complex interactions of multiple light sources in real-world scenes. The results indicate that higher-order statistics capture substantial deviations between illuminated and white-balanced images compared to lower-order statistics. This distinction is particularly evident across all three camera setups, highlighting the importance of incorporating higher-order statistical alignment for robust WB correction.

These findings validate the claim for the inadequacy of lower-order alignment for handling the non-Gaussian distortions introduced by multi-illuminant lighting. Higher-order statistics can provide a more nuanced representation of these variations by capturing the intricate chromatic effects that define multi-illuminant environments. By aligning the full feature distribution, EFDM can ensure robust WB correction by preserving both lower- and higher-order moments to deliver perceptually consistent results.

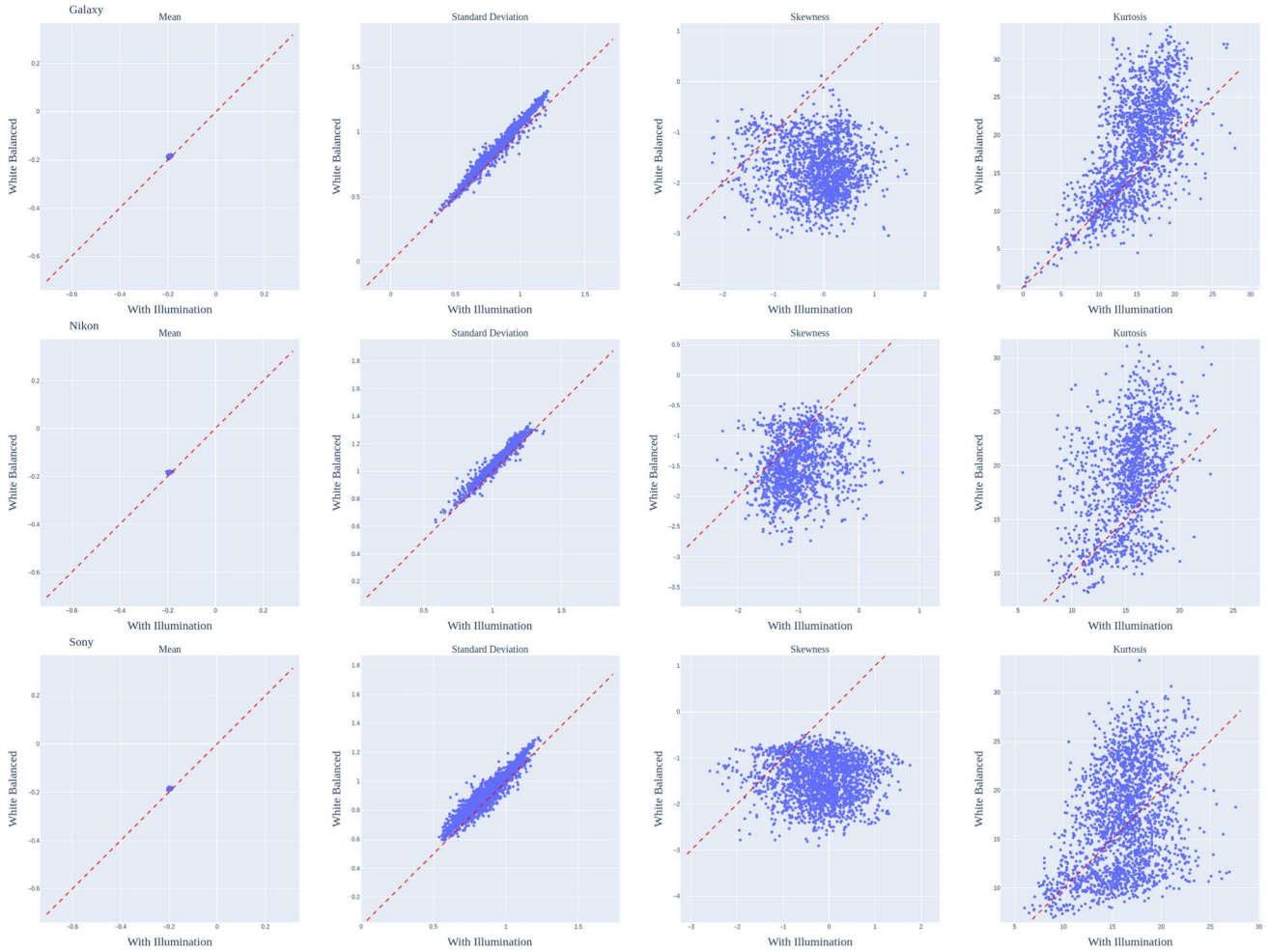


**Fig. 5** [CLS] token distribution statistics for illuminated vs. white balanced images across three cameras in the LSMI dataset

### 3.3.4 Feature distribution matching beyond attention mechanism

While attention mechanisms dynamically adjust feature relationships by reweighting pairwise dependencies, they do not explicitly enforce feature distribution alignment. Self-attention operates by computing similarity scores between features, enabling the model to emphasize relevant dependencies, but it does not directly constrain the statistical structure of learned feature maps.

In contrast, EFDM explicitly aligns feature distributions by sorting and matching feature values, ensuring that the predicted feature distribution retains statistical consistency with the reference. Unlike pixel-wise losses, which focus on intensity differences, or attention mechanisms, which refine feature importance, EFDM directly minimizes distributional discrepancies, affecting not just first-order statistics but also higher-order moments such as variance, skewness, and kurtosis. Our analysis further supports this distinction, demonstrating that higher-order moments effectively capture illumination-induced variations, and this makes them a more robust optimization target than direct intensity-based losses. Standard pixel-wise loss functions often fail to generalize across multi-illuminant scenarios, whereas EFDM ensures a consistent distributional alignment, leading to superior color correction performance.

Several prior works have explored the importance of feature distribution constraints in deep learning. For instance, Batch Normalization [54] stabilizes training by normalizing feature distributions, while Contrastive Learning [55–63] *indirectly* aligns representations through their statistical characteristics. More explicitly, optimal transport-based loss functions [58, 59] align distributions in structured prediction tasks. EFDM follows a similar philosophy, but operates at the feature level, providing a robust and computationally efficient alternative to direct pixel-wise losses.

## 4 Experimental results

This section outlines the experimental setup used to evaluate the proposed framework. We detail the dataset and preprocessing techniques, describe the architectural choices and hyperparameters for training, and specify the metrics used for quantitative and qualitative evaluations. These components are designed to rigorously assess the effectiveness and robustness of the proposed objective in various illumination scenarios.

### 4.1 Experimental setup

#### 4.1.1 Dataset and preprocessing

The experiments are conducted on the LSMI dataset [60], which provides a comprehensive benchmark for WB correction across diverse illumination conditions. The dataset includes samples captured using three distinct camera models: *Galaxy*, *Nikon*, and *Sony*. Each sample consists of ground truth white-balanced images, images captured under single illuminants, and those affected by multi-illuminant lighting scenarios. This dataset is particularly suited for evaluating the robustness of WB correction methods due to its diverse range of lighting conditions and camera-specific characteristics.

All images undergo a random resized crop to $256 \times 256$ pixels during training. The full scope of illumination settings, including single- and multi-illuminant samples, is utilized to increase the robustness of model. Furthermore, random illumination augmentation, as proposed in [60], is applied with a saturation range of 0.2 to 0.8 and a hue threshold of 0.25. This augmentation strategy was introduced to simulate real-world variations in color temperature and illumination conditions. By introducing controlled perturbations in saturation and hue, the model is exposed to a broader distribution of color shifts, and this improves its ability to handle both single- and mixed-illuminant conditions.

#### 4.1.2 Architecture & details

The proposed framework integrates the Feature Distribution Matching (FDM) loss function into the Uformer [18] architecture, a Vision Transformer-based model designed for image restoration tasks. Input images are first converted into the *UVL* color space before being processed by the network. The network outputs two channels, $U$ and $V$, corresponding to the chromatic components $R/G$ and $B/G$, respectively. These channels are then combined with the $G$ channel of the original input image to reconstruct the white-balanced *RGB* image, as defined by Eq. 3.

$$R = G \cdot e^{U}, \quad B = G \cdot e^{V}. \tag{3}$$

The $R$, $G$, and $B$ channels are normalized by dividing by the white level and subsequently clamped to the range [0, 1] to maintain intensity consistency in the reconstructed image. The reconstructed image is then processed using a pre-trained Vision Transformer model, which extracts the [CLS] token as its global representation. Similarly, the ground truth white-balanced image undergoes the same process to obtain its corresponding [CLS] token. The proposed FDM loss function is applied to align the distributions of the predicted and ground truth [CLS] tokens, ensuring precise matching of their empirical distributions across all statistical moments.

Furthermore, the attention mechanism of Uformer facilitates effective modeling of both local and global illumination variations, which may relax the complexities of multi-illuminant scenarios. To independently validate the effectiveness of the proposed loss objective, separate from Uformer's contribution, it is also evaluated on the UNet [61] architecture, a widely used CNN architecture for image processing. Both architectures employ FDM loss to align feature distributions across varying illumination conditions to ensure robust performance in challenging lighting environments.

The framework is optimized using the Adam optimizer [62] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The model is trained with a batch size of 32 and an initial learning rate of $2 \times 10^{-4}$, which is progressively reduced over the final 800 epochs to refine the convergence of the model. Training is carried out over 2, 400 epochs, ensuring sufficient iterations for the model to fully optimize its parameters.

The choice of Adam optimizer was motivated by its ability to efficiently adapt learning rates across different parameters, making it well-suited for tasks involving complex feature distributions. The learning rate decay in the later training stages was introduced as a stabilization mechanism to prevent large updates from destabilizing the model after reaching a refined feature alignment. These hyperparameters were determined on the basis of extensive experimental trials, balancing training efficiency and convergence speed.

The batch size of 32 was selected on the basis of available GPU resources, ensuring optimal memory utilization while maintaining stable optimization dynamics. Since batch size affects both learning rate selection and total training epochs, these values were chosen through empirical evaluations to achieve the best convergence behavior without introducing computational bottlenecks. Training is conducted on $2\times$ NVIDIA A100 GPUs, leveraging their high computational efficiency for processing large-scale datasets.

**Table 1** Benchmark results on the LSMI dataset for the Galaxy camera

| Model | Single | | Multi | | Mixed | | MSR |
|---|---|---|---|---|---|---|---|
| | Mean | Median | Mean | Median | Mean | Median | |
| Pix2Pix [65] | 6.53 | 2.17 | 4.28 | 2.63 | 5.66 | 2.44 | 0.66 |
| Gijsenij et al. [24] | 7.49 | 6.04 | 12.38 | 9.57 | 10.09 | 7.43 | 1.65 |
| Bianco et al. [6] | 4.15 | 3.30 | 5.56 | 4.33 | 4.89 | 3.83 | 1.34 |
| HDRNet [60, 64] | 2.85 | 2.20 | 3.13 | 2.70 | 3.06 | 2.54 | 1.10 |
| HDRNet [13, 64] | – | – | – | – | 3.06 | 2.54 | – |
| UNet [60, 61] | 2.95 | 1.86 | 2.35 | 2.00 | 2.63 | 1.91 | 0.80 |
| UNet [13, 61] | 2.85 | – | 2.55 | – | 2.68 | 2.17 | 0.90 |
| One-Net [63] | **1.19** | 0.75 | 2.16 | 1.53 | **1.57** | 0.93 | 1.82 |
| AID [13] | **1.19** | – | 2.03 | – | 1.66 | 1.41 | 1.71 |
| Uformer + FDM (**ours**) | 1.78 | 1.48 | **1.87** | 1.69 | 1.83 | 1.62 | **1.05** |

The Multi-to-Single Ratio reflects the robustness of the models in multi-illuminant scenarios

**Table 2** Benchmark results on the LSMI dataset for the Nikon camera

| Model | Single | | Multi | | Mixed | | MSR |
|---|---|---|---|---|---|---|---|
| | Mean | Median | Mean | Median | Mean | Median | |
| Pix2Pix [65] | 6.10 | 2.27 | 4.18 | 2.76 | 5.41 | 2.49 | 0.77 |
| Bianco et al. [6] | 3.18 | 2.61 | 4.65 | 4.19 | 3.93 | 3.48 | 1.18 |
| HDRNet [60, 64] | 2.76 | 2.43 | 3.20 | 3.01 | 2.99 | 2.61 | 1.07 |
| HDRNet [13, 64] | – | – | – | – | 2.99 | 2.61 | – |
| UNet [60, 61] | 1.51 | 1.14 | 2.36 | 1.84 | 1.95 | 1.45 | 1.21 |
| UNet [13, 61] | 1.49 | – | 2.30 | – | 1.92 | 1.54 | 1.20 |
| One-Net [63] | 1.27 | 0.67 | 1.99 | 1.43 | 1.53 | 0.85 | 1.30 |
| AID [13] | **1.11** | – | 2.26 | – | 1.71 | 1.34 | 1.32 |
| Uformer + FDM (**ours**) | 1.26 | 0.97 | **1.54** | 1.13 | **1.48** | 1.05 | **1.22** |

#### 4.1.3 Evaluation metrics

In this study, the primary evaluation metric is the Mean Absolute Error (MAE) between the corrected and ground truth images. In addition to MAE, we propose a novel evaluation metric, namely Multi-to-Single Ratio (MSR), which can be used to assess the generalization ability of the models across single- and multi-illuminant scenarios. Qualitative evaluations include visual comparisons of corrected images to highlight the perceptual improvements achieved by the proposed method.

The Multi-to-Single Ratio is defined as the ratio of the mean MAE in multi-illuminant scenarios to the mean MAE in single-illuminant scenarios across all samples in the given dataset. This metric evaluates a model's ability to generalize effectively to multi-illuminant conditions while avoiding overfitting to single-illuminant scenarios. An MSR value closer to 1 indicates superior adaptability, while a higher MSR reflects greater performance degradation between single- and multi-illuminant conditions.

The proposed MSR provides a deeper insight into the robustness of the model. Overfitting to single-illuminant scenarios often leads to significant performance degradation under multi-illuminant conditions, which are more representative of real-world settings. By maintaining a balance across these scenarios, our proposed objective demonstrates its ability to address practical challenges in WB correction.

#### 4.1.4 Complexity analysis

Since we integrate EFDM into a training objective rather than an architectural modification, it does not introduce additional computational complexity during inference or deployment. While inference time and model size depend on the chosen backbone architecture, a balanced comparison is provided by evaluating using EFDM as loss function with both CNN-based (i.e., UNet) and Transformer-based (i.e., Uformer) architectures. Among the compared methods, AID [13] employs an attention-based approach, making its computational characteristics similar to Uformer, while One-Net [63] and HDRNet [64] use CNN-based feature extraction, resembling UNet. This ensures that our experimental setup covers a range of architectures with comparable computational properties, demonstrating that EFDM integration does not impose additional inference costs beyond what is expected in existing approaches.

### 4.2 Experimental results

Tables 1, 2, and 3 present the benchmark results on the LSMI dataset [60], showcasing the performance of various

**Table 3** Benchmark results on the LSMI dataset for the Sony camera

| Model | Single | | Multi | | Mixed | | MSR |
|---|---|---|---|---|---|---|---|
| | Mean | Median | Mean | Median | Mean | Median | |
| Pix2Pix [65] | 4.08 | 1.72 | 4.37 | 3.26 | 4.20 | 2.20 | **1.07** |
| Bianco et al. [6] | 3.25 | 2.62 | 4.38 | 3.93 | 3.86 | 3.19 | 1.35 |
| HDRNet [60, 64] | – | – | – | – | 3.21 | 2.89 | – |
| HDRNet [13, 64] | 2.76 | 2.43 | 3.2 | 3.01 | 2.99 | 2.61 | **1.07** |
| UNet [60, 61] | 2.83 | 2.44 | 3.04 | 2.78 | 2.94 | 2.66 | **1.07** |
| UNet [13, 61] | 1.92 | – | 2.34 | – | 2.15 | 1.74 | 1.22 |
| One-Net [63] | 1.45 | 0.60 | 2.23 | 1.65 | 1.76 | 0.93 | 1.54 |
| AID [13] | **1.01** | – | 2.16 | – | 1.66 | 1.35 | 2.14 |
| Uformer + FDM (**ours**) | 1.52 | 1.39 | **1.67** | 1.57 | **1.61** | 1.47 | **1.10** |

methods, including our proposed approach, *Uformer with FDM Loss*, under single-, multi-, and mixed-illuminant scenarios for *Galaxy*, *Nikon*, and *Sony* cameras. The results demonstrate the effectiveness of integrating the EFDM-based loss function into the Uformer architecture for WB correction. Our proposed loss objective with Uformer architecture ensures robust performance under challenging multi-illuminant conditions while maintaining accuracy in single-illuminant scenarios.

#### 4.2.1 Benchmark on the LSMI dataset

In multi-illuminant scenarios, the Uformer with FDM Loss achieves state-of-the-art performance across all three camera setups, as demonstrated by both the mean and median MAE values. For the *Galaxy* camera, Uformer with FDM Loss records a mean MAE of 1.87 and a median of 1.69 in multi-illuminant scenarios, outperforming key competitors such as One-Net (2.16, 1.53). Similarly, for the *Nikon* camera, the proposed method achieves a mean MAE of 1.54 and a median of 1.13, surpassing One-Net (1.99, 1.43) and other notable methods. On the *Sony* camera, Uformer with FDM Loss achieves a mean MAE of 1.67 and a median of 1.57, which further solidifies its superiority under multi-illuminant conditions. The consistent performance across all three camera setups highlights the robustness of the proposed method in handling challenging lighting conditions. By using EFDM as the loss function to ensure the exact match of feature distributions, the method effectively models the interaction of multiple light sources, which results in superior perceptual quality and accurate color correction.

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

In contrast, Uformer with FDM Loss achieves an MSR value close to 1 across all configurations, which indicates its strong generalization ability. For the *Galaxy*, *Nikon*, and *Sony* cameras, the ratios are 1.05, 1.22 and 1.10, respectively. These values underscore the ability of the method to maintain robust performance in multi-illuminant conditions without sacrificing accuracy in single-illuminant scenarios.

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

The results highlight the importance of achieving a balance between single- and multi-illuminant performance for practical real-world applications. The ability of our proposed approach, Uformer with FDM Loss, to generalize across varied conditions without overfitting makes it a highly reliable solution for WB correction. Its superior performance in multi-illuminant conditions, coupled with an MSR value closer to 1, underscores its robustness and adaptability. This robustness can be attributed to the feature distribution matching mechanism, which directly aligns feature representations rather than relying on pixel-wise intensity comparisons or statistical normalization techniques. Compared methods such as HDRNet [64] and One-Net [63] optimize the illumination by minimizing direct pixel-level errors, which often leads to overfitting in single-illuminant settings and instability when illumination conditions vary significantly.

In contrast, EFDM preserves the overall structure of feature distributions, ensuring that learned representations remain consistent across different lighting conditions and preventing excessive adaptation to a specific illuminant type. This ability to maintain stable performance under

**Table 4** Ablation study on the proposed loss function using the Uformer architecture

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

**Table 5** Ablation study on the proposed loss function using the UNet architecture

| Camera | Loss | Single | | Multi | | Mixed | | MSR |
|---|---|---|---|---|---|---|---|---|
| | Function | Mean | Median | Mean | Median | Mean | Median | |
| | MSE | 2.95 | 1.86 | 2.35 | 2.00 | 2.63 | 1.91 | 0.80 |
| Galaxy | AdaIN-s | 2.64 | 1.99 | 2.35 | 2.02 | 2.48 | 1.98 | 0.89 |
| | FDM | **2.42** | 1.81 | **2.14** | 1.74 | **2.27** | 1.79 | 0.88 |
| | MSE | 1.51 | 1.14 | 2.36 | 1.84 | 1.95 | 1.45 | 1.21 |
| Nikon | AdaIN-s | **1.25** | 0.89 | **1.22** | 0.91 | **1.24** | 0.90 | 0.98 |
| | FDM | 1.40 | 1.17 | 1.89 | 1.33 | 1.66 | 1.25 | 1.14 |
| | MSE | 2.83 | 2.44 | 3.04 | 2.78 | 2.94 | 2.66 | 1.07 |
| Sony | AdaIN-s | 2.07 | 1.60 | 2.24 | 1.90 | 2.18 | 1.79 | 1.08 |
| | FDM | **1.96** | 1.63 | **2.10** | 1.74 | **2.04** | 1.67 | 1.07 |

multi-illuminant scenarios is particularly crucial, as common pixel-intensity-based loss functions struggle with the non-linear interactions of multiple light sources. The consistently low MSR values achieved further validate that EFDM's distribution matching prevents catastrophic degradation under complex illumination settings, making it a more reliable objective function for real-world applications.

#### 4.2.2 Ablation study

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

The results presented in Table 4 illustrate the significant advantages of employing the FDM loss function over the traditional pixel-wise mean squared error loss. The key strength of the FDM loss lies in its ability to perform exact feature distribution matching, which ensures that higher-order statistics, such as skewness and kurtosis, are effectively aligned between the predicted and ground truth images. This capability enables the model to better handle non-Gaussian and highly variable illumination conditions, which are common in multi- and mixed-illuminant scenarios.

For example, applying the FDM loss to the Uformer architecture significantly improves performance in multi-illuminant scenarios. For the *Galaxy* camera, the mean MAE improves from 2.03 with MSE loss to 1.87 with FDM loss. This demonstrates the robustness of our proposed loss function under complex lighting conditions. Similarly, for the *Nikon* camera, the mean MAE is reduced from 1.56 with MSE loss to 1.43 with FDM loss, which reflects improved mixed-illuminant performance.

Unlike MSE loss, which focuses solely on minimizing pixel-wise intensity differences, FDM loss captures intricate relationships among image regions by incorporating higher-order statistical insights derived from feature information. In combination with Uformer's attention mechanisms, which capture both global context and local relationships, our framework ensures spatial consistency, which is a critical requirement for handling uneven lighting variations across a scene. This enhanced capability enables the proposed loss objective to generalize effectively across diverse scenes without overfitting to specific illumination conditions, as evidenced by the balanced MSR values observed across the three camera setups.

The results in Table 5 demonstrate the versatility of the proposed FDM loss through its significant performance improvements within the UNet architecture. For the *Sony* camera, the multi-illuminant mean MAE decreases from 3.04 with MSE loss to 2.10 with FDM loss. Similarly, the mixed-illuminant mean MAE improves from 2.94 to 2.04. These results highlight the adaptability of FDM loss, effectively enhancing the performance of a convolutional network like UNet, despite its architectural limitations compared to Uformer.

Additionally, we evaluate the effectiveness of using AdaIN-stylized feature difference (AdaIN-s) as a loss function to enforce only first- and second-order feature alignment. Although AdaIN-s outperforms MSE in some cases, particularly on the Nikon camera, where it achieves the lowest multi-illuminant mean MAE, it remains less effective than FDM in other scenarios. This suggests that enforcing only mean and variance alignment provides improvements over traditional pixel-wise loss functions, but remains insufficient for robust WB correction under diverse illumination conditions compared to full distribution alignment.

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

#### 4.2.3 Qualitative results

The qualitative comparisons in Figs. 6, 7, and 8 illustrate the efficacy of our proposed approach, Uformer with FDM loss, in addressing the challenges posed by multi-illuminant scenarios in the LSMI dataset. Each figure comprises three samples, with six rows in total. The odd-numbered rows include the illuminated version of the scene, followed by WB correction results from three different methods (i.e., UNet with FDM Loss, Uformer with MSE Loss and Uformer with FDM Loss), and finally the ground truth white-balanced image in the last column. The even-numbered rows display the RAW image, the corresponding network outputs in the *UV* color space representing the illumination, and the ground truth illumination in the last column. Across the three camera setups (i.e., *Galaxy*, *Nikon*, *Sony*), these results consistently demonstrate that our method outperforms counterparts, including those utilizing traditional loss functions such as MSE and traditional architectures like UNet, in achieving perceptually accurate WB correction under diverse lighting conditions.

The introduction of FDM-based loss represents a significant advancement in performance by enabling exact matching of feature distributions for the [CLS] token, a critical representation of global image characteristics. By harnessing the [CLS] token's representational power in the global context, the FDM loss not only ensures spatial consistency but also preserves intricate lighting details and color fidelity, particularly in complex multi-illuminant scenarios.

In summary, our proposed approach effectively overcomes the limitations of traditional methods by delivering robust and reliable performance in diverse camera setups and challenging illumination conditions. These results highlight the practical applicability and versatility of our proposed method in real-world imaging scenarios.

#### 4.2.4 Comparison with state-of-the-art

To further evaluate the effectiveness of our proposed approach, we present a qualitative comparison with LSMI-U [60] and AID [13], as shown in Fig. 9. Given the lack of publicly available models for state-of-the-art methods, it was not possible to directly reproduce their results on the LSMI test set. Instead, we extracted qualitative results for AID from their original paper to enable a fair comparison. However, OneNet [63] does not provide test set comparisons in their qualitative results, making its inclusion infeasible.

The results highlight the differences in illumination estimation among the methods. LSMI-U, which serves as a baseline dataset-derived approach, demonstrates reasonable WB correction but exhibits residual color shifts in challenging multi-illuminant scenarios. AID, designed for illumination decomposition, estimates chromatic illumination components rather than optimizing for perceptual correction, leading to noticeable deviations in certain regions. By contrast, both variants of our method produce more consistent illumination maps with fewer artifacts, reinforcing the effectiveness of using EFDM as a training objective to enforce feature distribution alignment for robust WB correction.

Notably, in cases where the scene contains strong color casts, the proposed Uformer with FDM Loss model demonstrates improved adaptation to multi-illuminant conditions, as seen in the highlighted regions. The method effectively mitigates undesired tints and preserves global color consistency. However, slight deviations in illumination maps can still be observed, particularly in highly saturated regions, which may indicate potential areas for future improvement in illuminant adaptation.

#### 4.2.5 Failure case analysis

Although the proposed method demonstrates strong performance in WB correction across diverse scenarios, certain failure cases reveal limitations under complex illumination conditions. Figure 10 presents the cases where the method encounters challenging lighting conditions, which leads to deviations from the ground truth. In these cases, the predicted illumination maps diverge from the reference, particularly in regions with mixed or non-uniform lighting distributions. The corresponding corrected images exhibit

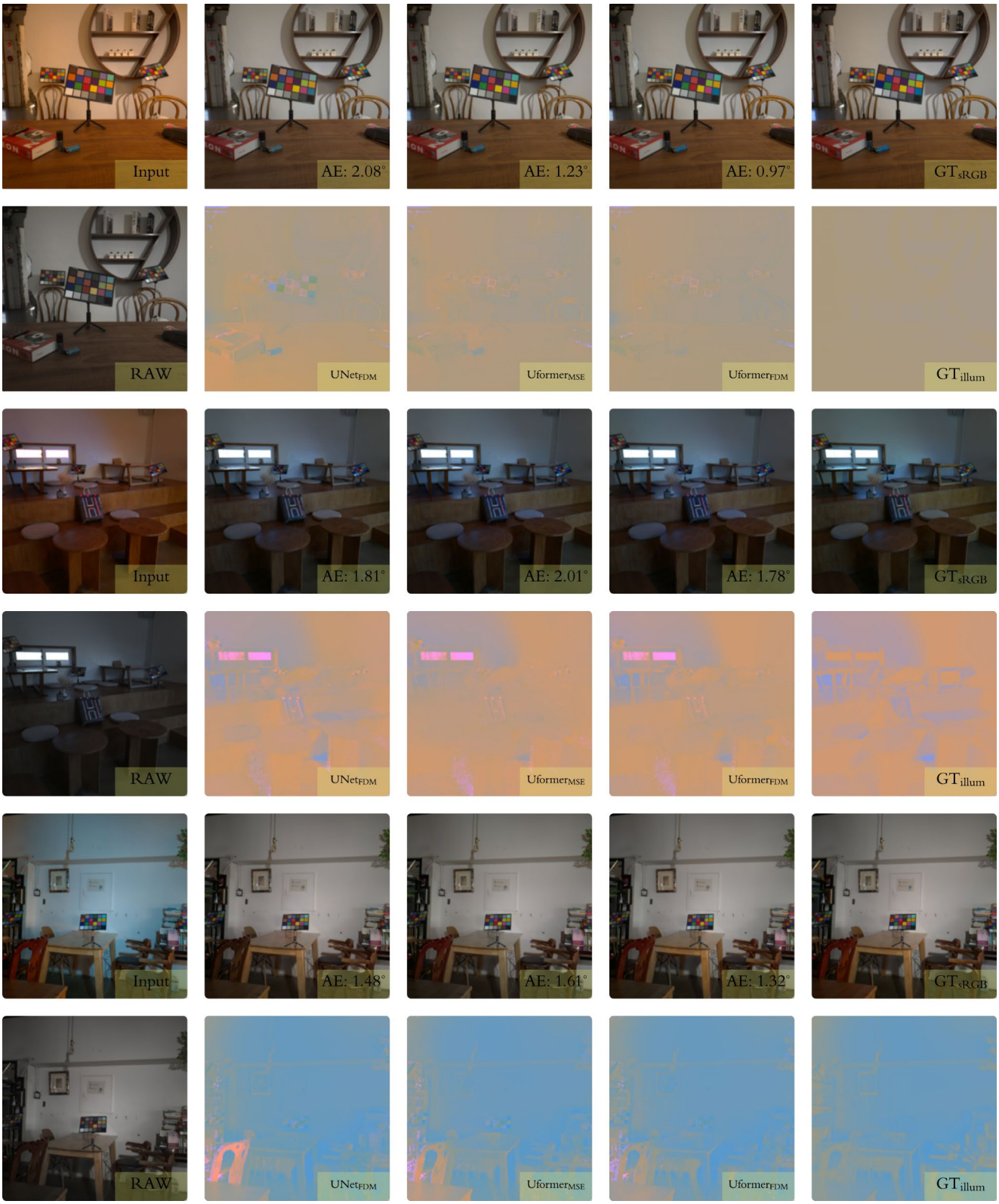


**Fig. 6** Visual comparison of WB correction outputs on the Galaxy camera from the LSMI dataset. Image indices: 312, 323, 896

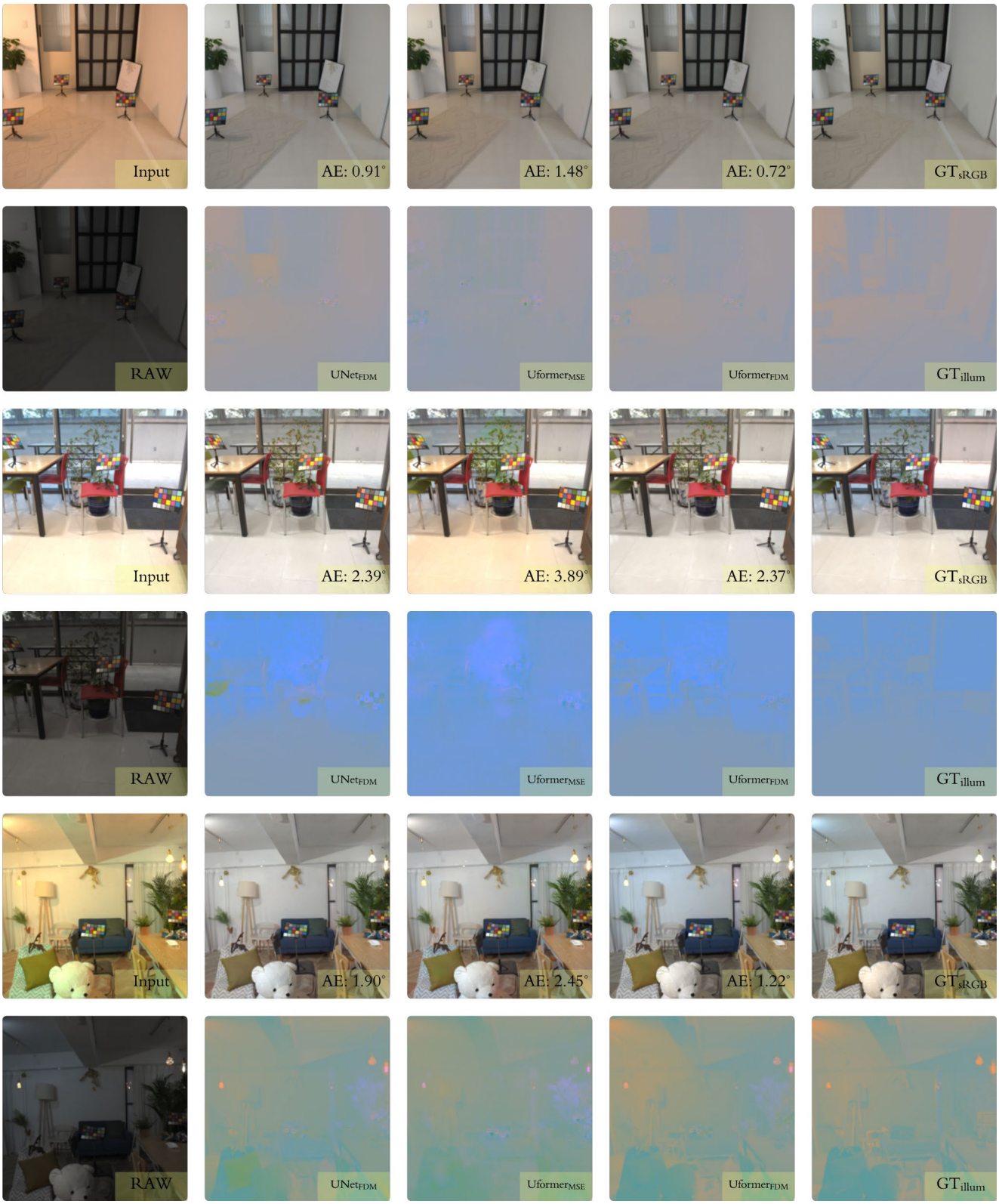


**Fig. 7** Visual comparison of WB correction outputs on the Nikon camera from the LSMI dataset. Image indices: 63, 221, 934

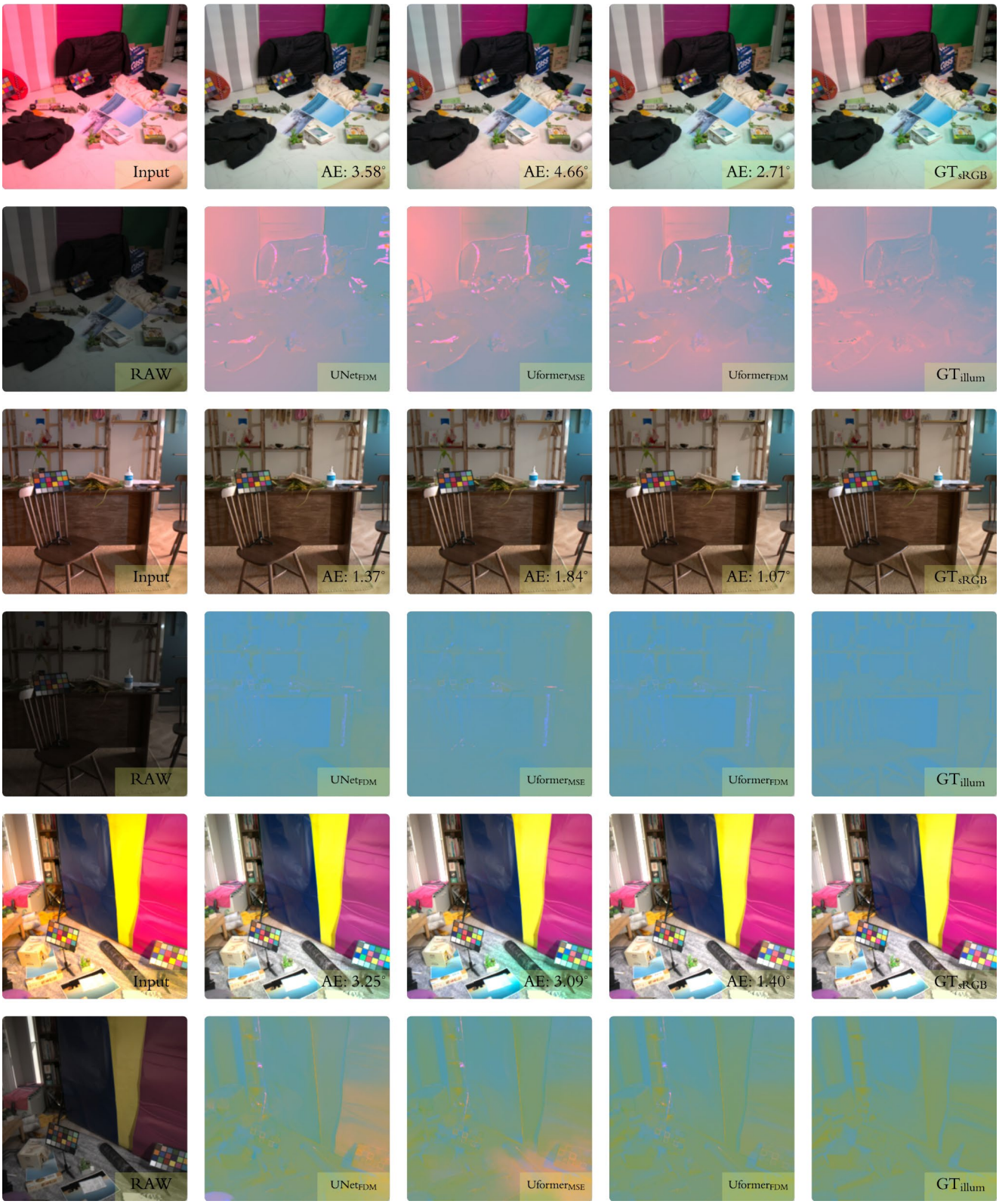


**Fig. 8** Visual comparison of WB correction outputs on the Sony camera from the LSMI dataset. Image indices: 790, 1202, 1314

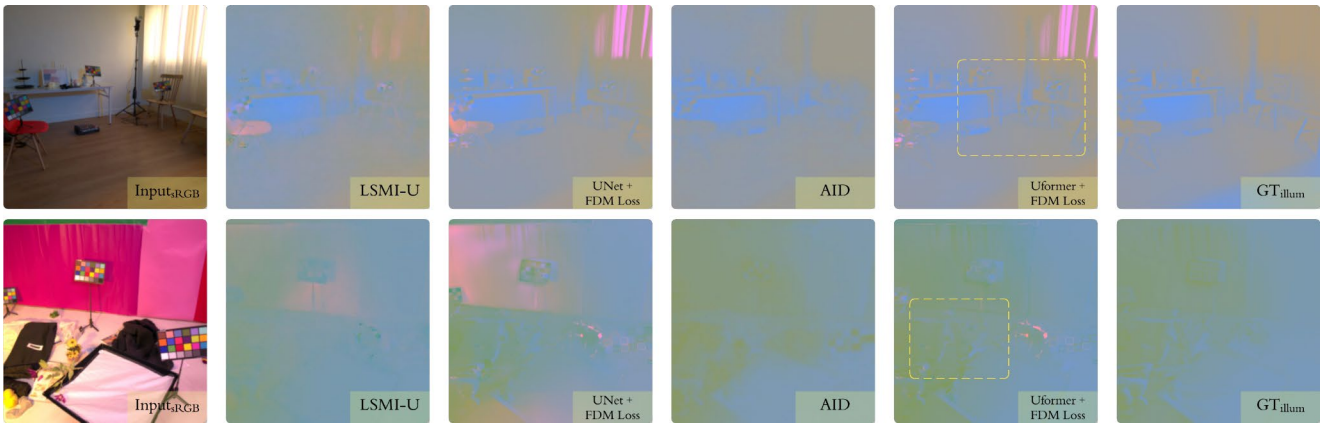


**Fig. 9** Qualitative comparison of illumination estimation results among LSMI-U [60], AID [13], and our proposed method. Image indices: 525 (Galaxy), 757 (Sony)

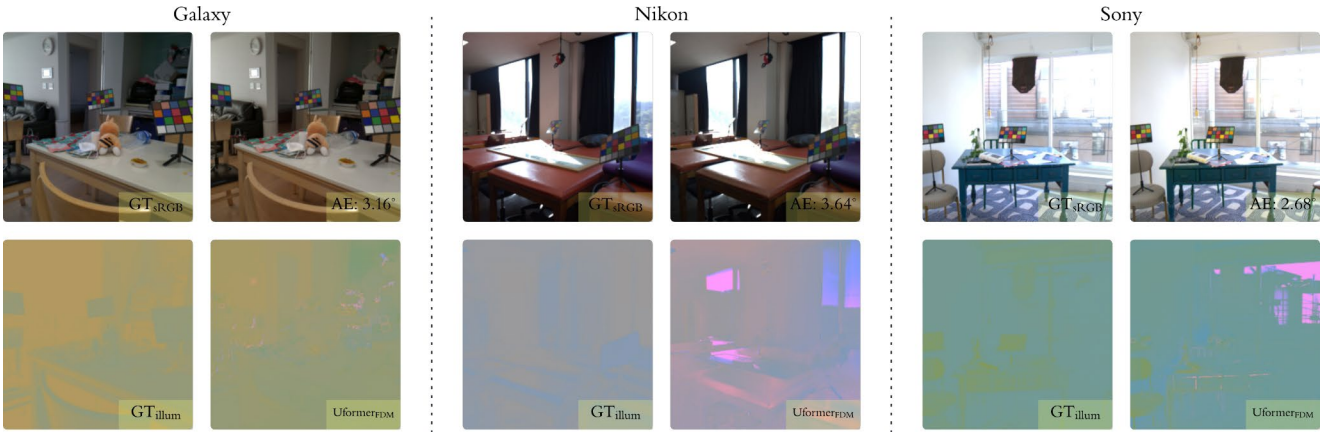


**Fig. 10** Illustration of failure cases observed in our proposed method. Image indices: 144, 20, 242

noticeable color shifts, and the angular error values remain higher than in other cases, which indicates the impact of these failures on overall performance.

One of the primary challenges observed is the difficulty in handling complex mixed illuminant regions, particularly in scenes where multiple light sources interact, such as a combination of natural daylight and artificial indoor lighting. Although the proposed method effectively aligns feature distributions in most cases, the predicted illumination maps in these scenarios can introduce artifacts. This suggests that EFDM may require further refinement to enforce consistent alignment when multiple dominant illuminants are present. As a result, corrected images may exhibit residual color shifts in highly complex lighting conditions. Additionally, in cases with strong directional illumination creating pronounced shadowed and brightly lit areas, the model can exhibit minor chromatic inconsistencies in the predicted illumination maps. This behavior does not necessarily indicate a failure to generalize, but rather highlights the challenges posed by extreme lighting variations, where global feature alignment alone may not fully capture local illumination nuances. While pixel-wise methods rely on intensity adjustments to correct for such variations, EFDM focuses on feature distribution alignment, which, although more robust to overall illumination shifts, does not explicitly enforce spatial coherence in extreme cases. Furthermore, environments with highly reflective surfaces or strong saturation can lead to slight deviations in the predicted illumination maps, particularly in regions where illumination discontinuities are highly localized. Despite these challenges, the method consistently maintains strong performance in diverse illumination conditions, and these cases suggest potential areas for further refinement, such as incorporating additional spatial priors to further enhance stability in extreme lighting scenarios.

## 5 Conclusion

This study introduces a novel approach to WB correction by leveraging feature distribution statistics as an optimization objective, addressing challenges posed by multi-illuminant

scenarios. By integrating EFDM loss into the Uformer architecture, the proposed method aligns higher-order statistical properties of feature distributions, which enables accurate and robust WB correction. Experimental results demonstrate state-of-the-art performance in multi-illuminant conditions across diverse camera setups, with consistent generalization reflected in balanced MSR values. The approach effectively balances global color consistency with local detail fidelity, and it outperforms traditional and modern methods under complex lighting scenarios. This work establishes a robust framework for WB correction and highlights the potential of distribution-based methodologies for broader image enhancement tasks, offering a foundation for future research into challenging real-world applications. Lastly, based on the insights gained from this study, future research will explore advanced methodologies to further enhance the robustness and scalability of distribution-based optimization for WB correction. Diffusion models [66, 67], which have shown great potential in modeling complex data distributions, will be investigated as a means to tackle intricate illumination scenarios. Additionally, frameworks such as flow matching [68] combined with neural ordinary differential equations (ODEs) [69] offer promising avenues for optimizing high-dimensional feature distributions.

**Funding** Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## References

1. Buchsbaum, G.: A spatial processor model for object colour perception. J. Franklin Inst. **310**(1), 1–26 (1980)
2. Afifi, M., Price, B., Cohen, S., Brown, M.S.: When color constancy goes wrong: Correcting improperly white-balanced images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1535–1544 (2019)
3. Bianco, S., Cusano, C., Schettini, R.: Color constancy using cnns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 81–89 (2015)
4. Barron, J.T.: Convolutional color constancy. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 379–387 (2015)
5. Shi, W., Loy, C.C., Tang, X.: Deep specialized network for illuminant estimation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer, pp. 371–387 (2016)
6. Bianco, S., Cusano, C., Schettini, R.: Single and multiple illuminant estimation using convolutional neural networks. IEEE Trans. Image Process. **26**(9), 4347–4362 (2017)
7. Hu, Y., Wang, B., Lin, S.: Fc4: Fully convolutional color constancy with confidence-weighted pooling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4085–4094 (2017)
8. Bianco, S., Cusano, C.: Quasi-unsupervised color constancy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12212–12221 (2019)
9. Afifi, M., Brown, M.S.: Deep white-balance editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1397–1406 (2020)
10. Lo, Y.-C., Chang, C.-C., Chiu, H.-C., Huang, Y.-H., Chen, C.-P., Chang, Y.-L., Jou, K.: CLCC: Contrastive learning for color constancy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8053–8063 (2021)
11. Afifi, M., Brubaker, M.A., Brown, M.S.: Auto white-balance correction for mixed-illuminant scenes. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1210–1219 (2022)
12. Li, S., Wang, J., Brown, M.S., Tan, R.T.: Mimt: multi-illuminant color constancy via multi-task local surface and light color learning. arXiv preprint arXiv:2211.08772 (2022)
13. Kim, D., Kim, J., Yu, J., Kim, S.J.: Attentive Illumination Decomposition Model for Multi-Illuminant White Balancing. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 25512–25521. IEEE Computer Society, Los Alamitos, CA, USA (2024). https://doi.org/10.1109/CVPR52733.2024.02410, https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.02410
14. Kinli, F., Ozcan, B., Kirac, F.: Instagram filter removal on fashionable images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 736–745 (2021)
15. Kınlı, F., Yılmaz, D., Özcan, B., Kıraç, F.: Modeling the lighting in scenes as style for auto white-balance correction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 4903–4913 (2023)
16. Zhang, Y., Li, M., Li, R., Jia, K., Zhang, L.: Exact feature distribution matching for arbitrary style transfer and domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8035–8045 (2022)
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=YicbFdNTTy
18. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17683–17693 (2022)
19. Brainard, D.H., Wandell, B.A.: Analysis of the retinex theory of color vision. JOSA A **3**(10), 1651–1661 (1986)
20. Finlayson, G.D., Trezzi, E.: Shades of gray and colour constancy. In: Color and Imaging Conference, vol. 12. Society of Imaging Science and Technology, pp. 37–41 (2004)
21. Van De Weijer, J., Gevers, T., Gijsenij, A.: Edge-based color constancy. IEEE Trans. Image Process. **16**(9), 2207–2214 (2007)
22. Xiong, W., Funt, B., Shi, L., Kim, S.-S., Kang, B.-H., Lee, S.-D., Kim, C.-Y.: Automatic white balancing via gray surface

identification. In: Color and Imaging Conference. Society of Imaging Science and Technology, vol. 15, pp. 143–146 (2007)
23. Ebner, M.: Color constancy based on local space average color. Mach. Vis. Appl. **20**(5), 283–301 (2009)
24. Gijsenij, A., Gevers, T., Van De Weijer, J.: Improving color constancy by photometric edge weighting. IEEE Trans. Pattern Anal. Mach. Intell. **34**(5), 918–929 (2011)
25. Gijsenij, A., Gevers, T., Van De Weijer, J.: Computational color constancy: survey and experiments. IEEE Trans. Image Process. **20**(9), 2475–2489 (2011)
26. Joze, H.R.V., Drew, M.S., Finlayson, G.D., Rey, P.A.T.: The role of bright pixels in illumination estimation. In: Color and Imaging Conference. Citeseer, vol. 2012, pp. 41–46 (2012)
27. Cheng, D., Prasad, D.K., Brown, M.S.: Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. JOSA A **31**(5), 1049–1058 (2014)
28. Qian, Y., Pertuz, S., Nikkanen, J., Kämäräinen, J.-K., Matas, J.: Revisiting gray pixel for statistical illumination estimation. arXiv preprint arXiv:1803.08326 (2018)
29. Forsyth, D.A.: A novel algorithm for color constancy. Int. J. Comput. Vision **5**(1), 5–35 (1990)
30. Finlayson, G.D., Hordley, S.D., Tastl, I.: Gamut constrained illuminant estimation. Int. J. Comput. Vision **67**, 93–109 (2006)
31. Finlayson, G., Hordley, S.: Improving gamut mapping color constancy. IEEE Trans. Image Process. **9**(10), 1774–1783 (2000)
32. Gijsenij, A., Gevers, T., Van De Weijer, J.: Generalized gamut mapping using image derivative structures for color constancy. Int. J. Comput. Vision **86**(2–3), 127 (2010)
33. Brainard, D.H., Freeman, W.T.: Bayesian color constancy. JOSA A **14**(7), 1393–1411 (1997)
34. Sapiro, G.: Color and illuminant voting. IEEE Trans. Pattern Anal. Mach. Intell. **21**(11), 1210–1215 (1999)
35. Finlayson, G.D., Hordley, S.D., Hubel, P.M.: Color by correlation: A simple, unifying framework for color constancy. IEEE Trans. Pattern Anal. Mach. Intell. **23**(11), 1209–1221 (2001)
36. Rosenberg, C., Hebert, M., Thrun, S.: Color constancy using kl-divergence. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001. IEEE, vol. 1, pp. 239–246 (2001)
37. Tsin, Y., Collins, R.T., Ramesh, V., Kanade, T.: Bayesian color constancy for outdoor object recognition. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. IEEE, vol. 1, p. I-I (2001)
38. Gehler, P.V., Rother, C., Blake, A., Minka, T., Sharp, T.: Bayesian color constancy revisited. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–8 (2008)
39. Gao, S., Han, W., Yang, K., Li, C., Li, Y.: Efficient color constancy with local surface reflectance statistics. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13. Springer, pp. 158–173 (2014)
40. Banić, N., Lončarić, S.: Green stability assumption: unsupervised learning for statistics-based illumination estimation. J. Imag. **4**(11), 127 (2018)
41. Afifi, M., Punnappurath, A., Finlayson, G., Brown, M.S.: As-projective-as-possible bias correction for illumination estimation algorithms. J. Opt. Soc. Am. A **36**(1), 71–78 (2018)
42. Qian, Y., Kamarainen, J.-K., Nikkanen, J., Matas, J.: On finding gray pixels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8062–8070 (2019)
43. Van De Weijer, J., Schmid, C., Verbeek, J.: Using high-level visual information for color constancy. In: 2007 IEEE 11th International Conference on Computer Vision, pp. 1–8 (2007). IEEE
44. Bianco, S., Ciocca, G., Cusano, C., Schettini, R.: Improving color constancy using indoor-outdoor image classification. IEEE Trans. Image Process. **17**(12), 2381–2392 (2008)
45. Lu, R., Gijsenij, A., Gevers, T., Nedović, V., Xu, D., Geusebroek, J.-M.: Color constancy using 3d scene geometry. In: 2009 IEEE 12th International Conference on Computer Vision. IEEE, pp. 1749–1756 (2009)
46. Rahtu, E., Nikkanen, J., Kannala, J., Lepistö, L., Heikkilä, J.: Applying visual object categorization and memory colors for automatic color constancy. In: Image Analysis and Processing–ICIAP 2009: 15th International Conference Vietri Sul Mare, Italy, September 8–11, 2009 Proceedings 15, pp. 873–882. Springer (2009)
47. Gijsenij, A., Gevers, T.: Color constancy using natural image statistics and scene semantics. IEEE Trans. Pattern Anal. Mach. Intell. **33**(4), 687–698 (2010)
48. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1501–1510 (2017)
49. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
50. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110–8119 (2020)
51. Erkol, M., Kınlı, F., Özcan, B., Kıraç, F.: [re] exact feature distribution matching for arbitrary style transfer and domain generalization. In: ML Reproducibility Challenge 2022 (2023)
52. Vaart, A.W.: Asymptotic Statistics vol. 3. Cambridge university press, Cambridge (2000)
53. Tumanyan, N., Bar-Tal, O., Bagon, S., Dekel, T.: Splicing VIT features for semantic appearance transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10748–10757 (2022)
54. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. PMLR, pp. 448–456 (2015)
55. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, pp. 297–304 (2010)
56. Ye, M., Zhang, X., Yuen, P.C., Chang, S.-F.: Unsupervised embedding learning via invariant and spreading instance feature. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6210–6219 (2019)
57. Le-Khac, P.H., Healy, G., Smeaton, A.F.: Contrastive representation learning: a framework and review. IEEE Access **8**, 193907–193934 (2020)
58. Cuturi, M.: Sinkhorn distances: lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
59. Peyré, G., Cuturi, M., et al.: Computational optimal transport: With applications to data science. Foundations and Trends ®. Mach. Learn. **11**(5-6), 355–607 (2019)
60. Kim, D., Kim, J., Nam, S., Lee, D., Lee, Y., Kang, N., Lee, H.-E., Yoo, B., Han, J.-J., Kim, S.J.: Large scale multi-illuminant (lsmi) dataset for developing white balance algorithm under mixed illumination. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2410–2419 (2021)
61. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-assisted intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, Proceedings, Part III 18. Springer, pp. 234–241 (2015)

62. Kingma, D.P.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
63. Domislović, I., Vršnjak, D., Subašić, M., Lončarić, S.: Color constancy for non-uniform illumination estimation with variable number of illuminants. Neural Comput. Appl. **35**(20), 14825–14835 (2023). https://doi.org/10.1007/s00521-023-08487-z
64. Li, J., Fang, P.: Hdrnet: Single-image-based hdr reconstruction using channel attention cnn. In: ICMSSP '19, pp. 119–124. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3330393.3330426 (2019)
65. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
66. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
67. Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-based style transfer with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10146–10156 (2023)
68. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
69. Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. Advances in neural information processing systems **31** (2018)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.