ReHiT: Retinex-guided Histogram Transformer for Mask-free Shadow Removal

Retinex-guided Histogram Transformer for Mask-free Shadow Removal

By

Seyed Amirreza Mousavi,

M.A.Sc (Electrical and Computer Engineering)

A Thesis Submitted to the School of Graduate Studies in the Partial Fulfillment of the Requirements for the Degree M.A.Sc.

August 19, 2025

McMaster University

M.A.Sc. (2025)

Hamilton, Ontario (Electrical and Computer Engineering)

TITLE: Retinex-guided Histogram Transformer for Mask-free Shadow Removal

AUTHOR: Seyed Amirreza Mousavi (McMaster University)

SUPERVISOR: Prof. Jun Chen

NUMBER OF PAGES: xii, 52

To my dear wife and parents

Abstract

While deep learning techniques have significantly advanced the field of shadow removal, a considerable number of current methods depend on shadow masks, which are often challenging to acquire accurately. This reliance on masks restricts their ability to generalize effectively to unconstrained real-world scenarios. To address this limitation, we introduce **ReHiT**, an efficient mask-free shadow removal framework that leverages a hybrid CNN-Transformer architecture, guided by the principles of Retinex theory. Our approach begins with a dual-branch pipeline designed to model the reflectance and illumination components of an image separately. Each of these components is then processed and restored by our novel Illumination-Guided Hybrid CNN-Transformer (IG-HCT) module. Furthermore, in addition to incorporating CNN-based blocks that excel at learning residual dense features and performing multi-scale semantic fusion, we have developed the Illumination-Guided Histogram Transformer Block (IGHB). This specialized block is designed to effectively handle the complexities of non-uniform illumination and spatially intricate shadow patterns. Comprehensive experiments conducted on several standard benchmark datasets demonstrate the superior performance of our proposed method compared to existing mask-free techniques. Notably, our solution achieves competitive results while boasting one of the smallest parameter counts and fastest inference speeds among the state-of-the-art models. This highlights the practical applicability of our method for real-world applications where computational resources may be constrained.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Jun Chen, for his exceptional guidance, encouragement, and invaluable support throughout the course of my research and the writing of this thesis. His profound knowledge, insightful feedback, and constant motivation have greatly contributed to both my academic growth and the successful completion of this work. I am sincerely grateful for the opportunities he has provided me and for the trust he placed in my abilities.

I would also like to extend my heartfelt thanks to my wife, whose unwavering love, patience, and encouragement have been my source of strength throughout this challenging journey. Her understanding and support during the long hours of research and writing were crucial in helping me persevere and achieve this milestone. Without her sacrifices and constant belief in me, this achievement would not have been possible.

Furthermore, I would like to thank Dr. Tim Davidson and Dr. Dongmei Zhao not only for being members of my defense committee but also for reviewing my thesis, and providing technical comments.

Contents

List of Figures

List of Tables

Abbreviations

CNN	Convolutional Neural Network
IG-HTBs	Illumination-Guided Histogram Transformer Blocks
DRDB	Dilated Residual Dense Block
SAM	Semantic-Aligned Scale-Aware Module
SL	Supervised Learning
UL	Unsupervised Learning
SSL	Semi-Supervised Learning
ST-CGAN	Stacked Conditional Generative Adversarial Networks
CGANs	Conditional Generative Adversarial Networks
MAdaIN	Masked Adaptive Instance Normalization
CPM	Contextual Patch Matching
CFT	Contextual Feature Transfer
ALN	Ambient Lighting Normalization
IG-HCT	Illumination-Guided Hybrid CNN-Transformer
IG-HTB	Illumination-Guided Histogram Transformer Block
IG-HSA	Illumination-Guided Histogram Self-Attention

FFN	Feed-Forward Network
LN	Layer Normalization
RDB	Residual Dense Block
MS-SSIM	Multi-Scale Structural Similarity Index
SSIM	Structural Similarity Index Measure
PSNR	Peak Signal-to-Noise Ratio
LPIPS	Learned Perceptual Image Patch Similarity
SOTA	State-of-the-Art
FLOPs	Floating Point Operations

Declaration of Authorship

I, Seyed Amirreza Mousavi, declare that this thesis titled, "Retinex-guided Histogram Transformer for Mask-free Shadow Removal" and the work presented in it are my own.

Chapter 1

Introduction and Problem Statement

1.1 Introduction

Shadows represent a prevalent visual phenomenon observed in natural environments, arising when a source of illumination is either partially or entirely blocked by physical objects present within the scene. The presence of shadows can introduce substantial complications and pose significant challenges for a diverse range of high-level computer vision applications, including, but not limited to, object tracking, object detection, and semantic segmentation [26, 13, 54]. As a direct consequence of these challenges, the task of shadow removal has evolved into a fundamental and critical problem within the field of image restoration.

Preceding the widespread adoption of deep learning methodologies [16, 15, 18, 21, 57], conventional techniques for shadow removal predominantly depended on manually designed discriminative prior knowledge. These priors were utilized to

identify and rectify shadows by analyzing image characteristics such as edges, intensity values, and geometric properties [30, 65]. Furthermore, physics-based illumination models [28] were frequently employed to estimate and subsequently compensate for the disparities in lighting conditions observed between regions affected by shadows and those that were not. Nevertheless, these earlier methodologies often encountered limitations and performed suboptimally when applied to complex real-world scenarios. This was largely attributable to the oversimplified assumptions inherent in these approaches and the intrinsic difficulty associated with modeling the intricate variations in illumination that occur in natural scenes.

In recent years, learning-based approaches [9, 24, 33, 36, 39, 49, 38, 45] have emerged as a dominant paradigm in shadow removal. These approaches effectively harness the considerable representational power inherent in deep neural network architectures. Both convolutional neural network (CNN)-based techniques [45] and methods employing Transformer architectures [20, 11] have shown noteworthy achievements in learning the intricate relationship between images containing shadows and their corresponding shadow-free versions through an end-to-end learning paradigm. These deep learning strategies can be broadly classified into two main groups: those that utilize masks [39] and those that operate without explicit masks for shadow removal [11, 45]. Mask-based methodologies leverage paired datasets consisting of images with shadows and their clean counterparts, in conjunction with explicit shadow masks that are either manually labeled or produced by pre-trained models, to direct the learning procedure. The integration of precise shadow masks enables these models to concentrate on acquiring the complex transformation between the regions affected by shadows and the clean areas, thereby

achieving cutting-edge performance. Nevertheless, this enhanced performance is accompanied by a significant drawback: these methods encounter difficulties in obtaining accurate shadow masks, particularly in intricate real-world scenarios where manual annotation can be laborious and automatic predictions may lack reliability.

On the other hand, current deep learning approaches often fail to fully incorporate the underlying physics of illumination and shadows. Many end-to-end models [27], though effective, struggle with generalization across diverse lighting conditions, leading to artifacts along shadow boundaries. Similarly, physics-based models [21, 57], despite leveraging certain illumination properties, rely on overly simplistic assumptions such as uniform lighting within shadow regions and basic linear transformations for illumination correction. These limitations motivate the development of more sophisticated shadow removal methods that can effectively handle intricate lighting variations and complex scene geometries.

Traditional Convolutional Neural Networks (CNNs) have been the cornerstone of many image restoration tasks, demonstrating remarkable success in various applications. However, their architectural design, primarily based on local receptive fields, inherently limits their ability to capture long-range dependencies and non-local self-similarity within an image. These are critical characteristics, as image restoration often necessitates understanding relationships between distant pixels or identifying repetitive patterns across an entire image to accurately infer missing or corrupted information. The emergence of the Transformer architecture, initially popularized in natural language processing and more recently adapted for computer vision as Vision Transformers (ViTs), offers a compelling solution

to these CNN limitations. Transformers excel at modeling long-range dependencies through their self-attention mechanism, which allows each pixel (or patch) to interact with every other pixel (or patch) in the input. This global connectivity is theoretically ideal for capturing the non-local information crucial for advanced image restoration.

However, the direct application of original ViT architectures to tasks like shadow removal presents a significant practical hurdle: computational complexity. The self-attention mechanism in standard Transformers involves calculating attention scores between all pairs of input tokens, leading to a computational cost that scales quadratically with the input spatial size $O((N^2)$, where N is the number of tokens/pixels). For high-resolution images, this computational burden becomes prohibitively expensive and memory-intensive, making it unaffordable for real-world applications or even large-scale research.

To tackle these challenges, we introduce **ReHiT**, an efficient two-branch mask-free shadow removal network based on illumination-guided hybrid CNN-Transformer architecture. We first extend and analyze Retinex theory [28] and develop a Retinex estimator to convert the input into two intermediate representations, each approximating the target reflectance and illumination map [14]. Second, we present a hybrid CNN-Transformer network, guided by Retinex information, as the core restoration framework of our method.

Within each block of this UNet encoder-decoder architecture, we develop the Illumination-Guided Histogram Transformer Blocks (IG-HTBs) to integrate the illumination guidance and employ the CNN-based Dilated Residual Dense Block

(DRDB) and Semantic-Aligned Scale-Aware Module (SAM) proposed in [59] multi-scale feature fusion. Figure 1.1 shows the result of our method applied to shadowed images.

Our contribution can be summarized as follows:

1. We introduce **ReHiT**, an efficient CNN-Transformer hybrid architecture for mask-free shadow removal.

2. Guided by the principles of Retinex theory, our approach employs a dual-branch restoration pipeline, where a hybrid CNN-Transformer network is responsible for the restoration process in each branch.

3. We develop an illumination-guided histogram Transformer specifically to perceive and recover the shadow regions within the primary restoration network.

4. Extensive experiments conducted across multiple established shadow removal benchmark datasets demonstrate the effectiveness and superiority of our proposed method

1.2 Thesis Structure

Within this thesis, Chapter 2 will provide a concise overview of established shadow removal models and relevant scholarly contributions to the field. Subsequently, Chapter 3 will present a detailed exposition of the architecture and operational principles of our novel dual-branch Retinex-guided Histogram Transformer model. Chapter 4 will be dedicated to the in-depth analysis of our proposed network and will feature a comparative evaluation of its performance against state-of-the-art

methodologies. Finally, Chapter 5 will serve as the concluding section, summarizing the key findings and contributions of this research.

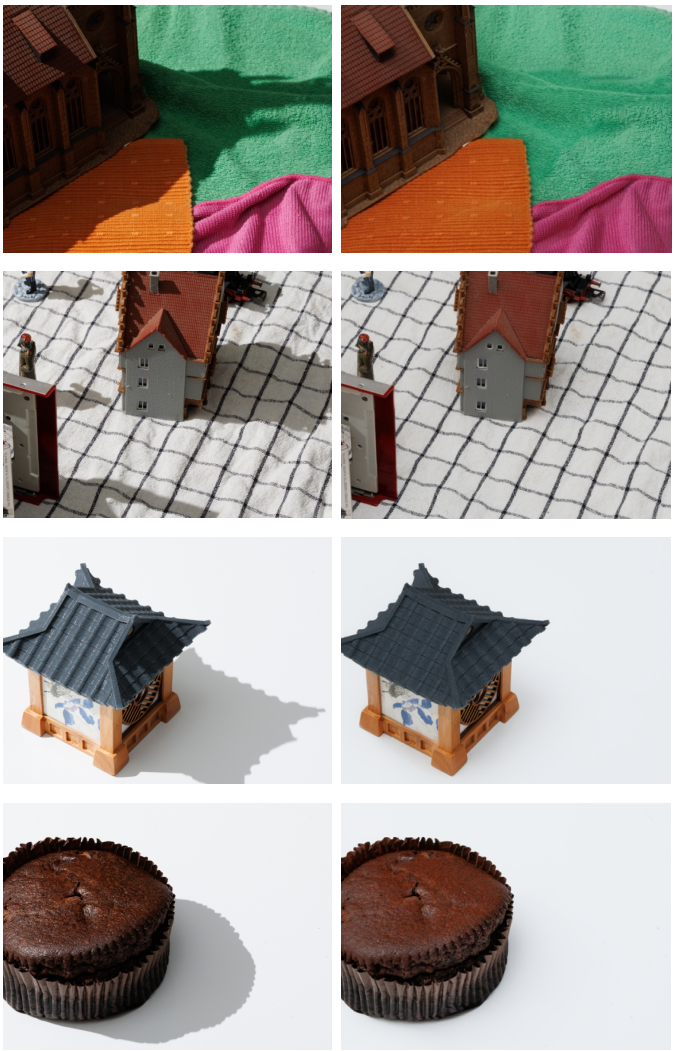

FIGURE 1.1: Visual result of our shadow removal model

Chapter 2

Related Works

Image shadow removal techniques generally fall into two categories: traditional and deep learning-based methods. A detailed examination of both approaches is presented below.

2.1 Traditional Image Shadow Removal

Early shadow removal methods [16, 15, 18, 21, 57] primarily leverage prior knowledge of an image's physical properties and the underlying principles of illumination to differentiate between shadowed and lit regions. These approaches often model shadow formation based on assumptions about lighting conditions, geometric relationships, and inherent image characteristics like gradients and color constancy.

One notable line of research focuses on illumination modeling and region pairing. Guo et al. [21] introduced a method for shadow detection and removal by identifying and analyzing "paired regions" within an image. Their core idea was to find corresponding shadowed and non-shadowed areas that share similar material

properties, allowing them to model the illumination change caused by the shadow. By establishing these relationships, they could then infer the intrinsic reflectance of the scene and reconstruct the shadow-free image by effectively "relighting" the shadowed regions. While innovative in its approach to leverage specific region relationships, the success of such methods can be sensitive to the accurate identification of these paired regions and the complexity of real-world illumination.

Another prominent strategy, exemplified by the work of Finlayson et a l.[16, 15], capitalizes on the principle of illumination invariance, often derived from gradient properties. Their methods, such as the one based on entropy minimization [15], aim to transform image data into a representation where the underlying material properties are separated from illumination changes (shadows). This is frequently achieved by exploiting the consistency of gradients or color ratios which are ideally invariant to changes in illumination. By identifying and manipulating these photometric cues, particularly around shadow boundaries, they seek to recover the original scene radiance. However, a significant limitation arises when the assumption of consistent gradient behavior is violated, often due to complex illumination interactions or variations in surface properties within the shadow, which can lead to undesirable artifacts like noticeable shadow boundary lines or color inconsistencies in the resulting shadow-free image.

To address the limitations of fully automatic approaches, some traditional methods incorporate user interaction. Gong and Cosker [18], for example, aimed to improve the robustness and accuracy of shadow removal, especially for "difficult shadow scenes," by integrating distinct forms of user-provided input. This human

guidance, such as scribbles indicating shadowed and lit regions, helps the algorithm better distinguish between genuine shadows and intrinsic scene properties (e.g., dark objects or textures), thereby enhancing the resilience of the removal process. While improving accuracy and handling challenging cases, the requirement for manual input can limit their applicability in large-scale or real-time scenarios where automation is paramount.

Furthermore, a broader category of traditional methods relies on analyzing various image features. Techniques rooted in color constancy, such as that explored by Zhao et al. [64], attempt to estimate the scene illuminant and then normalize the image colors to remove the influence of the light source, thereby making the colors intrinsic to the object regardless of shadow. This is based on the idea that human perception maintains object color despite changes in illumination. Other approaches leverage texture analysis and edge detection, as seen in the work of Wu et al. [3] and [52] . Shadows primarily alter illumination but ideally preserve texture details. By analyzing texture patterns, algorithms can identify regions where illumination changes (due to shadows) without significant changes in underlying texture. Similarly, edge detection can be used to identify shadow boundaries, which are often characterized by strong intensity gradients, and then smooth or remove these specific gradients while preserving true object edges. However, distinguishing between shadow edges and genuine object edges remains a persistent challenge for these methods, as both can manifest as strong intensity discontinuities.

2.2 Deep Learning-based Shadow Removal

This section will begin by formally defining the problem of single-image shadow removal. Following this, we will proceed to review and engage in a discussion of the current landscape of existing shadow removal methodologies. We begin by problem definition:

For a shadow image $I_s \in \mathbb{R}^{H\times W\times 3}$ with H, W as height and width respectively, we can model the process of shadow removal as:

$$\hat{I}_{sf} = f(I_s\ ;\ \theta), \tag{2.1}$$

Or:

$$\hat{I}_{sf} = f(I_s\ , M;\ \theta) \tag{2.2}$$

where $\hat{I}_{sf} \in \mathbb{R}^{H\times W\times 3}$ denotes the resulting shadow-free image after restoration, and f symbolizes the shadow removal network parameterized by a set of learnable parameters θ. To facilitate this process of shadow identification and restoration, an optional shadow mask $M \in \mathbb{R}^{H\times W}$ can be incorporated as supplementary information. This mask serves to indicate the regions within the image that are affected by shadows, and it can be derived either through manual annotation by a human expert or automatically detected by a pre-trained shadow detection model. In contrast to other image restoration tasks that typically deal with global image

degradation, shadow removal presents a unique challenge as a partial corruption problem. This necessitates not only the accurate identification of the image regions affected by shadows but also their subsequent restoration to a shadow-free state. Based on different learning strategies, we generally categorize existing image shadow removal methods into supervised learning, unsupervised learning and semi-supervised learning.

Supervised Learning (SL) In the context of shadow removal, supervised learning involves training a model using pairs of images depicting the same scene, one with shadows and the other without, captured under varying illumination. A groundbreaking early work leveraging deep learning, known as DeShadowNet[40], introduced an automated and end-to-end deep neural network designed to integrate the tasks of shadow detection, classification of umbra and penumbra regions within shadows, and the actual removal of shadows. This network directly learns the mapping function that transforms an image containing shadows into its corresponding shadow matte.

Unsupervised Learning (UL) Although supervised learning methodologies have demonstrated significant achievements across a range of applications, they are fundamentally dependent on the availability of substantial quantities of paired training data. The acquisition of such paired data can be a resource-intensive and time-consuming undertaking. Furthermore, the training of deep learning models on paired datasets is generally tailored to address specific tasks. Consequently, these models often encounter difficulties in generalizing and adapting effectively to novel or out-of-distribution scenarios without undergoing retraining.

Semi-Supervised Learning (SSL) In recent years, semi-supervised learning has emerged as an approach to leverage the strengths of both supervised learning and unsupervised learning. It leverages both paired data and unpaired data to boost the model performance and improve generalization ability.

2.2.1 Mask-based Image Shadow Removal

Over the past few years, a significant number of deep neural network models [9, 24, 33, 36, 39, 49] have been proposed to tackle the task of shadow removal. These methods commonly employ both supervised and unsupervised training strategies and can be broadly classified into two categories: mask-based and mask-free approaches. Gryka et al. [19] introduced a learning-based technique for automatic shadow removal, utilizing a supervised regression algorithm to effectively handle both umbra and penumbra shadow types. ST-CGAN [49] presented an integrated framework for shadow detection and removal by employing two stacked Conditional Generative Adversarial Networks (CGANs). Wan et al. [48] addressed the problem of inconsistent static styles between shadowed and shadow-free areas by proposing a style-guided network for shadow removal. S2Net [4] focused on leveraging semantic guidance and refinement to preserve the overall integrity of the image and utilized shadow masks to guide the shadow removal process, employing semantic-guided blocks to facilitate information transfer from non-shadowed to shadowed regions. He et al. [23] developed Mask-ShadowNet, which aims to maintain global illumination consistency through the use of Masked Adaptive Instance Normalization (MAdaIN) and adaptively refines features using alignment modules. Furthermore, FusionNet [17] employed fusion weight maps and a boundary-aware RefineNet to further minimize any residual shadow traces. However, a significant

limitation of these approaches is their strong dependence on the accuracy of the input shadow masks. The inherent complexity and diversity of real-world scenes often make it challenging to generate precise shadow masks, which can consequently impact the overall effectiveness and reliability of these methods.

2.2.2 Mask-free Image Shadow Removal

Mask-free methods have demonstrated greater adaptability and promise across a wider range of scenarios. CANet [7] incorporates a Contextual Patch Matching (CPM) module to locate corresponding patches in shadowed and non-shadowed areas and a Contextual Feature Transfer (CFT) mechanism to propagate contextual information between these regions. Vasluianu et al. [45] introduced Ambient Lighting Normalization (ALN) to improve image restoration under complex lighting conditions and proposed IFBlend, an advanced image enhancement framework that optimizes the joint entropy of the image and its frequency components, thereby enhancing visual quality without requiring explicit shadow localization. Le et al. [29] utilized an illumination model in conjunction with image decomposition techniques to effectively restore regions affected by shadows. Liu et al. presented a shadow-aware decomposition network designed to disentangle the illumination and reflectance components of an image, facilitating a more accurate reconstruction of the scene's lighting. This is further enhanced by a bilateral correction network, which refines the consistency of the lighting and restores textural details, resulting in a more natural and perceptually coherent output. ShadowRefiner [11] employs a UNet architecture built upon ConvNext [37, 68, 12], utilizing multi-scale ConvNext blocks as powerful encoders for learning robust latent feature representations.

Transformer-based Image Restoration. Transformer-based networks, which leverage self-attention mechanisms to capture complex relationships between different components, have demonstrated unparalleled efficacy in modeling long-range dependencies [2, 1, 46, 67, 14, 10, 31, 34, 56]. Their superior ability to understand contextual relationships has led to state-of-the-art performance in image restoration, surpassing traditional architectures in both accuracy and robustness. SwinIR [32], a widely recognized backbone for image restoration, is constructed using a series of residual Swin Transformer [35] blocks, leveraging hierarchical feature representation for enhanced performance. Building upon the Vision Transformer [47] framework, DehazeFormer [42] has been introduced to address the image dehazing task, demonstrating superior capability in atmospheric degradation removal. Guo et al. [20] propose Shadowformer to exploit non-shadow regions to help shadow region restoration. More recently, a lightweight transformer architecture [5] has been proposed for low-light image enhancement, effectively capturing illumination and reflectance characteristics to improve visual quality under challenging lighting conditions. In [11], a Fast-Fourier attention transformer structure is used in an encoder-decoder architecture to further refine image details and maintain color consistency after shadows are removed. Sun et al. [43] propose a histogram self-attention mechanism to categorize spatial elements into bins and allocate varying attention within and across bins.

Chapter 3

Research Methodology

This section details our novel approach to high-quality image shadow removal. Achieving robust shadow removal necessitates the use of advanced deep learning architectures capable of effectively extracting crucial features from shadowed images and modeling the complex relationship between these inputs and their shadow-free counterparts. The overarching architecture of our proposed method is visually represented in Figure Fig. 3.1. Our solution is built upon the foundational principles of Retinex theory and is implemented within an Illumination-Guided Shadow Removal framework. This framework features two distinct, parallel pathways [14], as discussed in (Sec. 3.1). Each pathway is specifically designed for the independent restoration of either the scene's reflectance map (representing the true colors of objects independent of illumination) or the illumination map (capturing the light distribution across the scene, including shadows). Retinex theory plays a pivotal role in our shadow removal process by providing a theoretical basis to disentangle the intrinsic reflectance properties of objects from variations in illumination. This fundamental separation allows our network to better understand and isolate

the shadow component. The initial decomposition provided by Retinex theory significantly aids the subsequent refinement process, which is performed by our novel Illumination-Guided Hybrid CNN-Transformer (IG-HCT) modules Sec. 3.2. Within each IG-HCT module, we integrate a key component: the Illumination-Guided Histogram Transformer Block (IG-HTB). This innovative block employs an illumination-guided, histogram-based self-attention mechanism, which allows it to adaptively focus on relevant features while considering the illumination context. The IG-HTB is strategically combined with a Convolutional Neural Network (CNN)-based Dilated Residual Dense Block (DRDB), designed for robust feature extraction and multi-scale contextual understanding, and a Semantic-aligned Scale-aware Module (SAM), which further refines features by aligning them with semantic information and handling variations across different scales. This synergistic combination of components within the IG-HCT modules is engineered to significantly enhance the overall shadow removal performance and fidelity of our proposed network Sec. 3.2.1.

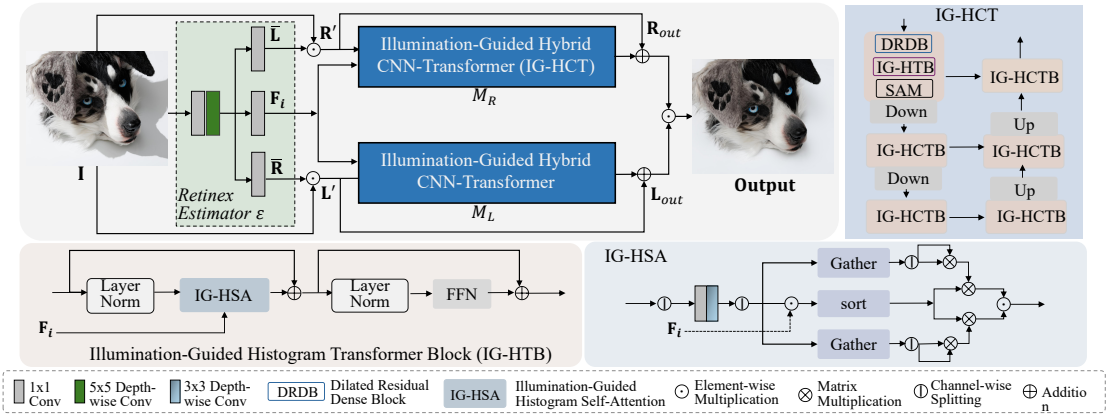

FIGURE 3.1: ReHiT Architecture

3.1 Dual-branch Retinex-based Pipeline

The Retinex theory can be expressed as $I_{GT} = R_{GT} \odot L_{GT}$, where I_{GT} is an ideal image without shadow, R_{GT} and L_{GT} represent the reflectance image and illumination map, respectively. However, a shadowed image I_{Sh} captured under non-ideal illumination conditions inevitably suffers from noise, color distortion, and constrained contrast. Therefore, as in [14], perturbations ($\hat{R}$ and $\hat{L}$) are introduced to model these shadowed images as:

$$\begin{aligned} I_{Sh} &= (R_{GT} + \hat{R}) \odot (L_{GT} + \hat{L}) \\ &= R_{GT} \odot L_{GT} + R_{GT} \odot \hat{L} + \hat{R} \odot L_{GT} + \hat{R} \odot \hat{L} \end{aligned} \tag{3.1}$$

To achieve satisfactory results, we simultaneously restore the reflectance and illumination components. This is done by element-wise multiplying both sides of Equation (3.1) by $\bar{L}$ and $\bar{R}$, respectively:

$$\begin{aligned} \mathrm{I_{Sh}} \odot \bar{L} = R' = \mathrm{R_{GT}} + \mathrm{R_{GT}} \odot \hat{L} \odot \bar{L} + \hat{R} + \hat{R} \odot \hat{L} \odot \bar{L}, \\ \mathrm{I_{Sh}} \odot \bar{R} = L' = \mathrm{L_{GT}} + \hat{R} \odot \mathrm{L_{GT}} \odot \bar{R} + \hat{L} + \hat{R} \odot \hat{L} \odot \bar{R}. \end{aligned} \tag{3.2}$$

After introducing $\bar{L}$ and $\bar{R}$ such that $\bar{L} \odot L_{GT} = 1$ and $\bar{R} \odot R_{GT} = 1$ and under the assumption that we can approximate $\bar{L}$ and $\bar{R}$ via Retinex estimator,

the results can be retrieved using deep learning networks by:

$$
\begin{aligned}
&(\bar{R}, \bar{L}, F_i) = \mathcal{E}(I_{Sh}), \\
&R' = I_{Sh} \odot \bar{L}, \;\; L' = I_{Sh} \odot \bar{R}, \\
&R_{out} = R' + \mathcal{M}_R(R'; F_i), \\
&L_{out} = L' + \mathcal{M}_L(L'; F_i), \\
&I_{out} = R_{out} \odot L_{out},
\end{aligned}
\tag{3.3}
$$

where $\mathcal{M}_R$ and $\mathcal{M}_L$ are networks utilized to predict the minus degradation in R' and L', and F_i serves as Retinex guidance information derived from I_{Sh}.

3.2 Illumination-Guided Hybrid CNN-Transformers (IG-HCT) Module

Our developed IG-HCT module is an encoder decoder architecture and serves as the $\mathcal{M}_R$ or $\mathcal{M}_L$ in Eq. 3.3. This module consists of three down-sampling and up-sampling levels. At each decoder level, the network produces intermediate results through a convolution layer and a pixelshuffle up-sampling operation, which are also supervised by the ground-truth, serving the purpose of deep supervision to facilitate training. Specifically, each encoder or decoder block, IG-HCTB (top right in Fig. 3.1), contains a Dilated Residual Dense Block (DRDB) [59] for refining the input features, an Illumination Guided Histogram Transformer Block (IG-HTB, introduced in the next subsection) to better capture dynamically distributed shadow-induced degradation, and a Semantic-aligned multi-scale module (SAM) [59] for extracting and dynamically fusing multi-scale features at the same

semantic level.

Dilated Residual Dense Block (DRDB) For each level $i \in \{1, 2, 3, 4, 5, 6\}$ (i.e., three encoder levels and three decoder levels), the input feature F_i is initially processed by a convolutional block. This convolutional block, specifically a dilated residual dense block, is designed to refine the input features. It integrates the structure of the residual dense block (RDB) [62, 25, 22] and incorporates dilated convolution layers [58] to effectively process the input features and produce refined output features. The refined feature representation is then fed to IG-HTB. More formally, if we denote the input feature to the $i-th$ level encoder or decoder as F_i^0, the sequence of cascaded local features generated from each layer within this block can be mathematically formulated as follows:

$$F_i^l = C^l([F_i^0, F_i^1, \ldots, F_i^{l-1}]), \quad (l = 1, 2, \ldots, L), \tag{3.4}$$

where $[F_i^0, F_i^1, \ldots, F_i^{l-1}]$ denotes the concatenation of all intermediate features inside the block before layer l, and C^l is the operator to process the concatenated features, consisting of a 3×3 convolution with dilation rate d_l, followed by a ReLU activation function. Subsequently, a 1×1 convolution is applied to ensure the output channel number matches that of the initial input feature F_i^0. Finally, we utilize a residual connection to produce the refined feature representation F_i^r, which can be formulated as:

$$F_i^r = F_i^0 + W([F_i^0, F_i^1, \ldots, F_i^k]), \tag{3.5}$$

where $W(\cdot)$ represents the 1×1 convolutional layer applied to the concatenated features. The refined feature representation F_i^r is then fed to our Illumination-Guided Histogram Transformer Block (IG-HTB).

3.2.1 Illumination-Guided Histogram Transformer Block (IG-HTB)

As the core element of our IG-HCT module, IG-HTB consists of two essential mechanisms: IG-HSA and FFN. These components are structured to engage with layer normalization and can be expressed as the following.

$$\begin{aligned} F_i &= F_{i-1} + \textit{IG-HSA}(LN(F_{i-1})), \\ F_i &= F_i + FFN(LN(F_i)), \end{aligned} \tag{3.6}$$

where $LN(\cdot)$ denotes layer normalization and F_i represents the feature at i-th level.

Illumination-guided Histogram Self-Attention To more effectively capture shadow-induced degradation that varies dynamically, we develop an illumination-guided Histogram Self-Attention (IG-HSA) mechanism. This layer incorporates a dynamic-range convolution process, which reorganizes the spatial arrangement of fractional features, along with a histogram self-attention mechanism that integrates both global and local dynamic feature aggregation. Traditional convolution, which primarily focuses on local information, does not naturally complement the self-attention mechanism's capability to model long-range dependencies. To address this limitation, we use a dynamic-range convolution approach that restructures input features before applying standard convolution operations. Moreover,

the illumination information extracted in Sec. 3.1 is integrated to modulate the attention calculation process.

Contrary to most existing vision Transformers [6, 50, 53, 60, 63], which leverage fixed range of attention which restricts the self-attention to span adaptively long range to associate desired features, we have noticed that shadow-induced degradation had better be assigned with various extent of attention. We thus propose a histogram self-attention mechanism to categorize spatial elements into bins and allocate varying attention within and across bins. For the sake of parallel computing, we set each bin contains identical number of pixels during implementation.

Semantic-aligned Multi-scale (SAM) Block The transformed feature representations generated from IG-HTB is given to the SAM [59] block to extract multi-scale features within the same semantic level i and allow them to interact and be dynamically fused, significantly improving the model's ability to handle shadow induced patterns. SAM encompasses two major modules pyramid context extraction and cross-scale dynamic fusion. To extract features at multiple scales, we employ a pyramid context extraction method. Starting with an initial feature map $F_r \in \mathbb{R}^{H\times W\times C}$, we generate a series of pyramid input features at progressively lower resolutions: F_r, $F_r{\downarrow} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$, and $F_r{\downarrow\downarrow} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$. These downsampled features are created using bilinear interpolation.

Each of these pyramid input features is then processed by a dedicated convolutional branch, each consisting of five convolutional layers. This process yields three corresponding pyramid outputs: Y_0, Y_1, and Y_2. This can be represented as:

$$\begin{aligned} Y_0 &= E_0(F_r), \\ Y_1 &= E_1(F_r{\downarrow}), \\ Y_2 &= E_2(F_r{\downarrow\downarrow}). \end{aligned} \tag{3.7}$$

Here, E_0, E_1, and E_2 are constructed using a dilated dense block followed by a 1×1 convolutional layer. To ensure that all three outputs have the same spatial dimensions as the original feature map ($H \times W \times C$), up-sampling operations are incorporated within E_1 and E_2.

Importantly, the internal architectures of E_0, E_1, and E_2 are identical, allowing their learnable parameters to be shared. This parameter sharing significantly reduces the total number of parameters, making the model more efficient. The performance gains observed are primarily attributable to the multi-scale pyramid architecture itself, rather than an increase in the number of learnable parameters.

After extracting pyramid features (Y_0, Y_1, Y_2), the cross-scale dynamic fusion module takes over to combine them. This module is designed to produce a fused, multi-scale feature representation for subsequent processing stages. The core idea behind this dynamic approach is that the scale of shadow patterns can differ significantly from image to image. Consequently, the importance of features extracted at various scales will also vary across different images. To address this, our module dynamically adjusts and adapts the fusion process for each individual image.

Specifically, we learn dynamic weights to intelligently fuse Y_0, Y_1, and Y_2. Given each pyramid feature $Y_i \in \mathbb{R}^{H \times W \times C}$ (where $i = 0, 1, 2$), we first apply global

average pooling across the spatial dimensions of each feature map. This step yields a 1D global feature $v_i \in \mathbb{R}^C$ for each scale, as shown in Eq. 3.8:

$$v_i = \frac{1}{H \times W} \sum_{s=1}^{H} \sum_{t=1}^{W} Y_i(s,t) \tag{3.8}$$

Next, these global features are concatenated along the channel dimension. A Multi-Layer Perceptron (MLP) module then learns the dynamic weights from this concatenated vector. The MLP consists of three fully connected layers and outputs $w_0, w_1, w_2 \in \mathbb{R}^C$, which are the dynamic weights used to fuse Y_0, Y_1, and Y_2. This process is described in Equation 5:

$$[w_0, w_1, w_2] = \text{MLP}([v_0, v_1, v_2]) \tag{3.9}$$

Finally, these input-adaptive fusion weights are used to channel-wise multiply and combine the pyramid features. The initial input feature F_r is then added to this fused representation to produce the final output of the Scale-Aware Module (SAM), denoted as F^{out}. This is shown in Eq. 3.10:

$$F^{\text{out}} = F^r + w_0 \odot Y_0 + w_1 \odot Y_1 + w_2 \odot Y_2 \tag{3.10}$$

Here, $\odot$ represents channel-wise multiplication. This F^{out} then proceeds to the next level (from level i to $i+1$) for further feature extraction and, ultimately, image reconstruction.

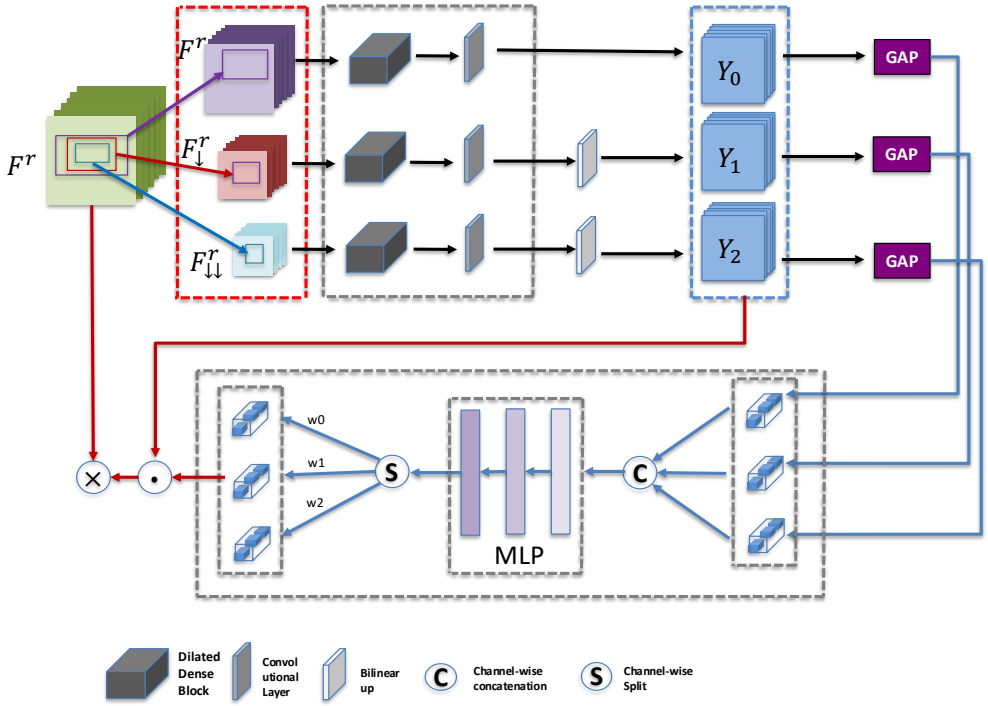

FIGURE 3.2: The architecture of SAM module [59]

3.3 Loss Function

Here we adopt a combination of Charbonnier loss, multi-scale SSIM loss and Perceptual loss.

Charbonnier Loss We employ the Charbonnier loss, which is mathematically defined as follows:

$$\mathcal{L}_{\text{charbonnier}} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\left\|\mathbf{I}_{gt}^{(i)} - \mathbf{I}_{c}^{(i)}\right\|^2 + \epsilon^2}, \tag{3.11}$$

where $\mathbf{I}_{gt}$ and $\mathbf{I}_c$ represent the ground truth and shadow-free images generated from our networks, respectively. In addition, ϵ is a small constant (e.g., 10^{-5}) used for stable and robust convergence, and n represents the total number of input images in a single iteration.

MS-SSIM Loss. Let $\mathcal{D}$ and $\mathcal{C}$ denote two windows of common size centered at pixel i in the shadowed image and the shadow-free image, respectively.The SSIM for pixel i can be computed by applying a Gaussian filter to $\mathcal{D}$ and $\mathcal{C}$, and compute the resulting means $\mu_{\mathcal{D}}, \mu_{\mathcal{C}}$, standard deviations $\sigma_{\mathcal{D}}, \sigma_{\mathcal{C}}$ and covariance $\sigma_{\mathcal{DC}}$.

$$\text{SSIM}(i) = \frac{(2\mu_{\mathcal{D}}\mu_{\mathcal{C}} + T_1)(2\sigma_{\mathcal{DC}} + T_2)}{(\mu_{\mathcal{D}}^2 + \mu_{\mathcal{C}}^2 + T_1)(\sigma_{\mathcal{D}}^2 + \sigma_{\mathcal{C}}^2 + T_2)} = l(i) \cdot s(i) \tag{3.12}$$

C_1, C_2 are two constants used to stabilize the division when the denominators are weak.

The MS-SSIM loss is computed using M levels of SSIM. Specifically, we have:

$$\mathcal{L}_{\text{MS-SSIM}} = 1 - \text{MS-SSIM} \tag{3.13}$$

where

$$\text{MS-SSIM} = l_M(i)^{\alpha} \cdot \prod_{m=1}^{M} cs_m(i)^{\beta_m} \tag{3.14}$$

with α and β_m being default parameters.

Perceptual Loss. We use the perceptual loss to optimize the visual effect. As illustrated in Eq. 3.15, C_l, H_l and W_l are the number of channels, height and width of the l-th feature map of the corresponding image, ϕ_l is the activation of the l-th layer. I_{gt} is the ground truth image and I_c is the shadow removed image.

$$L_p = \sum_{l=1}^{3} \left\| \frac{1}{C_l H_l W_l} \left(\phi_l(I_{gt}) - \phi_l(I_c)\right) \right\|_2^2 \tag{3.15}$$

We employ a pre-trained VGG16 [41] network as our loss network, specifically utilizing the features extracted from its first, second, and third convolutional layers to compute the perceptual loss.

Chapter 4

Experiments

4.1 Datasets

Our proposed method is evaluated using three established benchmark datasets. The first is the ISTD [49] dataset. The second is the Adjusted ISTD (ISTD+) dataset [29], which has undergone processing to minimize inconsistencies in illumination between the shadowed and shadow-free image pairs present in the original ISTD dataset. Lastly, we utilize the WSRD+ dataset [44], from which 1000 image pairs are used for training our model, and an additional 100 pairs are reserved for the purpose of validation.

4.2 Implementation details

In this section we provide the implementation details of our method. To enhance the robustness and generalization capability of our model, we employ several data augmentation techniques during training. These include random rotations by angles of 90°, 180°, or 270°, as well as flipping the input images both vertically and

horizontally with a certain probability. The spatial dimensions of the image crops used during training are set to 384×384 pixels, and the batch size for each training iteration is set to 4. The optimization of the model's trainable parameters is performed using the Adam optimization algorithm with its default hyper-parameter settings ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The initial learning rate is set to 1×10^{-4}, and it is progressively decreased over the course of training to a final value of 6.25×10^{-6} using a predefined learning rate schedule, a systematic rule or function that automatically adjusts the learning rate at specific intervals or after a certain number of training epochs. In addition to the commonly used L1 loss and multi-scale Structural Similarity Index Measure (SSIM) loss [66], we incorporate a structure loss [65] and additional constraints [14] to provide further guidance and supervision during the optimization process. The training is carried out on a NVIDIA GeForce RTX 3090Ti.

4.2.1 Evaluation Metrics

To evaluate the performance of different shadow removal techniques, we utilize a set of three quantitative evaluation metrics. These metrics include the Peak Signal-to-Noise Ratio (PSNR), which quantifies the pixel-level fidelity of the restored images; the Structural Similarity Index (SSIM) (SSIM) [51], designed to measure the perceived structural similarity between the restored and ground-truth images; and the Learned Perceptual Image Patch Similarity (LPIPS) [61], a metric that assesses the perceptual quality of the results by considering higher-level image features learned by a deep neural network. Together, these metrics provide a holistic assessment, considering both the pixel-wise accuracy and the subjective visual quality of the shadow removal outcomes.

Table 4.1: The ablation results on the WRSD+ dataset demonstrate that each component of our method contributes to the overall effectiveness in shadow removal.

Configurations	PSNR↑	SSIM↑	LPIPS↓
Full Model	**26.15**	**0.826**	**0.0860**
w/o dual-branch pipeline	25.86	0.818	0.0893
w/o IG-HTB	25.74	0.816	0.0915
w/o illumination in IG-HTB	25.97	0.821	0.0872

4.3 Ablation Study

In this section, we delve into a series of ablation studies performed on the WSRD+ dataset.

Importance of Dual-branch Retinex-based Pipeline To investigate the specific contribution of the dual-branch Retinex-based pipeline, as detailed in Section 3, we conducted an ablation study where this pipeline was removed from our complete method. To assess its impact on the overall performance Tab. 4.2, we directly applied our developed hybrid CNN-Transformer module to learn the mapping from shadowed input images to their corresponding clean, shadow-free versions. The quantitative results of this ablation experiment are presented in Tab. 4.1. Evidently, from the reported metrics the removal of the dual-branch Retinex-based pipeline results in a noticeable degradation of performance across all evaluation metrics. This significant drop in quantitative scores clearly highlights the importance of the dual-branch Retinex-based pipeline in achieving satisfactory shadow removal performance with our proposed method.

Contributions of IG-HTB and the Illumination Guidance To further demonstrate the effectiveness of our proposed Illumination-Guided Histogram Transformer Block (IG-HTB) and the utility of the illumination guidance mechanism within it, we conducted additional ablation experiments. By comparing the results presented in row 3 and row 4 of Tab. 4.1 against the performance of our complete model, we can observe the individual impact of these components. The quantitative improvements seen when the IG-HTB is incorporated and when the illumination guidance is effectively utilized within it strongly suggest that both the design of the IG-HTB and the integration of illumination guidance contribute to achieving a more favorable shadow removal performance.

4.4 Comparison to State-of-the-Art Methods

We evaluated the performance of our proposed method by comparing it against several existing state-of-the-art (SOTA) algorithms in the field. Our comparative analysis includes both mask-free methods, which do not rely on explicit shadow masks during inference, such as Refusion [38], DCShadowNet [27], ShadowRefiner [11], and IFBlend [45], as well as mask-based approaches, which utilize shadow masks, including ShadowFormer [20] and SADC [55]. This comprehensive comparison allows us to assess the effectiveness of our method relative to the current leading techniques in both categories of shadow removal.

4.4.1 Quantitative Results

As clearly shown in Tab.4.2, our proposed approach demonstrates superior performance among mask-free shadow removal methods across the three benchmark

TABLE 4.2: Quantitative comparisons with SOTA methods. Our ReHiT secures comparable performances to ShadowRefiner [11] and IFBlend [45], which incorporate large-scale pre-trained ConvNeXt [37] for transfer learning. Compared to mask-based methods, our ReHiT achieves comparable or even better performance (WSRD+ dataset). [Key: Best performance among mask-free models, Second-best performance among mask-free models, **Best performance among mask-based methods**, *: re-trained with officially released code.]

Methods	**Mask-free**	ISTD [49]			ISTD+ [29]			WSRD+ [44]		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DHAN [8]	No	24.86	0.919	0.0535	27.88	0.917	0.0529	22.39	0.796	0.1049
BMNet [69]	No	29.02	0.923	0.0529	31.85	0.932	0.0432	24.75	0.816	0.0948
FusionNet [17]	No	25.84	0.712	0.3196	27.61	0.725	0.3123	21.66	0.752	0.1227
SADC [55]	No	29.22	0.928	0.0403	—	—	—	—	—	—
ShadowFormer [20]	No	**30.47**	**0.928**	**0.0418**	**32.78**	**0.934**	**0.0385**	**25.44**	**0.820**	**0.0898**
DCShadowNet [27]	Yes	24.02	0.677	0.4423	25.50	0.694	0.4237	21.62	0.593	0.4744
Refusion [38]	Yes	25.13	0.871	0.0571	26.28	0.887	0.0437	22.32	0.738	0.0937
IFBlend* [45]	Yes	28.55	0.906	0.0558	30.87	0.916	0.0476	25.79	0.809	0.0905
ShadowRefiner [11]	Yes	28.75	0.916	0.0521	31.03	0.928	0.0426	26.04	0.827	0.0854
Ours (ReHiT)	Yes	28.81	0.914	0.0533	31.16	0.925	0.0442	26.15	0.826	0.0860

datasets: ISTD [49], ISTD+ [29], and WSRD+ [44]. Specifically, our method achieves higher PSNR values and comparable SSIM and LPIPS scores when compared to ShadowRefiner [11], which was the winning solution in the NTIRE 2024 Image Shadow Removal Challenge. It is noteworthy that while mask-based models often exhibit an inherent advantage due to their utilization of explicit shadow masks, our method attains comparable accuracy on both the ISTD and ISTD+ datasets without requiring any mask input during inference. Furthermore, on the WSRD+ dataset, where only estimated shadow masks are accessible, our method outperforms the best-performing mask-based model, ShadowFormer [20]. This result underscores the robust generalization capability of our approach when applied to more complex, real-world scenarios where perfect shadow masks may not

be available.

4.4.2 Qualitative Comparisons

Visual comparisons on ISTD dataset, ISTD+ dataset and WSRD+ dataset are reported in Fig. 4.1, 4.2, and 4.3, respectively. In Fig. 4.1 While DCShadow exhibits a strength in preserving image textures, it often fails completely removing shadows, leaving behind noticeable residual artifacts and discontinuities along shadow boundaries. In contrast, our proposed method, similar to ShadowRefiner, demonstrates a capability to effectively eliminate shadows without introducing such artifacts. Furthermore, our approach yields shadow removal results that appear more uniform and natural, successfully addressing both soft and hard shadows present in the input images while concurrently maintaining the integrity of the underlying textural details.

In Fig. 4.2, the presented results highlight that DCShadow, while achieving partial shadow removal, tends to leave behind discernible residual shadows and inconsistencies in illumination, particularly evident along the edges of shadow regions. Conversely, both ShadowRefiner and our proposed method demonstrate a greater ability to effectively eliminate shadows while introducing minimal visual artifacts, resulting in outputs that exhibit a higher degree of visual fidelity to the ground truth images. It is particularly noteworthy that our method achieves comparable performance to ShadowRefiner while utilizing only approximately 5% of the parameters of the ShadowRefiner model, indicating a significantly more parameter-efficient approach.

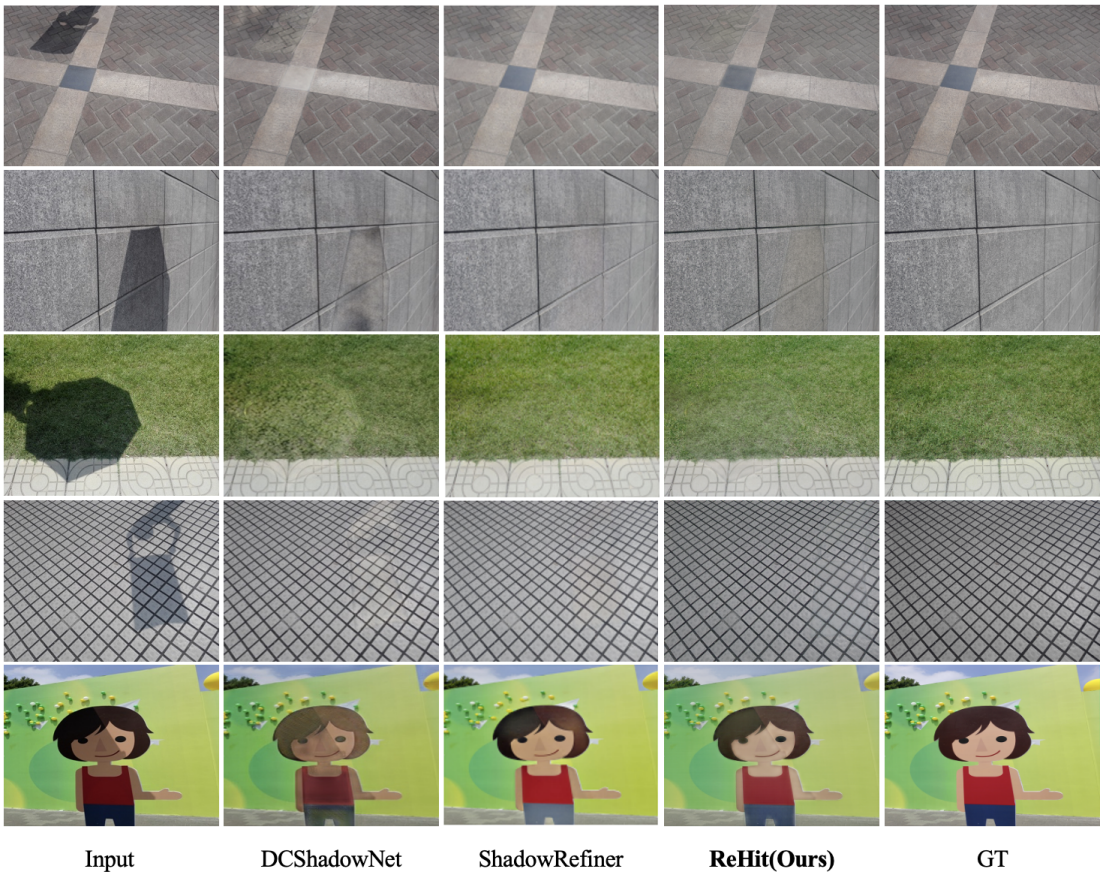

FIGURE 4.1: Qualitative comparisons on the ISTD dataset [49].

In Fig. 4.3 our proposed method exhibits a strong capability in effectively removing both subtle soft shadows and more distinct hard shadows from images. Crucially, this shadow removal is achieved while preserving the intricate structural and textural details present in the original scene. Notably, the image regions that were initially obscured by shadows are relit with a high degree of fidelity, and this relighting process does not introduce any visually discernible artifacts. The resulting shadow-free images demonstrate a consistent and natural illumination across the entire scene, with seamless transitions observed between the areas that were previously shadowed and those that were not. These qualitative observations

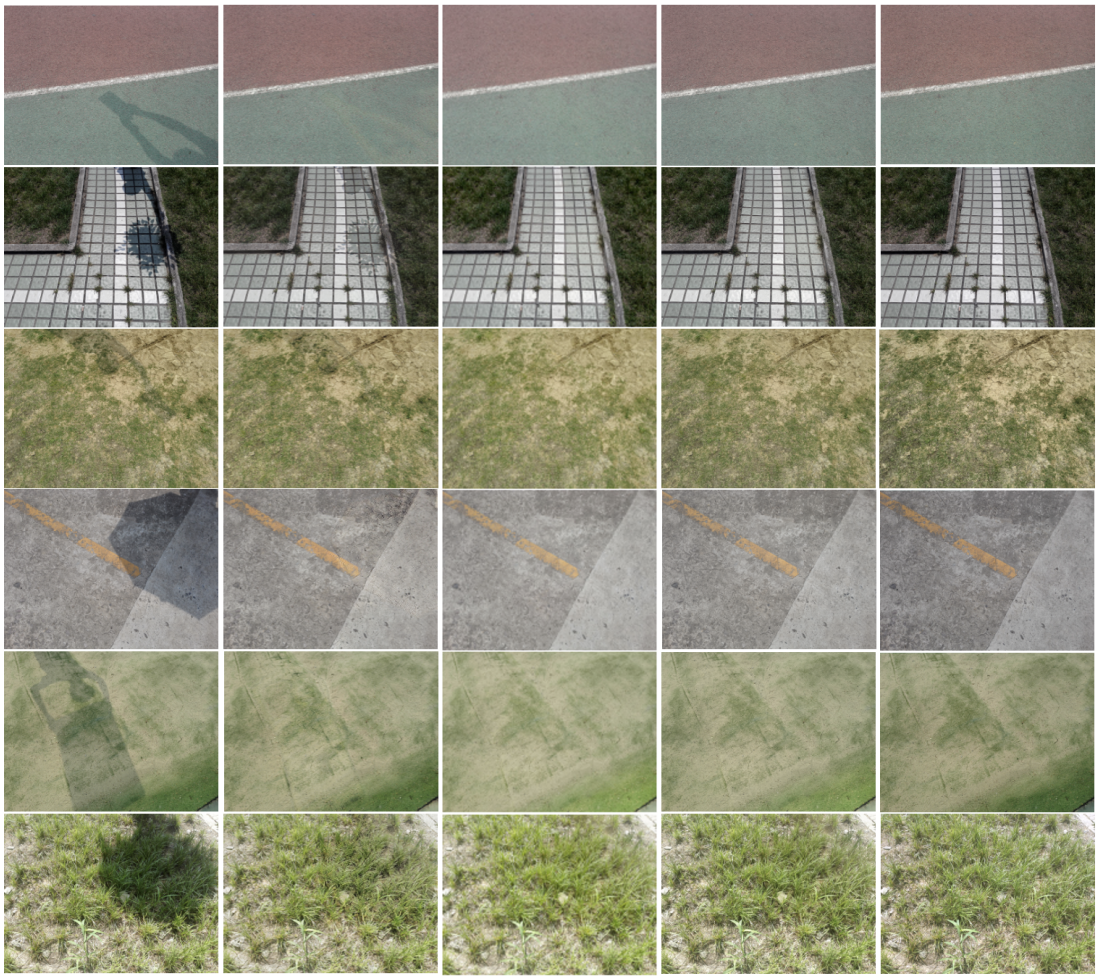

FIGURE 4.2: Qualitative comparisons on the ISTD+ dataset [29].

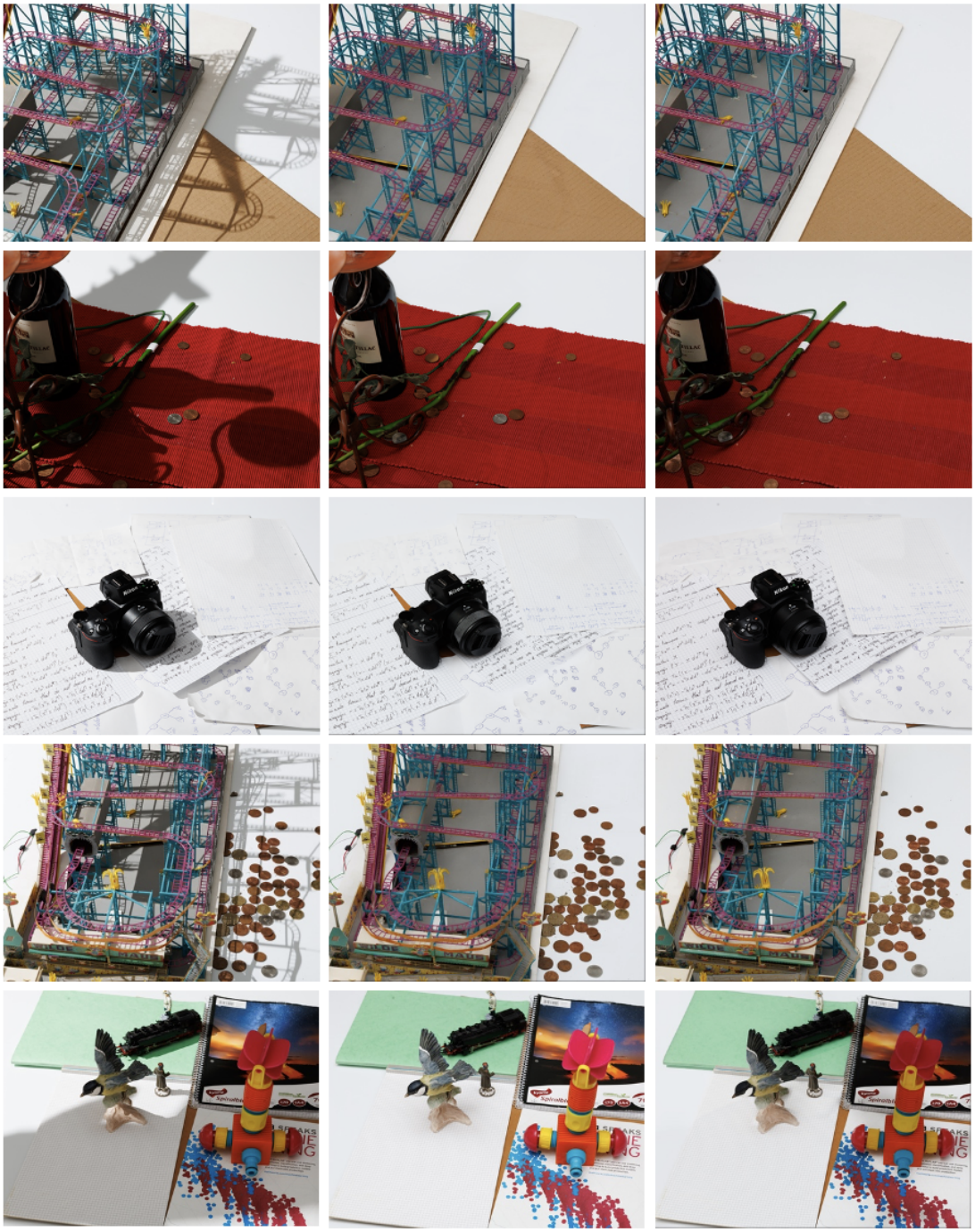

FIGURE 4.3: Our method delivers promising performance on the WSRD+ validation set [44].

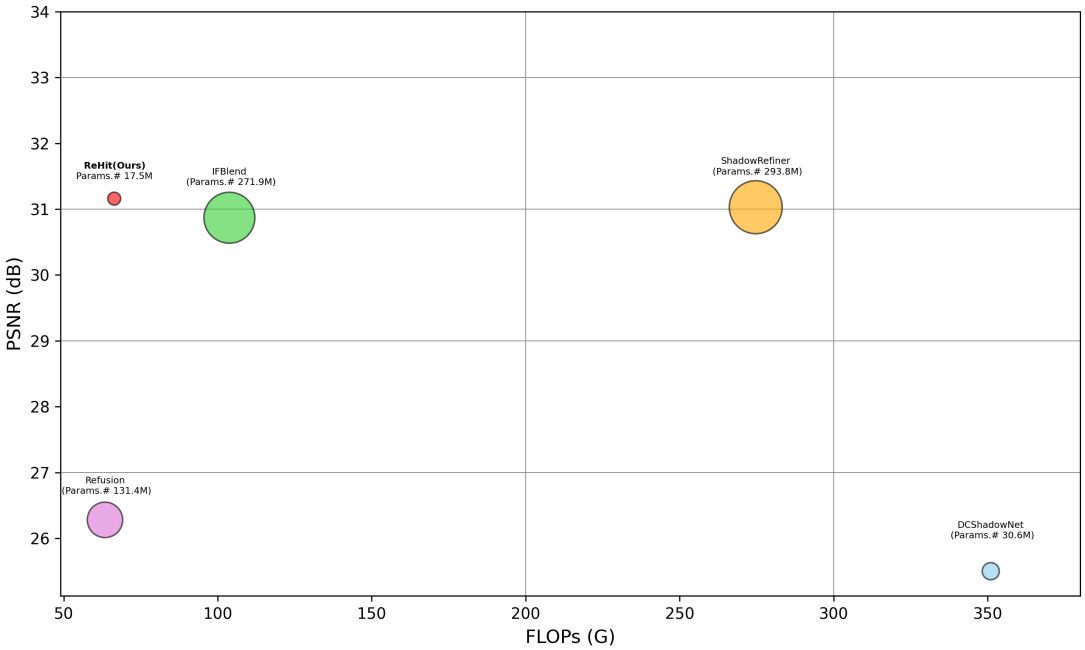

FIGURE 4.4: Comparison of computational cost of different methods. The x-axis and y-axis denote FLOPs (G) and PSNR (dB), respectively. The area of each circle represents the number of parameters.

underscore the robustness and strong generalization ability of our approach when applied to the complexities inherent in real-world shadow removal scenarios.

4.4.3 Computational cost

As shown in Fig. 4.4, our method strikes a sweet point of balancing the parameter number, computation cost, and shadow removal performance. Our method requires fewer parameters and FLOPs at inference, making it highly efficient. Combined with its competitive quantitative performance and the vivid, high-quality results restored by ReHiT in Fig.4.1, 4.2, and 4.3, this demonstrates that our

method delivers comparable or even better results with substantially lower computational overhead, indicating its practical value in real-world scenarios. This efficiency, coupled with its robust performance across diverse datasets, underscores its potential for deployment in resource-constrained environments while maintaining high-fidelity shadow removal.

Chapter 5

Conclusion

This thesis successfully developed a novel, lightweight, and mask-free framework for robust shadow removal, representing an advancement in image processing. Our approach integrates the strengths of Convolutional Neural Networks (CNNs) and Transformers, guided by the fundamental principles of Retinex theory, to effectively decompose images into their reflectance and illumination components. This decomposition is crucial for accurately modeling and correcting illumination inconsistencies caused by shadows.

A key innovation of our work is the Illumination-Guided Histogram Transformer, which proved instrumental in significantly enhancing the network's ability to intelligently handle complex shadow artifacts. This component directly addresses the nuances of scene-specific illumination variations, enabling a more precise and adaptive compensation process.

Extensive experimental validation across diverse benchmark datasets consistently demonstrated that our method achieves strong shadow removal performance, comparable to or surpassing existing state-of-the-art techniques. Critically, this high performance is attained while maintaining substantially lower model complexity and computational overhead. This efficiency makes our framework highly practical and suitable for real-world applications where resource constraints are a major consideration.

In essence, this thesis contributes a robust, efficient, and intelligent framework that pushes the boundaries of shadow removal. By offering a practical, mask-free, and computationally efficient solution, it paves the way for more resilient and adaptable computer vision systems across various domains, including autonomous driving, surveillance, and digital photography.

Bibliography

[1] C. O. Ancuti, C. Ancuti, F. A. Vasluianu, R. Timofte, Y. Liu, X. Wang, Y. Zhu, G. Shi, X. Lu, X. Fu, et al. NTIRE 2024 Dense and Non-Homogeneous Dehazing Challenge Report. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, 6453–6468.

[2] C. O. Ancuti, C. Ancuti, F. A. Vasluianu, R. Timofte, H. Zhou, W. Dong, Y. Liu, J. Chen, H. Liu, L. Li, et al. NTIRE 2023 HR Nonhomogeneous Dehazing Challenge Report. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, 1808–1825.

[3] E. Arbel and H. Hel-Or. Shadow Removal Using Intensity Surfaces and Texture Anchor Points. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 33(6) (2010), 1202–1216.

[4] Q. Bao, Y. Liu, B. Gang, W. Yang, and Q. Liao. S2Net: Shadow Mask-Based Semantic-Aware Network for Single-Image Shadow Removal. *IEEE Transactions on Consumer Electronics* 68(3) (2022), 209–220.

[5] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang. RetinexFormer: One-Stage Retinex-Based Transformer for Low-Light Image Enhancement. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, 12504–12513.

[6] X. Chen, H. Li, M. Li, and J. Pan. Learning a Sparse Transformer Network for Effective Image Deraining. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 5896–5905.

[7] Z. Chen, C. Long, L. Zhang, and C. Xiao. CANet: A Context-Aware Network for Shadow Removal. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, 4743–4752.

[8] X. Cun, C. M. Pun, and C. Shi. Towards Ghost-Free Shadow Removal via Dual Hierarchical Aggregation Network and Shadow Matting GAN. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 2020.

[9] B. Ding, C. Long, L. Zhang, and C. Xiao. ARGAN: Attentive Recurrent Generative Adversarial Network for Shadow Detection and Removal. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, 10212–10221.

[10] W. Dong, Y. Min, H. Zhou, and J. Chen. Towards Scale-Aware Low-Light Enhancement via Structure-Guided Transformer Design. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2025.

[11] W. Dong, H. Zhou, Y. Tian, J. Sun, X. Liu, G. Zhai, and J. Chen. ShadowRefiner: Towards Mask-Free Shadow Removal via Fast Fourier Transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, 6208–6217.

[12] W. Dong, H. Zhou, R. Wang, X. Liu, G. Zhai, and J. Chen. DehazeDCT: Towards Effective Non-Homogeneous Dehazing via Deformable Convolutional

Transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, 6405–6414.

[13] W. Dong, H. Zhou, and D. Xu. A New Sclera Segmentation and Vessels Extraction Method for Sclera Recognition. In: *2018 10th International Conference on Communication Software and Networks (ICCSN)*. 2018.

[14] W. Dong, H. Zhou, Y. Zhang, X. Liu, and J. Chen. ECMamba: Consolidating Selective State Space Model with Retinex Guidance for Efficient Multiple Exposure Correction. *Advances in Neural Information Processing Systems (NeurIPS)* 37 (2024), 53438–53457.

[15] G. D. Finlayson, M. S. Drew, and C. Lu. Entropy Minimization for Shadow Removal. *International Journal of Computer Vision (IJCV)* 85 (2009), 35–57.

[16] G. D. Finlayson, S. D. Hordley, C. Lu, and M. S. Drew. On the Removal of Shadows From Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 28 (2006), 59–68.

[17] L. Fu, C. Zhou, Q. Guo, F. J. Xu, H. Yu, W. Feng, Y. Liu, and S. Wang. Auto-Exposure Fusion for Single-Image Shadow Removal. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021.

[18] H. Gong and D. P. Cosker. Interactive Removal and Ground Truth for Difficult Shadow Scenes. *Journal of the Optical Society of America A* 33(9) (2016), 1798–1811.

[19] M. Gryka, M. Terry, and G. J. Brostow. Learning to Remove Soft Shadows. *ACM Transactions on Graphics (TOG)* 34(5) (2015), 1–15.

[20] L. Guo, S. Huang, D. Liu, H. Cheng, and B. Wen. ShadowFormer: Global Context Helps Image Shadow Removal. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 2023.

[21] R. Guo, Q. Dai, and D. Hoiem. Paired Regions for Shadow Detection and Removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 35(12) (2012), 2956–2967.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, 770–778.

[23] S. He, B. Peng, J. Dong, and Y. Du. MaskShadowNet: Toward Shadow Removal via Masked Adaptive Instance Normalization. *IEEE Signal Processing Letters* 28 (2021), 957–961.

[24] X. Hu, L. Zhu, C.-W. Fu, J. Qin, and P.-A. Heng. Direction-Aware Spatial Context Features for Shadow Detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, 7454–7462.

[25] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, 4700–4708.

[26] G. Huang, X. Wang, W. Wu, H. Zhou, and Y. Wu. Real-Time Lane-Vehicle Detection and Tracking System. In: *Chinese Control and Decision Conference (CCDC)*. 2016.

[27] Y. Jin, A. Sharma, and R. T. Tan. DC-ShadowNet: Single-Image Hard and Soft Shadow Removal Using Unsupervised Domain-Classifier Guided Network. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[28] E. H. Land. The Retinex Theory of Color Vision. *Scientific American* 237(6) (1977), 108–129.

[29] H. Le and D. Samaras. Shadow Removal via Shadow Image Decomposition. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2019.

[30] A. Lengyel, S. Garg, M. Milford, and J. C. van Gemert. Zero-Shot Day-Night Domain Adaptation with a Physics Prior. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, 4399–4409.

[31] X. Li, Y. Jin, X. Jin, Z. Wu, B. Li, Y. Wang, W. Yang, Y. Li, Z. Chen, B. Wen, R. Tan, R. Timofte, et al. NTIRE 2025 Challenge on Day and Night Raindrop Removal for Dual-Focused Images: Methods and Results. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2025.

[32] J. Liang, J. Cao, G. Sun, K. Zhang, L. V. Gool, and R. Timofte. SwinIR: Image Restoration Using Swin Transformer. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, 1833–1844.

[33] Y.-H. Lin, W.-C. Chen, and Y.-Y. Chuang. BEDSR-Net: A Deep Shadow Removal Network From a Single Document Image. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, 12902–12911.

[34] X. Liu, Z. Wu, F. A. Vasluianu, H. Yan, B. Ren, Y. Zhang, S. Gu, L. Zhang, C. Zhu, R. Timofte, et al. NTIRE 2025 Challenge on Low Light Image Enhancement: Methods and Results. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2025.

[35] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, 10012–10022.

[36] Z. Liu, H. Yin, Y. Mi, M. Pu, and S. Wang. Shadow Removal by a Lightness-Guided Network With Training on Unpaired Data. *IEEE Transactions on Image Processing (TIP)* 30 (2021), 1853–1865.

[37] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A ConvNet for the 2020s. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, 11976–11986.

[38] Z. Luo, F. K. Gustafsson, Z. Zhao, et al. Refusion: Enabling Large-Size Realistic Image Restoration with Latent-Space Diffusion Models. In: *CVPR Workshops (CVPRW)*. 2023.

[39] L. Qu, J. Tian, S. He, Y. Tang, and R. W. H. Lau. DeshadowNet: A Multi-Context Embedding Deep Network for Shadow Removal. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, 2308–2316.

[40] L. Qu, J. Tian, S. He, Y. Tang, and R. W. Lau. Deshadownet: A Multi-Context Embedding Deep Network for Shadow Removal. In: *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[41] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556* (2014).

[42] Y. Song, Z. He, H. Qian, and X. Du. Vision Transformers for Single Image Dehazing. *IEEE Transactions on Image Processing (TIP)* 32 (2023), 1927–1941.

[43] S. Sun, W. Ren, X. Gao, R. Wang, and X. Cao. Restoring Images in Adverse Weather Conditions via Histogram Transformer. In: *European Conference on Computer Vision (ECCV)*. Springer Nature Switzerland, 2024.

[44] F. A. Vasluianu, T. Seizinger, and R. Timofte. WsrD: A Novel Benchmark for High-Resolution Image Shadow Removal. In: *CVPR Workshops (CVPRW)*. 2023.

[45] F. A. Vasluianu, T. Seizinger, Z. Wu, R. Ranjan, and R. Timofte. Towards Image Ambient Lighting Normalization. In: *European Conference on Computer Vision (ECCV)*. Springer. 2024, 385–404.

[46] F. A. Vasluianu, T. Seizinger, Z. Zhou, Z. Wu, C. Chen, R. Timofte, W. Dong, H. Zhou, Y. Tian, J. Chen, et al. NTIRE 2024 Image Shadow Removal Challenge Report. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, 6547–6570.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention Is All You Need. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 30. 2017.

[48] J. Wan, H. Yin, Z. Wu, X. Wu, Y. Liu, and S. Wang. Style-Guided Shadow Removal. In: *European Conference on Computer Vision (ECCV)*. Springer Nature Switzerland, 2022, 361–378.

[49] J. Wang, X. Li, and J. Yang. Stacked Conditional Generative Adversarial Networks for Jointly Learning Shadow Detection and Shadow Removal. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[50] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li. Uformer: A General U-Shaped Transformer for Image Restoration. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 17683–17693.

[51] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing (TIP)* 13(4) (2004), 600–612.

[52] M. Wu, R. Chen, and Y. Tong. Shadow Elimination Algorithm Using Color and Texture Features. *Computational Intelligence and Neuroscience* (2020).

[53] J. Xiao, X. Fu, A. Liu, F. Wu, and Z. J. Zha. Image De-Raining Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45(11) (2022), 12978–12995.

[54] D. Xu, W. Dong, and H. Zhou. Sclera Recognition Based on Efficient Sclera Segmentation and Significant Vessel Matching. In: *The Computer Journal*. 2022.

[55] Y. Xu, M. Lin, H. Yang, F. Chao, and R. Ji. Shadow-Aware Dynamic Convolution for Shadow Removal. In: *Pattern Recognition*. 2024.

[56] K. Yang, J. Cai, L. Ouyang, F. A. Vasluianu, R. Timofte, J. Ding, H. Sun, L. Fu, J. Li, C. M. Ho, Z. Meng, et al. NTIRE 2025 Challenge on Single Image Reflection Removal in the Wild: Datasets, Methods and Results. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2025.

[57] Q. Yang, K. H. Tan, and N. Ahuja. Shadow Removal Using Bilateral Filtering. *IEEE Transactions on Image Processing (TIP)* 21(10) (2012), 4361–4368.

[58] F. Yu and V. Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv Preprint arXiv:1511.07122* (2015).

[59] X. Yu, P. Dai, W. Li, L. Ma, J. Shen, J. Li, and X. Qi. Towards Efficient and Scale-Robust Ultra-High-Definition Image Demoiréing. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022, 646–662.

[60] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang. Restormer: Efficient Transformer for High-Resolution Image Restoration. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 5728–5739.

[61] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[62] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual Dense Network for Image Super-Resolution. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, 2472–2481.

[63] H. Zhao, Y. Gou, B. Li, D. Peng, J. Lv, and X. Peng. Comprehensive and Delicate: An Efficient Transformer for Image Restoration. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 14122–14132.

[64] Y. Zhao, C. Elliott, H. Zhou, and K. Rafferty. Pixel-Wise Illumination Correction Algorithms for Relative Color Constancy Under the Spectral Domain. In: *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. 2018.

[65] H. Zhou, W. Dong, and J. Chen. LITA-GS: Illumination-Agnostic Novel View Synthesis via Reference-Free 3D Gaussian Splatting and Physical Priors. *arXiv preprint arXiv:2504.00219* (2025).

[66] H. Zhou, W. Dong, X. Liu, S. Liu, X. Min, G. Zhai, and J. Chen. Glare: Low-Light Image Enhancement via Generative Latent Feature Based Codebook Retrieval. In: *European Conference on Computer Vision (ECCV)*. Springer. 2024, 36–54.

[67] H. Zhou, W. Dong, X. Liu, Y. Zhang, G. Zhai, and J. Chen. Low-Light Image Enhancement via Generative Perceptual Priors. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. Vol. 39. 10. 2025, 10752–10760.

[68] H. Zhou, W. Dong, Y. Liu, and J. Chen. Breaking Through the Haze: An Advanced Non-Homogeneous Dehazing Method Based on Fast Fourier Convolution and ConvNeXt. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, 1895–1904.

[69] Y. Zhu, Z. Xiao, Y. Fang, X. Fu, Z. Xiong, and Z. J. Zha. Efficient Model-Driven Network for Shadow Removal. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 2022.