HomoFormer: Homogenized Transformer for Image Shadow Removal

Jie Xiao[1] Xueyang Fu[1†] Yurui Zhu[1] Dong Li[1] Jie Huang[1] Kai Zhu[2] Zheng-Jun Zha[1]
[1]University of Science and Technology of China [2]Alibaba Group
ustchbxj@mail.ustc.edu.cn xyfu@ustc.edu.cn

Abstract

The spatial non-uniformity and diverse patterns of shadow degradation conflict with the weight sharing manner of dominant models, which may lead to an unsatisfactory compromise. To tackle with this issue, we present a novel strategy from the view of shadow transformation in this paper: directly homogenizing the spatial distribution of shadow degradation. Our key design is the random shuffle operation and its corresponding inverse operation. Specifically, random shuffle operation stochastically rearranges the pixels across spatial space and the inverse operation recovers the original order. After randomly shuffling, the shadow diffuses in the whole image and the degradation appears in a homogenized way, which can be effectively processed by the local self-attention layer. Moreover, we further devise a new feed forward network with position modeling to exploit image structural information. Based on these elements, we construct the final local window based transformer named HomoFormer for image shadow removal. Our HomoFormer can enjoy the linear complexity of local transformers while bypassing challenges of non-uniformity and diversity of shadow. Extensive experiments are conducted to verify the superiority of our HomoFormer across public datasets. Code is available at https://github.com/jiexiaou/HomoFormer.

1. Introduction

Shadow is ubiquitous in images captured under natural scenes when the light sources are partially or fully blocked. However, shadow not only impairs the visual quality of images but imposes severe limitations on various subsequent downstream vision tasks, e.g., object tracking [40] and detection [36], face recognition [56], etc. Hence, it is significant to restore clean images from their shaded counterparts.

One of main obstacles of image de-shadowing is that the spatial distribution of shadow degradation is non-uniform and the patterns of shadow are diverse. These characteristics make shadow very hard to be modeled by dominant models, such as convolutional neural networks (CNNs) and window based Transformers [34]. It is caused by the inherent weight sharing property of those models. The build-in weight sharing property determines that they have to utilize a single set of parameters to cover shadow cases of complex degraded degree, which may lead to an unsatisfactory compromise. Fig. 1 illustrates this idea schematically.

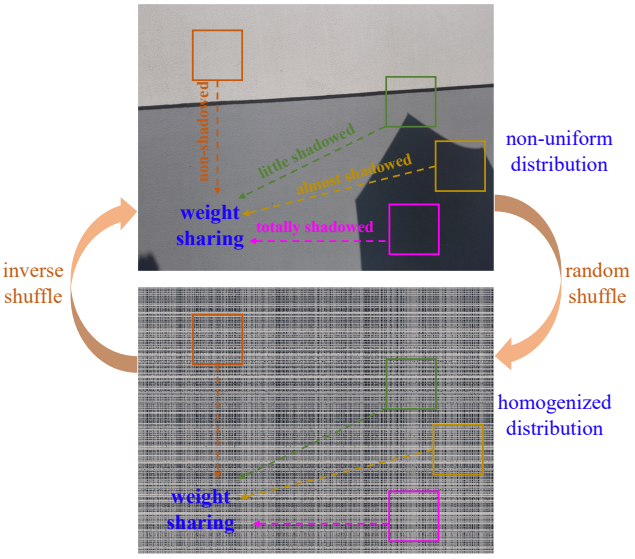


Figure 1. Schematic illustration of challenges posed by non-uniform distribution of shadow. Non-uniformity imposes a constraint to weight sharing models, where they struggle to seek for a *compromise* among regions of various degenerated degrees. Random shuffle creates a homogenized distribution, laying the foundation for the weight-sharing local self-attention.

To overcome this challenge, one straightforward solution is to choose advanced models, which are capable of modeling interactions with spatial heterogeneity. In other words, the desired model is anticipated to take an adaptive action relying on the concrete content of shadows. A competitive candidate is vanilla vision transformers [4, 10], which, by leveraging the global self-attention, are capable of process-

† Corresponding author.

ing images adaptively. Although vanilla vision transformers can tackle with this heterogeneity issue to some degree, their application is still limited by the *quadratic* complexity with respect to the input resolution, which is usually high (e.g., 840×640) for this task. Local window based Transformers [34] can process images efficiently in linear complexity while they suffer from the weight sharing when dealing with non-uniform shadow degradation. In this work, we turn to explore another side of the coin: *is it possible to homogenize the non-uniform distribution instead of passively choosing more sophisticated models for adaptation?*

To achieve this goal, we wish to directly transform the original non-uniform shadow by some function, implemented by a dedicated operation pair $\mathcal{S}(\cdot)$ and $\mathcal{I}_{\mathcal{S}}(\cdot)$. $\mathcal{S}(\cdot)$ is used to project the non-uniform distribution to a homogenized space while $\mathcal{I}_{\mathcal{S}}(\cdot)$ is the exact inverse operation of $\mathcal{S}(\cdot)$ to project it back. We provide a potential solution that $\mathcal{S}(\cdot)$ is the random shuffle operation and $\mathcal{I}_{\mathcal{S}}(\cdot)$ is the corresponding inverse shuffle operation. Specifically, through stochastic rearrangement across spatial space, each pixel will be allocated to any position with the equal probability, accomplishing the purpose of homogenizing the non-uniform distribution. On the other hand, the random rearrangement of pixels thoroughly destroys the semantic information (see Fig. 1, image semantics are lost after shuffling). Consequently, after series of calculations in the homogenized space, it is necessary to project inversely to align with the original space. Fortunately, the random shuffle operation can be inverted exactly. We can readily invert the random shuffle to reconstruct the image semantics by recovering original relative positions between pixels. Notably, random shuffle operation and its inverse are extremely cheap to be implemented, without introducing additional parameters or FLOPs.

With the random shuffle operation and its inverse operation, we have assess to a homogenized space *without* any information loss, eliminating the constraint to models with weight sharing property (see Fig. 1). For now, we come to the stage of considering a concrete model for image de-shadowing. A subtle issue is that since random shuffle destroys the relative position relationships, the desired layer working in the homogenized space should not rely on position-based information to model relationships. Given these considerations, the desired layer is implemented as the local self-attention [34] without utilize position encoding. We can move the responsibility of modeling structural information to subsequent feed forward network (FFN) layer [43]. Accordingly, we introduce a local window Transformer, dubbed *HomoFormer*, as the overall model. HomoFormer not only enjoys linear complexity to input resolution, strong representation of transformers but also bypasses the challenge of modeling non-uniformly distributed shadow degradation. We conduct extensive and comprehensive experiments to verify the superiority (Secs. 4.2 and 4.3) as well as explain the behavior (Sec. 4.4) of our HomoFormer.

In summary, the main contributions of this work include:

- We analyse the challenge of modeling non-uniformly distributed shadow degradation and provide a fresh perspective to this problem: homogenizing the non-uniform distribution.
- We design the random shuffle and inverse shuffle, a complementary operation pair without any loss of information, to accomplish the homogenization.
- We elaborate a local window Transformer named *HomoFormer* which can process image with linear complexity while avoiding to suffer from modeling non-uniform and diverse shadow.
- Extensive and comprehensive experiments on benchmark datasets are conducted to verify and further explain the superiority of our HomoFormer.

2. Related Works

2.1. Image Shadow Removal

Classic approaches [11, 12, 42, 52] for shadow removal often leverage various handcrafted priors, e.g., illumination [48, 55], regions [17], density [1], or user interaction [14]. Recently, with the splendor development of deep learning, learning based methods have also achieved brilliant progress for image shadow removal. For instance, DeshadowNet [38] aggregates context information by fusing multi-level features to predict a shadow matte for shadow removal. Hu *et al.* [19, 20] leverage a direction-aware spatial context for detecting and removing shadows. MaskshadowGAN [21] proposes a framework that estimates the shadow mask from the input shadow image and sequently utilizes the masks as guidance for the shadow generation to establish the cycle-consistency constraints. Cun *et al.* [8] exploit the contextual features by stacking dilated convolutions. Chen *et al.* [6] attempt to remove shadows by transfer the contextual information from non-shadow regions to shadow regions. Fu *et al.* [13] formulate the shadow removal task as a multiple exposure images fusion problem. DC-ShadowNet [23] integrates the domain classifiers and the physics-based losses to achieve the unpaired shadow removal. G2R [35] and BMNet [57] both introduce the shadow generation process for boosting the performance of shadow removal. Several works [9, 33] also employ generative adversarial networks to enhance the reality of shadow removed results or unpaired data training. SP+M-Net [26] and EMDN [58] both attempt to propose the reasonable shadow illumination models for shadow removal. Guo *et al.* [15] make use of channel attention to exploit the global contextual correlation between shadow and non-shadow regions. Wan *et al.* [44] propose a style-guided

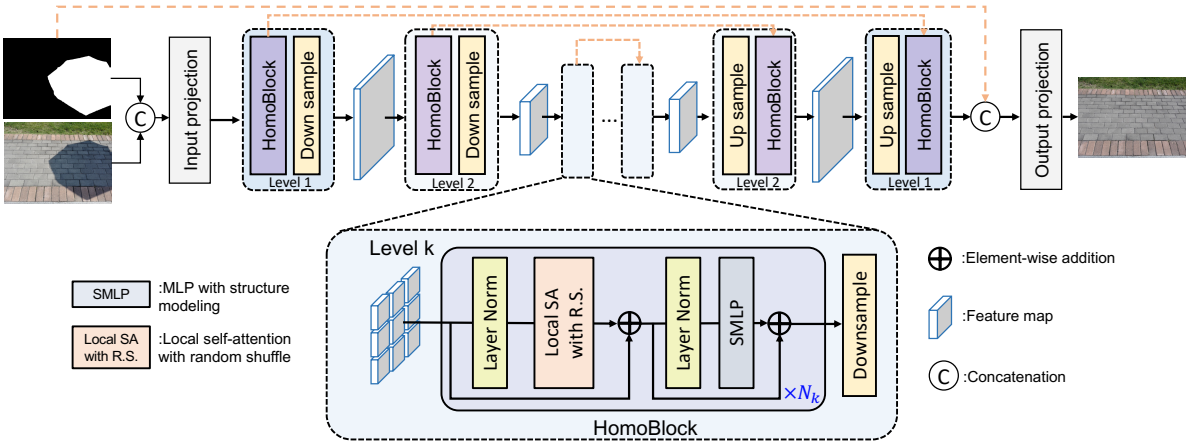


Figure 2. The overall architecture of our proposed HomoFormer. The core of HomoFormer is to use random shuffle to homogenize the original image space and employ local self-attention to model interactions in the homogenized space.

shadow removal network for better image-style consistency after shadow removal.

2.2. Vision Transformer

Vision Transformers [10, 32, 34, 37, 47] have gained glorious achievements in vision community. ViT [10] firstly treated image patches as token sequence and applied the vanilla transformer on it for image classification. Swin Transformer [34] brought in the locality and hierarchy prior to self-attention and adopted the shifted window self-attention to establish a efficient architecture for various tasks, including image classification, object detection and semantic segmentation. A few transformers [4, 31, 46, 49, 50, 53] have also arose for various low-level vision tasks. Nevertheless, most of them adopts the vanilla self-attention or shifted window self-attention, suffering from either huge complexity or modeling non-uniformly distributed shadow degradation. For image shadow removal, ShadowFormer [15] adopts channel attention to aggregate global context instead of spatial attention to avoid expensive complexity. In comparison with ShadowFormer, our motivation is to primarily focus on the non-uniform issue of shadow and our HomoFormer still adopts the classic paradigm of "self-attention→MLP" of vision transformers.

2.3. Image Shuffle Strategy

Shuffling pixels is a common technique in computer vision but we shed new light on its potential for shadow removals. Kang et al. [24] proposes to shuffle pixels in a local patch as a training regularization. Except for distinct motivation, HomoFormer extends the shuffling range to the whole image. Pixel shuffle also plays a role in upsampling/downsampling to reshape features [29, 41]. It is used to exchange information between channel and space and does not randomly disrupt the spatial rearrangements of pixels. Recently, to capture non-local interactions, Xiao et al. [51] proposes random shuffle to replace shifted window strategy of Swin Transformer. Different from [51], the motivation of HomoFormer is to create a homogenized space, which is compatible with the weight-sharing mechanism. Besides, we employ random shuffle throughout the network rather than replace the shifted window strategy and develop the separate SMLP module to model structural information.

3. Method

We first introduce the dedicate random shuffle operation and inverse shuffle operation in Sec. 3.1. Then in Sec. 3.2, we recall the formulation of local self-attention and then integrate these two shuffle operations with local self-attention to establish a elaborate layer calculating on the homogenized space. FFN with structure modeling is proposed in Sec. 3.3. Last, we present the overall Transformer model for image shadow removal in Sec. 3.4.

3.1. Random Shuffle and Inverse Shuffle

Shadow degradation is non-uniformly distributed across spatial space, which is undesirable for dominant models with weight sharing property. To tackle with the non-uniformity issue, we present two key operations to homogenize the distribution: the random shuffle operation $\mathcal{S}(\cdot)$ and corresponding inverse shuffle operation $\mathcal{I}_{\mathcal{S}}(\cdot)$. Random shuffle is responsible for stochastically permuting the elements of input while inverse shuffle corresponds to recovering the original order. Formally, suppose that X is the input, m is the list of index of X and $\mathtt{m}$ is the random permutation of m, we have the definitions for random shuffle operation:

$$\mathcal{S}(X)_{\mathtt{m}} = X_m. \tag{1}$$

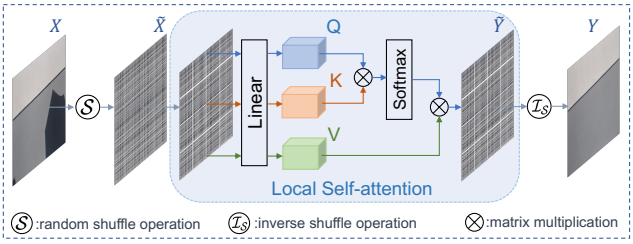


Figure 3. Computational graph of the proposed local self-attention with random shuffle.

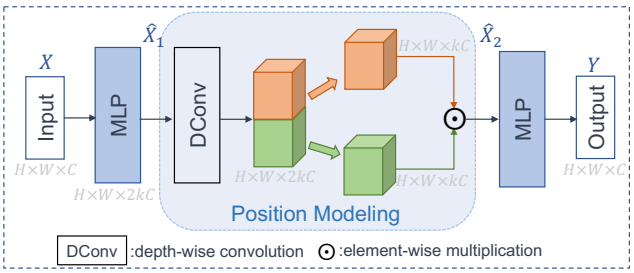


Figure 4. Computational graph of the proposed SMLP.

After random shuffle, each pixel has an equal probability of appearing in any position. Consequently, random shuffle can play a key role in homogenizing the non-uniformly distributed shadow degradation, eliminating the constraint to weight sharing models.

On the other hand, random shuffle thoroughly destroys the image semantics, which is closely related to the order of pixels. Therefore, after possessing in the homogenized space, it is necessary to invert the shuffle projection to align with original feature. Motivated by that, we elaborate the inverse shuffle operation, which has the definition:

$$\mathcal{I}_{\mathcal{S}}(\mathcal{S}(X)) = X. \tag{2}$$

As analyzed above, the inverse shuffle operation is able to offset the stochastic reordering of random shuffle operation. Besides, since only rearrangement of elements is involved, the shuffle operation pair is extremely efficient to be implemented on modern accelerators, without incurring extra parameters or FLOPs.

3.2. Local Self-Attention with Random Shuffle

Self-attention [43] can be summarized as mapping a query and a set of key-value pairs to an output, where the query, keys and values are obtained from linear projections of input. The formulation of self-attention is expressed by

$$\mathrm{SA}(X) = \mathrm{Softmax}\left(\frac{XW^Q(XW^K)^T}{\sqrt{d_k}}\right)XW^V. \tag{3}$$

W^Q, W^K, and W^V are parameter matrices for query, key and value, respectively. Self-attention is often used to model long-range interactions. However, its complexity in both time and memory is quadratic with respect to the token number (the resolution of input image for vision tasks). The quadratic complexity incurs prohibitively huge burden in both computation and memory cost when self-attention is directly applied to image shadow removal, since the input resolution is often high. Swin Transformer [34] adopts local self-attention instead of global self-attention, which reduces significantly to linear complexity to input resolution. Specifically, local attention first partitions the input into non-overlapping windows, and restricts the computing of self-attention within local windows. The weights for processing those non-overlapping windows are shared. The mathematical formulation of local self-attention is

$$\mathrm{LSA}(X) = \mathrm{SA}\left(\mathrm{Par}(X)\right), \tag{4}$$

where $\mathrm{Par}(\cdot)$ denotes the partition function. Self-attention can be implemented as the multi-head version to boost its expressiveness [10, 34, 43]. For simplicity, we here take single head case as example without loss of generality.

Despite its high efficiency, the weight sharing property of local self-attention makes it undesirable for image deshadowing due to the spatial non-uniformity and diversity of shadow. Given the random shuffle and inverse shuffle operation, we can overcome this difficulty by integrating the shuffle pair with local self-attention. Fig. 3 presents the detailed computational graph. Specifically, the input X is first stochastically rearranged by random shuffle operation $\mathcal{S}(\cdot)$, resulting a homogenized version of input denoted by $\tilde{X}$. The homogenized input is fed to local self-attention, yielding the homogenized output $\tilde{Y}$. Lastly, the homogenized output is recovered to the original order of ultimate output Y by inverse shuffle operation $\mathcal{I}_{\mathcal{S}}(\cdot)$. The above procedure is formulated by

$$\tilde{X} = \mathcal{S}(X), \tilde{Y} = \mathrm{LSA}(\tilde{X}), Y = \mathcal{I}_{\mathcal{S}}(\tilde{Y}). \tag{5}$$

A potential concern may be whether the stochasticity in pixel's position brought by random shuffle will affect the training stability of local self-attention. Fortunately, an important property of self-attention is that it is equivariant to reordering [7], that is, it gives the same output regardless of how the input tokens are shuffled. Therefore, self-attention can serve as a desired layer processing the homogenized feature without being interfered by stochasticity of random shuffle. A distinction with original local self-attention [34] is that relative position encoding is discarded, leaving structural information unexplored [2]. The reason is straightforward: after randomly shuffling, the position is no longer reliable. But this problem can be readily solved by moving the function of modeling structure-based interactions to the subsequent feed forward network.

3.3. FFN with Structure Modeling

Followed by the classic paradigm of Transformer, we wish to design the basic transformer building block by employing this customized self-attention followed by feed forward network (FFN). Given the consideration that our customized self-attention cannot exploit structure-based information to explore structural information, we move the responsibility of modeling structure-based interactions to the FFN. Typically, FFN in Transformer is implemented by two multilayer perceptrons (MLPs). We here resort to FFN to model structure-based interactions. Reminder that CNNs allocate weight according to the relative position, thus acting as a simple layer to model structure-based interactions. Hence, we design a simple FFN by inserting depth-wise convolution between MLPs, which we call SMLP. Fig. 4 shows the procedure. Formally, SMLP computes as

$$\begin{aligned} \hat{X}_1 &= \mathrm{MLP}(X), \\ \hat{X}_2 &= \mathrm{DConv}_1(\hat{X}_1) \odot \mathrm{DConv}_2(\hat{X}_1), \\ Y &= \mathrm{MLP}(\hat{X}_2), \end{aligned} \tag{6}$$

where $\mathrm{DConv}_{1,2}(\cdot)$ denotes depth-wise convolution and $\odot$ denotes the element-wise multiplication. Note that we discard the GELU activation since it is redundant when working with element-wise multiplication [5], which is also validated by our ablation study.

3.4. HomoFormer

Given the elemental blocks of local self-attention with random shuffle and structure MLPs, we are now ready to construct the ultimate Transformer (i.e., HomoFormer) by integrating the basic building block in the widely used UNet [22, 39] architecture. This process is straightforward and the overall architecture is illustrated in Fig. 2. The loss function adopted for training our HomoFormer is the single Charbonnier loss [3] instead of complex hybrid loss [13, 57], whose mathematical expression is

$$L(I', I) = \sqrt{||I' - I||^2 + \epsilon^2}, \tag{7}$$

where I' and I are the output and shadow-free image respectively. The constant ϵ is empirically set to 10^{-3} for numerical stability.

4. Experiments

4.1. Experimental Settings

Random factors. Since random shuffle operations introduce stochasticity, we run evaluation for five times, calculate quantitative scores, and report the mean score [1], which is the default configuration (denoted by "HomoFormer" in tables). Besides, given these random factors, it is expected for the model to average it outputs as the final prediction, *i.e.*, marginalizing the random factors from Bayesian perspective. However, random shuffle operations are independent for each self-attention layer, resulting in exponential number of combinations. Therefore, we approximate the expected prediction using Monte Carlo averaging. The number of Monte Carlo samples is set to 8, which is marked by "HomoFormer+" in tables. We will further study the effect of the sample number in ablation study.

[1] In practice, the standard deviation is much smaller than the precision of reported mean. Hence, we overlook the standard deviation.

Datasets. Shadow removal experiments are conducted on the two representative benchmark datasets. (i) Adjusted ISTD (ISTD+) dataset [26] comprises of 1870 images triples (shadow images, shadow-free images, and shadow mask), which is divided into 1330 training triplets and 540 testing triplets. Compared with ISTD dataset [45], ISTD+ dataset [26] reduces the illumination inconsistency between the shadow and shadow-free image of ISTD by the image processing algorithm. Due to the repeated data, we move evaluation results on ISTD dataset to Supplementary Material; (ii) SRD dataset [38] is composed of 2680 training pairs and 408 testing pairs. Due to the lack of ground truth shadow masks in SRD, we directly utilize the public SRD shadow masks provided by DHAN [8] for the training and testing phase.

Evaluation metrics. To compare with other methods quantitatively, following the previous methods [13, 18, 23, 45], we utilize the root mean square error (MAE) in the LAB color space between the estimated images and ground truth shadow-free images. For the MAE metric, the lower values mean more faithful restoration, thus better results. Moreover, we also adopt the classic Peak Signal-to-Noise Ratio (PSNR) and structural similarity (SSIM) criterion to evaluate the performance of various methods in the RGB space. For the PSNR and SSIM metrics, higher values indicate better results. For consistent comparison, we resize estimated shadow-free images to the resolution of 256×256 to obtain quantitative results.

4.2. Comparison on the ISTD+ Dataset

We report the MAE score in comparison with other state-of-the-art methods on the ISTD+ Dataset [26] in Tab. 2. Twelve previous SOTA methods are included, ranging from traditional shadow removal method: Guo *et al.* [18], to recent deep learning based methods: DeshadowNet [38], ST-CGAN [54], ShadowGAN [21], SP+M-Net [26], Param+M+D-Net [27], G2R [35], Fu *et al.* [13], Jin *et al.* [23], BMNet [57], SG-ShadowNet [44], and ShadowFormer [15]. To guarantee fair comparison, the results of these compared methods are provided by the authors or

Table 1. Quantitative comparisons with the SOTA methods on the SRD dataset [38]. The best and the second results are **boldfaced** and underlined, respectively.

Method	Shadow Region			Non-Shadow Region			All Region		
	PSNR ↑	SSIM ↑	MAE ↓	PSNR ↑	SSIM ↑	MAE ↓	PSNR ↑	SSIM ↑	MAE ↓
Input images	18.96	0.871	36.69	31.47	0.975	4.83	18.19	0.829	14.05
Guo *et al.* [18] (TPAMI'12)	-	-	29.89	-	-	6.47	-	-	12.60
DeshadowNet [38] (CVPR'17)	-	-	11.78	-	-	4.84	-	-	6.64
DSC [20] (TPAMI'19)	30.65	0.960	8.62	31.94	0.965	4.41	27.76	0.903	5.71
DHAN [8] (AAAI'20)	33.67	0.978	8.94	34.79	0.979	4.80	30.51	0.949	5.67
Fu *et al.* [13] (CVPR'21)	32.26	0.966	8.55	31.87	0.945	5.74	28.40	0.893	6.50
Jin *et al.* [23] (ICCV'21)	34.00	0.975	7.70	35.53	0.981	3.65	31.53	0.955	4.65
BMNet [57] (CVPR'22)	35.05	0.981	6.61	36.02	0.982	3.61	31.69	0.956	4.46
SG-ShadowNet [44] (ECCV'22)	-	-	7.53	-	-	2.97	-	-	4.23
ShadowFormer [15] (AAAI'23)	36.91	**0.989**	5.90	36.22	**0.989**	3.44	32.90	0.958	4.04
ShadowDiffusion [16] (CVPR'23)	38.72	0.987	4.98	37.78	0.985	3.44	34.73	0.970	3.63
Li *et al.* [30] (ICCV'23)	**39.33**	0.985	6.09	35.61	0.967	2.97	33.17	0.941	3.83
HomoFormer(ours)	38.81	0.987	**4.25**	39.45	0.988	2.85	35.37	0.972	3.33
HomoFormer+(ours)	38.64	0.987	4.33	**40.04**	**0.989**	**2.76**	**35.50**	**0.972**	**3.29**

Table 2. Quantitative comparisons with the SOTA methods on the ISTD+ datasets. The best and the second results are **boldfaced** and underlined, respectively.

Region / Method	Shadow MAE↓	Non-Shadow MAE↓	All MAE↓
Input images	40.2	2.6	8.5
Guo *et al.* [18] (TPAMI'12)	22.0	3.1	6.1
DeshadowNet [38] (CVPR'17)	15.9	6.0	7.6
ST-CGAN [54] (CVPR'18)	13.4	7.7	8.7
ShadowGAN [21] (ICCV'19)	12.4	4.0	5.3
SP+M-Net [26] (ICCV'19)	7.9	3.1	3.9
Param+M+D-Net [27] (ECCV'20)	9.7	3.0	4.0
G2R [35] (CVPR'21)	7.3	2.9	3.6
Fu *et al.* [13] (CVPR'21)	6.5	3.8	4.2
Jin *et al.* [23] (ICCV'21)	10.3	3.5	4.6
BMNet [57] (CVPR'22)	5.6	2.5	3.0
SG-ShadowNet [44] (ECCV'22)	5.9	2.9	3.4
ShadowFormer [15] (AAAI'23)	5.2	2.3	2.8
ShadowDiffusion [16] (CVPR'23)	4.9	2.3	2.7
Li *et al.* [30] (ICCV'23)	5.9	2.9	3.3
HomoFormer(ours)	**5.0**	2.3	2.7
HomoFormer+(ours)	**5.0**	**2.2**	**2.6**

Table 3. Abalation study on the ISTD+ datasets.

Region / Method	Shadow MAE↓	Non-Shadow MAE↓	All MAE↓
w/o random shuffle	6.4	2.5	3.0
w/o structure	5.6	2.5	2.9
Ours (default)	5.0	2.3	2.7

obtained from the original paper. As shown in Tab. 2, our HomoFormer achieves lower MAE score than all previous SOTA methods, for example with the gain of 0.4 compared with BMNet [57], suggesting that our method is capable of restoring the clean image more faithfully. Moreover, with Monte Carlo averaging, our method (HomoFormer+) obtains better results in both shadow and non-shadow region. Fig. 5 shows that compared with other methods, our HomoFormer produces results with less boundary artifacts. Supplementary material provides extended results on SBU dataset [28] to further support its generalization.

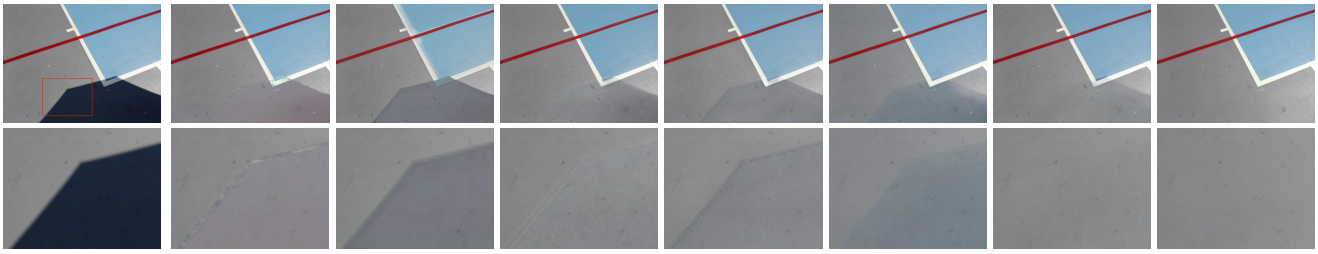
4.3. Comparison on SRD Dataset

In Table Tab. 1, we report the quantitative comparisons in terms of PSNR/SSIM/MAE with other SOTA methods on the SRD dataset [38], including Guo et al. [18], DeshadowNet [38], DSC [20], DHAN [8], Fu *et al.* [13], Jin *et al.* [23], BMNet [57], SG-ShadowNet [44] ShadowFormer [15], Li et al. [30] and ShadowDiffusion [16]. Our method also achieves the best de-shadowing performance with the lowest MAE and the highest PSNR/SSIM values. Compared with ShadowFormer [15], the PSNR value of our method is improvod from 32.90 dB to 35.37 dB. Besides, we also provide the visual comparisons in Fig. 6. We can observe that HomoFormer can remove shadow with less artifacts left.

4.4. Ablation Study and Analysis

To verify our choice and further promote understanding for our method, we conduct ablation experiments and analysis based on the ISTD+ Dataset [26].

Effect of random shuffle. To study the effect of random shuffle, we design a model variant by removing the random shuffle and inverse shuffle. As shown in Tab. 3, MAE increases on shadowed (+1.4), non-shadowed (+0.2) and all regions (+0.3), which suggests that random shuffle helps in removing shadow degradation clearly and maintaining the non-shadowed region faithfully. The underlying reason is that random shuffle homogenizes the non-uniformed distri-



(a) input (b) SP+M-N (c) Jin *et al.* (d) BMNet (e) SG-ShadowNet (f) ShadowFormer (g) HomoFormer (h) GT

Figure 5. Visual comparisons with state-of-the-art methods on the ISTD+ dataset [26]. First row: the full-size evaluations. Second row: the enlarged region. Our HomoFormer obtains visually pleasant results with less artifacts.

(a) input (b) DSC (c) DHAN (d) BMNet (e) SG-ShadowNet (f) HomoFormer (g) GT

Figure 6. Visual comparisons with state-of-the-art methods on the SRD dataset [38].

bution of degradation, which is beneficial for the weight-sharing local self-attention.

To take a further step towards understanding the behavior of random shuffle, we visualize intermediate features of our default HomoFormer and the variant without random shuffle. Fig. 7 shows some representative examples from the first encoding layer and last decoding layer. Generally, without random shuffle, local self-attention has to utilize a single parameter set to reach a compromise among regions with various shaded degrees (see Fig. 1). We can observe this effect from experiments: for features from the first encoding layer (column $2-3$ in Fig. 7), the highlighted region discards nearly all textures without random shuffle, which is harmful for recovering faithful results. For the features from the last decoding layer, we wish they should be as clean as possible, *i.e.* they should contain no artifacts brought by degradation, since they will be used to construct the final clean output. Observing the features from the last decoding layer, we find that HomoFormer with random shuffle produces much more visually pleasing results. These observations together lead to the conclusion that random shuffle is beneficial for extracting more effective features (more textures or less artifacts in our case).

Uncertainty can predict errors. Uncertainty can play a significant role for computer vision [25]. For shadow removal, uncertainty can estimate the degree of *confidence* about the prediction. The presented HomoFormer provides a natural approach to estimate its uncertainty due to its inherent random shuffle behaviour. For example, we can evaluate an image multiple times and compute the standard deviation as the uncertainty. Fig. 9 suggests that uncertainty computing *without* resorting to groundtruth image can predict where errors are prone to take place, which is of practical significance for real-world scenarios.

Effect of structure-modeling in FFN. Since we discard the structural information in local self-attention, we move the responsibility of modeling structural information to FFN. We validate the effectiveness by removing depth-wise convolution in FFN. Tab. 3 suggests that the absence of position modeling impairs the performance significantly.

Effect of the number of Monte Carlo samples. To marginalize random factors, we leverage Monte Carlo averaging to approximate the expectation. In theory, as the sampled number tends to infinity, the average gets close to the true expectation. We investigate this property on SRD dataset. Fig. 8 reveals that the performance is first promoted and then converged as the sample number increases. The sampled number of 8, which we adopt as the default value, can produce promising results while saving computations as much as possible.

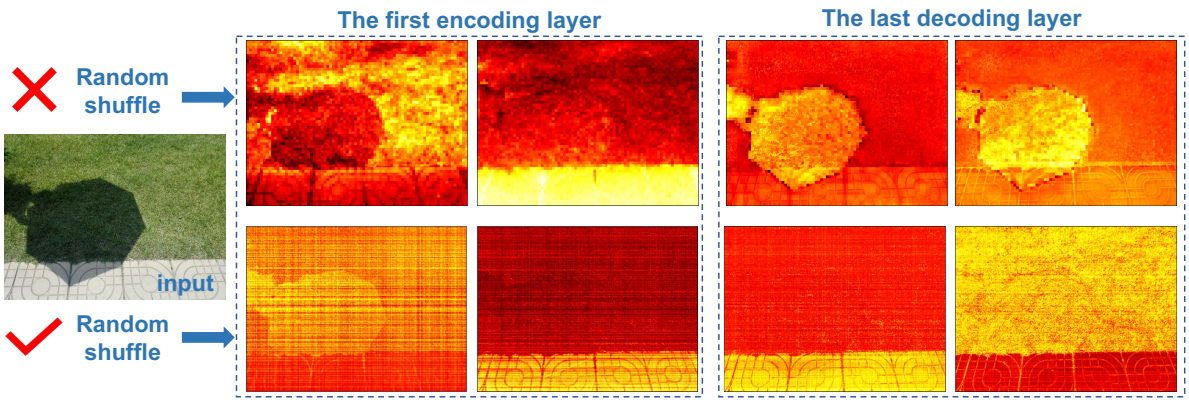


Figure 7. Visualization of features from the first encoding layer and last decoding layer. Features of the first row are taken from the variant model without random shuffle and the second row corresponds to features from the default HomoFormer model. Compared with the variant without random shuffle, features from the default HomoFormer contain more detailed textures in the shallow layer (column 2-3) and faithful contents in the deep layer (column 4-5).

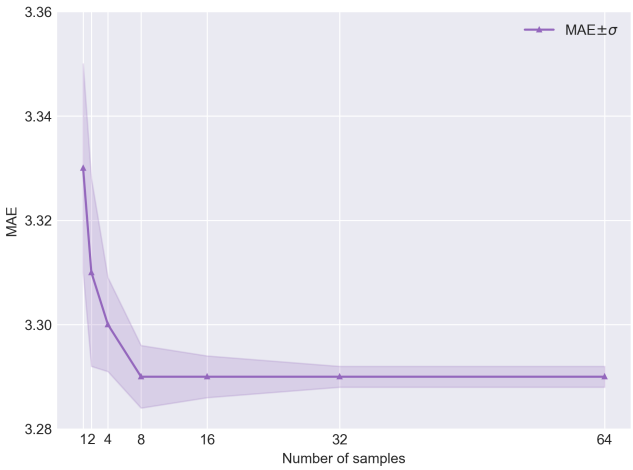
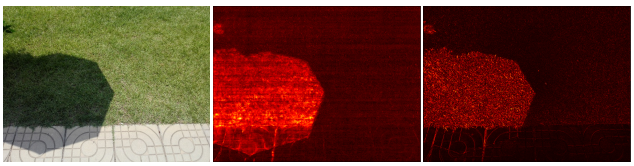


Figure 8. Effect of the number of Monte Carlo samples.

5. Discussions and Future Works

In this paper, we integrate the invertible shuffle operation with local self-attention, providing a fresh perspective to the challenge of modeling complex shadow degradation with non-uniform distribution and diverse pattern. Expect the motivation of being compatible with weight-sharing, the effectiveness of HomoFormer can be explained from the view of global interactions. We assume that pixels within an image are related. Random shuffle can pull two pixels into a single window regardless their distance. Hence, local SA on a shuffled image is equivalent to capturing sparse global interactions from that image, providing rich information for model to learn. Besides, since non-uniformity and diversity are not unique to shadow degradation, we are also excited about the future of homogenization in more general image restoration tasks, such as image inpainting, which is also our future work.



(a) Input (b) Uncertainty distribution (c) Error distribution

Figure 9. Uncertainty originating from random shuffle can be used to predict the error distribution (*i.e.*, $|I-I'|$) between the evaluated clean image I' and shadow-free image I.

6. Conclusion

In this paper, we provide a fresh perspective to tackle with the issue of modeling complex shadow degradation with non-uniform distribution and diverse pattern. By the elaborate random shuffle and inverse shuffle operation pair, the non-uniform distribution is homogenized, laying the foundation for effectively modeling the complex degradation with weight sharing models. Based on that, we establish a novel local window based transformer named HomoFormer for image shadow removal. Our HomoFormer can enjoy the efficient linear complexity to input resolution as well as overcome the challenge of modeling non-uniformly distributed shadow degradation. We conduct extensive and comprehensive experiments to validate and understand the proposed method.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants 62225207 and 62276243.

References

[1] Masashi Baba, Masayuki Mukunoki, and Naoki Asada. Shadow removal from a real image based on shadow density. In ACM SIGGRAPH 2004 Posters. 2004.

[2] Irwan Bello. Lambdanetworks: Modeling long-range interactions without attention. In ICLR, 2021.

[3] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In ICIP, 1994.

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

[5] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In ECCV, 2022.

[6] Zipei Chen, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Canet: A context-aware network for shadow removal. In ICCV, 2021.

[7] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In ICLR, 2020.

[8] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In AAAI, 2020.

[9] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Argan: Attentive recurrent generative adversarial network for shadow detection and removal. In ICCV, 2019.

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

[11] Graham D Finlayson, Steven D Hordley, Cheng Lu, and Mark S Drew. On the removal of shadows from images. TPAMI, 2005.

[12] Graham D Finlayson, Mark S Drew, and Cheng Lu. Entropy minimization for shadow removal. IJCV, 2009.

[13] Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang. Auto-exposure fusion for single-image shadow removal. In CVPR, 2021.

[14] Han Gong and Darren Cosker. Interactive removal and ground truth for difficult shadow scenes. JOSA A, 2016.

[15] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps image shadow removal. In AAAI, 2023.

[16] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In CVPR, 2023.

[17] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. TPAMI, 2012.

[18] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. TPAMI, 2013.

[19] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In CVPR, 2018.

[20] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection and removal. TPAMI, 2019.

[21] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. In ICCV, 2019.

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In CVPR, 2017.

[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 ICCV, 2021.

[24] Guoliang Kang, Xuanyi Dong, Liang Zheng, and Yi Yang. Patchshuffle regularization. arXiv preprint arXiv:1707.07103, 2017.

[25] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In NeurIPS, 2017.

[26] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In ICCV, 2019.

[27] Hieu Le and Dimitris Samaras. From shadow segmentation to shadow removal. In ECCV, 2020.

[28] Hieu Le and Dimitris Samaras. Physics-based shadow image decomposition for shadow removal. TPAMI, 2022.

[29] Wooseok Lee, Sanghyun Son, and Kyoung Mu Lee. Ap-bsn: Self-supervised denoising for real-world images via asymmetric pd and blind-spot network. In CVPR, 2022.

[30] Xiaoguang Li, Qing Guo, Rabab Abdelfattah, Di Lin, Wei Feng, Ivor Tsang, and Song Wang. Leveraging inpainting for single-image shadow removal. 2023.

[31] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In ICCV Workshops, 2021.

[32] Jing Lin, Yuanhao Cai, Xiaowan Hu, Haoqian Wang, Youliang Yan, Xueyi Zou, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. Flow-guided sparse transformer for video deblurring. In ICML, 2022.

[33] 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 CVPR, 2020.

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In ICCV, 2021.

[35] Zhihao Liu, Hui Yin, Xinyi Wu, Zhenyao Wu, Yang Mi, and Song Wang. From shadow generation to shadow removal. In CVPR, 2021.

[36] S. Nadimi and B. Bhanu. Physical models for moving shadow and object detection in video. TPAMI, 2004.

[37] Tam Minh Nguyen, Tan Minh Nguyen, Dung DD Le, Duy Khuong Nguyen, Viet-Anh Tran, Richard Baraniuk, Nhat Ho, and Stanley Osher. Improving transformers with probabilistic attention keys. In ICML, 2022.

[38] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson W. H. Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In CVPR, 2017.

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In MICCAI, 2015.

[40] Andres Sanin, Conrad Sanderson, and Brian C Lovell. Improved shadow removal for robust person tracking in surveillance scenarios. In ICPR, 2010.

[41] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In CVPR, 2016.

[42] Yael Shor and Dani Lischinski. The shadow meets the mask: Pyramid-based shadow removal. In Computer Graphics Forum. Wiley Online Library, 2008.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NeurIPS, 2017.

[44] Jin Wan, Hui Yin, Zhenyao Wu, Xinyi Wu, Yanting Liu, and Song Wang. Style-guided shadow removal. In ECCV, 2022.

[45] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In CVPR, 2018.

[46] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general u-shaped transformer for image restoration. In CVPR, 2022.

[47] Haixu Wu, Jialong Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Flowformer: Linearizing transformers with conservation flows. In ICML, 2022.

[48] Chunxia Xiao, Ruiyun She, Donglin Xiao, and Kwan-Liu Ma. Fast shadow removal using adaptive multi-scale illumination transfer. In CGF, 2013.

[49] Jie Xiao, Xueyang Fu, Feng Wu, and Zheng-Jun Zha. Stochastic window transformer for image restoration. In NeurIPS, 2022.

[50] Jie Xiao, Xueyang Fu, Aiping Liu, Feng Wu, and Zheng-Jun Zha. Image de-raining transformer. PAMI, 2023.

[51] Jie Xiao, Xueyang Fu, Man Zhou, Hongjian Liu, and Zheng-Jun Zha. Random shuffle transformer for image restoration. In ICML, 2023.

[52] Qingxiong Yang, Kar-Han Tan, and Narendra Ahuja. Shadow removal using bilateral filtering. TIP, 2012.

[53] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In CVPR, 2022.

[54] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In CVPR, 2019.

[55] Ling Zhang, Qing Zhang, and Chunxia Xiao. Shadow remover: Image shadow removal based on illumination recovering optimization. TIP, 2015.

[56] Wuming Zhang, Xi Zhao, Jean-Marie Morvan, and Liming Chen. Improving shadow suppression for illumination robust face recognition. TPAMI, 2018.

[57] Yurui Zhu, Jie Huang, Xueyang Fu, Feng Zhao, Qibin Sun, and Zheng–Jun Zha. Bijective mapping network for shadow removal. In CVPR, 2022.

[58] Yurui Zhu, Zeyu Xiao, Yanchi Fang, Xueyang Fu, Zhiwei Xiong, and Zheng-Jun Zha. Efficient model-driven network for shadow removal. In AAAI, 2022.