Shadow Removal via Global Residual Free Unet and Shadow Generation

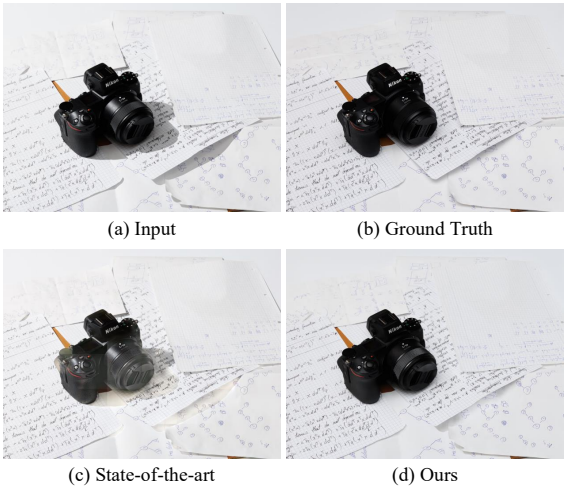
Dong Li, Xin Lu, Yurui Zhu, Xi Wang, Jie Xiao, Yunpeng Zhang, Xueyang Fu*, Zheng-Jun Zha
University of Science and Technology of China
dongli6@mail.ustc.edu.cn, xyfu@ustc.edu.cn

Abstract

Existing shadow removal methods face ch confronted with real-world scenes, parti dealing with complex background informat resolution images. To address these issues, shadow removal framework based on shad and Global Residual Free Unet (GRFUnet), w shadow removal from both data and network For data enhancement, we train a generati network for producing shadow masks, which tiplied with clean images to obtain new shado additional shadow data produced by this str ation approach allows for ample constraini work in color aspects, thereby enhancing col in the images. In terms of network archite sign the Global Residual Free Unet that emp lutional framework to sequentially conduct S tion and Channel Evolution for effective feature extraction. Moreover, we eliminate the commonly used Global residual connections in image restoration, as we discern their ineffectiveness for the non-additive task of shadow removal. Through this methodology, we achieve excellent shadow removal results, both qualitatively and quantitatively. Our method achieves the highest PSNR in the NTIRE24-Image Shadow Removal Challenge and achieves commendable and balanced outcomes across two tracks—placing third in the fidelity track and fourth in the perceptual track.

1. Introduction

Shadow removal is an important research direction in computer vision. Shadows are typically observed in various natural scenes when light sources are partially or completely obstructed by objects. Inevitably, the presence of shadows in images reduces the perceptual quality of background information. Consequentially, image shadows present a series of challenges to subsequent advanced visual tasks such as object tracking [32] and detection [30], semantic segmentation [44]. The underlying shadow formation model essentially describes a highly complex phenomenon, where the intensity and shape of the shadows are determined by multiple factors, including the attributes of the light source, the geometric shape of shadow-casting light, and the surface properties of the shadow-casting object. Such a complex yet significant system necessitates extensive research and shadow removal has continued to garner significant attention in recent years.



(a) Input (b) Ground Truth (c) State-of-the-art (d) Ours

Figure 1. Visualization comparison between the previous state-of-the-art method [16] and our method. Our method demonstrates superior background reconstruction and color consistency when dealing with complex scenes at high resolutions.

Current methods for image shadow removal can be broadly divided into two categories: traditional methods based on physical models and solutions based on deep learning. Early works typically rely on the physical characteristics of the deterministic shadow formation model. These works explore and utilize various physical priors, such as illumination [38], image gradients [15], and regions [18, 36] *etc*. While these methods have advanced the understanding of the physics behind shadow removal, they struggle to cope with real-world scenarios due to the high complexity of real-world shadow formation models.

*Corresponding author.

Subsequently, with the advancement of deep learning and its success in computer vision [9, 12, 25, 43, 45, 46], this data-driven approach has been introduced into the field of shadow removal and has achieved accomplishments. For instance, Qu *et al*. [31] presented an automatic, end-to-end deep neural network (DeshadowNet) that utilizes a context architecture embedded with information from three different perspectives to predict the output shadow mask. Hu *et al*. [20] introduced an orientation-aware spatial attention module and a growing dilated convolution strategy for shadow removal, effectively using multiple contextual features. Guo *et al*. proposed a novel unsupervised shadow removal solution based on diffusion [17], modeling the shadow, non-shadow, and their boundary regions separately. These solutions have fostered development in the shadow removal community and continuously inspire future research. However, existing methods still encounter challenges in some real-world scenarios. Firstly, existing models often exhibit noticeable inconsistencies in color between the shadow-removed and naturally clear regions when dealing with complex images. Moreover, in cases where shadows are unevenly distributed and concentrated, models not only fail to obtain clean images but might even damage the contour information of background areas [45]. Additionally, most current models perform poorly in high-resolution image shadow removal tasks.

To mitigate these challenges, we propose a shadow removal method based on shadow generation and Global Residual Free Unet. We find the cause of color inconsistency between shadow-removed and naturally clear regions is primarily due to the limited number of shadow images corresponding to the same clean scene, making it difficult to sufficiently constrain the network in terms of color restoration. Addressing this issue, we introduce an effective shadow generation strategy. We train a generative adversarial network to produce shadow masks, subsequently multiplying the masks with clean images to generate the final shadow images. This strategy differs from directly generating shadow images, allowing for a more structured and stable training process. Regarding the shadow removal network, our goal is to design a simple yet effective shadow removal network that can process high-resolution images with minor computational expense. To this end, we devise the Global Residual Free Unet (GRFUnet), employing the classic U-shaped design. Emulating transformers, the foundational blocks of GRFUnet sequentially perform Spatial Interaction and Channel Evolution for effective feature extraction, albeit employing a convolutional architecture. In the foundational block, we develop the Simple Gate Module (SGM) and Pooling Attention Module (PAM) to facilitate feature extraction. Specifically, the former replaces Gated Linear Units [8] (GLU) to retain information according to temporal position to enhance performance, while the latter is a simple and efficient attention mechanism that introduces global information into the image feature map. Besides, we discarded the globally-used residual connections commonly adopted in image restoration fields, finding them to be ineffective for non-additive shadow removal tasks. In this way, we achieve effective high-resolution shadow removal. As shown in Figure 1, our method not only accurately removes shadows but also achieves improved color consistency in the restored images. In summary, our contributions are as follows:

- We develop a shadow generation method based on GAN and multiplicative operations, effectively alleviating the color inconsistency problem between the shadow-removed and naturally clear regions.
- We propose the Global Residual Free Unet, a simple yet efficient shadow removal network, which tackles high-resolution image shadow removal tasks effectively with a reasonable overall structure and efficient modules.
- Experiments on the validation and test datasets provided in the NTIRE 24 Image Shadow Removal Challenge [35] demonstrate the effectiveness of our method. Additionally, our method achieves **the highest PSNR** in the challenge and obtains excellent and balanced results across two tracks (ranking third in the fidelity track and fourth in the perception track).

2. Related work

2.1. Shadow removal

In the field of computer vision, shadow removal remains a pivotal task that has been assiduously explored over the years. The persistence of shadows in digital imagery poses a considerable challenge, often obscuring essential details and affecting the visual fidelity of the scene. A diverse array of methods has been brought to bear on this task, with varying degrees of sophistication and success. Early endeavors in shadow removal often relied on handcrafted features, leveraging insights from image gradients, regional illumination discrepancies, and occasionally user input to distill shadow-free representations of images. Pioneers in this realm, such as Finlayson et al. [10, 11], exploited the gradient consistency principle to reconstruct shadow-obscured visuals. Concurrently, approaches by Guo et al. [18] and Gong et al. [13] proffered robust algorithms hinging on relative lighting conditions and user-guided inputs, respectively. The advent of Deep Neural Networks (DNNs) heralded a renaissance in this domain, greatly augmented by the availability of extensive, publicly curated datasets. Emblematic of this revolution, techniques like DeshadowNet [31], CANet [4], and DSC [21] ingeniously integrated context embedding, spatial attention mechanisms, and transfer learning to bridge the gap between shadowed and non-shadowed regions—exemplifying the breadth of deep learn-

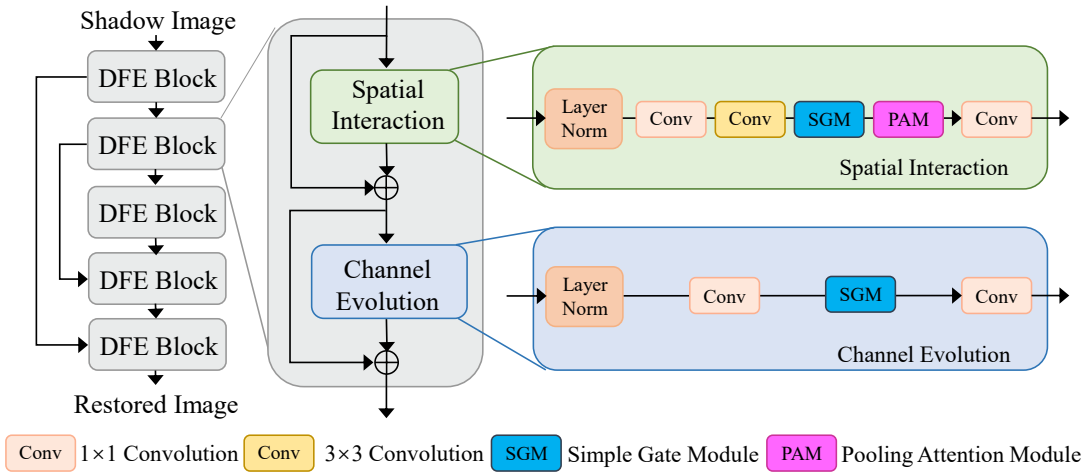


Figure 2. Global Residual Free Unet. Due to the non-additive nature of image shadows, we opt not to use global residual connections.

ing's prowess. More contemporary methods underscore the significance of retaining high-quality ground truth as an instrumental guide. Innovative networks, such as ShadowFormer [16], have adeptly capitalized on transformer architecture to imbibe and extrapolate global contextual cues in an end-to-end manner, showcasing a growing trend towards transformer-based solutions. The orchestration of shadow removal has also begun to intersect with the physical realm, with some researchers opting to model physical illumination in tandem with shadow matte prediction. Notably, Le's [24] work accentuates this intersection, harnessing a linear transformation model to parse the interplay between shadowed and non-shadowed regions. Despite these achievements, the reliance on large volumes of paired shadow and shadow-free images for supervised learning introduces certain limitations. To mitigate these, unsupervised methods deploying Generative Adversarial Networks (GANs) have emerged, training models on unpaired datasets, albeit often with less satisfactory restorative outcomes [23, 29].

2.2. Shadow generation

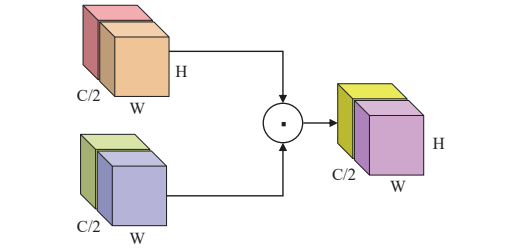
The generation of photorealistic shadows has attracted remarkable interest over recent years, complementing shadow removal research to achieve realistic scene relighting in computer vision. Earlier shadow generation approaches were focused on creating convincing shadows for virtual entities within a scene [28, 41]. The integration of shadow generation mechanisms within shadow removal networks has been a progressive stride, with generative adversarial networks (GANs) playing a pivotal role in this advancement. G2R-ShadowNet [29] is imbued with a shadow generation segment utilizing GANs [14] to innovate the way shadows are cast over shadow-free areas. This aspect of work draws from precedents set by G2R [29] and DHAN [6], whereby both synthesized numerous pseudo shadow cases to aid network training. Similarly, Mask-ShadowGAN [22, 29], inspired by the foundational work of CycleGAN [1], hones the shadow synthesis through adversarial learning, ensuring a semblance between the artificially crafted shadows and their real-world counterparts. The advancement in shadow generation research is evident in the realm of augmented reality as well. Works such as ARShadowGAN [28] and those by Zhang et al. [41] showcase GANs' capacity to cast virtual shadows that coalesce with the environmental lighting and context. These practices historically depend on fully-supervised learning, utilizing paired datasets that span shadow, non-shadow, and mask dimensions.

3. Methodology

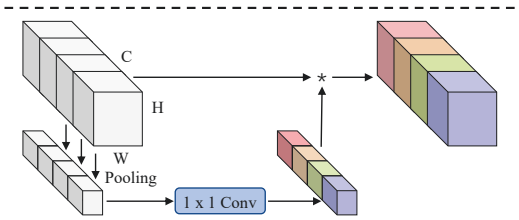
3.1. Shadow generation

A pivotal challenge in the domain of shadow removal is the task of rectifying color inconsistencies between areas where shadows have been removed and those that are naturally clear. This issue primarily arises due to the limited quantity of shadow images corresponding to the same clean scene, leading to insufficient constraints for the network in terms of color restoration. To overcome this issue, we implement a data augmentation strategy designed to enrich our dataset with additional pairs of shadow-impacted images I_{shadow} and shadow-free images I_{clear}. Our original dataset comprises 1,000 paired training images. From these, we extract shadow masks M and input them into our network alongside clean images, facilitating the synthesis of new shadow images I'_{shadow}. This approach significantly broadens the network's capacity to handle diverse scenes.

To further refine the generation of shadow images, we employed a Generative Adversarial Network (GAN) with a specific focus on training a generator. We started by extract-



(a) Simple Gate Module



(b) Pooling Attention Module

Figure 3. Two custom-designed modules in the proposed network. (a) Simple Gate Module. (b) Pooling Attention Module.

ing shadow masks from a collection of 1,000 training images, resulting in 1,000 shadow masks. These masks were subsequently shuffled and paired with clean images before being input into the generator. For training consistency, we resized all images and masks to 1280×960, ensuring the use of full, resized images for training. Our approach does not involve directly generating shadow images; instead, the network outputs shadow matting, by multiplying the shadow matting with the clean images, we are able to obtain the final shadow images, which as shown in Eq.1

$$I'_{shadow} = G(I_{clear}, M) \cdot I_{clear}, \tag{1}$$

where $G(\cdot)$ is the generator. To further refine our network's performance and ensure training stability, we have implemented a paired-task training approach. This methodology is particularly beneficial as it allows for a more structured and stable training process. By employing this strategy, we utilize three specific loss functions to optimize our network: the L1 loss, Perceptual loss, and GAN loss. For the L1 loss, we measure the pixel-by-pixel similarity between the generated shadow images and their corresponding real shadow counterparts, as follows:

$$L_1 = \frac{1}{n}\sum_{i=1}^{n} |I'_{shadow} - I_{shadow}|, \tag{2}$$

where n is the total number of the training images. For the Perceptual loss, we employ VGG19 as the feature extractor, leveraging features from five distinct scales, as follows:

$$L_p = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{5} \lambda_k \|VGG_k(I'_{shadow}) - VGG_k(I_{shadow})\|_1, \tag{3}$$

where k represents different layers. To guarantee that the distribution of the generated shadow images closely aligns with that of the real shadow images, we incorporate GAN loss as a constraint mechanism within our network:

$$\begin{aligned} L_{GAN} = \sum [&\log\left(D\left(I_{clear}, I'_{shadow}\right)\right) \\ &- \log\left(D\left(I_{clear}, I_{shadow}\right)\right)] \end{aligned} \tag{4}$$

where D represents the Patch Discriminator from CycleGAN, allowing the discriminator to focus more on the local details of the image. Additionally, the two inputs in D are concatenated along the channel dimension. The final objective of our method can be summarized as follows: $L_1 + \lambda L_{VGG} + \alpha L_{GAN}$. Empirically, we set α to 0.1 and β to 0.05. We trained the generator for a total of 60 epochs.

3.2. Network Architecture

Our goal is to architect a simple yet efficient network for shadow removal. To balance network simplicity with effectiveness, we employ a classic U-shaped network architecture with skip connections, named Global Residual Free Unet (GRFUnet), as illustrated in Figure 2. The GRFUnet consists of several Deeper Feature Extraction (DFE) blocks that utilize a two-stage structure akin to that of a transformer. Within a DFE block, an image sequentially undergoes Spatial Interaction and Channel Evolution for extensive information interchange. We have actualized these processes through custom-designed modules such as the Pooling Attention Module (PAM) and Simple Gate Module (SGM), alongside foundational elements like convolutions and layer normalization. During Spatial Interaction, an image is first subjected to Layer normalization to stabilize the training—a technique inspired by numerous state-of-the-art methods. Subsequently, the image proceeds through a 1×1 convolution followed by a 3×3 convolution to extract spatial features. Thereafter, the SGM and PAM are applied for additional refinement, as shown in Figure 3. As for Channel Evolution, Layer Norm, 1×1 convolution, and the SGM are employed to process the image.

The skip connections within the U-net and between the functional blocks ensure a smoother network training process. It is noteworthy that we refrain from using global residuals typically employed in image restoration, which has enabled us to achieve enhanced performance. While each element in our design is trivial in isolation, the combination of these elements yields a robust baseline that secures the best PSNR on the NTIRE 2024 Image Shadow Removal Challenge test set. Furthermore, we opt for a CNN-based network over transformers to conserve computational resources: high-resolution images impose an excessive computational load on transformers.

3.3. SGM and PAM

Simple Gate Module (SGM). We observe that many state-of-the-art methods [27, 33, 40] implement Gated Linear

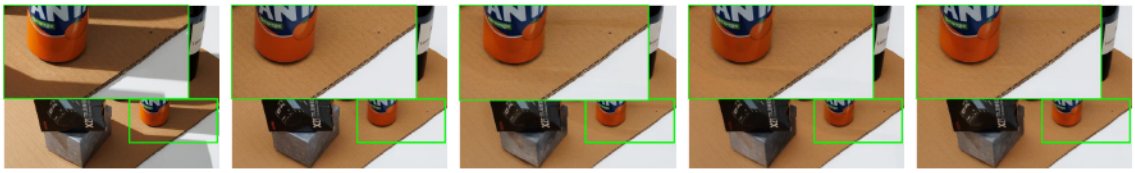


Figure 4. Visualization of various variants of our method.

Units [8] (GLU), which can be expressed as:

$$\mathrm{GLU}(\mathrm{X}) = f(X) \odot \sigma(g(X)), \tag{5}$$

where X denotes the input image feature, $f(\cdot)$ and $g(\cdot)$ are linear transformers, $\odot$ is element-wise multiplication, and σ represents a non-linear activation function. However, incorporating GLU into our network would inadvertently escalate its complexity. NAFNet [3] has recognized that the GLU itself embodies nonlinearity and is not contingent on σ: even with the omission of σ, the operation $Gate(X) = f(X) \odot g(X)$ still retains nonlinearity. Given this insight, we propose a simplified version of GLU, where a feature map is divided into two halves across the channel dimension and then multiplicatively combined, as shown in Figure 3, known as the Simple Gate Module (SGM). SGM achieves an efficient GLU-like operation through element-wise multiplication. The process can be written as follows:

$$\mathrm{SGM}(X) = X_1 \odot X_2, \tag{6}$$

where X_1 and X_2 are image feature maps of the same size.

Pooling Attention Module (PAM). The attention mechanism is a widely adopted method in the field of computer vision, with numerous variations, including the self-attention mechanism in transformers. Striving for efficiency, we seek an attention approach that is amenable to high-resolution images. [40] modifies the spatial-wise attention to channel-wise, which circumvents the computational burden while preserving global information in each feature. This can be considered a distinct variant of channel attention [19]. Inspired by [40] and NAFNet [3], we realize that ordinary channel attention fulfills our criteria of computational efficiency and incorporation of global information into feature maps. Hence, we contemplate integrating channel attention into our network. Typically, channel attention first compresses spatial information into the channel space then employs multi-layer perceptrons to compute the channel attentions, which are subsequently used to weight the feature maps [3]. This process can be written as

$$\mathrm{CA}(X) = X * \sigma(W_2 max(0, W_1 pool(X))), \tag{7}$$

where X represents the input feature map, *pool* denotes the global average pooling operation, σ stands for the nonlinear activation function Sigmoid, and W1,W2 are fully connected layers separated by ReLU.

Accounting for the presence of an activation function and two linear layers in Equation 7, we propose a simplification analogous to that applied to Equation 5, paring it down to involve only a pooling layer and a linear operation. The resulting simplified module is our proposed Pooling Attention Module (PAM), which captures the essence of channel attention, specifically the aggregation of global information and channel interaction. The operation of PAM can be expressed as

$$\mathrm{PAM}(X) = X * \mathrm{Conv}_{1\times1}(pool(X)), \tag{8}$$

where $\mathrm{Conv}_{1\times1}(\cdot)$ denotes a 1×1 convolution, and *pool* represents global average pooling.

3.4. Global Residual Free

Global residual connections are a common network design technique in the image restoration field, facilitating the network's direct learning of the residual between degraded and clean images. Typically, this technique is quite effective, as learning the residual information is simpler than learning how to recover the clean image directly from the degraded one. However, we have discovered that in our task-shadow removal—global residual connections prove to be unnecessary. As illustrated in Figure 4, the restoration result without global residual connections surpasses the result with large skip connections. We conjecture that this may be due to the fact that image shadows do not constitute additive degradation, which complicates the task of learning the residual between clean and shadowed images.

3.5. Optimization

The training of our GRFUnet is divided into three stages (see Sec 4.2 for details), wherein the first two stages employ the Charbonnier loss [2] and a loss based on frequency domain information. The Charbonnier loss is designed to ensure image fidelity and can be computed as follows:

$$\mathcal{L}_c = \frac{1}{n}\sum_{n=1}^{n}\sqrt{\|\mathbf{I}_{gt} - \mathbf{I}_{out}\|^2 + \epsilon^2}, \tag{9}$$

where $\mathbf{I}_{gt}$ and $\mathbf{I}_{out}$ represent the ground truth and shadow-free images predicted by the network, respectively. ϵ is seen as a tiny constant (*e.g.*, 10^{-5}) for stable and robust convergence, and n represents the total number of input images in a single iteration.

In addition to spatial domain constraints, we have imposed a loss function in the frequency domain to enhance the recovery of frequency domain information. Specifically, we utilize the FFT loss based on the Fast Fourier Transform, which can be expressed as

$$\mathcal{L}_f = \frac{1}{n}\sum_{n=1}^{n} \left\| \mathcal{F}\left(\mathbf{I}_{gt}\right) - \mathcal{F}\left(\mathbf{I}_{out}\right) \right\|_1, \tag{10}$$

where the $\mathcal{F}(\cdot)$ represents Fourier Transform.

During the first and second stages of training, we employ the two aforementioned losses as constraints, which can be expressed as follows:

$$\mathcal{L}_{stage1,2} = \mathcal{L}_c + \lambda \mathcal{L}_f, \tag{11}$$

where λ denotes the balanced weight. In particular,λ is set to 0.02 empirically.

In the third stage of training, to further enhance the quality of image restoration, we utilize the SSIM loss. The SSIM loss relies on the structural similarity index, a metric that quantifies the similarity between two images by considering a combination of luminance, contrast, and structural similarity. The SSIM loss can be calculated as:

$$\mathcal{L}_{SSIM} = \frac{\left(2\mu_{\mathbf{I}_{gt}}\mu_{\mathbf{I}_{out}} + \eta_1\right)\left(2\sigma_{\mathbf{I}_{gt}\mathbf{I}_{out}} + \eta_2\right)}{\left(\mu^2_{\mathbf{I}_{gt}} + \mu^2_{\mathbf{I}_{out}} + \eta_1\right)\left(\sigma^2_{\mathbf{I}_{gt}} + \sigma^2_{\mathbf{I}_{out}} + \eta_2\right)}, \tag{12}$$

where μ and σ denote the mean and standard deviation of image intensities, respectively. The $\sigma_{\mathbf{I}_{gt}\mathbf{I}_{out}}$ represents the covariance between the two images. And the constants η_1 and η_2 are included to prevent division by zero. Additionally, the third stage of training also employs the Charbonnier loss. The loss for the third stage can be expressed as:

$$\mathcal{L}_{stage3} = \mathcal{L}_c + \psi \mathcal{L}_{SSIM}, \tag{13}$$

where ψ is the parameter to balance the two terms in the loss function and we empirically set it to 0.02 as default.

4. Experiments

4.1. Dataset

In our study, we employ the NTIRE 2024 Image Shadow Removal Challenge dataset, building upon the WSRD dataset [34] from prior iterations. The WSRD dataset signifies a leap forward in exploring complex shadow interactions, featuring surfaces of increased complexity and detailed content. The NTIRE 2024 version includes a preprocessing stage of image alignment using homography estimation, and strategically excludes 25 samples from the WSRD testing split, preserving only the samples that pose the greatest challenge. Distributed through the official NTIRE channel, this dataset encompasses 1000 training images, 100 validation images, and 75 test images, each with a resolution of 1440 x 1920 pixels in RGB format.

Furthermore, we integrate the NTIRE 2023 dataset [34] into our training regime, amalgamating it with 24 training samples from the previous year to amplify the diversity and complexity, thereby bolstering the robustness of our model against a gamut of shadow scenarios.

To enrich our training resources further, we include images with shadows generated by traditional methods and those synthesized by GANs [7] for specific scenarios. By treating the initially restored images as pseudo-labels added to the training set, we extract shadow masks from the original dataset that contains pairs of shadow and non-shadow images. These masks are then fed into the network along with clean images, enabling the generation of new shadow images and facilitating the training process focused on shadow synthesis and subsequent removal.

4.2. Implementation Details

Our training approach progressively sharpens the learning process, scaling up patch sizes while decreasing learning rates for optimization, applying gradient accumulation for larger patches, and incorporating generated shadow data to improve shadow removal performance. Details for each stage are provided.

Stage 1: Utilizing the Adam optimizer, we commence with a batch size of 4 and an initial patch size of 512×512. The learning rate begins at 4×10^{-4} and is modulated using the Cosine Annealing scheme over the course of 1000 epochs. This initial stage is carried out on an NVIDIA 4090 device, with the finest resultant model serving as the foundation for the subsequent stage.

Stage 2: We continue employing the Adam optimizer; however, we reduce the batch size to 1 and amplify the patch size to 960×960. Starting with a learning rate of 4×10^{-5}, we apply the Cosine Annealing scheme through 300 epochs. Performed on the NVIDIA 4090 device and incorporating gradient accumulation, the premier model obtained at this juncture initializes the following stage.

Stage 3: Transitioning to the SGD optimizer, we maintain a batch size of 1 while significantly expanding the patch size to 1920×1920. The learning rate is set at 2×10^{-5} and undergoes adjustment in accord with the Cosine Annealing scheme, spanning 200 epochs. This final stage is executed on an NVIDIA A40 device and persistently employs gradient accumulation.

When it comes to the testing phase, we employ the fine-tuned model to attain the optimal state of performance. To further elevate the model's output, we engage an input-ensemble strategy. The testing procedure is meticulously carried out on an NVIDIA A40 device.

Table 1. Results of top six methods on **the fidelity track**. Our method achieves **the highest PSNR** and ranks third overall among 18 teams.

Team Name	PSNR↑	SSIM↑	LPIPS↓	Params.(M)	Rank
LUMOS	$24.78_{(2)}$	$0.832_{(2)}$	$0.110_{(4)}$	23	$1_{/18}$
Shadow_R	$24.58_{(3)}$	$0.832_{(1)}$	$0.098_{(2)}$	376	$2_{/18}$
ShadowTech Innovators (Ours)	$24.81_{(1)}$	$0.832_{(3)}$	$0.111_{(5)}$	26	$3_{/18}$
LVGroup_HFUT	$24.35_{(4)}$	$0.823_{(6)}$	$0.082_{(1)}$	17	$4_{/18}$
USTC_ShadowTitan	$24.04_{(5)}$	$0.827_{(4)}$	$0.104_{(3)}$	83	$5_{/18}$
GGBond	$23.87_{(6)}$	$0.824_{(5)}$	$0.127_{(6)}$	8.895	$6_{/18}$

Table 2. Results of the top six methods on **the perceptual track**. Our method achieves the highest PSNR and a fourth-place ranking in perceptual performance among nineteen teams.

Team Name	PSNR↑	MOS↑	Rank
Shadow_R	$24.59_{(3)}$	$7.750_{(1)}$	$1_{/19}$
LVGroup_HFUT	$24.23_{(4)}$	$7.519_{(2)}$	$2_{/19}$
USTC_ShadowTitan	$24.05_{(5)}$	$7.444_{(3)}$	$3_{/19}$
Ours	$24.81_{(1)}$	$7.436_{(4)}$	$4_{/19}$
GGBond	$23.05_{(6)}$	$7.400_{(5)}$	$5_{/19}$
PSU Team	$22.22_{(12)}$	$7.400_{(6)}$	$6_{/19}$

Table 3. Ablation results on ntire24 validation dataset using different variants of our method.

Models	**PSNR↑**	**SSIM↑**
w/o shadow generation	25.499	0.836
w global residual	25.967	0.843
Ours	**26.565**	**0.844**

4.3. Results of Challenge

Table 1 presents our method's comparison against the top six teams in the fidelity track of the NTIRE 2024 Image Shadow Removal Challenge [35]. Our approach achieves **the highest PSNR** and is competitive in other metrics as well, such as SSIM, where it is only 7.77E-04 lower than the first place and merely 2.73E-06 lower than the second. Our comprehensive ranking in the fidelity track is third. Besides, we also secure a commendable position in the perceptual track, attaining fourth place as shown in Table 2.

4.4. Ablation Study

To validate the effectiveness of our proposed Global Residual Free Unet (GRFUnet) and shadow generation technique, we performed assessments on the validation and test sets of NTIRE 2024, utilizing Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) as objective quality metrics. Specifically, our ablation study is structured into three main parts: a) Forgoing the use of shadow generation (w/o shadow generation); b) Utilizing global residual connections (w global residual); c) Employing the proposed approach, which entails the absence of global residuals and the incorporation of shadow generation (Ours).

Qualitative and quantitative comparisons are illustrated in Figure 4 and Table 3, respectively. It can be seen that the proposed method achieves superior color consistency compared to the variants without shadow generation, owing to the generated shadows which facilitate the network's ability to robustly constrain color restoration. Additionally, compared to the variant with global residuals, the proposed method exhibits enhanced detail preservation, attributable to the disuse of global residual connections which are typically ill-suited for non-additive shadow tasks. Regarding quantitative outcomes, the proposed method outperforms both variants in terms of PSNR and SSIM, which emphatically validates the efficacy of our design.

Table 4. Quantitative comparison of our method with existing methods on the ntire24 validation dataset.

Method	**PSNR↑**	**SSIM↑**
ShadowFormer [16]	22.907	0.819
SwinIR [26]	23.257	0.814
ShuffleFormer [39]	24.724	0.821
Ours	**26.565**	**0.844**

Table 5. Quantitative comparison of our method with existing shadow removal methods on ISTD [37] dataset.

Method	**PSNR↑**	**SSIM↑**
DSC [20]	26.62	0.845
DHAN [5]	27.21	0.921
SpA-Former [42]	27.73	0.931
ShadowFormer [16]	30.47	0.935
Ours	**30.91**	**0.938**

4.5. Comparisons

Structural Similarity Our proposed method is compared against several state-of-the-art image shadow removal approaches, including ShadowFormer [16], SwinIR [26], and ShuffleFormer [39]. We employ Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) as the evaluation metrics, with higher values indicating superior performance. To ensure a fair assessment, we utilize the same training data and methodology for all compared methods and evaluate their performances on the same test set (ntire24-valid). Comparative visual results are presented in

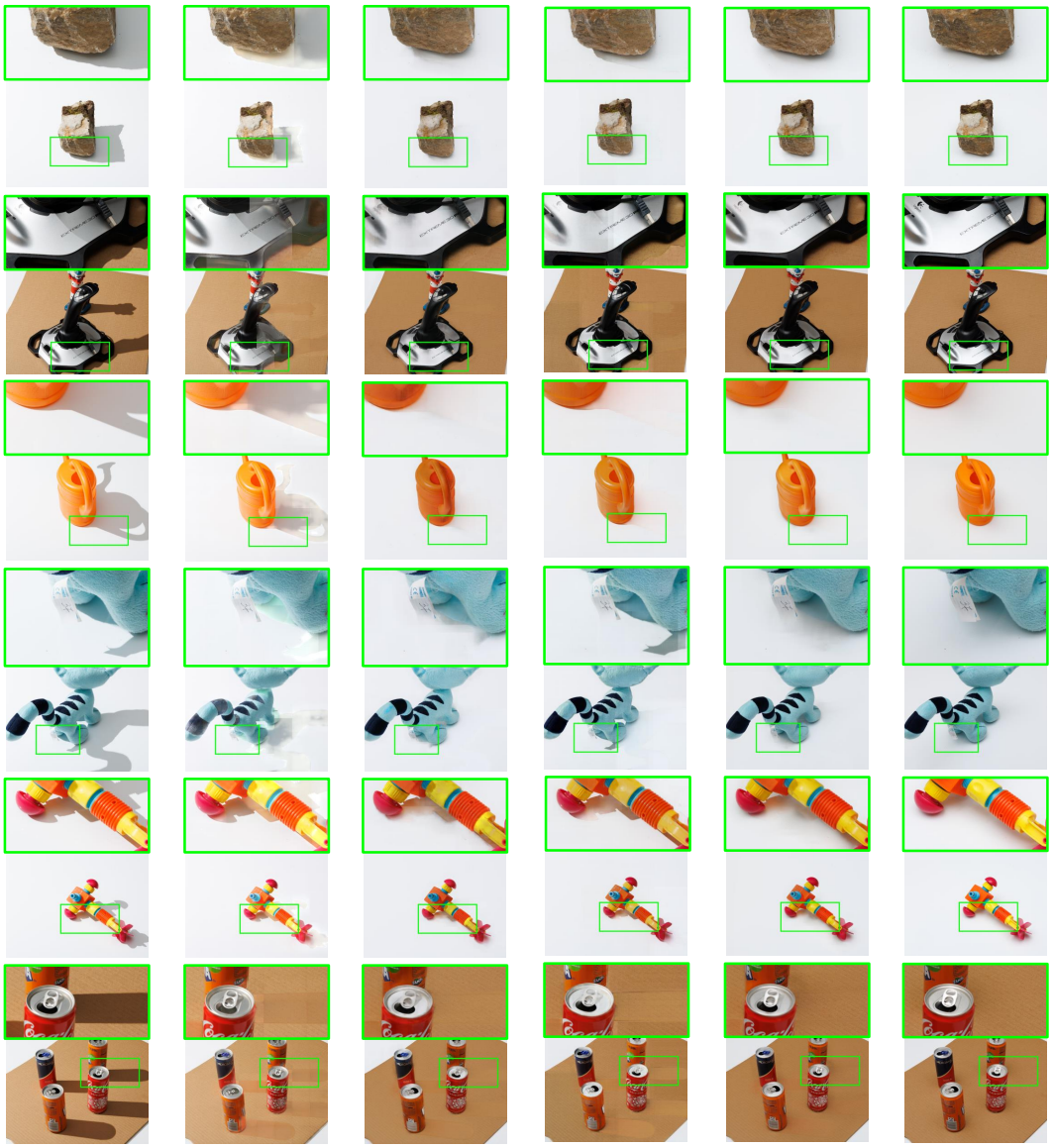


Figure 5. Visualization comparison results of shadow removal on the ntire24 validation dataset. (a) to (c) are the predicted results from SOTA methods: ShadowFormer [16], SwinIR [26], ShuffleFormer [39], respectively.

Figure 5, while a quantitative comparison is provided in Table 4. Our method can be seen to possess superior fidelity capabilities and is more adept at managing complex background information. Moreover, we conduct quantitative comparisons on the ISTD dataset with previous methods of shadow removal. As shown in Table 5, our method continues to demonstrate superior performance."

5. Conclusion

In this paper, our study introduces a novel framework for shadow removal, combining shadow generation with Global Residual Free Unet (GRFUnet) to tackle complex and high-resolution images. Our use of GAN-generated shadow masks enhances color consistency, and our modified UNet architecture performs effective feature extraction. Besides, considering the non-additive attributes of image shadows, we eliminate global residual connections to further improve performance. Validated by the NTIRE 24 Image Shadow Removal Challenge results, our method not only achieves the highest PSNR but also ranks highly in both fidelity and perception tracks. This work contributes to the domain of image shadow removal and provides insights for future development in the field.

References

[1] Amjad Almahairi, Sai Rajeshwar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *International conference on machine learning*, pages 195–204. PMLR, 2018. 3

[2] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st International Conference on Image Processing*, pages 168–172 vol.2, 1994. 5

[3] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33. Springer, 2022. 5

[4] Zipei Chen, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Canet: A context-aware network for shadow removal. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4743–4752, 2021. 2

[5] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan, 2019. 7

[6] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10680–10687, 2020. 3

[7] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10680–10687, 2020. 6

[8] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017. 2, 5

[9] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Argan: Attentive recurrent generative adversarial network for shadow detection and removal. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10212–10221, 2019. 2

[10] Graham D Finlayson, Steven D Hordley, Cheng Lu, and Mark S Drew. On the removal of shadows from images. *IEEE transactions on pattern analysis and machine intelligence*, 28(1):59–68, 2005. 2

[11] Graham D Finlayson, Mark S Drew, and Cheng Lu. Entropy minimization for shadow removal. *International Journal of Computer Vision*, 85(1):35–57, 2009. 2

[12] Xueyang Fu, Wu Wang, Yue Huang, Xinghao Ding, and John Paisley. Deep multiscale detail networks for multiband spectral image sharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2090–2104, 2021. 2

[13] Han Gong and Darren Cosker. Interactive removal and ground truth for difficult shadow scenes. *JOSA A*, 33(9): 1798–1811, 2016. 2

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3

[15] Maciej Gryka, Michael Terry, and Gabriel J Brostow. Learning to remove soft shadows. *ACM Transactions on Graphics (TOG)*, 34(5):1–15, 2015. 1

[16] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps image shadow removal. *arXiv preprint arXiv:2302.01650*, 2023. 1, 3, 7, 8

[17] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13045–13054, 2023. 2

[18] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2956–2967, 2012. 1, 2

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5

[20] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7454–7462, 2018. 2, 7

[21] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection and removal. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2795–2808, 2019. 2

[22] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2472–2481, 2019. 3

[23] Yeying Jin, Aashish Sharma, and Robby T Tan. Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5027–5036, 2021. 3

[24] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8578–8587, 2019. 3

[25] Dong Li, Jiaying Zhu, Menglu Wang, Jiawei Liu, Xueyang Fu, and Zheng-Jun Zha. Edge-aware regional message passing controller for image forgery localization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8222–8232, 2023. 2

[26] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1833–1844, 2021. 7, 8

[27] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE Transactions on Image Processing*, 2024. 4

[28] Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, Xinzhi Dong, and Chunxia Xiao. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8139–8148, 2020. 3

[29] Zhihao Liu, Hui Yin, Yang Mi, Mengyang Pu, and Song Wang. Shadow removal by a lightness-guided network with training on unpaired data. *IEEE Transactions on Image Processing*, 30:1853–1865, 2021. 3

[30] Zhihao Liu, Hui Yin, Xinyi Wu, Zhenyao Wu, Yang Mi, and Song Wang. From shadow generation to shadow removal. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4925–4934, 2021. 1

[31] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4067–4075, 2017. 2

[32] Andres Sanin, Conrad Sanderson, and Brian C. Lovell. Improved shadow removal for robust person tracking in surveillance scenarios. In *2010 20th International Conference on Pattern Recognition*, pages 141–144, 2010. 1

[33] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5769–5780, 2022. 4

[34] Florin-Alexandru Vasluianu, Tim Seizinger, and Radu Timofte. Wsrd: A novel benchmark for high resolution image shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1825–1834, 2023. 6

[35] Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Zongwei Wu, Cailian Chen, Radu Timofte, et al. NTIRE 2024 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 2, 7

[36] Tomas F Yago Vicente, Minh Hoai, and Dimitris Samaras. Leave-one-out kernel optimization for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):682–695, 2017. 1

[37] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1788–1797, 2018. 7

[38] Chunxia Xiao, Ruiyun She, Donglin Xiao, and Kwan-Liu Ma. Fast shadow removal using adaptive multi-scale illumination transfer. In *Computer Graphics Forum*, pages 207–218. Wiley Online Library, 2013. 1

[39] Jie Xiao, Xueyang Fu, Man Zhou, HongJiang Liu, and Zhengjun Zha. Random shuffle transformer for image restoration. In *International Conference on Machine Learning*, 2023. 7, 8

[40] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 4, 5

[41] Shuyang Zhang, Runze Liang, and Miao Wang. Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media*, 5:105–115, 2019. 3

[42] Xiaofeng Zhang, Yudi Zhao, Chaochen Gu, Changsheng Lu, and Shanying Zhu. Spa-former:an effective and lightweight transformer for image shadow removal. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2023. 7

[43] Jiaying Zhu, Dong Li, Xueyang Fu, Gang Yang, Jie Huang, Aiping Liu, and Zheng-Jun Zha. Learning discriminative noise guidance for image forgery detection and localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7739–7747, 2024. 2

[44] Yurui Zhu, Xueyang Fu, Chengzhi Cao, Xi Wang, Qibin Sun, and Zheng-Jun Zha. Single image shadow detection via complementary mechanism. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 6717–6726, New York, NY, USA, 2022. Association for Computing Machinery. 1

[45] Yurui Zhu, Jie Huang, Xueyang Fu, Feng Zhao, Qibin Sun, and Zheng–Jun Zha. Bijective mapping network for shadow removal. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5617–5626, 2022. 2

[46] Yurui Zhu, Zeyu Xiao, Yanchi Fang, Xueyang Fu, Zhiwei Xiong, and Zhengjun Zha. Efficient model-driven network for shadow removal. In *AAAI Conference on Artificial Intelligence*, 2022. 2