This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

ShadowHack: Hacking Shadows via Luminance-Color Divide and Conquer

Jin Hu* Mingjia Li* Xiaojie Guo†
Tianjin University
{jinhu, mingjiali}@tju.edu.cn, xj.max.guo@gmail.com

Abstract

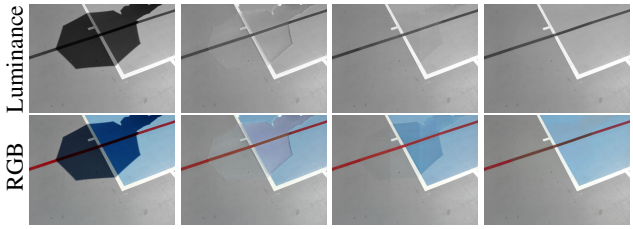
*Shadows introduce challenges such as reduced brightness, texture deterioration, and color distortion in images, complicating a holistic solution. This study presents **ShadowHack**, a divide-and-conquer strategy that tackles these complexities by decomposing the original task into luminance recovery and color remedy. To brighten shadow regions and repair the corrupted textures in the luminance space, we customize LRNet, a U-shaped network with a rectified attention module, to enhance information interaction and recalibrate contaminated attention maps. With luminance recovered, CRNet then leverages cross-attention mechanisms to revive vibrant colors, producing visually compelling results. Extensive experiments on multiple datasets are conducted to demonstrate the superiority of ShadowHack over existing state-of-the-art solutions both quantitatively and qualitatively, highlighting the effectiveness of our design. Our code is available at https://github.com/lime-j/ShadowHack.*

> *Shadow is a color as light is, but less brilliant; light and shadow are only the relation of two tones.*
> — *Paul Cézanne*

1. Introduction

Shadows are the absence of light, an everyday phenomenon that often shapes the aesthetic beauty of nature. However, shadows obscure the actual shape of instances, diminishing their visibility. Very likely, they deteriorate the performance of various machine vision tasks, such as object detection [21, 42], face recognition [62], and semantic segmentation [63], to name just a few. Consequently, shadow removal has become a critical research topic.

Recent advancements in deep learning have greatly improved the efficiency and accuracy of shadow removal. Specifically, DeshadowNet [43] marked a pioneering step with its end-to-end Convolutional Neural Network (CNN) approach, while subsequent GAN-based methods [22, 36, 48, 61] further advanced the field by adopting adversarial learning frameworks, enabling effective unpaired training. Additionally, Wan *et al.* [47] introduced style transfer, and Li *et al.* [30] incorporated pre-trained inpainting priors to better restore textural information. Transformer-based architectures, such as ShadowFormer [13], HomoFormer [53], and RASM [32], have since been developed to restore shadow images more effectively, while diffusion-based approaches [14, 24, 41, 49] have also demonstrated promising results, particularly in complex scenarios where previous generative models often struggled.



(a) Input (b) SG [47] (c) HF [53] (d) Ours

Figure 1. Visual comparison between our method and state-of-the-art methods in both Luminance and RGB spaces.

Despite these advancements, the odyssey of removing shadows is still far from its end. As evidenced in Fig. 1, even state-of-the-art models fail to recover color and underlying texture accurately. The difficulty stems from the complex and intertwined nature of shadow degradation, which involves not only *reduced brightness* in shadow (umbra) regions [40], but also *texture deterioration* from the imaging process [30, 34], and *color distortion* from surface material properties and environmental color influences [7, 28, 47], an example of which is in Fig. 2. This complexity naturally suggests breaking down the original task, with its entangled interferences, into isolated and simpler sub-problems, each targeting a specific aspect and addressed independently.

Following the above line of thought, some works [16, 27, 28] attempted to estimate light and color biases as scalars. Insights can also be drawn from the low-light enhancement literature, where the Retinex decomposition [1, 26] is commonly employed [2, 17, 18, 54]. Under the Retinex theory,

*Equal Contribution.

†Corresponding author. This work was supported by National Natural Science Foundation of China under Grant nos. 62372251 and 62072327.

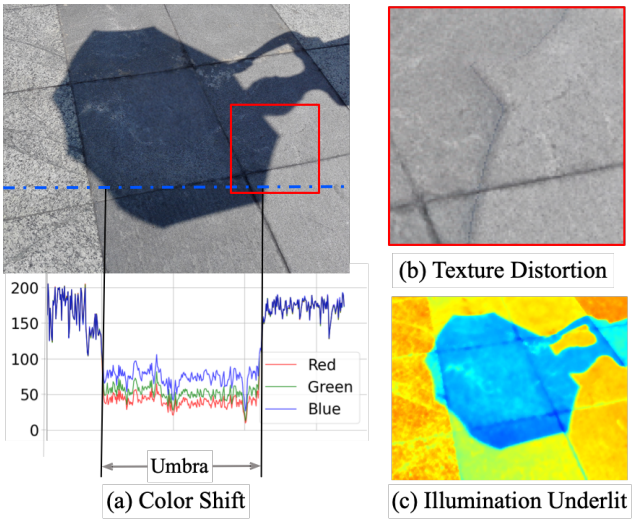


Figure 2. Example of the complex degradation in a shadow image. a) Color shift in the umbra region; (b) Structural texture deterioration in the penumbra transition area; (c) Attenuated illumination levels across the shadow (Blue indicates lower illumination).

an image can be decomposed into a structured illumination and a colored texture-rich counterpart. However, these approaches encounter challenges due to *the inherent inhomogeneity of shadows*. In other words, shadow removal extends beyond low-light enhancement and brings additional color biases to the reflectance map. This realization underscores the need for more sophisticated solutions to effectively disentangle the complex degradations introduced by shadows.

This work invents ShadowHack, an effective shadow removal framework that individually tackles illumination and color components. ShadowHack starts with a simple yet highly effective hack for shadow decomposition. Unlike conventional methods that decouple illumination from color and texture, we alternatively separate color from illumination and texture. Once the input images are decomposed into luminance and color components, two dedicated networks are established, *i.e.*, the Luminance Restoration Network (LRNet) and the Color Regeneration Network (CRNet). LRNet, a lightweight transformer-based restoration model, focuses on recovering illumination and texture. To enable an effective reference for shadow region refinement, we introduce the RecTified Attention (RTA) Module, which employs a rectified mechanism on window-based attentions to support high-quality luminance restoration.

Afterward, the processed luminance serves as a guide for querying accurate color information in shadow regions from the non-shadow areas. We inject color information via an ImageNet-pre-trained color encoder, which integrates color data into the intermediate features of the skip connections of our CRNet. This is facilitated by cross-attention modules that align the color in the shadow region with patches of similar luminance. Through this sequential processing, ShadowHack achieves effective shadow removal, supported by extensive experimental results. The primary contributions of this study are summarized as follows:

- Based on an in-depth analysis of shadow properties and the limitations in current end-to-end decomposition methods, we introduce ShadowHack, a novel shadow removal framework that breaks shadow removal into two distinct subtasks: luminance recovery and color regeneration.
- We identified the distraction phenomenon in the attention matrix of shadow removal networks, and propose RecTified Attention (RTA) to regain focus in attention layers with marginal computational overhead, enabling precise restoration of irregular texture and illumination.
- We design two specialized networks, LRNet and CRNet, tailored for luminance refinement and color regeneration. Together, these networks allow our model to achieve accurate and robust results, setting new state-of-the-art performance across benchmarking datasets.

2. Related Work

Shadow Removal. Traditional methods, such as [57] and [45], leveraged principles of physics and mathematical statistics on factors like illumination [10, 60], color [25, 51], region [16], and gradient [11, 12]. However, these methods often struggle with complex real-world scenes, where shadows interact in nuanced ways with surrounding colors and textures. Deep learning has recently made significant strides in the field [27, 43, 48]. For instance, Qu *et al.* [43] proposed to extract both global and semantic features from images, marking an early attempt to integrate deep learning into shadow removal. Later, Wang *et al.* [48] proposed a multi-task framework that performs shadow detection and removal concurrently, enhancing overall accuracy. Generative Adversarial Networks (GANs) have also been applied to this task. As a representative, RIS-GAN [61] incorporates residual and illumination features to refine shadow regions, while other works [22, 36] use unsupervised GANs to generate shadows for achieving cycle consistency. Li *et al.* [30] adopted the concept of inpainting, simultaneously using a shadow image encoder and a generative-pretrained image encoder to extract features. Zhu *et al.* [65] employed normalizing flow for a reversible link between shadow generation and removal. Additionally, Le *et al.* [28] proposed a linear model that adjusts the global lightness of shadow regions through physical modeling. More recently, diffusion models [46] have shown even greater generative capabilities than GANs. Guo *et al.* [14] pioneered the use of diffusion models for shadow removal, refining shadow masks throughout the process. Liu *et al.* [34] applied Retinex decomposition [18, 26] and used diffusion models to refine textures under the guidance of the obtained illumination. However, a primary drawback of these diffusion models is their intensive computational cost.

ViT-based Restoration. Vision Transformers (ViTs) have demonstrated impressive performance across vision tasks, owing to their powerful attention mechanisms that can effectively capture long-range dependencies [3, 9, 19, 35, 52]. In image restoration, IPT [4] was among the first ViT-based solutions. Unfortunately, its dependence on large-scale paired datasets for training restricts its applicability when data is scarce. Subsequent models have sought to mitigate this limitation with more efficient architectures. Uformer [50] combined UNet with a window attention, yielding strong results, while Restormer [59] introduced transposed attention for efficient channel-wise processing. Further, X-Restormer [5] extended Restormer by integrating the OCA [6] module, providing a generalized framework for image restoration through rigorous analysis. Specific to the domain of shadow removal, ShadowFormer [13] introduced an attention mechanism that facilitates interaction between shadow and non-shadow regions, while HomoFormer [53] improved the transformer by integrating a shuffling mechanism into its transformer-based encoder. RASM [32] refreshed the window-based attention with a sliding-window mechanism, achieving more accurate results. More recently, OmniSR [55] leveraged DINO features and depth maps to incorporate semantic and geometric attention into self-attention, expanding its perception of image content and ability to distinguish shadows. Even though, as discussed in Sec. 1, there remains significant room for the community to improve the effectiveness of shadow removal.

3. Problem Analysis

Degradation in umbra primarily encompasses reduced brightness, color distortion, and texture degradation (*e.g.*, noise and blurriness), as illustrated in Fig. 2. To better understand the types of involved degradation, we begin by formulating shadow effects via a modified image formation model as proposed in [1, 28, 45]. In this model, the observed pixel intensity $I(x, \lambda)$ at a given spatial coordinate x and wavelength λ is expressed as the product of shading $S(x, \lambda)$ and albedo $A(x, \lambda)$ as follows:

$$I(x, \lambda) = S(x, \lambda) \cdot A(x, \lambda). \tag{1}$$

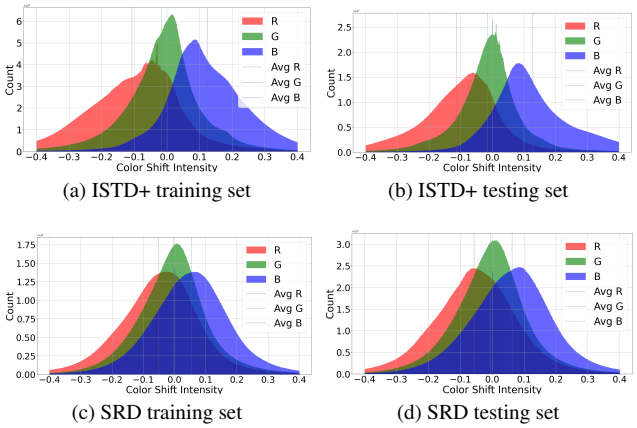
In scenes with a single dominant light source (*e.g.*, sunlight in outdoor settings), the shading for a pixel x can be decomposed into direct $S_d(x, \lambda)$ and ambient $S_a(x, \lambda)$ components:

$$I(x, \lambda) = (S_d(x, \lambda) + S_a(x, \lambda)) \cdot A(x, \lambda). \tag{2}$$

However, when an object occludes the primary light source, shadow regions occur, and the pixel intensity becomes:

$$I(x, \lambda) = a(x) \cdot S_a(x, \lambda) \cdot A(x, \lambda), \tag{3}$$

where $a(x) \in [0, 1]$ designates a spatially varying attenuation function. This non-uniform attenuation of light forms



(a) ISTD+ training set (b) ISTD+ testing set

(c) SRD training set (d) SRD testing set

Figure 3. Color bias in shadow regions between the reflectances of shadow and shadow-free images by Retinex decomposition [18].

shadow. This involves not only illumination variance but also challenging inconsistency in texture[31], an example of which is illustrated in Fig. 2 (b) and (c). To obtain an RGB image[1] $I_{RGB} := \{I_R, I_G, I_B\}$, the observed color intensity at each pixel is integrated across wavelengths as follows:

$$I_k(x) = \int_{\lambda_{\min}}^{\lambda_{\max}} S(x, \lambda) \cdot A(x, \lambda) \cdot P_k(\lambda)\, d\lambda, k \in \{R, G, B\} \tag{4}$$

where $P_k(\lambda)$ is the spectral sensitivity function of the camera for the k-channel and $[\lambda_{\min}, \lambda_{\max}]$ defines the visible spectrum range. Previous shadow removal and low-light enhancement methods [15, 17, 18, 34] suggest that it would be neat to decompose lightness from the shadow image following the Retinex theory. In these works, the image is assumed to be decomposed into a one-channel illumination map $L(x)$ and an illumination-invariant counterpart (reflectance) $R_k(x)$. For sRGB space and white-balanced images, the operation can be formulated as:

$$I_k(x) \approx L(x) \cdot R_k(x) = L(x) \int_{\lambda_{\min}}^{\lambda_{\max}} A(x, \lambda) \cdot P_k(\lambda)\, d\lambda. \tag{5}$$

We argue that shadow posits not just illumination map changes, but shifts that variants across the color. To quantify this, we plot the per-channel color difference between the input and ground-truth images in the reflectance map derived from a Retinex-based illumination estimator (Fig.3). The reflectance decomposition exhibits substantial color bias, particularly in the red and blue channels, violating Retinex's assumption of illumination-invariant reflectance. This color shift cannot be adequately addressed by previous Retinex-inspired decomposition methods [15, 34], as their core premise which reflectance remains uncontaminated by illumination effects are broken down in shadow regions.

[1]Here, we omit the impact from the ISP pipeline.

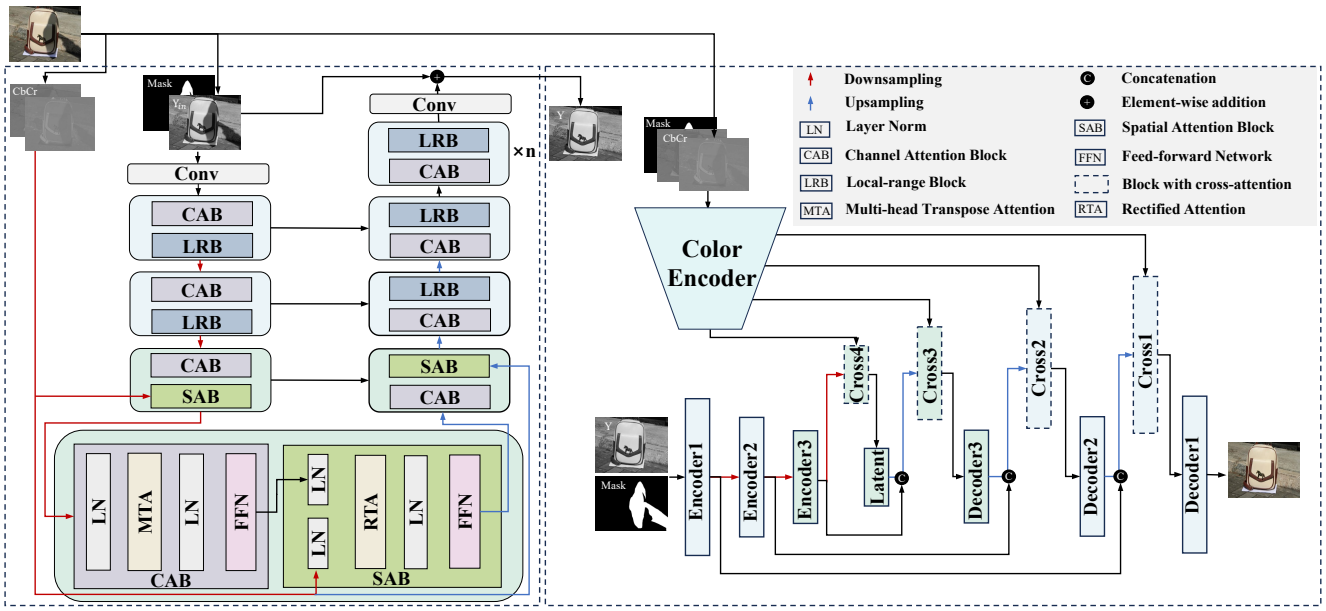


Figure 4. The overall architecture of ShadowHack comprises two core components: (a) a luminance restoration network, which focuses on recovering illumination and texture, and (b) a color regeneration network dedicated to recalibrating harmonized colors for natural outputs.

We propose a novel approach to address the triple degradation of reduced brightness, texture deterioration, and color distortion in a divide-and-conquer manner. The key insight is identifying a decomposition scheme that cleanly separates brightness and texture information from color information. An ideal decomposition would isolate a grayscale component capturing illumination and textural details with minimum color interference, alongside a pure color component that can be independently corrected. This separation would enable a targeted restoration process, which first recovers accurate textural details and then uses this restored foundation as a reliable reference for precise color regeneration.

YCbCr color space provides an optimal lossless decomposition for this task. The Y channel, which predominantly weights the green channel (0.587G compared to 0.299R and 0.114B), is significantly less contaminated by the color bias that predominantly affects red and blue channels in shadow regions. Additionally, the Y channel is specifically designed to preserve perceptually important texture details that align with human visual sensitivity. By first restoring this luminance component, we can recover accurate illumination and texture information without interference from color distortions. The restored luminance then serves as a geometric and semantic guide for subsequent color regeneration, ensuring spatial coherence between brightness and chrominance adjustments. This sequential enables dedicated networks to specialize in luminance refinement (LRNet) and chroma harmonization (CRNet) without mutual interference.

4. Methodology

A viable solution involves separating an image into texture I_t and color I_c components for texture restoration and color regeneration, respectively. We achieve this through an invertible decoupling function $D(\cdot)$, which splits an RGB image I into a luminance component I_y and a chrominance component I_c:

$$(I_t, I_c) = \mathcal{D}(I). \tag{6}$$

We adopt the conversion from RGB to YCbCr as $\mathcal{D}(I)$, where I_t corresponds to the Y (luminance) channel, and I_c comprises the Cb and Cr (chrominance) channels. This decomposition isolates brightness and texture degradation in I_t while retaining color distortions in I_c. The inverse function $\mathcal{D}^{-1}$ reconstructs the final RGB image from processed components. This manner enables a divide-and-conquer strategy for shadow removal:

$$\hat{I} = \mathcal{D}^{-1}\mathcal{C}(\mathcal{R}(I_t), I_c), \tag{7}$$

where $\mathcal{R}(\cdot)$ denotes the luminance restoration network (LRNet) and, $\mathcal{C}(\cdot)$ represents the color regeneration network (CRNet). $\hat{I}$ is the shadow-free output, and I_c serves as a conditioning variable for color regeneration. Additionally, $\mathcal{D}^{-1}(\cdot)$ is the inverse procedure of $\mathcal{D}(\cdot)$.

As in Fig. 4, ShadowHack operates in three stages: 1) *Decomposition.* The RGB image is decomposed via $\mathcal{D}(\cdot)$ into luminance I_t and color I_c components; 2) *Luminance Restoration.* LRNet processes I_t to recover illumination and texture, yield in $\hat{I}_t = \mathcal{R}(I_t)$; 3) *Color Regeneration.*

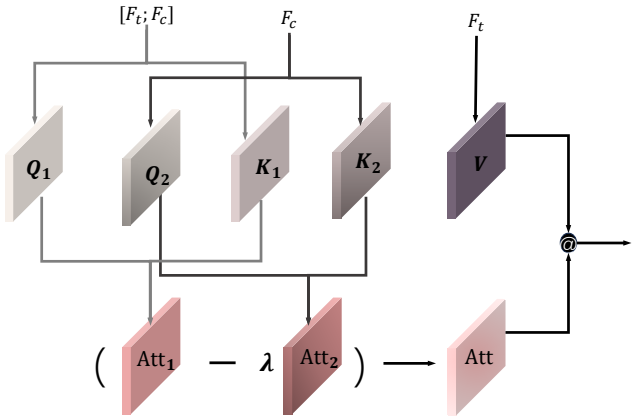


Figure 5. The rectified attention module. We compute two sets of attention. In Att_1, both texture (I_t) and color (I_c) are involved while in Att_2, we calculate attention with only color components. The difference $Att_1 - \lambda Att_2$ delivers more accurate interactions

CRNet harmonizes colors using I_c and $\hat{I}_t$. Finally, the estimated luminance and color components are recombined using $\mathcal{D}^{-1}(\cdot)$ to produce the final shadow-free RGB result. Below, we detail LRNet and CRNet.

4.1. Luminance Restoration Network

To eliminate shadows in I_t, enhancing the brightness of the shadow areas and restoring their textures is essential. As illustrated in Fig. 4 (a), the shallow two stages and the refinement module in the luminance restoration network (LRNet) are constructed using a Local Range block (LRB), a Multi-head Transpose Attention (MTA) module, and a Feed-Forward Network (FFN), while a RecTified Attention (RTA) module replaces LRB in the deeper stages. The LRB is designed to extract local texture information through a depth-wise convolution block, capturing fine-grained details. The MTA, which lies in the channel attention block (CAB), following previous works [5, 59], reallocates the weights of channels to enhance the relevant features. The FFN further processes these extracted features to improve texture restoration. Finally, the RTA module, introduced in the deeper layers, captures more global information by focusing on regions close to shadow boundaries, enabling the network to refine luminance restoration in complex shadow regions.

Rectified Attention Module. In the shadow removal process, achieving robust enhancement in shadow regions requires restoration efforts to reference better surrounding, related non-shadow areas. Therefore, we replace the Local Range Block (LRB) with self-attention modules [20, 32] in the deeper layers to capture more distant information. Since the goal of LRNet is to cure shadow-contaminated texture and lightness, the feature from the deep layer (i.e., blocks with SAB equipped) becomes less and less discriminative to shadow. As a result, the attention matrix in deep layers, suffers from a distraction. In the first column of Fig. 6, the

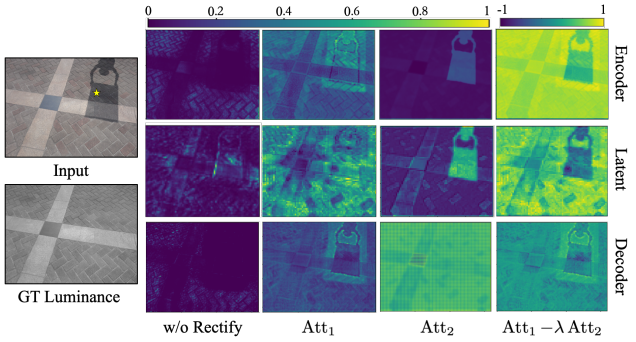


Figure 6. An example of our rectified attention module. The star indicates the position of the query token. We exhibit the attention map from the first head of the encoder, latent, and decoder layers. The final attention is obtained by subtracting Att_2 from Att_1. The attention matrix without rectify, Att_1 and Att_2 are normalized to [0, 1], while their differential result is in [-1, 1].

model tends to focus on only a limited number of tokens which may be inaccurate for shadow area refinement. We address this via a RecTified Attention (RTA) mechanism to rectify the original attention weight that is re-injecting contaminated chrominance channels I_c, by combining them into Q and K calculations (Please notice that the V is still clear without chromance), resulting in Att_1. Since this operation re-integrates color bias to the calculation, inspired by [58], we calculate a second set with only I_c as Att_2. By subtracting the two sets, LRNet has the flexibility to get rid of chrominance information if needed.

To be specific, we obtain the features F_t corresponding to I_t through convolution with downsampling, and compute the softmax attention map for F_t and F_c, as shown in Fig. 5. The query, key, and value matrices are formed via:

$$\begin{aligned} Q_1 &= [F_t; F_c]W_1^Q,\ K_1 = [F_t; F_c]W_1^K, \\ Q_2 &= F_c W_2^Q,\ K_2 = F_c W_2^K,\ V = F_t W^V, \end{aligned} \tag{8}$$

where W matrices are learnable projection parameters and ";" refers to the concatenation of the two features. Next, the above two sets of queries and keys are used to compute their respective attentions simply by:

$$\mathrm{Att}_i = \mathrm{Softmax}\left(\frac{Q_i K_i^{\mathrm{T}}}{\sqrt{d}} + B\right),\ i \in \{1, 2\}, \tag{9}$$

where B denotes the relative position bias. This allows us to apply the rectified attention mechanism as follows:

$$\mathrm{RTA}(F_t, F_c) = (\mathrm{Att}_1 - \lambda \cdot \mathrm{Att}_2)\, V, \tag{10}$$

where λ is a reweighting factor. As illustrated in Fig. 6, such a rectified attention mechanism creates a strong affinity with well-lit reference, urging the model to focus on relevant content. As validated in Sec. 5.3, our rectified attention can focus on more non-shadow area information to transfer robust clues for shadow removal.

Datasets	Method	Shadow Region (S)			Non-Shadow Region (NS)			All Image (ALL)		
		PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓
ISTD+	DHAN [8] (AAAI'20)	33.08	0.988	9.49	27.28	0.972	7.39	25.78	0.958	7.74
	G2R [36] (CVPR'21)	33.88	0.978	8.71	35.94	0.977	2.81	30.85	0.946	3.78
	BMNet [65] (CVPR'22)	38.17	0.991	5.72	37.95	0.986	2.42	34.34	0.974	2.93
	SG-ShadowNet [47] (ECCV'22)	36.80	0.990	5.93	35.57	0.978	2.92	32.46	0.962	3.41
	ShadowFormer [13] (AAAI'23)	39.67	0.992	5.21	38.82	0.983	2.30	35.46	0.973	2.80
	ShadowDiffusion [14] (CVPR'23)	39.82	-	4.90	38.90	-	2.30	35.72	-	2.70
	Li *et al.* [30] (ICCV'23)	38.46	0.989	5.93	37.27	0.977	2.90	34.14	0.960	3.39
	Liu *et al.* [34] (AAAI'24)	38.04	0.990	5.69	39.15	0.984	2.31	34.96	0.968	2.87
	RASM [32] (MM'24)	**40.73**	**0.993**	**4.41**	39.23	0.985	2.17	36.16	0.976	2.53
	Homoformer [53] (CVPR'24)	39.47	**0.993**	4.72	38.73	0.984	2.23	35.34	0.975	2.64
	OmniSR+GM [55] (AAAI'25)	-	-	-	-	-	-	34.20	0.973	-
	ShadowHack (Ours)	40.56	**0.993**	4.46	**39.66**	**0.987**	**2.09**	**36.31**	**0.977**	**2.48**
SRD	DHAN [8] (AAAI'20)	33.67	0.978	8.94	34.79	0.979	4.80	30.51	0.949	5.67
	BMNet [65] (CVPR'22)	35.05	0.981	6.61	36.02	0.982	3.61	31.69	0.956	4.46
	SG-ShadowNet [47] (ECCV'22)	-	-	7.53	-	-	2.97	-	-	4.23
	ShadowFormer [13] (AAAI'23)	36.91	0.989	5.90	36.22	0.989	3.44	32.90	0.958	4.04
	ShadowDiffusion [14] (CVPR'23)	38.72	0.987	4.98	37.78	0.985	3.44	34.73	0.970	3.63
	Li *et al.* [30] (ICCV'23)	39.33	0.985	6.09	35.61	0.967	2.97	33.17	0.940	3.83
	Liu *et al.* [34] (AAAI'24)	36.51	0.983	5.49	37.71	0.986	3.00	33.48	0.967	3.66
	DeS3 [24] (AAAI'24)	37.45	0.984	5.88	38.12	0.988	2.83	34.11	0.968	3.72
	RASM [32] (MM'24)	37.91	0.988	5.02	38.70	0.992	2.72	34.46	0.976	3.37
	Homoformer [53] (CVPR'24)	38.81	0.987	4.25	39.45	0.988	2.85	35.37	0.972	3.33
	OmniSR+GM [55] (AAAI'25)	-	-	-	-	-	-	34.56	0.977	-
	ShadowHack (Ours)	**39.47**	**0.991**	**4.18**	**40.06**	**0.994**	**2.38**	**35.94**	**0.982**	**2.90**

Table 1. Quantitative results on the ISTD+ and SRD datasets. The best results are in **bold**, while the second-best ones are underlined.

4.2. Color Regeneration Network

After restoring texture and lighting information in the luminance channel, the next task is to accurately regenerate color. Inspired by exemplar-based colorization models [23, 29, 39], we leverage contextual cues and condition-based guidance to improve color fidelity and ensure alignment with the semantic content of the images. Concretely, we propose a dual-encoder network, termed CRNet, which integrates a multi-scale color feature extractor into the U-shaped architecture of LRNet. The color encoder is a ConvNext-v2 atto model [37] pre-trained on ImageNet-21k with 2M parameters, which takes the Cb and Cr channels as input and extracts color features across multiple scales, as illustrated in Fig. 4 (b). To achieve precise color restoration, the network shall reference regions with similar/identical material properties and accurate color, using them as reliable guides for color regeneration. The outputs generated by LRNet provide qualified indexing information to facilitate this purpose. Therefore, we introduce an attention-based mechanism where luminance features serve as the query and key to compute similarity, while color features act as the value, aggregating essential color information. This approach injects color features into the skip connections of the U-shaped backbone, remarkably enriching the network's color regeneration capability.

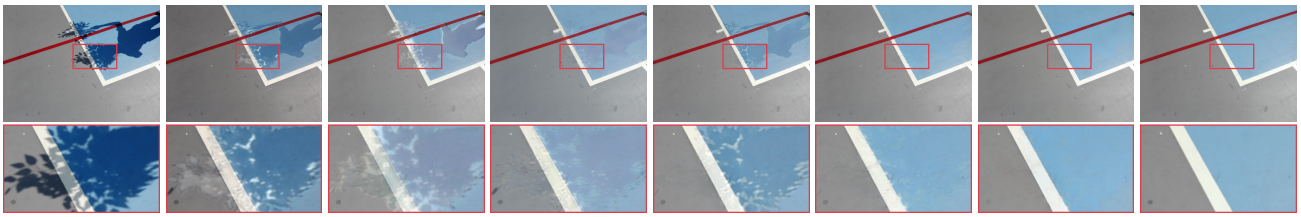
Multiple Checkpoint Training. We observe that the luminance network often yields better performance on the training set, which suggests the color regeneration network faces challenges in adapting to imperfect luminance outputs encountered during testing. To address this issue, we employ a simple yet effective way to overcome this issue, where multiple early-stage checkpoints from the luminance network training are randomly selected as candidates. This technique, applied only during training, strengthens the robustness of the color network without incurring additional computational costs at inference time.

5. Experimental Validation

5.1. Experimental Setup

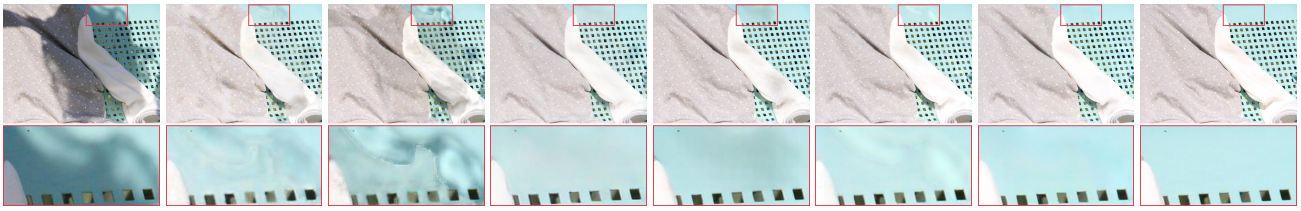
Implementation details. Our models are implemented in PyTorch, and the entire training process utilizes the AdamW optimizer [38] with momentum set to (0.9, 0.999) and a weight decay of 10^{-2}. The initial learning rate is set to 2×10^{-4}. During training, we randomly crop the training data into patches of size 384×384. For the base network of each stage, we use a four-layer encoder-decoder structure (from L1 to L4), consisting of three downsampling and three upsampling operations. The dimension of the first layer is 32. Our RTA modules are integrated into L3 and L4, with an overlapping ratio set to 0.5 and a rectified mechanism λ_0 value of 0.7. For training, we use L1 and VGG loss terms as employed in previous works [32, 34]. For detailed information on the implementation of each stage and training setup, please refer to the supplementary material.

Datasets. We conduct our experiments on the following benchmark datasets for shadow removal: 1) Adjusted ISTD (ISTD+) dataset [27], which refines the ISTD dataset [48]



(a) Input (b) BMNet [65] (c) SG [47] (d) SF [13] (e) HF [53] (f) RASM [32] (g) Ours (h) GT

Figure 7. Visual comparison on a sample from the ISTD+ dataset.



(a) Input (b) BMNet [65] (c) SG [47] (d) DMTN [33] (e) DeS3 [24] (f) HF [53] (g) Ours (h) GT

Figure 8. Visual comparison on a sample from the SRD dataset.

Metric	RGB Input	YCbCr Input	Retinex-based	Ours
PSNR↑	36.16	35.62	35.85	**36.31**
RMSE↓	2.53	2.62	2.63	**2.48**

Table 2. Ablation study of the proposed decoupling framework on the ISTD+ dataset.

by correcting the illumination discrepancies between the ground-truth and input images, includes 1330 training and 540 testing triplets (shadow image, shadow-free image, and shadow mask); 2) SRD dataset [43] consists of 2680 training and 408 testing pairs. We adopt the predicted masks provided by DHAN [8] for training and testing following previous works [32, 53, 65]; 3) UIUC dataset [16] provides 76 images with shadow mask for testing; and 4) UCF dataset [64] provided by Jin *et al.* [24] contains shadow images without shadow-free references.

Evaluation metrics. To evaluate the performance of our proposed framework, we utilize the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Root Mean Square Error (RMSE). PSNR and SSIM are both calculated in the RGB space, with higher values indicating better results. For RMSE, which is calculated in the LAB color space, the lower values indicate better results. To be consistent with previous works [32, 53, 65], we inference the full-resolution image and then evaluate the images at a resolution of 256×256 .

5.2. Comparison with Prior Arts

We compare our method with the state-of-the-art shadow removal approaches on benchmark datasets, including DHAN [8], G2R [36], BMNet [65], SG-ShadowNet [47], ShadowFormer [13], ShadowDiffusion [14], Li *et al.* [30], Liu *et al.* [34], DeS3 [24], Homoformer [53], RASM [32]



(a) Input (b) RGB Input (c) w/o rectify (d) Ours

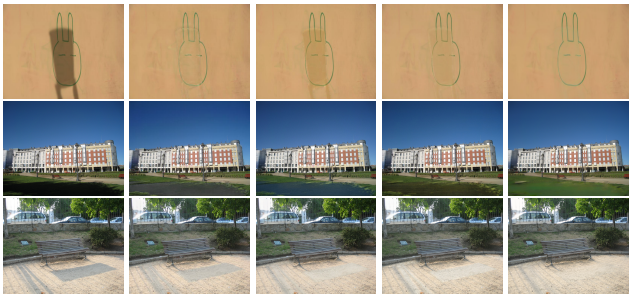
Figure 9. Ablation study on rectified attention. The results are displayed in the luminance space.

and OmniSR+GM [55] (GM means the version with shadow mask), as shown in Tab. 1. For a fair comparison, we use results published by the authors. Regarding missing data, we utilize available images or checkpoints provided by the authors and evaluate them via the same metrics.

Our proposed method achieves the highest scores in most metrics on the ISTD+ dataset compared to other state-of-the-art methods. Specifically, our model attains a PSNR of 36.31 for all image regions, surpassing the second-highest result of 36.16 by 0.15. The high PSNR highlights our model's ability to recover texture details effectively, demonstrating superior reconstruction quality. Additionally, the RMSE value of 2.48 is the lowest, which indicates that our recoloring module plays a crucial role in accurately restoring color information and reducing errors across the entire image. From the magnified areas in Figs. 7 and 8, it can be observed that our method not only effectively corrects color bias but also provides superior edge correction compared to other methods. The second part of Tab. 1 presents a comparison of our results on the SRD dataset against other state-of-the-art methods. It can be observed that our method achieves a significant improvement in RMSE (-0.43), further demonstrating the substantial benefit of the decoupling approach for color recovery.

Method	Shadow PSNR ↑	Non-Shadow PSNR ↑	All PSNR ↑
$[F_t]$	39.96	41.27	36.60
$[F_t; F_c]$	40.17	41.29	36.76
$[F_t]\&[F_t]$	40.02	41.18	36.58
$[F_t; F_c]\&[F_t; F_c]$	40.35	41.37	36.87
$[F_t; F_c]\&[F_c]$	**40.36**	**41.40**	**36.90**

Table 3. Ablation study of the proposed rectified attention module on the SRD dataset. We calculate PSNRs on the luminance.



(a) Input (b) SG [47] (c) SF [13] (d) HF [53] (e) Ours

Figure 10. Visual comparisons on the UIUC and UCF datasets.

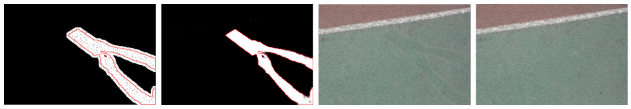
5.3. Ablation Study

Effect of the decoupling framework. To validate the effectiveness of our decoupling framework, we fine-tune our baseline model structure to adapt to RGB and YCbCr input/output for end-to-end shadow removal and compare its performance with our decoupling framework. As shown in Tab. 2, our decoupling approach significantly improves both PSNR and RMSE. Moreover, based on Retinex decomposition, we separately recover L and R, and compare it with our decoupling method. Detailed architectures are provided in the supplementary material. This improvement indicates that this strategy benefits texture recovery and color correction.

Effect of rectified attention module. We demonstrate the effectiveness of the rectified attention module in Tab. 3, by conducting experiments on rectification mechanisms. Furthermore, extra experiments are conducted on different feature combinations of rectified approaches to explore more suitable configurations. $[X]$ & $[Y]$ in Tab. 3 means performing a subtraction calculation of corresponding maps for rectification between X and Y. $[X]$ alone denotes only one pair of Q and K is used, where both inputs are X. Clearly, leveraging color information, even with inherent bias, can allow shadow regions to better reference surrounding areas, resulting in significant performance improvements. As for potential color bias, rectification refines the attention map for greater precision. Furthermore, we find that $[F_t; F_c]\&[F_c]$ slightly outperforms $[F_t; F_c]\&[F_t; F_c]$, and thus we ultimately choose $[F_t; F_c]\&[F_c]$. Fig. 9 visualizes the effectiveness of rectification, which has a more precise result with less edge remaining.

Method	Shadow RMSE ↓	Non-Shadow RMSE ↓	All RMSE ↓
w/o cross-att.	5.78	2.64	3.39
w/o multi checkpt.	5.01	2.59	3.28
Ours	**4.18**	**2.38**	**2.90**

Table 4. Ablation study of the CRNet on the SRD dataset.



(a) dirty mask (b) refined mask (c) w/o refinement (d) w refinement

Figure 11. Robustness analysis under corrupted mask. The accurate mask is highlighted in red. For fairness, we only involve mask refinement network here, and it is not involved in any comparision.

Effect of the color restoration network. To validate the effectiveness of the key component and design in the CRNet, we replace the cross-attention-based color feature injection with concatenation along the feature channel. As demonstrated in Tab. 4, if cross-attention is replaced, the accuracy has dropped by a large margin, indicating the effectiveness of the cross-attention. Furthermore, the model suffers from a misalignment between training and testing luminance results if no previous checkpoints are involved.

5.4. Robustness Analysis

Generalization to other datasets. To demonstrate the generalization capability of ShadowHack, we evaluate it on the UIUC [16] and UCF [64] datasets with our ISTD+ pre-trained model. The results are presented in Fig. 10, where ShadowHack achieves the cleanest shadow removal with minimal color bias. More comparisons are available in the supplementary material.

Robustness to inaccurate masks. To get rid of the performance degradation brought by inaccurate masks, we propose a mask refine network, which corrects the mask for better generalization. After our remediation, the result is recovered from inaccurate enlightening, as in Fig. 11 shows. More details are available in the supplementary material.

6. Conclusion

This work proposed a novel framework, *ShadowHack*, which decouples images into luminance and color components, enabling targeted processing for more effective shadow removal. This approach can well address limitations in existing end-to-end models by enhancing the use of color information to improve both texture and luminance recovery. To further boost shadow removal, a rectified attention mechanism has been designed to boost the reference ability of self-attention. Additionally, the color regeneration network has demonstrated its critical role in accurately correcting image colors. Extensive experimental results have underscored the superior performance and robustness of our model.

References

[1] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. vis. syst*, 2(3-26):2, 1978. 1, 3

[2] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *ICCV*, pages 12470–12479, 2023. 1

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9630–9640, 2021. 3

[4] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, pages 12299–12310, 2021. 3

[5] Xiangyu Chen, Zheyuan Li, Yuandong Pu, Yihao Liu, Jiantao Zhou, Yu Qiao, and Chao Dong. A comparative study of image restoration networks for general backbone network design. *arXiv preprint arXiv:2310.11881*, 2023. 3, 5

[6] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *CVPR*, pages 22367–22377, 2023. 3

[7] Michael E Churma. Blue shadows: physical, physiological, and psychological causes. *Applied optics*, 33(21):4719–4722, 1994. 1

[8] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting GAN. In *AAAI*, pages 10680–10687, 2020. 6, 7

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[10] Graham D. Finlayson, Steven D. Hordley, and Mark S. Drew. Removing shadows from images. In *ECCV (4)*, pages 823–836, 2002. 2

[11] Graham D. Finlayson, Steven D. Hordley, Cheng Lu, and Mark S. Drew. On the removal of shadows from images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(1):59–68, 2006. 2

[12] Graham D. Finlayson, Mark S. Drew, and Cheng Lu. Entropy minimization for shadow removal. *Int. J. Comput. Vis.*, 85(1): 35–57, 2009. 2

[13] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps shadow removal. In *AAAI*, pages 710–718, 2023. 1, 3, 6, 7, 8, 2, 4

[14] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *CVPR*, pages 14049–14058, 2023. 1, 2, 6, 7

[15] Lanqing Guo, Chong Wang, Wenhan Yang, Yufei Wang, and Bihan Wen. Boundary-aware divide and conquer: A diffusion-based solution for unsupervised shadow removal. In *ICCV*, pages 12999–13008. IEEE, 2023. 3

[16] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):2956–2967, 2013. 1, 2, 7, 8

[17] Xiaojie Guo and Qiming Hu. Low-light image enhancement via breaking down the darkness. *Int. J. Comput. Vis.*, 131(1): 48–66, 2023. 1, 3, 2

[18] Xiaojie Guo, Yu Li, and Haibin Ling. LIME: low-light image enhancement via illumination map estimation. *IEEE Trans. Image Process.*, 26(2):982–993, 2017. 1, 2, 3

[19] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *CVPR*, pages 6185–6194, 2023. 3

[20] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *CVPR*, pages 6185–6194, 2023. 5

[21] Jun-Wei Hsieh, Wen-Fong Hu, Chia-Jung Chang, and Yung-Sheng Chen. Shadow elimination for effective moving object detection by gaussian shadow modeling. *Image Vis. Comput.*, 21(6):505–516, 2003. 1

[22] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. In *ICCV*, pages 2472–2481, 2019. 1, 2

[23] Zhitong Huang, Nanxuan Zhao, and Jing Liao. Unicolor: A unified framework for multi-modal colorization with transformer. *ACM Trans. Graph.*, 41(6):205:1–205:16, 2022. 6

[24] Yeying Jin, Wei Ye, Wenhan Yang, Yuan Yuan, and Robby T. Tan. Des3: Adaptive attention-driven self and soft shadow removal using vit similarity. In *AAAI*, pages 2634–2642, 2024. 1, 6, 7, 2, 5

[25] Salman Hameed Khan, Mohammed Bennamoun, Ferdous Ahmed Sohel, and Roberto Togneri. Automatic shadow detection and removal from a single image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(3):431–446, 2016. 2

[26] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977. 1, 2

[27] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *ICCV*, pages 8577–8586, 2019. 1, 2, 6

[28] Hieu Le and Dimitris Samaras. Physics-based shadow image decomposition for shadow removal. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12):9088–9101, 2022. 1, 2, 3

[29] Junsoo Lee, Eungyeup Kim, Yunsung Lee, Dongjun Kim, Jaehyuk Chang, and Jaegul Choo. Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence. In *CVPR*, pages 5800–5809, 2020. 6

[30] Xiaoguang Li, Qing Guo, Rabab Abdelfattah, Di Lin, Wei Feng, Ivor W. Tsang, and Song Wang. Leveraging inpainting for single-image shadow removal. In *ICCV*, pages 13009–13018, 2023. 1, 2, 6, 7

[31] Feng Liu and Michael Gleicher. Texture-consistent shadow removal. In *ECCV (4)*, pages 437–450. Springer, 2008. 3

[32] Hengxing Liu, Mingjia Li, and Xiaojie Guo. Regional attention for shadow removal. In *ACM Multimedia*, 2024. 1, 3, 5, 6, 7, 2, 4

[33] Jiawei Liu, Qiang Wang, Huijie Fan, Wentao Li, Liangqiong Qu, and Yandong Tang. A decoupled multi-task network for

shadow removal. *IEEE Trans. Multim.*, 25:9449–9463, 2023. 7, 1, 5, 6

[34] Yuhao Liu, Zhanghan Ke, Ke Xu, Fang Liu, Zhenwei Wang, and Rynson W. H. Lau. Recasting regional lighting for shadow removal. In *AAAI*, pages 3810–3818, 2024. 1, 2, 3, 6, 7

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 9992–10002, 2021. 3

[36] Zhihao Liu, Hui Yin, Xinyi Wu, Zhenyao Wu, Yang Mi, and Song Wang. From shadow generation to shadow removal. In *CVPR*, pages 4927–4936, 2021. 1, 2, 6, 7

[37] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11966–11976, 2022. 6, 2

[38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[39] Peng Lu, Jinbei Yu, Xujun Peng, Zhaoran Zhao, and Xiaojie Wang. Gray2colornet: Transfer more colors from reference image. In *ACM Multimedia*, pages 3210–3218, 2020. 6

[40] David K. Lynch. Shadows. *Appl. Opt.*, 54(4):B154–B164, 2015. 1

[41] Kangfu Mei, Luis Figueroa, Zhe Lin, Zhihong Ding, Scott Cohen, and Vishal M Patel. Latent feature-guided diffusion models for shadow removal. In *WACV*, pages 4313–4322, 2024. 1

[42] Sohail Nadimi and Bir Bhanu. Physical models for moving shadow and object detection in video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(8):1079–1087, 2004. 1

[43] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson W. H. Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *CVPR*, pages 2308–2316, 2017. 1, 2, 7

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI (3)*, pages 234–241. Springer, 2015. 2

[45] Yael Shor and Dani Lischinski. The shadow meets the mask: Pyramid-based shadow removal. *Comput. Graph. Forum*, 27 (2):577–586, 2008. 2, 3

[46] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265, 2015. 2

[47] Jin Wan, Hui Yin, Zhenyao Wu, Xinyi Wu, Yanting Liu, and Song Wang. Style-guided shadow removal. In *ECCV (19)*, pages 361–378, 2022. 1, 6, 7, 8, 3, 4, 5

[48] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *CVPR*, pages 1788–1797, 2018. 1, 2, 6

[49] Tianyu Wang, Jianming Zhang, Haitian Zheng, Zhihong Ding, Scott Cohen, Zhe Lin, Wei Xiong, Chi-Wing Fu, Luis Figueroa, and Soo Ye Kim. Metashadow: Object-centered shadow detection, removal, and synthesis. In *CVPR*, pages 28252–28262, 2025. 1

[50] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, pages 17662–17672, 2022. 3

[51] Tai-Pang Wu and Chi-Keung Tang. A bayesian approach for shadow extraction from a single image. In *ICCV*, pages 480–487, 2005. 2

[52] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *CVPR*, pages 4784–4793, 2022. 3

[53] Jie Xiao, Xueyang Fu, Yurui Zhu, Dong Li, Jie Huang, Kai Zhu, and Zheng-Jun Zha. Homoformer: Homogenized transformer for image shadow removal. In *CVPR*, pages 25617–25626, 2024. 1, 3, 6, 7, 8, 2, 4, 5

[54] Han Xu, Hao Zhang, Xunpeng Yi, and Jiayi Ma. Cretinex: A progressive color-shift aware retinex model for low-light image enhancement. *Int. J. Comput. Vis.*, 132(9):3610–3632, 2024. 1

[55] Jiamin Xu, Zelong Li, Yuxin Zheng, Chenyu Huang, Renshu Gu, Weiwei Xu, and Gang Xu. Omnisr: Shadow removal under direct and indirect lighting. *arXiv preprint arXiv:2410.01719*, 2024. 3, 6, 7, 2

[56] Han Yang, Tianyu Wang, Xiaowei Hu, and Chi-Wing Fu. Silt: Shadow-aware iterative label tuning for learning to detect shadows from noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12687–12698, 2023. 2

[57] Qingxiong Yang, Kar-Han Tan, and Narendra Ahuja. Shadow removal using bilateral filtering. *IEEE Trans. Image Process.*, 21(10):4361–4368, 2012. 2

[58] Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. *arXiv preprint arXiv:2410.05258*, 2024. 5, 1

[59] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5718–5729, 2022. 3, 5

[60] Ling Zhang, Qing Zhang, and Chunxia Xiao. Shadow remover: Image shadow removal based on illumination recovering optimization. *IEEE Trans. Image Process.*, 24(11): 4623–4636, 2015. 2

[61] Ling Zhang, Chengjiang Long, Xiaolong Zhang, and Chunxia Xiao. RIS-GAN: explore residual and illumination with generative adversarial networks for shadow removal. In *AAAI*, pages 12829–12836, 2020. 1, 2

[62] Wuming Zhang, Xi Zhao, Jean-Marie Morvan, and Liming Chen. Improving shadow suppression for illumination robust face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41 (3):611–624, 2019. 1

[63] Wei Zhou, Julie Stephany Berrio, Stewart Worrall, and Eduardo M. Nebot. Automated evaluation of semantic segmentation robustness for autonomous driving. *IEEE Trans. Intell. Transp. Syst.*, 21(5):1951–1963, 2020. 1

[64] Jiejie Zhu, Kegan G. G. Samuel, Syed Zain Masood, and Marshall F. Tappen. Learning to recognize shadows in monochromatic natural images. In *CVPR*, pages 223–230, 2010. 7, 8

[65] Yurui Zhu, Jie Huang, Xueyang Fu, Feng Zhao, Qibin Sun, and Zheng-Jun Zha. Bijective mapping network for shadow removal. In *CVPR*, pages 5617–5626, 2022. 2, 6, 7, 1, 3, 4, 5