RMMamba: Randomized Mamba for Remote Sensing Shadow Removal

Jun Chu, Kaichen Chi, and Qi Wang, *Senior Member, IEEE*

***Abstract*—Remote sensing shadow removal task aims to effectively restore key regions of an image obscured by shadows. However, the spatial non-uniformity of shadow distribution presents significant challenges to this task. To address this issue, we propose RMMamba, a shadow removal network based on the SS2D architecture. The core concept of RMMamba involves balancing the spatial non-uniformity of shadow distribution, thereby optimizing the utilization efficiency of non-shadow pixel information across different windows. Specifically, RMMamba employs a random pixel shuffling operation to evenly disperse pixels from shadow regions with pronounced spatial non-uniformity into non-shadow areas, ensuring a more balanced spatial distribution of shadow and non-shadow pixels within each window. Subsequently, a shared weight local State Space Model (SS2D) is employed to integrate non-shadow pixel features uniformly distributed around shadow pixels, consequently effectively relighting shadow pixels. Reverse shuffling operations are then applied to restore the processed image to its original pixel order. Coupled with CP-FFN, a lightweight feedforward network incorporating color priors, RMMamba effectively restores color in shadow regions. More importantly, given the difficulty of acquiring remote sensing shadow samples with corresponding ground truth and shadow masks, we leverage the game GTA to control its shadow renderer and create SRGTA, a synthetic fully supervised dataset, hence providing a new benchmark for the performance evaluation of remote sensing shadow removal algorithms. Extensive experiments conducted on SRGTA and UAV-SC have demonstrated the outstanding performance of RMMamba. The code and SRGTA dataset are publicly available at https://github.com/xgd-cj/RMMamba.**

***Index Terms*—Remote sensing, shadow removal, state space model, random pixel shuffling.**

I. Introduction

SHADOWS are illumination degradation phenomena due to light occlusion [1], [2]. In high-resolution remote sensing images, the presence of shadows can significantly disrupt key information, thereby posing substantial obstacles to downstream remote sensing analysis tasks such as object detection [3]–[8], change detection [9], and object counting [10]. Consequently, shadow removal is an important research direction in remote sensing.

Traditional shadow removal methods typically rely on hand-crafted prior features, such as illumination intensity [11], gradient [12], and region [13]. However, these methods exhibit limited generalization capability and poorly adapt to varying illumination conditions. With the rapid advancement of deep learning, Convolutional Neural Network (CNN) and Transformer-based methods have become mainstream. Some pioneering perspectives include multi-task learning [14], [15], shadow generation [16], image decomposition [17], *etc*. Nevertheless, CNN-based methods struggle to fully learn the shadow-to-shadow-free mapping, often resulting in artifacts. Nowadays, numerous Transformer solutions adopt a window-based architecture, wherein weights are shared across different windows. Given the highly uneven spatial distribution of shadows, this results in significant variations in how different windows utilize information from shadow-free regions, leading to noticeable fluctuations in shadow removal performance.

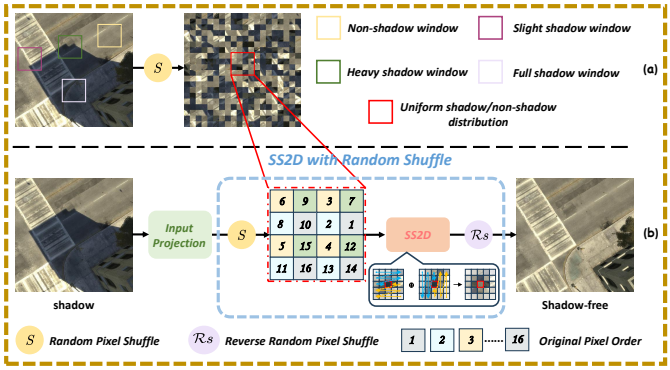

Fig. 1. Schematic of our basic idea. RMMamba employs a random shuffling operation to significantly enhance the information utilization of shadow-free pixels within the window, coupled with the SS2D scanning strategy to effectively handle shadow pixels.

In this work, we propose the SS2D shadow removal network based on a random pixel shuffling strategy to address the aforementioned challenges. As illustrated in Fig. 1 (a), the spatially uneven distribution of shadows leads to varying degrees of information utilization for shadow-free pixels within different processing windows. To enhance the effectiveness of the window-based weight-sharing mechanism, we employ an innovative randomized shuffling operation, which redistributes the pixels of the input image randomly, ensuring that each pixel has an equal probability of being reassigned to any location. This process uniformly disperses localized shadow regions across shadow-free areas, creating a consistent spatial distribution of shadow and shadow-free pixels, enhancing the effectiveness of shared-weight window-based processing. Furthermore, the SS2D feature extraction framework replaces the Transformer's self-attention mechanism with a state-space model (SSM) of linear complexity, enabling more efficient long-sequence modeling while reducing computational over-

This work was supported in part by the National Natural Science Foundation of China under Grant 62471394 and Grant U21B2041.

The authors are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: junchu@mail.nwpu.edu.cn, chikaichen@mail.nwpu.edu.cn, crabwq@gmail.com). Corresponding author: Qi Wang.

head and memory consumption, simultaneously enhancing local pattern learning and ensuring a more stable training process [18]. Subsequently, we incorporate a feedforward network that integrates prior information. Combining color channel information with a lightweight depthwise separable convolution module, RMMamba improves its capacity to restore shadow region colors and maintain image texture details, concurrently reducing extra training and computational overhead. More importantly, high-quality datasets are crucial for evaluating model performance, yet the difficulty of collecting shadow-free images for shadow scenes in natural environments makes existing remote sensing shadow removal datasets often inadequate for training and evaluating fully supervised models. To address this limitation, we leverage the popular game GTA, modifying its internal shadow renderer to control the spatial distribution of shadows. This allows us to generate a synthetic, fully supervised dataset for remote sensing shadow removal, facilitating model performance evaluation. In summary, our main contributions are as follows:

- **Perspective contribution.** We advance shadow removal beyond traditional localized processing, achieving more effective intra-window information utilization through global pixel distribution optimization. By reorganizing pixel distribution, we eliminate boundaries between shadow and shadow-free regions, enabling dynamic integration of shadow-free information across the image, fundamentally addressing challenges caused by uneven shadow distribution.
- **Technical contribution.** We propose RMMamba, based on random pixel shuffling, for remote sensing shadow removal and design a feedforward depthwise separable convolutional network incorporating color prior information to further restore and preserve image color and texture details.
- **Practical contribution.** We construct the synthetic fully supervised dataset SRGTA for evaluating remote sensing shadow removal methods. RMMamba demonstrates superior shadow removal quality on SRGTA and UAV-SC.

II. Related work

In this section, we first discuss representative methods for shadow removal, followed by an analysis of existing benchmarks in the field of shadow removal and their construction methodologies.

A. Shadow Removal

Early shadow removal methods typically achieve shadow elimination by exploring various physical shadow properties. To address the challenge of progressive shadow removal, Finlayson *et al.* [12] leveraged 1D grayscale for pixel-level illumination invariance, 2D chromaticity for unified relighting, and edge restoration for 3D full-color recovery. Movia *et al.* [19] integrated a shadow removal algorithm with Procrustes analysis to restore brightness. Silva *et al.* [20] presented a CIELCh-based method which employed multi-level thresholding and morphological operations for shadow detection. Building upon this, illumination ratios were then applied for shadow removal. However, traditional shadow removal methods often struggle to achieve satisfactory processing results and robust generalization performance, as their reliance on physical modeling is heavily dependent on scene-specific assumptions.

Recently, thanks to the emergence of large-scale training datasets [21], [22], deep learning-based methods have become the mainstream approach for shadow removal. Chi *et al.* [23] combined Retinex decomposition with contour and gradient regularization for illumination fidelity. Such a manner integrated neurophysiological knowledge into neural networks for efficient and interpretable shadow removal. To reduce boundary artifacts and illumination inconsistencies, Guo *et al.* [24] proposed a multi-scale channel attention framework with a shadow interaction module, modeling shadow and non-shadow region correlations. For effective handling of soft and hard shadows, Jin *et al.* [25] developed an unsupervised domain classifier-guided shadow removal network, integrating shadow/shadow-free classifiers with physics-based losses. Liu *et al.* [26] presented a weakly supervised method, employing a shadow generation sub-network to create paired data for shadow removal. However, these methods struggle with uneven shadow spatial distribution, failing to effectively eliminate artifacts.

B. Benchmarks

Natural scenes. Shadow removal datasets for natural scenes have been extensively developed. SRD [27], the first large-scale shadow removal dataset, comprises 3,088 shadow and shadow-free image pairs. ISTD [14] and ISTD+ [17] both consist of shadow images, shadow-free images, and shadow masks, comprising 1,330 training images and 540 testing images derived from 135 unique scenes. USR [22] is tailored for unpaired shadow removal tasks. SBU-Timelapse [28] includes 50 videos of static scenes with moving shadows. In addition, some datasets cater specifically to document shadow removal and facial shadow removal, such as SDSRD [29], SD7K [30], UCB [31], and PSE [32].

Remote sensing scenes. Capturing precisely corresponding shadow-free images in remote sensing scenes is challenging. This difficulty arises from the uncontrolled illumination and object conditions, unlike the controlled environments typical of natural scenes. At present, UAV-SC [33] is the only fully supervised shadow removal dataset available for remote sensing. Although UAV-SC fills the gap in this field, it still contains certain imperfections due to difficulties in controlling shadow-free conditions.

Synthetic Scenes. Given the difficulty of acquiring shadow-free samples in the real world, researchers have increasingly adopted synthetic image generation techniques. SVSRD-85 [34] is a synthetic video shadow removal dataset derived from the GTAV environment, comprising 85 videos (4,250 frames), with each frame paired with its corresponding shadow-free image. GTAV [35], also rendered from the GTAV environment, is a synthetic dataset comprising 5,723 pairs of shadow and shadow-free images. Regrettably, SVSRD-85 lacks detailed documentation on its creation methodology and remains unavailable to the public. In addition, GTAV is confined to data

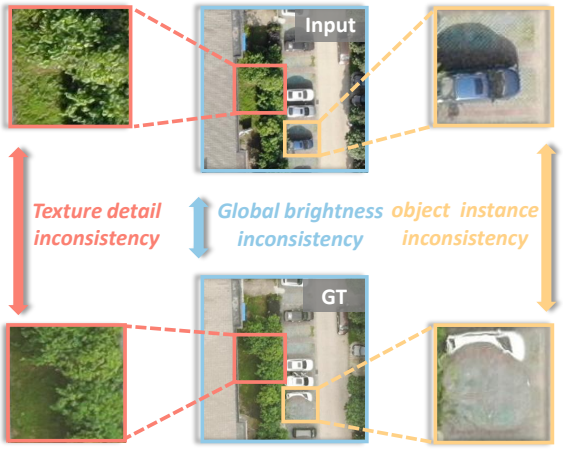

Fig. 2. Illustration of inherent issues in the UAV-SC.

collection for natural scenes and has not extended its exploration to remote sensing scenarios. Consequently, leveraging synthetic technologies to construct high-quality, fully supervised remote sensing datasets for shadow removal remains an unresolved yet profoundly significant research endeavor.

III. SRGTA

We construct SRGTA, the first fully supervised synthetic dataset specifically designed for remote sensing imagery shadow removal, effectively addressing the limitations of existing datasets in terms of data collection and annotation accuracy. SRGTA comprises 1,000 high-resolution triplets (1920×1080), each consisting of a shadow image, a shadow-free image, and a shadow mask, providing a reliable platform for evaluating algorithmic performances. Example images from the dataset are illustrated in Fig. 3.

A. Data Collection

As the only fully supervised dataset in the field of remote sensing shadow removal, UAV-SC still exhibits several inherent limitations in data collection process. First, acquisition conditions cannot be identical in two independent acquisition processes (shadow image and shadow-free image). This leads to a significant inconsistency in object instances between the input images and the ground truth images. Second, although the research team performed post-processing operations such as brightness enhancement on the ground truth images captured under overcast conditions, the global brightness differences between the input images and the ground truth images remain substantial. Moreover, the post-processing inevitably led to the loss of local texture details. The issues of inconsistent object instances, global brightness differences, and local texture distortion in UAV-SC are illustrated in Fig. 2.

The GTA (Grand Theft Auto) engine, with its powerful shadow rendering capabilities, provides reliable technical support for constructing high-quality image pairs. During the data collection process, we achieve precise control over the data through the following three key techniques: First, we utilize a physics modifier to adjust the game engine's parameters, successfully transitioning the camera perspective from

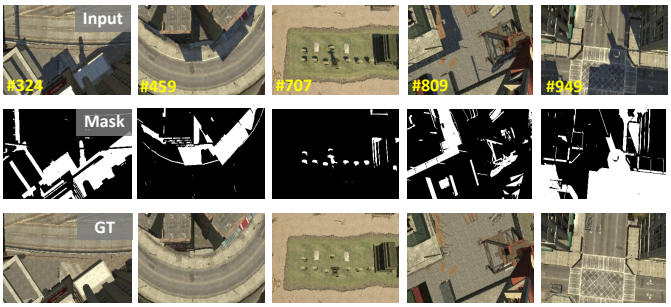

Fig. 3. Example images from SRGTA.

ground level to an aerial viewpoint and achieving a top-down angle suitable for remote sensing observations. Second, we employ scripts to edit and remove dynamic objects within the game scenes, effectively addressing the common issue of inconsistent object instances in traditional data collection methods and ensuring the static stability of the scenes. Finally, under the same observation perspective, we precisely control the parameters of the shadow renderer to obtain high-quality image pairs where the only variable is the presence or absence of shadows.

As shown in Fig. 4 (b), SRGTA consists of buildings, vehicles, and other typical scenes, encompassing 34 major object categories. This multi-scenario, multi-category data distribution not only ensures dataset diversity but also poses a challenge for performance testing.

B. Mask Processing

Fig. 4 (a) illustrates the data collection and mask processing pipeline. After obtaining the image pairs, we generate precise and detailed binary shadow masks through the following processing steps.

First, the shadow and shadow-free images are converted to grayscale, and the absolute difference between the two is calculated. We subsequently apply Gaussian blur to reduce the undesired noise present in the initial difference image:

$$F_G(x,y) = \sum_{i,j} F(x-i, y-j) \cdot G(i,j) \tag{1}$$

where $F(\cdot)$ represents the absolute difference at the corresponding pixel, $G(\cdot)$ represents the Gaussian blur kernel, $F_G(\cdot)$ represents the image after Gaussian blur, (x, y) represents the image pixel, and (i, j) represents the Gaussian kernel element.

We binarize the blurred difference image using a threshold to obtain the initial shadow mask. Subsequently, We perform meticulous morphological operations on the mask to further eliminate small noise points and smooth contours:

$$\begin{aligned} D(x,y) &= \max_{(i,j)\in K} \text{mask}(x-i, y-j) \\ E(x,y) &= \min_{(i,j)\in K} \text{mask}(x-i, y-j) \end{aligned} \tag{2}$$

where $D(\cdot)$ represents the dilation operation, $E(\cdot)$ represents the erosion operation, and K represents the structural element for morphological operations. Specifically, we use the closing

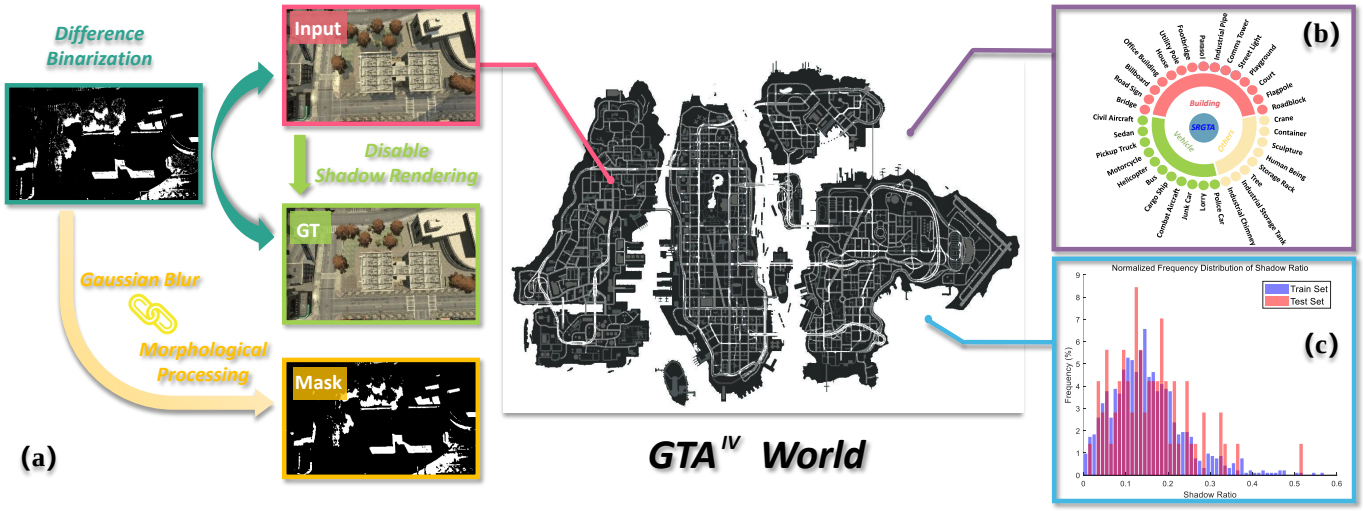

Fig. 4. Detailed information of SRGTA. (a) The creation process of SRGTA. (b) The collection scenarios and main object categories in SRGTA. (c) The proportional distribution of shadow areas in SRGTA.

operation (dilation followed by erosion) to fill small holes in the initial mask, and then apply the opening operation (erosion followed by dilation) to effectively remove discrete noise points in the mask, ultimately yielding a high-quality mask image. As shown in Fig. 4 (c), the shadow area proportions in SRGTA span a wide range, a characteristic that contributes to more stable learning outcomes during model training.

IV. Methodology

In this section, we first describe the principles of RMMamba's core components. Subsequently, RMMamba's overall architecture is presented.

A. R.S.R Module

The highly non-uniform spatial distribution of shadows leads to significant disparities in the utilization of non-shadow pixel information among different windows. To address this, we propose a local window SS2D feature extraction method that combines random pixel shuffling and reverse restoration. As illustrated in Fig. 5 (c), the process begins with randomly shuffling the pixels of the entire image, which ensures a uniform distribution of shadow and non-shadow pixels across different windows. Subsequently, the SS2D operation is applied. It effectively processes shadow pixels within each window by leveraging uniformly distributed non-shadow pixels. Finally, reverse restoration operation recovers the original semantic information, which was disrupted by the random shuffling. The workflow of the R.S.R Module can be expressed as:

$$\begin{aligned} X_s &= \mathrm{win}\left(S\left(X\right)\right) \\ \tilde{X}_s &= SS2D\left(X_s\right) \\ X_{out} &= R_S\left(\mathrm{mer}(\tilde{X}_s)\right) \end{aligned} \tag{3}$$

where $S(\cdot)$ represents the random pixel shuffling, $\mathrm{win}(\cdot)$ represents the windowed feature map operation, $\mathrm{mer}(\cdot)$ represents the window merging operation, and $R_S(\cdot)$ represents the reverse restoration of the shuffled pixels. It is notable that the implementations of $\mathrm{win}(\cdot)$ and $\mathrm{mer}(\cdot)$ align with the window partitioning and patch merging operations in Swin Transformer [36], respectively. Moreover, during the windowed feature map operation, we fix the window size to 8×8 to achieve an optimal receptive field for the model.

We elaborate on the specific implementations of $S(\cdot)$ and $R_S(\cdot)$. Given the input feature $X \in \mathbb{R}^{H\times W\times C}$, we generate random permutations $\pi_H \in S_H$ and $\pi_W \in S_W$ (S_H denotes the permutation group of H elements and S_W the permutation group of W elements) to produce the shuffled feature:

$$S\left(X\right) = X[\pi_H, \pi_W, C] \tag{4}$$

The inverse permutations π_H^{-1} and π_W^{-1} are then applied to the window-merged feature map to reconstruct the original spatial configuration:

$$R_S\left(\mathrm{mer}(\tilde{X}_s)\right) = \mathrm{mer}(\tilde{X}_s)[\pi_H^{-1}, \pi_W^{-1}, C] \tag{5}$$

B. CP-FFN

Regions after shadow removal may suffer from artifact residues and color distortion. Moreover, the aforementioned R.S.R Module does not explicitly explore structural information in the image. To address this, we design CP-FFN (Color-Prior Feed-Forward Network) following the R.S.R Module, a feed-forward network that incorporates prior information from the original image to correct color abnormalities in shadow-free regions while simultaneously mining structural details. As illustrated in Fig. 5, we encode the relative salience of each pixel's RGB channels by quantifying their intensity, providing a quantitative representation of the image's color feature. The color prior feature is formulated as follows:

$$P\left(x,y\right) = \left[P_R\left(x,y\right), P_G\left(x,y\right), P_B\left(x,y\right)\right] \tag{6}$$

where P denotes the derived color prior feature map. P_R, P_G, P_B are the rank-ordered values for the R-, G-, and B-channels at corresponding pixel positions in the original image. The three color channel values for each pixel are sorted by

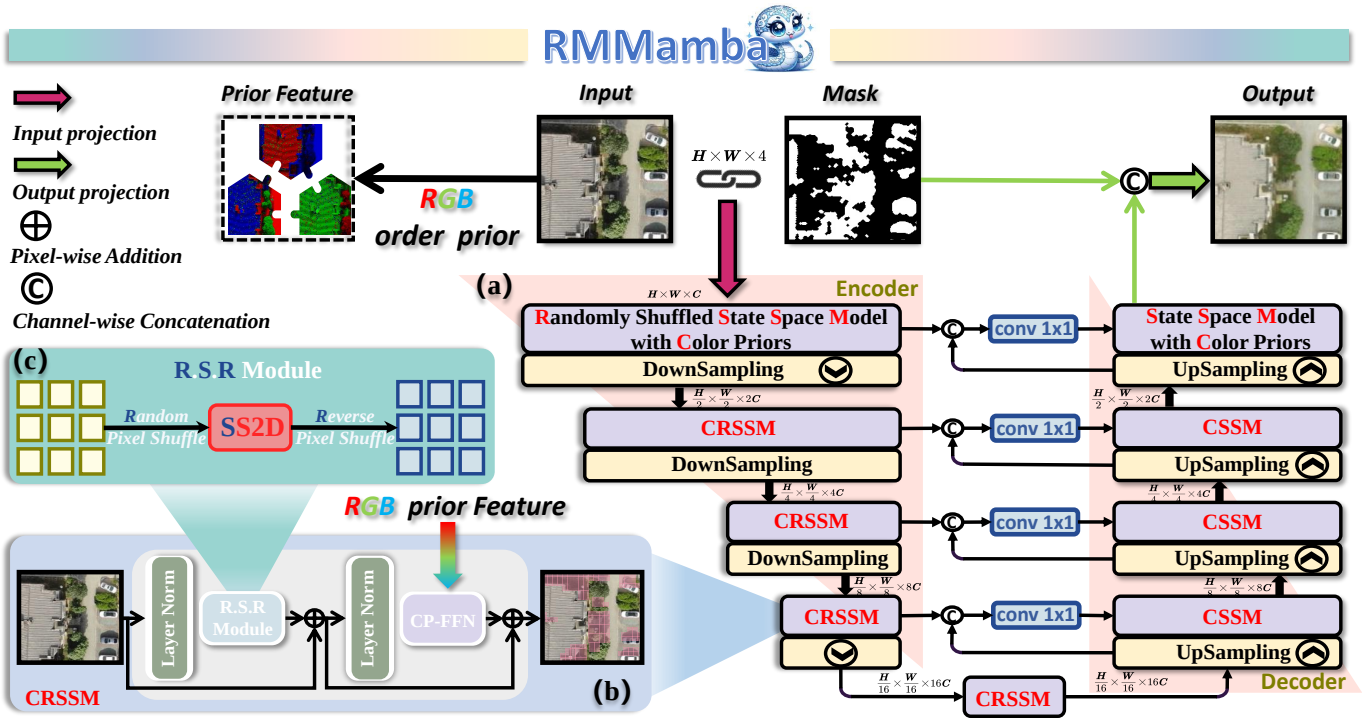

Fig. 5. Schematic diagram of RMMamba architecture. (a) Overall framework of RMMamba. The proposed model consists of multiple Randomly Shuffled State Space Models with Color Priors (CRSSM), following a U-Net structure for shadow removal. (b) Schematic diagram of the CRSSM Module. (c) Schematic diagram of the R.S.R Module.

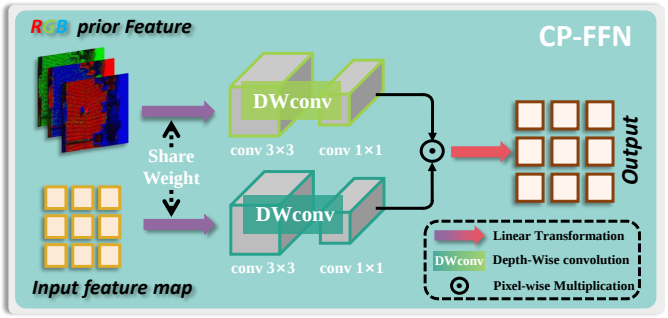

Fig. 6. Schematic diagram of the CP-FFN. CP-FFN utilizes dual-path depth-wise separable convolutions to separately aggregate color prior information and structural image features.

magnitude, and all sorted values are then normalized to the range of [-1, 1].

Subsequently, as illustrated in Fig. 6, color prior features and input features undergo a shared linear projection. The projected features are then fed into two independent depthwise separable convolution (DW) groups: DW_1 extracts color prior information, while DW_2 captures structural details through convolutional operations. The outputs are combined via pixel-wise multiplication to generate the final CP-FFN feature maps. This process can be expressed as:

$$\begin{aligned} \hat{P} &= DW_1\left(L_T\left(P\right)\right) \\ \hat{X} &= DW_2\left(L_T\left(X\right)\right) \\ Y &= \hat{P} \odot \hat{X} \end{aligned} \tag{7}$$

where $L_T(\cdot)$ represents the linear transformation, $DW(\cdot)$ represents the depthwise separable convolution, and $\odot$ represents the pixel-wise multiplication operator.

C. Overall Pipeline

The integration of the aforementioned R.S.R Module with CP-FFN through layer normalization and residual connections yields the CRSSM (Randomly Shuffled State Space Model with Color Priors). We then embed the CRSSM within a U-Net architecture, where skip connections between encoder and decoder pathways aggregate multi-level spatial and semantic features to strengthen shadow representation learning. The network output concatenates with the mask before undergoing output projection to produce the restored image. This framework constitutes our proposed RMMamba, as illustrated in Fig. 5. Notably, the decoder's feature processing stages exclude the random shuffling operation. To streamline RMMamba training, we adopt a simplified $\mathcal{L}_1$ loss between output and ground truth images:

$$\mathcal{L}_1\left(Y, \bar{Y}\right) = \sqrt{\|Y - \bar{Y}\|^2 + \alpha} \tag{8}$$

where Y represents the output image, $\bar{Y}$ represents the ground truth image, and α is fixed at 10^{-6} to ensure numerical stability during training. The parameter α serves as a core smoothing factor within the loss function. By introducing a positive bias, α effectively stabilizes the gradient-based optimization process. Consequently, the choice of α directly influences the balance among the loss function's smoothness, convergence stability, and sensitivity to error penalties. In this study, following the approach of [37], the value of α is ultimately set to 10^{-6}.

Fig. 7. Qualitative comparison on SRGTA.

TABLE I
QUANTITATIVE COMPARISONS ON SRGTA. "↑" REPRESENTS THAT LARGER SCORES ARE BETTER, WHILE "↓" REPRESENTS THAT LOWER SCORES ARE BETTER. "★" REPRESENTS THAT THE CODE IS NOT PUBLICLY AVAILABLE AND IS IMPLEMENTED BY OURSELVES. THE BEST SCORE IS IN RED, THE SECOND-BEST SCORE IS IN BLUE, AND THE THIRD-BEST SCORE IS IN GREEN.

Methods	RMSE(↓)			PSNR(↑)			SSIM(↑)		
	S.	N.S.	All	S.	N.S.	All	S.	N.S.	All
Silva (ISPRS)	33.5334	1.2507	7.7219	20.9973	36.9557	20.8220	0.9028	0.9942	0.8616
Gong (BMVC)	22.1542	7.4113	10.3665	24.1434	22.1625	19.3873	0.9456	0.9207	0.8447
Mask-ShadowGAN (ICCV)	23.7432	3.3467	7.4352	22.6094	33.1764	21.9479	0.9414	0.9819	0.8933
DC-ShadowNet (ICCV)	14.7434	3.2595	5.5615	27.1863	32.0873	25.5366	0.9697	0.9826	0.9402
LG-ShadowNet (TIP)	19.6012	3.3681	6.6220	24.1294	30.4691	22.9625	0.9463	0.9771	0.8988
G2R-ShadowNet (CVPR)	17.6342	1.2204	4.5106	26.0084	37.4944	25.5632	0.9497	0.9950	0.9287
DMTN (TMM)	14.3663	5.0243	6.8967	27.7200	30.0299	25.3523	0.9761	0.9753	0.9358
ST-CGAN (CVPR) ★	16.1323	5.9712	8.0080	26.4471	28.3581	23.9163	0.9589	0.9545	0.8918
ShadowFormer (AAAI)	21.2913	18.0981	18.7382	29.5884	24.7623	23.2589	0.9754	0.8948	0.8566
TBRNet (TNNLS)	9.5906	2.6792	4.0646	30.4391	34.1810	28.6021	0.9860	0.9878	0.9651
RASM (ACMMM)	8.9523	3.5145	4.6046	33.0155	36.6047	31.1646	0.9890	0.9939	0.9772
RMMamba	4.6074	0.8507	1.6037	37.7661	41.3771	35.8565	0.9959	0.9974	0.9910

V. EXPERIMENT

A. Experimental Setting

Implementation Details. RMMamba is implemented using PyTorch and trained for 500 epochs on an NVIDIA RTX 3090 GPU. The initial learning rate is set to 2×10^{-4} and gradually decrease to 1×10^{-6} using a cosine annealing schedule. The Adam optimizer is employed to optimize the training process. We have examined the implications of random seed initialization in pixel shuffling. Our trials have revealed that maintaining a fixed seed has no significant effect on the shadow removal quality. Therefore, we do not fix the random seed throughout both the training and inference stages.

Benchmarks. The performance of the shadow removal models is evaluated on two datasets. Our proposed SRGTA comprises 1,000 triplets (shadow image, shadow-free image, and shadow mask), which are divided into 930 training samples and 70 testing samples. UAV-SC [33] consists of 5,771 pairs of images (shadow image and shadow-free image), with a split of 5,741 training samples and 30 testing samples. For SRGTA, since its scene categories are quite evenly distributed, we simply use random sampling to split the data. For UAV-SC, we directly proceed with the experiments using the officially released split. Notably, since UAV-SC does not provide shadow masks, we employ a pre-trained BDRAR [38] to predict the masks for subsequent experiments. The BDRAR model employs ResNeXt101 as the backbone network. To ensure experimental reproducibility, we adopt the official default model version, parameter settings, and pre-trained weights for shadow mask generation.

Competitor. We conduct qualitative and quantitative comparisons of RMMamba with the following methods: Sliva [20], Gong [39], Mask-ShadowGAN [22], DC-ShadowNet [25], LG-ShadowNet [40], G2R-ShadowNet [26], DMTN [15], ST-CGAN [14], ShadowFormer [24], TBRNet [41], and RASM [42]. All competitors were re-trained on SRGTA and UAV-SC, achieving their best quantitative scores.

Quantitative Metrics. We quantitatively evaluate the performance of the models using Root Mean Squared Error (RMSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM). Specifically, a lower RMSE score indicates better shadow removal performance, while higher PSNR and SSIM scores signify superior shadow removal results. Furthermore, to comprehensively assess the quality of the restored images, we calculate these metrics within the shadow regions (S.), non-shadow regions (N.S.), and the entire image (All).

B. Qualitative Comparison

Qualitative Results on SRGTA. We first present visual comparisons on the SRGTA in Fig. 7. Traditional methods not only fail to improve the brightness of shadow regions but also alter the intrinsic colors of areas such as buildings. Furthermore, Mask-ShadowGAN [22], DC-ShadowNet [25], LG-ShadowNet [40], and G2R-ShadowNet [26] partially restore

Fig. 8. Qualitative comparison on UAV-SC.

TABLE II
QUANTITATIVE COMPARISONS ON UAV-SC. "↑" REPRESENTS THAT LARGER SCORES ARE BETTER, WHILE "↓" REPRESENTS THAT LOWER SCORES ARE BETTER. "★" REPRESENTS THAT THE CODE IS NOT PUBLICLY AVAILABLE AND IS IMPLEMENTED BY OURSELVES. THE BEST SCORE IS IN RED, THE SECOND-BEST SCORE IS IN BLUE, AND THE THIRD-BEST SCORE IS IN GREEN.

Methods	RMSE(↓)			PSNR(↑)			SSIM(↑)		
	S.	N.S.	All	S.	N.S.	All	S.	N.S.	All
Silva (ISPRS)	32.0735	18.5151	21.3105	22.8892	18.1030	16.2615	0.8899	0.8036	0.6870
Gong (BMVC)	26.0877	18.8191	20.3164	25.1442	18.1943	16.8288	0.9179	0.7930	0.6989
Mask-ShadowGAN (ICCV)	19.7937	17.5229	17.9907	27.5591	20.4367	19.1223	0.9447	0.8132	0.7360
DC-ShadowNet (ICCV)	16.1668	13.6317	14.1539	29.2334	22.4016	21.0528	0.9540	0.8469	0.7837
LG-ShadowNet (TIP)	23.0471	16.6574	17.9737	25.4449	19.6337	18.0468	0.9351	0.8348	0.7624
G2R-ShadowNet (CVPR)	22.0546	18.6261	19.3324	24.6921	17.9896	16.8316	0.9015	0.8190	0.7184
DMTN (TMM)	11.5474	9.9616	10.2882	31.5792	24.1271	22.9987	0.9660	0.8682	0.8175
ST-CGAN (CVPR) ★	9.3289	9.7077	9.6296	32.6684	24.3316	23.4039	0.9707	0.8786	0.8326
ShadowFormer (AAAI)	10.2398	10.0976	10.7881	31.7308	24.1267	23.0978	0.9692	0.8761	0.8293
TBRNet (TNNLS)	11.4553	10.9248	11.0341	31.1372	23.5854	22.5042	0.9617	0.8294	0.7702
RASM (ACMMM)	10.7721	9.3528	9.6452	32.8280	24.9515	23.8625	0.9721	0.9087	0.8678
RMMamba	9.0785	8.8784	8.9196	33.0767	25.2915	24.2230	0.9755	0.9078	0.8705

the illumination in shadow areas. However, these unsupervised and weakly-supervised strategies struggle to adapt to shadows' highly stochastic distribution, causing significant artifacts in their results. DMTN [15], ST-CGAN [14], and TBR-Net [41] cause shifts in global brightness and the intrinsic colors of objects. Although ShadowFormer [24] effectively removes shadows, it severely distorts the color information of the image. Furthermore, RASM [42] exhibits substantial color errors in the shadow removal regions. In contrast, RMMamba achieves the best shadow removal performance and effectively recovers the color characteristics of the shadow regions.

Qualitative Results on UAV-SC. Fig. 8 presents visual comparisons on the UAV-SC. In addition to the persistent issues observed in the results of the competing methods, such as the failure to restore brightness in shadow regions, the generation of artifacts, the alteration of intrinsic colors in shadow areas, and global brightness shifts, DMTN [15] and TBR-Net [41] also produce undesirable stripe artifacts in regions like rooftops. Furthermore, the results of RASM [42] exhibit a significant loss of structural and textural details. In contrast, RMMamba, benefiting from the meticulously designed CP-FFN, effectively removes shadows while preserving the color information and structural-textural details of the image.

C. Quantitative Comparison

Table I presents the performance scores of RMMamba and the competitors on the SRGTA dataset. RMMamba outperforms all competing methods. Compared to the best-performing competitor, RMMamba achieves performance gains of 60.54%, 15.06%, and 1.41% in terms of RMSE, PSNR, and SSIM, respectively. Table II presents the performance scores of RMMamba and the competitors on the UAV-SC, where RMMamba also achieves near state-of-the-art performance. Compared to the best-performing competitor, RMMamba achieves performance gains of 7.37%, 1.51%, and 0.31% in terms of RMSE, PSNR, and SSIM, respectively.

TABLE III
QUANTITATIVE SCORES OF THE ABLATION STUDY. "↑" REPRESENTS THAT LARGER SCORES ARE BETTER, WHILE "↓" REPRESENTS THAT LOWER SCORES ARE BETTER. THE BEST SCORE IS IN RED.

Baselines	UAV-SC		
	RMSE(↓)	PSNR(↑)	SSIM(↑)
w/o pixel shuffle	10.1744	22.2175	0.8632
w/o prior	9.9203	24.0153	0.8658
w/o CP-FFN	11.1034	23.5217	0.8419
RMMamba	8.9196	24.2230	0.8705
Baselines	**SRGTA**		
	RMSE(↓)	PSNR(↑)	SSIM(↑)
w/o pixel shuffle	2.0183	33.5588	0.9862
w/o prior	1.7474	35.8327	0.9813
w/o CP-FFN	2.6316	32.0401	0.9801
RMMamba	1.6037	35.8565	0.9910

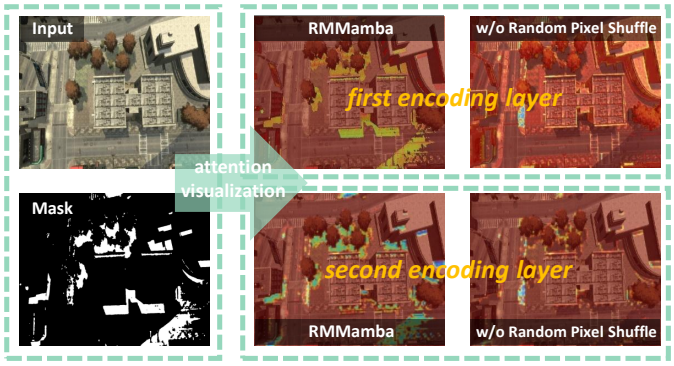

Fig. 9. Ablation study toward the random pixel shuffling strategy.

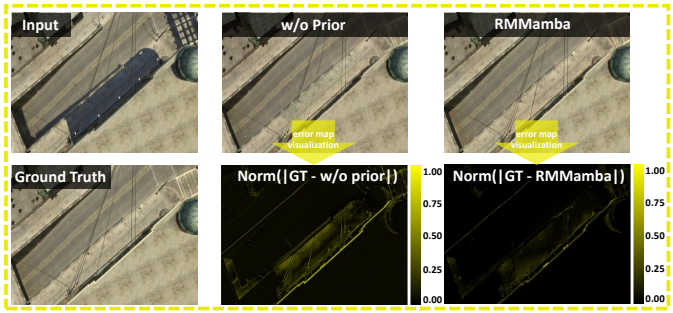

Fig. 10. Ablation study toward the RGB order prior.

The quantitative comparison results fully demonstrate the effectiveness of the well-designed RMMamba. Furthermore, by comparing the quantitative results across the two benchmarks, we observe significant variations in the relative performance of different competing methods. This highlights the necessity of employing diverse benchmarks for a comprehensive evaluation of method performance and underscores the practical significance of our proposed SRGTA.

D. Ablation Study

We conduct a series of ablation experiments, specifically examining the impact of the random pixel shuffling strategy, color prior information, and the CP-FFN.

- w/o Random Pixel Shuffle refers to RMMamba without the random pixel shuffling step, relying solely on SS2D for feature processing in each feature extraction and representation stage.
- w/o Prior refers to RMMamba without the RGB order prior. In the original network, the CP-FFN receives two inputs: the current input feature map and the prior guidance information. In the ablation experiment, the RGB order prior is replaced by the current feature map, so both inputs to the CP-FFN are the input feature map.
- w/o CP-FFN refers to RMMamba without the CP-FFN feed-forward network. After the R.S.R Module processes the feature map, it undergoes Layer Normalization and a residual connection before being directly outputted.

Table III presents a quantitative comparison of the performance of each ablated model. Furthermore, the effectiveness of the random pixel shuffling strategy, the impact of color prior information, and the contribution of the CP-FFN feed-forward network are illustrated in Fig. 9, 10, and 11, respectively. The following conclusions can be drawn from the ablation experiments:

- As demonstrated in Table III, RMMamba achieve the optimal performance scores in comparison to other ablated models, indicating the necessity of the combined effect of the random pixel shuffling strategy, color prior information, and the CP-FFN feed-forward network.
- As illustrated in Fig. 9, the ablated model w/o Random Pixel Shuffle exhibits limitations in adapting to the non-uniform spatial distribution of shadows, consequently hindering its ability to rapidly focus on shadow regions during the encoding stage. In contrast, RMMamba, benefiting from the random pixel shuffling strategy, can allocate optimal processing weights between shadow and non-shadow areas.
- As depicted in Fig. 10, the ablated model w/o Prior causes noticeable color deviations in shadow removal regions. Conversely, RMMamba, leveraging the RGB order prior information, demonstrates a superior ability to restore the inherent color characteristics of shadow regions, thereby achieving minimal color deviation.
- We employ the Canny operator to extract image contours and conduct a visualization comparison, as illustrated in Fig. 11. The processing results of the ablated model w/o CP-FFN exhibit a significant loss of texture details, whereas the processing results of RMMamba demonstrate the highest similarity to the ground truth image contours, achieving minimal texture information distortion. This indicates that the meticulously designed CP-FFN is capable of efficient structural modeling.

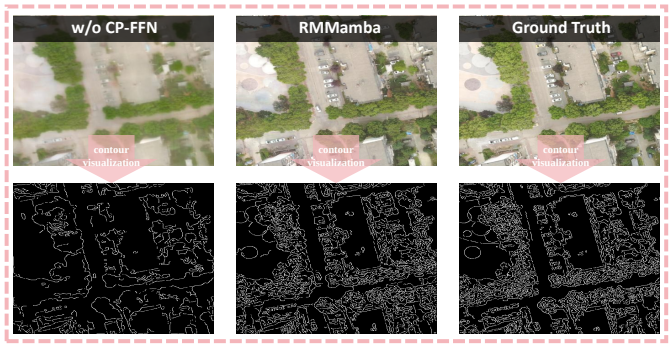

Fig. 11. Ablation study toward the CP-FFN.

E. Fourier Domain Adaptation Experiment

To further explore the potential of SRGTA in cross-domain applications, we draw inspiration from the method proposed by [34] and employ Fourier Domain Adaptation (FDA) [43] to validate the cross-domain image processing capability of the RMMamba trained with SRGTA.

Fig. 12 illustrates the principle of FDA in mitigating domain disparity. The low-frequency components of an image primarily encompass holistic information, smooth regions, color distribution, and lower-level textures, which are typically more

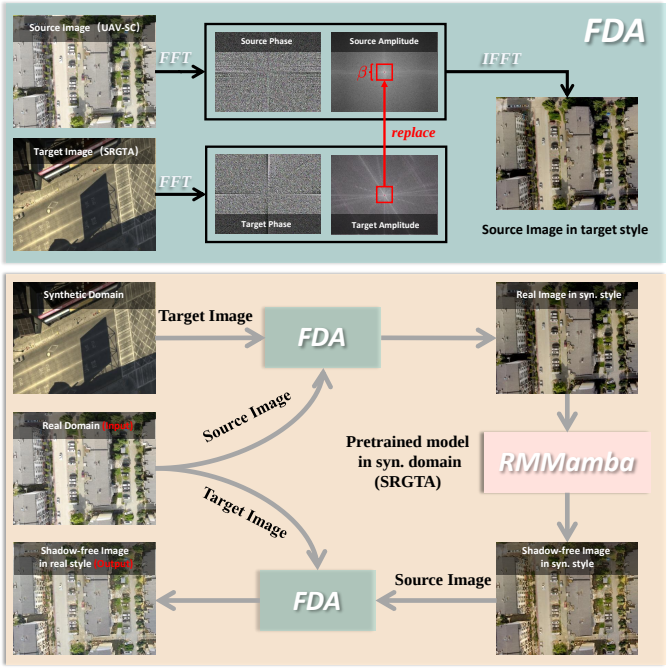

Fig. 12. Our cross-domain experiments employed FDA.

Fig. 13. Cross-domain shadow removal results.

strongly associated with image style. We perform Fast Fourier Transform (FFT) on real-world images (source domain) and SRGTA-synthesized images (target domain). Subsequently, the low-frequency amplitudes of the source images are replaced with those of the target images, with the degree of replacement controlled by parameter β, while preserving the phase of the source images. Following this, the source images are reconstructed via Inverse FFT (IFFT), yielding images imbued with the SRGTA style. Finally, the RMMamba model, pre-trained on SRGTA, is employed to perform shadow removal on these transformed images, and the results are then converted back to the real-world style using FDA to obtain the final shadow-removed images.

FDA is highly suitable for cross-domain tasks due to its minimal computational requirements. It exclusively involves basic FFT, IFFT, and low-frequency signal swapping, completely eschewing deep learning networks. This allows for extremely high computation speed coupled with very low resource consumption. FDA was performance-tested on a Windows 64-bit environment, featuring an Intel(R) Core(TM) i7-14700F processor (2.10 GHz) and 16.0 GB of RAM. For the evaluation, the input consisted of a 512×512 source image and a 1920×1080 target image. The results show that a single FDA processing took 0.2163 seconds and incurred an additional RAM memory overhead of only 47.09 MB during runtime. This demonstrates the FDA's high operational speed and minimal memory footprint under this configuration.

It's worth noting that the selection of the β parameter is crucial for the quality of cross-domain images. Both excessively large or small β values can lead to poor visual results in the transferred images, potentially causing damage to high-frequency image details. Through experimentation, we found that setting β between 0.001 and 0.01 yielded optimal results, achieving style transfer while maximally preserving the original image's high-frequency details. To ensure experimental reproducibility, we standardized β to 0.004.

As depicted in Fig. 13, the application of this straightforward, training-free cross-domain strategy yield significant shadow removal results, demonstrating the robust domain generalization capability of the model trained on SRGTA. Crucially, our FDA experiments serve as a proof-of-concept, demonstrating SRGTA's promising potential for cross-domain applications. Therefore, we do not quantify the domain gap between SRGTA and real-world data in this study. Detailed quantitative analysis will be a key focus in our future work.

VI. Conclusion

In this paper, we propose RMMamba, a shadow removal network leveraging random pixel shuffling to address the non-uniform shadow spatial distribution in remote sensing images. Specifically, RMMamba utilizes a random pixel shuffling operation to balance the spatial distribution of shadow and non-shadow pixels. Concurrently, we design a Color Prior Feed-forward Network to further restore the color and texture details of the shadow removal results. More importantly, we construct a novel fully-supervised synthetic remote sensing shadow removal dataset, SRGTA. Furthermore, the effectiveness of RMMamba's core components and the potential of SRGTA for cross-domain applications are also validated through experiments.

References

[1] Q. Wang, K. Chi, W. Jing, and Y. Yuan, "Recreating brightness from remote sensing shadow appearance," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–11, 2024.
[2] K. Chi, J. Li, W. Jing, Q. Li, and Q. Wang, "Neural implicit fourier transform for remote sensing shadow removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–10, 2024.
[3] C. Yang, M. Chen, Z. Xiong, Y. Yuan, and Q. Wang, "Cm-net: Concentric mask based arbitrary-shaped text detection," *IEEE Trans. Image Process.*, vol. 31, pp. 2864–2877, 2022.
[4] C. Yang, M. Chen, Y. Yuan, and Q. Wang, "Text growing on leaf," *IEEE Trans. Multimedia*, vol. 25, pp. 9029–9043, 2023.
[5] C. Yang, M. Chen, Y. Yuan, and Q. Wang, "Zoom text detector," *IEEE Trans. Neural Netw. Learn. Syst.*, 2023.
[6] C. Yang, M. Chen, Y. Yuan, and Q. Wang, "Reinforcement shrink-mask for text detection," *IEEE Trans. Multimedia*, vol. 25, pp. 6458–6470, 2022.
[7] C. Yang, X. Han, T. Han, H. Han, B. Zhao, and Q. Wang, "Edge Approximation Text Detector," *IEEE Trans. Circuits Syst. Video Technol.*, 2025.
[8] C. Yang, B. Zhao, Q. Zhou, and Q. Wang, "MMO-IG: Multi-Class and Multi-Scale Object Image Generation for Remote Sensing," *IEEE Trans. Geosci. Remote Sens.*, 2025.
[9] W. Jing, K. Chi, Q. Li, and Q. Wang, "ChangeRD: A registration-integrated change detection framework for unaligned remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 220, pp. 64–74, 2025.

[10] H. Guo, J. Gao, and Y. Yuan, "Balanced density regression network for remote sensing object counting," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, 2024.
[11] M. Gryka, M. Terry, and G. J. Brostow, "Learning to remove soft shadows," *ACM Trans. Graph.*, vol. 34, pp. 1–15, 2015.
[12] G. D. Finlayson, S. D. Hordley, C. Lu, and M. S. Drew, "On the removal of shadows from images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, pp. 59–68, 2006.
[13] R. Guo, Q. Dai, and D. Hoiem, "Paired regions for shadow detection and removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, pp. 2956–2967, 2012.
[14] J. Wang, X. Li, and J. Yang, "Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1788–1797.
[15] J. Liu, Q. Wang, H. Fan, W. Li, L. Qu, and Y. Tang, "A decoupled multi-task network for shadow removal," *IEEE Trans. Multimedia*, vol. 25, pp. 9449–9463, 2023.
[16] N. Inoue and T. Yamasaki, "Learning from synthetic shadows for shadow detection and removal," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, pp. 4187–4197, 2020.
[17] H. Le and D. Samaras, "Shadow removal via shadow image decomposition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 8578–8587.
[18] Y. Liu *et al.*, "Vmamba: Visual state space model," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 103031–103063, 2025.
[19] A. Movia, A. Beinat, and F. Crosilla, "Shadow detection and removal in RGB VHR images for land use unsupervised classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 119, pp. 485–495, 2016.
[20] G. F. Silva, G. B. Carneiro, R. Doth, L. A. Amaral, and D. F. G. de Azevedo, "Near real-time shadow detection and removal in aerial motion imagery application," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 104–121, 2018.
[21] Q. Yang, K.-H. Tan, and N. Ahuja, "Shadow removal using bilateral filtering," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4361–4368, 2012.
[22] X. Hu, Y. Jiang, C.-W. Fu, and P.-A. Heng, "Mask-shadowgan: Learning to remove shadows from unpaired data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 2472–2481.
[23] K. Chi, S. Guo, J. Chu, Q. Li, and Q. Wang, "RSMamba: Biologically plausible retinex-based Mamba for remote sensing shadow removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–10, 2025.
[24] L. Guo, S. Huang, D. Liu, C. Hao, and B. Wen, "Shadowformer: global context helps shadow removal," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 1, pp. 710–718, 2023.
[25] Y. Jin, A. Sharma, and R. T. Tan, "Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 5027–5036.
[26] Z. Liu, H. Yin, X. Wu, Z. Wu, Y. Mi, and S. Wang, "From shadow generation to shadow removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 4927–4936.
[27] L. Qu, J. Tian, S. He, Y. Tang, and R. W. H. Lau, "Deshadownet: A multi-context embedding deep network for shadow removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4067–4075.
[28] H. Le and D. Samaras, "Physics-based shadow image decomposition for shadow removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, pp. 9088–9101, 2021.
[29] Y. H. Lin, W. C. Chen, and Y. Y. Chuang, "BEDSR-Net: A deep shadow removal network from a single document image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020.
[30] Z. Li, X. Chen, C.-M. Pun, and X. Cun, "High-resolution document shadow removal via a large-scale real-world dataset and a frequency-aware shadow erasing net," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 12415–12424.
[31] X. Zhang *et al.*, "Portrait shadow manipulation," *ACM Trans. Graph.*, vol. 39, pp. 78–1, 2020.
[32] J. Lyu, Z. Wang, and F. Xu, "Portrait eyeglasses and shadow removal by leveraging 3d synthetic data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 3429–3439.
[33] S. Luo *et al.*, "An evolutionary shadow correction network and a benchmark UAV dataset for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023.
[34] Z. Chen, L. Wan, Y. Xiao, L. Zhu, and H. Fu, "Learning physical-spatio-temporal features for video shadow removal," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, pp. 5830–5842, 2024.
[35] O. Sidorov, "Conditional gans for multi-illuminant color constancy: Revolution or yet another approach?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 1748–1758.
[36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10012–10022.
[37] J. Xiao, X. Fu, Y. Zhu, D. Li, J. Huang, K. Zhu, and Z.-J. Zha, "Homoformer: Homogenized transformer for image shadow removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 25617–25626.
[38] L. Zhu *et al.*, "Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 121–136.
[39] H. Gong and D. P. Cosker, "Interactive shadow removal and ground truth for variable scene categories," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2014.
[40] Z. Liu, H. Yin, Y. Mi, M. Pu, and S. Wang, "Shadow removal by a lightness-guided network with training on unpaired data," *IEEE Trans. Image Process.*, vol. 30, pp. 1853–1865, 2021.
[41] J. Liu, Q. Wang, H. Fan, J. Tian, and Y. Tang, "A shadow imaging bilinear model and three-branch residual network for shadow removal," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 15857–15871, 2024.
[42] H. Liu, M. Li, and X. Guo, "Regional attention for shadow removal," in *Proc. ACM Int. Conf. Multimedia*, 2024, pp. 5949–5957.
[43] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 4085–4095.

Jun Chu received the B.E. degree in automation from Northwestern Polytechnical University, Xi'an, China, in 2024. He is currently pursuing the M.S. degree with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include deep learning and computer vision.

Kaichen Chi received the B.E. degree in electronic and information engineering and the M.E. degree in communication and information system from Liaoning Technical University, Huludao, China, in 2019 and 2022 respectively. He is currently working toward the Ph.D. degree in the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include image processing and deep learning.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition and remote sensing. For more information, visit the link (https://crabwq.github.io/).