I^2HDiffuser: Image Illumination Harmonization Meets the Diffusion Model

Zhongyun Bao
tantouxy@163.com
School of Computer and Information
Anhui Polytechnic University, China

Gang Fu
xyzgfu@gmail.com
College of Computer Science and Software Engineering
Shenzhen University, China

Jianchi Sun
sunjc0306@whu.edu.cn
School of Computer Science
Wuhan University, China

Jing Zhou
zhoujing@whu.edu.cn
School of Computer Science
Wuhan University, China

Ziqi Yu
ziqiyu@whu.edu.cn
School of Computer Science
Wuhan University, China

Chunxia Xiao*
cxxiao@whu.edu.cn
School of Computer Science
Wuhan University, China

(a) Input (b) SGDiffusion (c) CFDiffusion (d) FHSG (e) Ours

Figure 1: Visual comparison results of different methods for illumination-shadow consistency generation of composite images.

ABSTRACT

Recently, since diffusion models show great potential in image generation, many pretrained diffusion models based image composition methods have been proposed for image illumination harmonization. However, they mainly face two key challenges: 1) the effective preservation of foreground appearance (i.e., content structure and texture details, etc); 2) Reasonable generation of the foreground casting shadow. To this end, we propose a novel Image Illumination Harmonization Diffusion model called I^2HDiffuser to achieve image illumination harmonization with high-fidelity foreground appearance and reasonable cast shadows. I^2HDiffuser mainly consists of frequency domain feature enhancement branch (FDFEB) and illumination-shadow consistency generation branch (ISCGB). Specifically, FDFEB first introduces the Wavelet Transform Module (WTM) for decomposing composite image features into low-frequency (i.e., illumination features, etc) and high-frequency (i.e., texture and content structure features, etc) components using the Haar wavelet transform. Then the Multi-Condition Guidance Mechanism (M-CGM) is proposed to interact these components as prior conditions, which are further injected into the ISCGB with a noise-to-denoise process for guiding high-fidelity content and background illumination-aware foreground regeneration. Meanwhile, a shadow mask step-wise iterative optimization strategy is introduced to the ISCGB to explicitly provide a reasonable shadow generation space for foreground objects. Extensive experiments on public image harmonization datasets **DESOBAv2** and **iHarmony4** and real illumination harmonization dataset **IH-SG** show that the I^2HDiffuser achieves the superiority.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Image illumination harmonization, shadow generation, diffusion model

ACM Reference Format:
Zhongyun Bao, Gang Fu, Jianchi Sun, Jing Zhou, Ziqi Yu, and Chunxia Xiao. 2025. I^2HDiffuser: Image Illumination Harmonization Meets the Diffusion Model. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (MM '25)*. ACM, New York, NY, USA, 10 pages. https://doi.org/https://doi.org/10.1145/3746027.3755314

1 INTRODUCTION

Image composition [1, 4, 37] is a fundamental task and widely used in computer vision [2, 53, 54, 57, 59] and augmented reality (AR)[44], which targets combining foreground and background

*Corresponding author

MM '25, October 27-31, 2025, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/https://doi.org/10.1145/3746027.3755314

scene from different illumination conditions to produce a composite image. However, the composite image usually has an inharmony appearance between the foreground and background due to different capture conditions, which usually leads to unsatisfactory visual effect and greatly affect the user sense of reality. Apparently, image illumination harmonization, aiming at achieving the seamless illumination blending between the foreground object and background scene of the composite image, is really an important and challenging task.

Existing deep learning-based image harmonization methods [2, 5, 8, 9, 12, 13, 18, 24, 29, 30, 33, 36, 41, 45, 46, 50] have developed various techniques from different perspectives (including image appearance, illumination, stylization, etc.) to address the illumination mismatch between foreground and background and achieved satisfactory results. However, they mainly focus on the foreground region and rarely consider the potential impact of foreground shadows, resulting in extremely unrealistic composition results. Recently, since diffusion models [19, 33, 42] show great potential in generating realistic images, many pretrained diffusion models based image composition methods [47, 52, 55, 56, 60] have been proposed for addressing image illumination consistency issues and achieved remarkable progress.

Among them, the methods [47, 52] use user-specified bounding box to aggregate a foreground object and a background scene from different illumination conditions to produce a realistic composite image. Similarly, the method [56] designs a two-stage fusion strategy and leverages the aligned foreground embedding map for feature modulation within diffusion model to achieve image illumination harmonization task. Besides, the methods [55, 60] take into account the controllability of foreground content for achieving image global illumination consistency. However, these methods still face two key challenges:1) the high-fidelity of local appearance of the foreground (e.g., semantic mismatch of local details); 2) the reasonable and accurate generation of foreground shadows matching the background illumination conditions.

Given these two key challenges, we propose a novel Image Illumination Harmonization Diffusion method named I^2HDiffuser, which is capable of achieving image illumination harmonization with the high-fidelity foreground appearance and reasonable shadows, as shown in Figure 1. I^2HDiffuser consists of the FDFEB providing strong guidance condition information and the ISCGB generating predetermined targets. Specifically, FDFEB takes a composite image as input and obtains the corresponding low-frequency (i.e, illumination features, etc) and high-frequency (i.e., foreground texture and content structure features, etc) components using the Haar wavelet transform [15]. The ISCGB, a conditional diffusion model, aims to take a noisy composite image as input and output a illumination harmonization result. Based on these two networks, a Multi-Condition Guidance Mechanism (M-CGM) is further proposed to realize the effective guidance of FDFEB to ISCGB.

Besides, since existing methods [47, 52, 56] use specified bounding box to constrain foreground position, the limited space without a clear target direction greatly limits the ability of foreground shadow generation. To effectively address the issue of reasonable generation of foreground shadows, our intuitive insight is to provide the ISCGB with a reasonable shadow generation space. Thus, we first define the shadow generation space as an shadow mask optimization problem , and further formulate the shadow generation problem as to jointly track image global illumination harmonization and refined foreground shadow mask. Specifically, given a coarse shadow mask prior produced by the pretrained model [20], a shadow mask step-wise iterative optimization strategy is introduced to the ISCGB. Note that, shadow mask optimization is designed as an auxiliary task of the ISCGB to progressively refine the shadow mask, enabling our model to generate more accurate and reasonable foreground shadow.

Our contributions are summarized as follows:

- We propose a novel Image Illumination Harmonization Diffusion model named I^2HDiffuser, which consists of frequency domain feature enhancement branch (FDFEB) and illumination-shadow consistency generation branch (ISCGB) to achieve image illumination harmonization with the high-fidelity foreground appearance and reasonable casting shadows.
- A Multi-Condition Guidance Mechanism (M-CGM) is proposed to effectively inject the prior conditions of FDFEB into ISCGB for guiding the generation of high-quality illumination harmonization images.
- A shadow generation space step-wise iterative optimization strategy is introduced to ISCGB for dynamically updating the target space of generating foreground cast shadow.

2 RELATED WORK

2.1 Image Illumination Harmonization

Previous image illumination harmonization works mainly consist of two categories: 1) traditional image illumination harmonization methods aim to adjust foreground to background appearance by leveraging low-level image representations, such as color distribution [39, 40, 51] and gradient information [22, 49]. 2) Some recent contributions [2, 5, 6, 10, 23, 24, 29, 33, 36, 45, 48, 61] are based on deep learning method, which design different end-to-end deep network structures to better understand image illumination harmonization from different perspectives.

The methods [9, 18, 46] used various attention modules to separately handle foreground and background, or modeled the relation between foreground and background to achieve image illumination harmonization. For example, the method [9] designed the channel-wise and spatial-wise attention mechanism to further improve the visual quality of harmonized results. The methods [2, 7, 8, 17, 30, 33] defined image illumination harmonization task as domain translation or style transfer problems. Among them, Cong *et al.* [8] defined the image harmonization as a domain adaptation problem and successfully used the enhanced U-Net generator to achieve notable performance. Bao *et al.*[2] used background shading stylization to guide the foreground illumination regeneration for effectively achieving global illumination harmonization.

Some methods [11–13] utilized Retinex theory [27] to achieve image illumination harmonization task by decomposing an image into reflectance map and lighting map. Besides, the method [38] proposed to use global information to guide foreground feature transformation and further transfer the foreground-background relation from real images to composite images for image illumination harmonization. Zhang *et al.*[56] presented a controllable image composition method that unifies four tasks in one diffusion

model: image blending, image harmonization, view synthesis, and generative composition.

2.2 Diffusion Models

Diffusion models are a class of deep generative models, which have recently have shown remarkable performance in image generation [19, 31] and become the new state-of-the-art (SOTA) generation models. Unsurprisingly, they have also shown great potential and been successfully applied to various CV tasks, including image editing [16, 21, 35, 58], super-resolution, inpainting [34, 43] and translation [25]. For the super-resolution through repeated refinement, they use DDPM [19] to make conditional image generation, and achieve image super resolution via a stochastic iterative denoising strategy. Rombach *et al.* [42] apply diffusion models in the latent space of powerful pretrained autoencoders , and turn diffusion models into powerful and flexible generators for general conditioning inputs via introducing cross-attention layers into the model architecture.

Besides, Lugmayr *et al.*[34] design an denoising strategy by resampling iterations for better conditioning the images and achieving high-quality and diverse output images for any inpainting form. Song *et al.* [47] presented the first diffusion model-based framework for generative object compositing that can handle mul-tiple aspects of compositing such as viewpoint, geometry, illumination and shadow. Recently, Zhou *et al.* [60] and Yu *et al.* [55] proposed the difussion model-based image illumination-shadow consistency generation. They all used the predicted coarse shadows and global image as prior conditions for controlling and guiding foreground illumination-shadow generation. However, they still cannot guarantee the high fidelity presentation of the local appearance of the foreground and the generation of casting shadows with reasonable shapes, resulting in a lack of realism in the results.

3 PROPOSED METHOD

3.1 Problem Formulation

Given one quaternion input (I, $\tilde{I}$, M, M_p) consists of a composite image $\tilde{I} \in \mathbb{R}^{H\times W\times 3}$ with H and W representing its height and width respectively, the corresponding ground truth image $I \in \mathbb{R}^{H\times W\times 3}$, the foreground mask $M \in \mathbb{R}^{H\times W\times 1}$ indicating the region to be harmonized and an estimated coarse shadow mask $M_p \in \mathbb{R}^{H\times W\times 1}$ indicating the foreground shadow generation region. Our goal is to train a generative model G, which is able to generate an illumination harmonization image $\hat{I}$ with the high fidelity foreground appearance and the reasonable shadow, expecting to be as close to I as possible. We thus formulate our generative model as $\hat{I} = G(\tilde{I}, M, M_p)$.

3.2 I^2HDiffuser Overview

We propose a novel Image Illumination Harmonization Diffusion model called I^2HDiffuser for achieving image illumination seamless integration. As illustrated in Figure 2, our I^2HDiffuser consists of frequency domain feature enhancement branch (FDFEB) containing the Wavelet Transform Module (WTM) and the Multi-Condition Guidance Mechanism (M-CGM) and illumination-shadow consistency generation branch (ISCGB), which collaborate with each other to complete our image illumination harmonization task.

WTM. As part of FDFEB (Figure 3), WTM first employs the auto-encoder based architecture[28] and takes the composite image $\tilde{I} \in \mathbb{R}^{C\times H\times W}$ with the corresponding foreground mask $M \in \mathbb{R}^{H\times W\times 1}$ as input to perform new spatial feature F extraction. Then, we further introduce wavelet transform module (WTM) based on the Haar wavelet transform [15] to convert spatial domain feature F into four frequency domain components:

$$A, H, V, D = HWT(F), \tag{1}$$

where, A, H, V, D present the low-frequency component, horizontal high-frequency component, vertical high-frequency component, and diagonal high-frequency component, respectively. The HWT presents the Haar wavelet transform, which is widely adopted in real-world applications due to its simplicity and computational efficiency. Further, we concatenate the high-frequency components H, V, D to form a high-frequency signal. Finally, the high-frequency signal and the low-frequency component are respectively followed by non-linear operations for feature mapping and obtaining ultimate high-frequency features (i.e., containing foreground texture and content structure features, et) F_h and low-frequency features (i.e, containing illumination features, etc) F_l:

$$F_l, F_h = (\theta(\delta_{1\times 1}(A)), \theta(\delta_{1\times 1}(Cat(H, V, D)))), \tag{2}$$

where the dimensions of F_h and F_l are $(C, H/2, W/2)$, $\theta(\cdot)$ and $\delta_{K\times K}$ present the batch normalization calculation and a convolution kernel size of $K \times K$, respectively.Note that the low-frequency features F_l and high-frequency features F_h play different roles in ISCGB, respectively. Among them, F_h is mainly used to enhance foreground content and structure details, ensuring that the high fidelity foreground appearance is generated in ISCGB. Meanwhile, F_l is employed to guide ISCGB to generate foreground illumination consistent with the background illumination.

ISCGB. The ISCGB, as shown in Figure 2 (the bottom branch), is designed as a controllable conditional generative model (IHDM) based on diffusion models. The ISCGB is to generate the global illumination harmonization result under the guidance of F_h and F_l from the WTM and the coarse foreground mask M_p predicted by the pretrained model [20]. Note that the M_p is mainly used to provide a initial coarse foreground shadow generation space for the input image, which is gradually optimized to achieve a coarse-to-fine result during the denoising process and effectively guides foreground shadow generation.

In ISCGB, our constructed denoising U-Net ϵ_{θ_t} at the step t first takes the time step t and the concatenation of the output features $F^{output}_{z_{t+1}}$ of the denoised U-Net $\epsilon_{\theta_{t+1}}$ and a shadow mask auxiliary feature M_{fa} as input $F^{input}_{z_t}$. Then the $F^{input}_{z_t}$ is passed through the encoder of ϵ_{θ_t} to produce intermediate output denoising feature F_{z_t}, which is followed by the WTM for generating frequency domain components $F_{z_t h}$ and $F_{z_t l}$. These components are further aggregated with F_h and F_l through the M-CGM to obtain a new denoising feature F_{out} with enhanced foreground content and sensitive background illumination. Finally, F_{out} is fed into the decoder of ϵ_{θ_t} to generate ultimate output $F^{output}_{z_t}$, which contains the gradually refined illumination harmonization image X_t and foreground shadow mask M_t.

Note that, we use a foreground shadow mask prediction head to predict the refined mask $\mathbf{M}_t$, which is connected to the last layer of

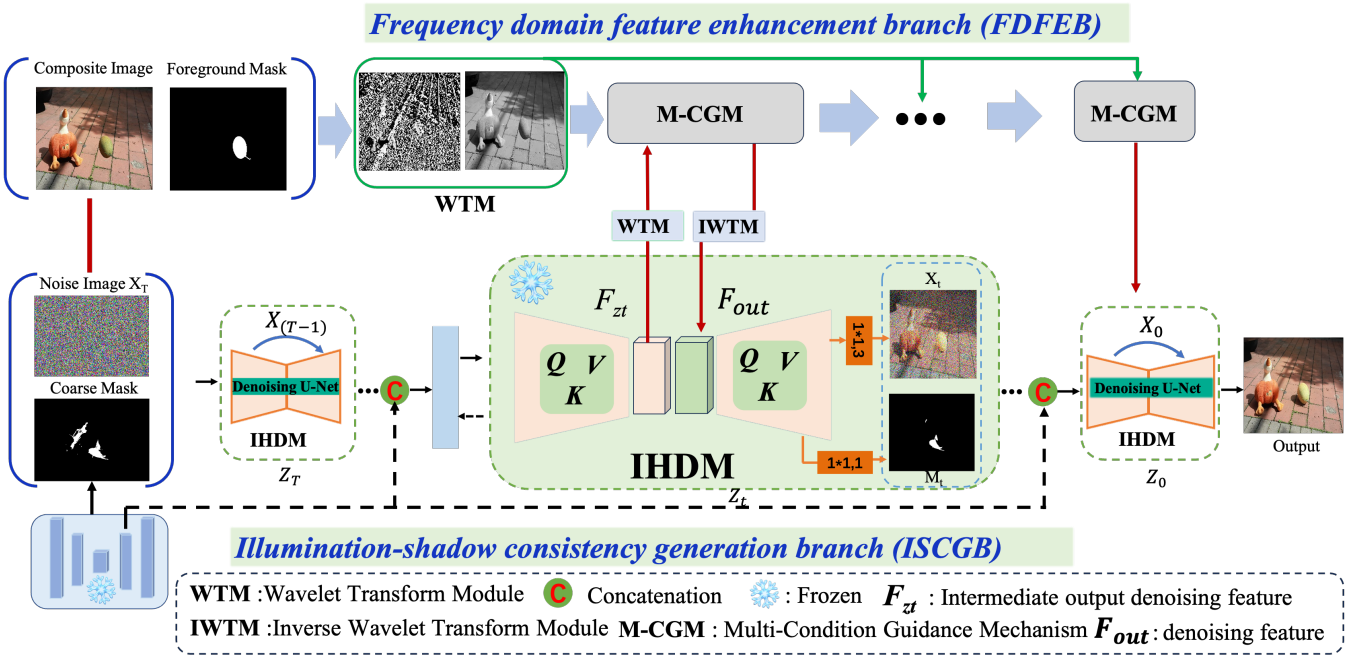

Figure 2: The overview of our I^2HDiffuser. It mainly consists of a frequency domain feature enhancement branch (FDFEB) and an illumination-shadow consistency generation branch (ISCGB).

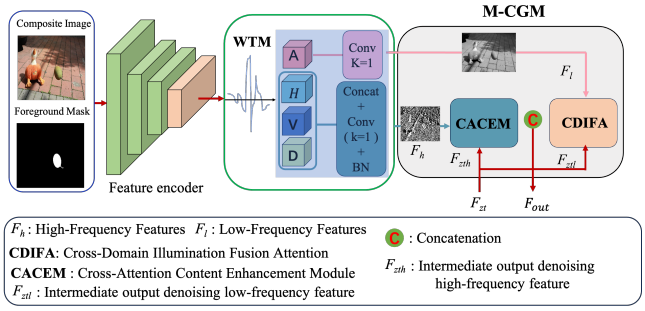

Figure 3: The structure of the FDFEB.

ϵ_θ and consists of one 1×1 convolution layer followed by a sigmoid function. Also, we adopt the corresponding ground truth shadow mask $\mathbf{M}_{GT}$ to further supervise by $\mathcal{L}_{\mathcal{FM}}$,

$$\mathcal{L}_{\mathcal{FM}} = \left\|\mathbf{M}_t - \mathbf{M}_{GT}\right\|_2^2, \tag{3}$$

where $\mathbf{M}_{GT}$ is obtained by binarizing the residual map between the composite image $\tilde{I}$ and the corresponding ground truth image I, i.e., setting $|\tilde{I} - I| > 0.5$ as 1 and others as 0.

In total, we can learn our controllable conditional illumination harmonization diffusion model (IHDM) by optimizing the objective function,

$$\mathcal{L}_{\mathcal{IHDM}} = \left\|\epsilon - \epsilon_{\theta_t}(X_t, t, F_l, F_h, M_p)\right\|_2^2 + \lambda \mathcal{L}_{\mathcal{FM}}, \tag{4}$$

where λ is the weighting coefficient to balance the influence of each term.

M-CGM. Multi-Condition Guidance Mechanism (M-CGM), as illustrated in Figure 3 (M-CGM), which aims to inject two different conditions F_h and F_l extracted by the WTM into the U-Net of our diffusion model to respectively preserve the foreground content details and guide the foreground illumination to be consistent with the background.

Specifically, M-CGM consists of two parts: 1) high fidelity foreground content enhancement module; 2) background illumination-aware guidance module. With the high-frequency features F_h and low-frequency features F_l of the WTM, F_h is first concatenated with the denoising feature $F_{z_t h}$ to form a new content feature F_{new}. To improve the capability of foreground content detail preservation, we design a cross-attention content enhancement module (CACEM), as shown in Figure 4, to achieve high fidelity foreground content aggregation. We separately employ F_{new} and F_h as Key and Value, while $F_{z_t h}$ is employed as Query. The cross-attention results in a global content-aware feature W, which is chunked into two parts in a channel-wise manner and further passed through the average pooling layer to produce two related content collaboration vectors W_1 and W_2. Finally, we separately conduct a channel-wise multiplication on F_h and $F_{z_t h}$ using W_1 and W_2, and further weight them to obtain output I_a with strengthen the content details.

To effectively exploit background illumination to guide harmonious foreground illumination generation, we further design a background illumination-aware cross-domain illumination fusion attention (CDIFA), as shown in Figure 5, to achieve illumination sensitive feature aggregation. CDIFA takes the $F_{z_t l}$ and the F_l as inputs, which are respectively fed into different weight filters composed of global average pooling and fully connected layers. Then, weight

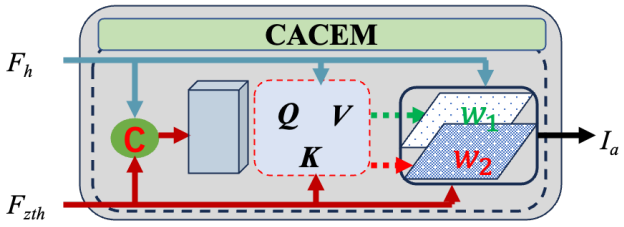

Figure 4: The structure of the CACEM.

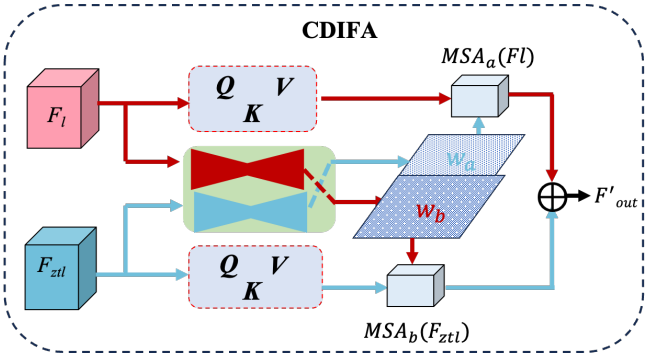

Figure 5: The structure of the CDIFA.

filters encode specific weight information for each channel and output corresponding weights W_a and W_b for a single channel. Meanwhile, the $F_{z_t l}$ and the F_l are also input into multi-head attention mechanisms, which are used to encode specific low-frequency information and obtain corresponding multi-channel features $MSA_a(F_l)$ and $MSA_b(F_{z_t l})$. Finally, these features are cross multiplied and summed to generate the final aggregated feature $F^{'}{}_{out}$:

$$F^{'}{}_{out} = MSA_a(F_l) \odot W_a + MSA_b(F_{z_t l}) \odot W_b. \tag{5}$$

With the $F^{'}{}_{out}$ and I_a, we aggregate them and output the corresponding spatial features F_{out} through inverse wavelet transform module (IWTM).

4 EXPERIMENTS

4.1 Implementation Details

Our I^2HDiffuser is implemented by PyTorch, which is trained using one RTX 3090Ti GPU. We use Adam optimizer with the momentum as (0.9, 0.999). The training epoch and initial learning rate are set as 1000 and 3×10^{-5}, respectively. We split the 15,000 quadruplets into 14,200 quadruplets for training and 800 quadruplets for testing. There is no crossover between our training dataset and testing dataset. Besides, we set the $\lambda = 0.5$ in our experiments.

4.2 Dataset and Evaluation Metrics

We conduct sufficient experimental verification and comparison on the following two datasets.

IH-SG Dataset. The IH-SG [60], a high-quality real-world outdoor illumination harmonization dataset, obtains 15000 quadruplets in total, each with a naive composite image, the corresponding masks of the foreground object and background object-shadow, and ground truth image. Besides, to facilitate the training and testing of our network, we set the resolution of our dataset to 256×256 for our experiments.

iHarmony4 dataset. The iHarmony4 dataset [8] consists of 4 sub-datasets: HCOCO, HAdobe5k, HFlickr, and Hday2night, each of which contains synthesized composite images, foreground masks of composite images, and corresponding real images.

DESOBAv2 dataset. The DESOBAv2 dataset [32] was originally used to shadow generation task, which has 21,575 images with 28,573 object-shadow pairs. In this paper, we improve this dataset that meets our global illumination editing task by perturbing the foreground appearance illumination of each image.

Evaluation Metrics. Following[2, 14], we evaluate the global image illumination harmonization results adopting two commonly-used metrics, i.e., Relative Mean Square Error (RMSE), Structural Similarity Index Measure (SSIM). Besides, foreground Mean Square Error (fMSE) and foreground Structural Similarity Index Measure (fSSIM) are also used to evaluate the harmonized foregrounds, which compute MSE and SSIM values between foreground regions of input and corresponding ground truth. Among them, the smaller fMSE, RMSE, and larger fSSIM, SSIM represent the better results.

Compared Methods. To prove the effectiveness of our method, we compare our method with five state-of-the-art image illumination harmonization methods, which include four global illumination harmonization methods DIH-GAN [2], ObjectStitch [47], FHSG[60], CFDiffusion[55], and one foreground cast shadow generation method SGDiffusion [32] .

4.3 Comparison with State-of-the-art Methods

Quantitative comparison. From Table 1 reporting the quantitative comparison results of state-of-the-art image illumination harmonization methods and our I^2HDiffuser, we can see that our method achieves the best quantitative results on all these four evaluation metrics on the IH-SG dataset. Especially compared to ObjectStitch [47] and SGDiffusion [32], our I^2HDiffuser largely outperforms them on SSIM, indicating that our generated global illumination harmonization image is closer to ground truth image. Besides, although DIH-GAN[2], CFDiffusion [55] and FHSG [60] also consider both foreground region illumination and shadows, the corresponding quantitative results show that its generalization ability on real datasets is far inferior to our method. Apparently, the significant advantage of our method on the real-world dataset is mainly attributed to the powerful generation ability of our diffusion model and the effective collaboration between the FDFEB and ISCGB.

Qualitative comparison. Figure 6 visualizes the qualitative results of our method and state-of-the-art baseline methods on real IH-SG dataset. We can see that our method achieves the best visual effect with consistent foreground illumination and more reasonable shadows. Note that SGDiffusion [32] ignores the foreground region illumination consistency generation and produces the undesirable appearance that does not match the background illumination, seriously effecting the quality of the generated result. Although ObjectStitch [47], DIH-GAN[2], CFDiffusion [55] and FHSG [60] all consider foreground illumination and casting shadow generation, they fail to be effectively generalized to real-world

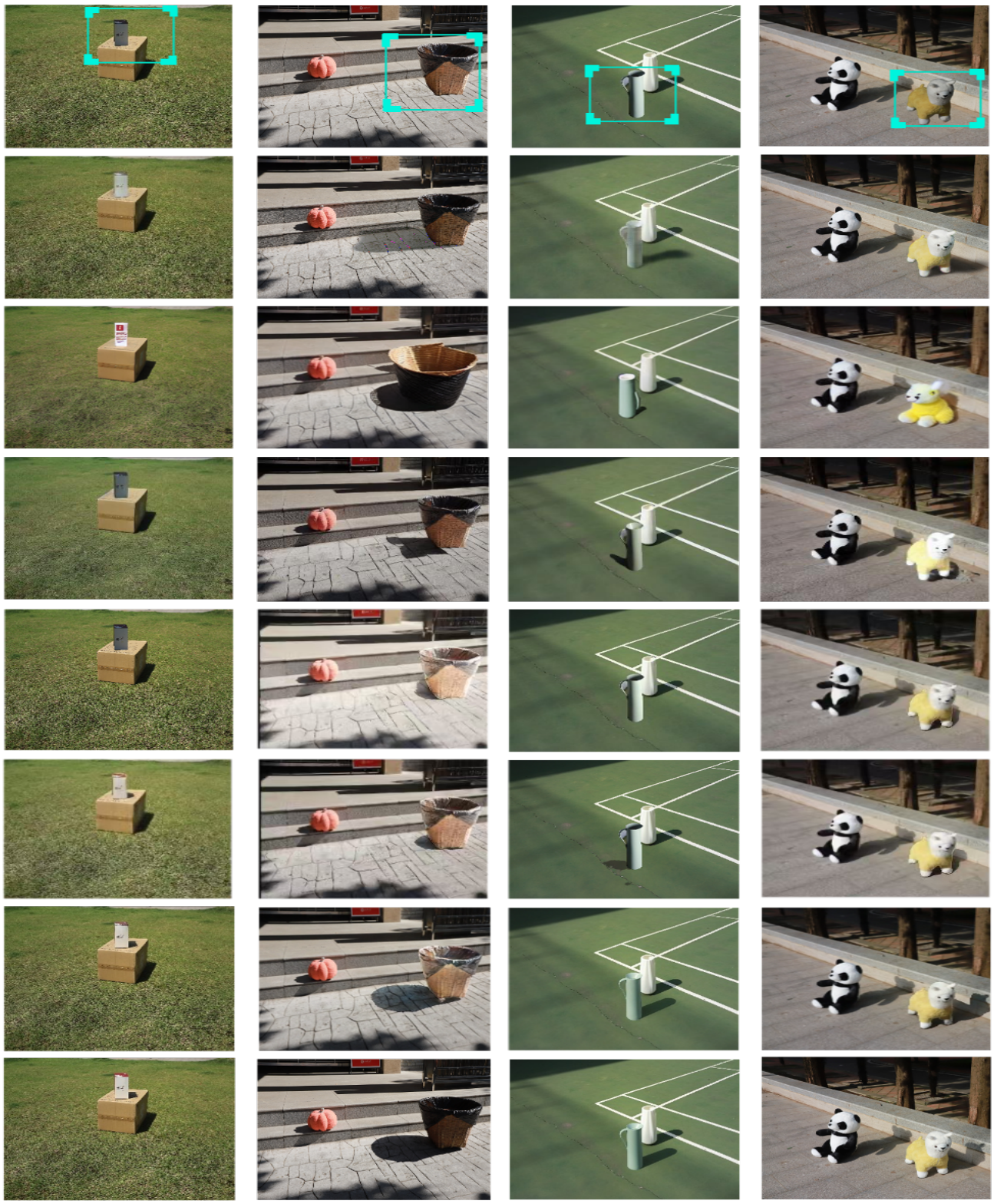

Figure 6: Visual comparison of our method with state-of-the-art methods on real-world scenes. From the first to eighth rows are the Input, the results of methods DIH-GAN, ObjectStitch, SGDiffusion, CFDiffusion, FHSG, I^2HDiffuser, and GT, respectively.

Table 1: Results of quantitative comparison on the testing set of IH-SG. "↑" indicates the higher the better, and "↓" indicates the lower the better. The best results are marked in bold.

Method	RMSE ↓	SSIM ↑	fMSE ↓	fSSIM ↑
SGDiffusion [32]	6.724	0.825	915.412	0.809
ObjectStitch [47]	9.345	0.798	1128.446	0.752
DIH-GAN [2]	6.658	0.847	518.465	0.876
CFDiffusion [55]	5.126	0.917	367.919	0.937
FHSG [60]	5.248	0.923	374.89	0.935
Ours	**4.987**	**0.942**	**342.153**	**0.952**

Table 2: Results of quantitative comparison on iHarmony4 and DESOBAv2 (formating iHarmony4 / DESOBAv2), respectively. "↑" indicates the higher the better, and "↓" indicates the lower the better. The best results are marked in bold.

Method	RMSE ↓	SSIM ↑	fMSE ↓	fSSIM ↑
SGDiffusion [32]	9.702 / 6.986	0.805 / 0.812	1615.433 / 1322.141	0.823 / 0.912
ObjectStitch [47]	8.973 / 6.742	0.842 / 0.853	1125.124 / 1024.145	0.845 / 0.897
DIH-GAN [2]	5.826 / 5.453	0.893 / 0.885	792.432 / 698.154	0.897 / 0.903
CFDiffusion [55]	4.902 / 5.214	0.918 / 0.927	527.234 / 367.919	0.911 / 0.924
FHSG [60]	4.910 / 5.032	0.920 / 0.923	582.143 / 425.286	0.917 / 0.920
Ours	**4.253 / 4.956**	**0.951 / 0.947**	**415.256 / 347.244**	**0.946 / 0.938**

dataset. Apparently, due to the lack of effective shadow generation space and foreground appearance content constraint, ObjectStitch [47] produced unrealistic illumination-shadow consistency results. DIH-GAN[2], CFDiffusion [55] and FHSG [60] strongly rely on shadow clues of real occluders in the background scene. They generated illumination-shadow results are highly dependent on the quality of the estimated foreground shadow.

Figure 7: Two testing cases of different methods on the public iHarmony4 dataset. From left to right are composite images, the results of DucoNet [48], PCT-Net [10] and the I^2HDiffuser, and ground truth images, respectively.

In contrast, our I^2HDiffuser effectively and intuitively achieves global illumination harmonization results, which utilizes collaboration between the FDFEB and the ISCGB and the effective guidance of the foreground shadow generation space.

Besides, we also provide some visual testing cases of different methods on public **iHarmony4** dataset (Figure 7), **DESOBAv2** dataset (Figure 1) and randomly captured **daily life real-world image** (Figure 8) . The quantitative comparision results on iHarmony4 and DESOBAv2 datasets are also reported in Table 2. It can be seen that our method outperforms the best existing related methods in both the generation of appearance illumination of foreground

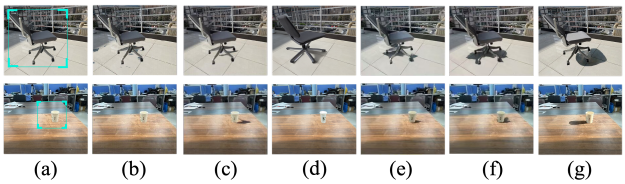

Figure 8: Two testing cases from daily life scene.From left to right are composite images, the results of DIH-GAN [2], ObjectStitch [47], SGDiffusion [32], CFDiffusion [55], FHSG [60] and the I^2HDiffuser, respectively.

region and foreground cast shadows, achieving impressive image illumination-shadow consistency results, which fully demonstrates the effectiveness of our I^2HDiffuser.

4.4 User Study on Real Composite Images

We also conduct a user study as done in [2, 38] to further evaluate the performance of our method and other four competitive methods. We first collect 200 extra real-world composite images outside the IH-SG dataset. For each composite image, we can acquire five different illumination harmonization results by using different methods (four baselines and our method). Then, we thus judge the realism of illumination harmonization results based on subjective visual effects. Specifically, we recruit 90 participants from different professions and ask them to select the more harmonious result from an image pair each time. Finally, we take the collected results to calculate the global ranking of all methods using the Bradley-Terry (B-T) model [3, 26]. The B-T scores are reported in Table 3, we can see that our I^2HDiffuser obtain the highest B-T score, which fully proves the superiority of our method on real datasets.

Table 3: B-T scores of different methods on 200 real composite images.

	Comp	SGDiffusion	ObjectStitch	DIH-GAN	CFDiffusion	FHSG	Ours
B-Tscores	0.041	0.203	0.306	0.241	0.289	0.276	0.324

4.5 Ablation Study

To prove the effectiveness of each design choice in our I^2HDiffuser, we further conduct ablation study by modifying the I^2HDiffuser architecture to evaluate the performance of different design choices. We mainly conduct experiments from the following five aspects: 1) Removing CACEM (i.e., w/o CACEM) for researching the impact of foreground content on the preservation of local detail content; 2) Removing CDIFA (i.e., w/o CDIFA) to study the guidance of background illumination on the generation of foreground illumination consistency; 3) removing Haar wavelet transform (w/o WT) to study the importance of frequency domain feature operations on the whole image illumination consistency generation; 4) investigating the interaction between CDIFA and CACEM in the task of generating global illumination consistency, including all using CACEM (all CACEM), all using CDIFA (all CDIFA) and exchanging them (CDIFA & CACEM). 5) Removing shadow generation space optimization strategy (i.e., w/o SGSO) to study the performance of foreground shadow generation.

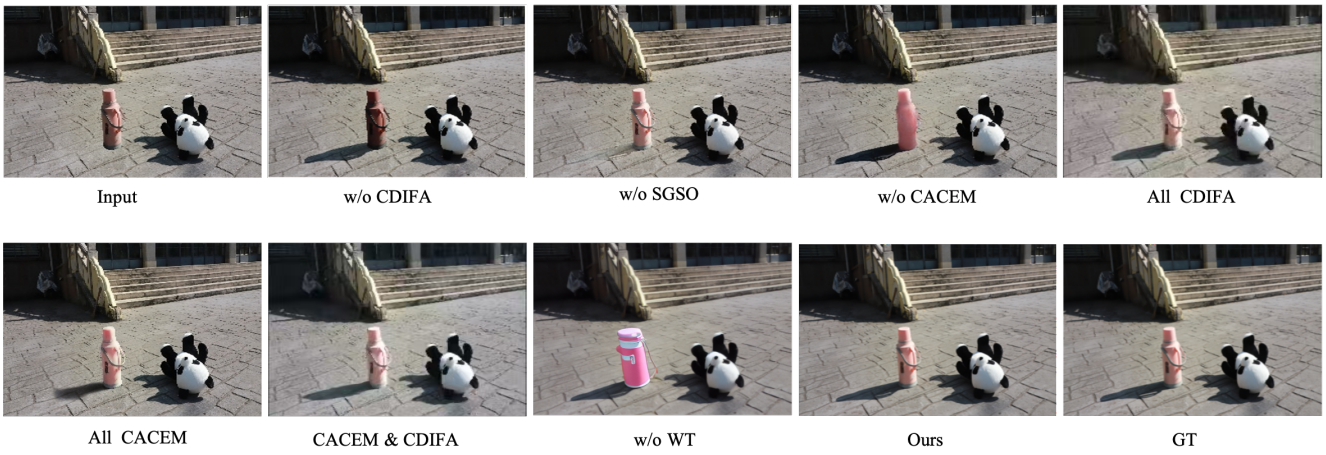

Figure 9: The visual results of the ablation study, including the input, the results of the w/o CDIFA, w/o SGSO, w/o CACEM, all CDIFA, all CACEM, CACEM & CDIFA, w/o WT, the I^2HDiffusion, and ground truth, respectively.

Table 4: Quantitative results of the ablation study on the IH-SG dataset. "↑" indicates the higher the better, and "↓" indicates the lower the better. The best results are marked in bold.

Method	RMSE ↓	SSIM ↑	fMSE ↓	fSSIM ↑
w/o CACEM	8.573	0.812	1224.536	0.796
w/o CDIFA	6.432	0.867	712.253	0.851
all CACEM	7.443	0.853	912.245	0.843
all CDIFA	8.432	0.847	1018.143	0.825
CDIFA & CACEM	7.456	0.819	611.254	0.806
w/o WT	8.154	0.806	978.543	0.779
w/o SGSO	5.124	0.893	542.142	0.898
Ours	**4.987**	**0.942**	**342.153**	**0.952**

Figure 9 and Table 4 report the qualitative and quantitative results of our ablation study, respectively. From Figure 9, we can see that w/o CACEM results in a mismatch foreground local appearance, i.e. incorrect generation of foreground local detail content, this is mainly because our method lacks strong enhancement of the foreground content priors to preserve local detail information. Similarly, when we perform the w/o CDIFA, due to the lack of effective guidance from background illumination, the generated foreground illumination fails to match the background illumination significantly, resulting in an inharmonious result as shown in the second column of the first row in Figure 9. Besides, the result of CACEM & CDIF also proves the uniqueness of the design of them.

Besides, w/o SGSO also produces an unrealistic result without reasonable shadow (the third column of the first row in Figure 9), which indirectly proves the effectiveness of our proposed shadow generation space optimization strategy. It is worth noting that when we remove the entire wavelet transform branch, i.e. w/o WT, we can see that the whole generated foreground appearance lacks realism due to the lack of controllable content constraints and illumination guidance during the generation process. In contrast, our I^2HDiffuser achieves the best illumination harmonization result with controllable foreground appearance and reasonable shadows. From Table 4, it can also be seen that our method obtains the best values on all four quantitative metrics, which quantitatively proves the superiority of our I^2HDiffuser.

Limitations. Although our I^2HDiffuser can produce impressive image illumination-shadow consistency results in real-world images, it is necessary to point out that our method faces challenges in generating complex non-planar casting shadows, which requires more geometry information as assistance. Besides, real-time video illumination-shadow consistency processing is also a great challenge for our I^2HDiffuser.

5 CONCLUSION AND FUTURE WORK

In this work, we propose a novel Image Illumination Harmonization Diffusion model called I^2HDiffuser, which consists of the FDFEB and ISCGB, and achieves image illumination harmonization with high fidelity foreground appearance and reasonable shadows.

Specifically, FDFEB first introduces the Wavelet Transform Module (WTM), which decomposes composite image features into low-frequency (i.e, containing illumination, etc) and high-frequency (i.e., containing texture and content structure features, etc) components using the Haar wavelet transform. With these components, a Multi-Condition Guidance Mechanism (M-CGM) is proposed to aggregate them as prior conditions, which are passed through the inverse wavelet transform (IWT) and further injected into the ISCGB based on diffusion models for generating global illumination harmonization result with high fidelity content details and background illumination-aware. Meanwhile, a shadow mask step-wise iterative optimization strategy is introduced to ISCGB for dynamically guiding foreground shadows generation, achieving global illumination harmonization results.

In the future, we will further expand our method to the multi-foreground objects illumination harmonization and video global illumination harmonization tasks.

ACKNOWLEDGMENTS

This work is partially supported by the National Natural Science Foundation of China (No. 62372336, No. 61972298 and No. 62402324).

REFERENCES

[1] Zhongyun Bao, Gang Fu, Zipei Chen, and Chunxia Xiao. 2024. Illuminator: Image-based illumination editing for indoor scene harmonization. *Computational Visual Media* 10, 6 (2024), 1137–1155.

[2] Zhongyun Bao, Chengjiang Long, Gang Fu, Daquan Liu, Yuanzhen Li, Jiaming Wu, and Chunxia Xiao. 2022. Deep Image-Based Illumination Harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18542–18551.

[3] Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39, 3/4 (1952), 324–345.

[4] Bor Chun Chen and Andrew Kae. 2019. Toward Realistic Image Compositing With Adversarial Learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] Haoxing Chen, Zhangxuan Gu, Yaohui Li, Jun Lan, Changhua Meng, Weiqiang Wang, and Huaxiong Li. 2023. Hierarchical dynamic image harmonization. In *Proceedings of the 31st ACM International Conference on Multimedia*. 1422–1430.

[6] Wenyan Cong, Junyan Cao, Li Niu, Jianfu Zhang, Xuesong Gao, Zhiwei Tang, and Liqing Zhang. 2021. Deep Image Harmonization by Bridging the Reality Gap. (2021).

[7] Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. 2021. Bargainnet: Background-guided domain translation for image harmonization. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[8] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. 2020. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8394–8403.

[9] Xiaodong Cun and Chi-Man Pun. 2020. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Processing* 29 (2020), 4759–4771.

[10] Julian Jorge Andrade Guerreiro, Mitsuru Nakazawa, and Björn Stenger. 2023. PCT-Net: Full Resolution Image Harmonization Using Pixel-Wise Color Transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5917–5926.

[11] Zonghui Guo, Zhaorui Gu, Bing Zheng, Junyu Dong, and Haiyong Zheng. 2022. Transformer for image harmonization and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[12] Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. 2021. Image harmonization with transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 14870–14879.

[13] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. 2021. Intrinsic image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16367–16376.

[14] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. 2021. Intrinsic Image Harmonization. In *Computer Vision and Pattern Recognition*.

[15] Alfred Haar. 1911. Zur theorie der orthogonalen funktionensysteme. *Math. Ann.* 71, 1 (1911), 38–53.

[16] Roy Hachnochi, Mingrui Zhao, Nadav Orzech, Rinon Gal, Ali Mahdavi-Amiri, Daniel Cohen-Or, and Amit Haim Bermano. 2023. Cross-domain Compositing with Pretrained Diffusion Models. *arXiv preprint arXiv:2302.10167* (2023).

[17] Yucheng Hang, Bin Xia, Wenming Yang, and Qingmin Liao. 2022. Scs-co: Self-consistent style contrastive learning for image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19710–19719.

[18] Guoqing Hao, Satoshi Iizuka, and Kazuhiro Fukui. 2020. Image Harmonization with Attention-based Deep Feature Modulation.. In *BMVC*, Vol. 2. 4.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[20] Yan Hong, Li Niu, and Jianfu Zhang. 2022. Shadow generation for composite image in real-world scenes. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 914–922.

[21] Jaeseok Jeong, Mingi Kwon, and Youngjung Uh. 2023. Training-free Style Transfer Emerges from h-space in Diffusion models. *arXiv preprint arXiv:2303.15403* (2023).

[22] Jiaya Jia, Jian Sun, Chi-Keung Tang, and Heung-Yeung Shum. 2006. Drag-and-drop pasting. *ACM Transactions on graphics (TOG)* 25, 3 (2006), 631–637.

[23] Yifan Jiang, He Zhang, Jianming Zhang, Yilin Wang, Zhe Lin, Kalyan Sunkavalli, Simon Chen, Sohrab Amirghodsi, Sarah Kong, and Zhangyang Wang. 2021. SSH: A Self-Supervised Framework for Image Harmonization. (2021).

[24] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson WH Lau. 2022. Harmonizer: Learning to perform white-box image and video harmonization. In *European Conference on Computer Vision*. Springer, 690–706.

[25] Gihyun Kwon and Jong Chul Ye. 2022. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264* (2022).

[26] Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang. 2016. A comparative study for single image blind deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1701–1709.

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

[28] Hui Li and Xiao-Jun Wu. 2024. CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion* 103 (2024), 102147.

[29] Jingtang Liang, Xiaodong Cun, Chi-Man Pun, and Jue Wang. 2022. Spatial-separated curve rendering network for efficient and high-resolution image harmonization. In *European Conference on Computer Vision*. Springer, 334–349.

[30] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. 2021. Region-aware adaptive instance normalization for image harmonization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9361–9370.

[31] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. 2022. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778* (2022).

[32] Qingyang Liu, Junqi You, Jianting Wang, Xinhao Tao, Bo Zhang, and Li Niu. 2024. Shadow Generation for Composite Image Using Diffusion model. *arXiv preprint arXiv:2403.15234* (2024).

[33] Lingxiao Lu, Jiangtong Li, Junyan Cao, Li Niu, and Liqing Zhang. 2023. Painterly Image Harmonization using Diffusion Model. In *Proceedings of the 31st ACM International Conference on Multimedia*. 233–241.

[34] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11461–11471.

[35] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073* (2021).

[36] Quanling Meng, Liu Qinglin, Zonglin Li, Xiangyuan Lan, Shengping Zhang, and Liqiang Nie. 2024. High-Resolution Image Harmonization with Adaptive-Interval Color Transformation. *Advances in Neural Information Processing Systems* 37 (2024), 13769–13793.

[37] Li Niu, Wenyan Cong, Liu Liu, Yan Hong, Bo Zhang, Jing Liang, and Liqing Zhang. 2021. Making Images Real Again: A Comprehensive Survey on Deep Image Composition. (2021).

[38] Li Niu, Linfeng Tan, Xinhao Tao, Junyan Cao, Fengjun Guo, Teng Long, and Liqing Zhang. 2023. Deep image harmonization with globally guided feature transformation and relation distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7723–7732.

[39] Francois Pitie, Anil C Kokaram, and Rozenn Dahyot. 2005. N-dimensional probability density function transfer and its application to color transfer. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, Vol. 2. IEEE, 1434–1439.

[40] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. 2001. Color transfer between images. *IEEE Computer graphics and applications* 21, 5 (2001), 34–41.

[41] Mengwei Ren, Wei Xiong, Jae Shin Yoon, Zhixin Shu, Jianming Zhang, HyunJoon Jung, Guido Gerig, and He Zhang. 2024. Relightful harmonization: Lighting-aware portrait background replacement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6452–6462.

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[43] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–10.

[44] Shiyuan Shen, Zhongyun Bao, Wenju Xu, and Chunxia Xiao. 2025. Illumidiff: indoor illumination estimation from a single image with diffusion model. *IEEE transactions on visualization and computer graphics* (2025).

[45] Xintian Shen, Jiangning Zhang, Jun Chen, Shipeng Bai, Yue Han, Yabiao Wang, Chengjie Wang, and Yong Liu. 2023. Learning Global-aware Kernel for Image Harmonization. *arXiv preprint arXiv:2305.11676* (2023).

[46] Konstantin Sofiiuk, Polina Popenova, and Anton Konushin. 2021. Foreground-aware semantic representations for image harmonization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1620–1629.

[47] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. 2023. ObjectStitch: Object Compositing With Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18310–18319.

[48] Linfeng Tan, Jiangtong Li, Li Niu, and Liqing Zhang. 2023. Deep Image Harmonization in Dual Color Spaces. In *Proceedings of the 31st ACM International Conference on Multimedia*. 2159–2167.

[49] Michael W Tao, Micah K Johnson, and Sylvain Paris. 2013. Error-tolerant image compositing. *International journal of computer vision* 103 (2013), 178–189.

[50] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. 2017. Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3789–3797.

[51] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly Rushmeier. 2012. Understanding and improving the realism of image composites. *ACM Transactions on graphics (TOG)* 31, 4 (2012), 1–10.

[52] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. 2023. Paint by Example: Exemplar-Based Image Editing With Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18381–18391.

[53] Hanning Yu, Wentao Liu, Chengjiang Long, Bo Dong, Qin Zou, and Chunxia Xiao. 2021. Luminance Attentive Networks for HDR Image and Panorama Reconstruction. (2021).

[54] Jiangjian Yu, Ling Zhang, Qing Zhang, Qifei Zhang, Daiguo Zhou, Chao Liang, and Chunxia Xiao. 2024. Portrait shadow removal using context-aware illumination restoration network. *IEEE Transactions on Image Processing* (2024).

[55] Ziqi Yu, Jing Zhou, Zhongyun Bao, Gang Fu, Weilei He, Chao Liang, and Chunxia Xiao. 2024. CFDiffusion: Controllable Foreground Relighting in Image Compositing via Diffusion Model. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24)*. Association for Computing Machinery, New York, NY, USA, 3647–3656. https://doi.org/10.1145/3664647.3681283

[56] Bo Zhang, Yuxuan Duan, Jun Lan, Yan Hong, Huijia Zhu, Weiqiang Wang, and Li Niu. 2023. ControlCom: Controllable Image Composition using Diffusion Model. arXiv:2308.10040 [cs.CV]

[57] Ling Zhang, Zhenyu Li, Liang Cheng, Qing Zhang, Zheng Liu, Xiaolong Zhang, and Chunxia Xiao. 2025. DLIENet: A lightweight low-light image enhancement network via knowledge distillation. *Pattern Recognition* (2025), 111777.

[58] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2022. Inversion-based creativity transfer with diffusion models. *arXiv preprint arXiv:2211.13203* (2022).

[59] Hongsheng Zheng, Wenju Xu, Zhenyu Wang, Xiao Lu, and Chunxia Xiao. 2024. Facial Highlight Removal With Cross-Context Attention and Texture Enhancement. *IEEE Transactions on Circuits and Systems for Video Technology* (2024).

[60] Jing Zhou, Ziqi Yu, Zhongyun Bao, Gang Fu, Weilei He, Chao Liang, and Chunxia Xiao. 2024. Foreground Harmonization and Shadow Generation for Composite Image. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24)*. Association for Computing Machinery, New York, NY, USA, 8267–8276. https://doi.org/10.1145/3664647.3681355

[61] Ziyue Zhu, Zhao Zhang, Zheng Lin, Ruiqi Wu, Zhi Chai, and Chun Le Guo. 2022. Image Harmonization by Matching Regional References. (2022).