Document Shadow Removal via Pixel-Adaptive Illumination Residual Learning

Yuanzhen Li[1], Leran Ye[1], Fubao Yang[1], Hongzhi Liu[1], Fengli Yang[2], Xiaoxiao Wang[3], Yue Zhao[1*]
[1] School of Mathematics and Statistics, Yunnan University, China.
[2] School of Microelectronics and Communication Engineering, Chongqing University, China.
[3] School of Statistics and Mathematics, Yunnan University of Finance and Economics, China.
20249047@ynu.edu.cn, cqu_yfl@cqu.edu.cn, xiaoxiaowang6@126.com, zhao6685@yeah.net

Abstract

Existing document shadow removal methods primarily rely on background images as guidance, which leads the network to misidentify shadows as background, resulting in residual shadows and background color distortion. To address this, we propose a novel framework that reformulates document deshadowing as a pixel-level adaptive illumination residual learning task. Our method employs residual learning to predict localized illumination adjustments (rather than directly generating pixel values). We innovatively integrate a dual-guidance mechanism that combines background images with difference maps, enhanced by Fast Fourier Transform (FFT) for global feature extraction, while introducing a shadow probability map to suppress interference in non-shadow regions. First, we extract multi-scale features from input images using a shared encoder, then decode them to generate shadow probability maps. Next, we predict illumination residuals by fusing encoded features with FFT-transformed representations. Finally, we reconstruct shadow-free images using the illumination residual model. Experimental result demonstrates that our method not only significantly outperforms state-of-the-art methods but also effectively enhances text-edge sharpness and background uniformity.

***Keywords:** Document image deshadowing, Shadow-illumination residual model, Physics-informed loss, Fast fourier transform.*

1. Introduction

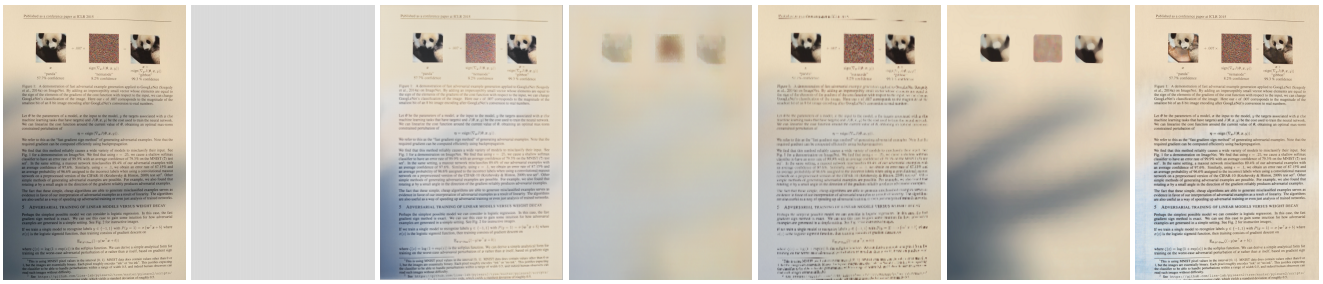
A wide range of everyday documents—including textbooks, newspapers, brochures, and receipts—are commonly digitized for electronic archiving or digital sharing. Given the widespread use and portability of smartphones, mobile devices have become the primary tool for capturing document images. However, when obstructions interfere with lighting during capture [14], the resulting images often contain shadows. These low-brightness regions can significantly degrade image clarity and impair text recognition accuracy. Removing shadows from documents is essential for improving text readability and accessibility, particularly in applications such as document analysis, optical character recognition (OCR), and assistive technologies for the visually impaired.

Although significant progress has been made in natural image shadow removal [43, 10, 42, 4, 41, 7], existing methods often fail to perform well on document images due to their inherent differences. While natural images prioritize shadow-free background content, document images are primarily concerned with textual information. Directly applying natural image processing techniques, including both traditional algorithms and deep learning approaches, typically leads to suboptimal results. Therefore, effective document shadow removal must account for the distinct structural properties of document images, such as text layout, contrast, and edge sharpness.

Shadow removal methods for document images can be primarily categorized into two approaches: traditional optimization methods [14, 30, 32] and deep learning methods [3, 39, 19, 18, 38]. Traditional optimization methods typically formulate optimization algorithms based on physical illumination models [14, 30, 32]. However, due to the overly idealized assumptions of illumination models [27], these methods struggle to accurately characterize complex shadow variations in real-world scenarios, thus exhibiting significant limitations in practical application.

For document image shadow removal, deep learning methods [19, 39, 18, 38] overcome the reliance on physical model assumptions inherent in traditional optimization approaches through end-to-end feature learning. These methods primarily utilize document background color as guidance for shadow removal [19, 39, 38], which can lead to artifact, text blurring, and distortion of background texture and text, as illustrated in Figure 1. The reason lies in the inaccurate shadow-free background estimation by BEDSR-Net [19] and BGShadowNet [39], which adversely affects

*Corresponding author: zhao6685@yeah.net



(a) (b) (c) (d) (e) (f) (g)

Figure 1 The issues with background-guided document shadow removal methods: (a) input image; (b) and (c) respectively represent the global background estimated by BEDSR-Net [19] and its shadow-removed image; (d) and (e) respectively represent the local background estimated by BGShadowNet [39] and its shadow-removed image; (f) and (g) respectively represent the local background estimated by DocRes [38] and its shadow-removed image.

the final shadow removal results. In DocRes [38], shadowed regions are mistakenly interpreted by the network as shadow-free document background color.

Background images can provide rich visual information for neural networks, but inaccurate background information may also interfere with the network's image understanding. Compared to existing methods [19, 39], DocRes [38] generates background images by filtering shadow images, which can more accurately reflect background information and shadow characteristics, thereby reducing the network's estimation error. However, color shadows in the background may be misinterpreted by the network as background color, leading to incorrect inference results. Therefore, background images are a double-edged sword, and we do not rely entirely on them as guidance information.

We found that the difference map between shadow images and background images in DocRes [38] can effectively eliminate shadow information while well preserving textual content. Based on this finding, we combine both background images and difference maps as guiding information for the shadow removal network, which enables precise localization of shadow regions while completely preserving document information.

Previous studies [5, 16] typically modeled the relationship between shadowed and non-shadowed images as a linear transformation. However, these approaches assume globally uniform illumination in shadowed regions and consequently fail to adapt to scenarios involving non-uniform shadows or multi-source lighting interference. To address this limitation, Self-ShadowGAN [11] proposes a pixel-adaptive shadow relighting model that employs a deep network to predict relighting coefficients.

Directly predicting illumination adjustment parameters constitutes a severely ill-posed problem, as infinitely many parameter combinations could produce similar correction effects for the same shadow region, preventing the optimization process from converging to physically plausible solutions. More critically, such global operations tend to corrupt high-frequency structural details like text strokes or fine textures, resulting in blurred reconstructions or artifacts. Furthermore, the network must simultaneously address two coupled objectives—shadow region localization and illumination parameter estimation—where any misjudgment in non-shadow areas becomes amplified through multiplicative operations, causing global illumination distortion. The wavelength-dependent attenuation characteristics of shadows also make direct parameter regression fail to maintain physical consistency across color channels. When the objective function contains flat regions, the training process easily gets trapped in local optima.

We propose a pixel-adaptive shadow-illumination residual model that transforms shadow removal into a local illumination correction task to overcome the limitations of existing approaches. This method innovatively incorporates a shadow probability guidance mechanism, which accurately identifies shadow regions while effectively suppressing erroneous adjustments in non-shadow areas. Through residual learning, the model stably learns fine-grained illumination adjustments, avoiding the optimization challenges inherent in direct prediction. By integrating edge-aware constraints, the approach achieves natural shadow boundary transitions while preserving high-frequency image details.

The framework of our method is illustrated in Fig. 3, which consists of two modules: shadow mask probability map estimation and illumination residual coefficient estimation, all trained in an end-to-end manner. Given a single shadowed document image, we first employ filtering operators to derive a background image and a difference map between the shadowed image and background. These three images (original, background, and difference map) are processed through a weight-shared encoder to extract multi-scale features, followed by a specially designed decoder with skip connections to generate the shadow mask probability map. For illumination residual prediction, we simi-

larly adopt an encoder-decoder architecture to estimate the illumination residual coefficients. Specifically, the shadow probability map undergoes feature encoding while both the shadowed image and difference map are transformed via Fast Fourier Transform (FFT). The encoded features and FFT-processed features are then fused to create enhanced representations, which are fed into the decoder to output the illumination residual coefficients. Finally, we apply our proposed illumination residual model to reconstruct the shadow-free document image.

The core purpose of the shadow probability map as an intermediate output is to explicitly localize shadow regions to guide the network in focusing on shadow processing, while avoiding excessive manipulation of non-shadow areas to preserve texture details.

In conclusion, the main contributions of this work can be outlined as follows:

- We propose a pixel-adaptive shadow-illumination residual model that transforms shadow removal into a local illumination correction task.
- We integrate background images and difference maps (the pixel-wise difference between shadow images and background images) as dual guidance information, replacing traditional approaches that rely solely on background images.
- We design an end-to-end dual-module network that integrates shadow probability map estimation and illumination residual prediction, while incorporating Fast Fourier Transform (FFT) to enhance feature representation.

2. Related Work

2.1. Natural image shadow removal

Shadow removal plays a pivotal role in enhancing visual quality and ensuring the reliability of downstream tasks, with critical applications in document digitization, medical imaging, and autonomous driving. As a key research area in computer vision, shadow removal techniques have advanced significantly—from traditional optimization-based methods [1, 41, 6] to data-driven deep learning approaches [37, 22, 26, 42, 23]. These advancements have led to robust solutions for real-world challenges.

Traditional methods typically leverage prior knowledge to formulate physical models for restoring illumination in shadowed regions [34, 9, 6]. However, due to variations in lighting conditions, these approaches often produce noticeable shadow boundary artifacts. Deep learning-based methods learn high-dimensional feature mappings from shadowed to shadow-free images by constructing large-scale annotated shadow datasets. Qu *et al*. [26] constructed a shadow removal dataset (SRD) and proposed an end-to-end network to recover a shadow-free image from a single shadow image. Wang *et al*. [31] proposed a conditional generative adversarial network to jointly learn shadow detection and shadow removal. ShadowDiffusion [8] introduces a dynamic mask-aware diffusion model to jointly pursue a shadow-free image and refined shadow mask, based shadow degradation as a prior. Diff-Shadow [24] proposes a globally-guided diffusion architecture, which employs a parallel UNet design and leverages the correlation between local branches and global non-shadow regions, achieving shadow removal. Omnisr [35] designs a shadow removal network integrating semantic-geometric priors via concatenation and attention mechanisms.

Although existing methods perform well in shadow removal for natural images, they often exhibit significant performance degradation in document scenarios due to fundamental differences in texture structure (e.g., regular text/tables vs. complex natural scenes) and shadow properties (hard-edged cast shadows vs. soft gradient shadows) between document and natural images. Unlike natural images, the core requirement of document shadow removal lies in accurately restoring occluded text/pattern information—any residual shadows or over-processing may lead to text illegibility or structural information loss.

2.2. Shadow removal for document image

Document shadow removal has consistently attracted significant attention and research [15, 40, 25, 36]. Water-Filling [14] proposes a digital document shadow removal method based on watershed transformation and diffusion equations, which reconstructs uniformly illuminated document images by simulating a water immersion process. Wang *et al*. [30] and Liu *et al*. [20] estimated the global background color to guide the shadow removal. Bako *et al*. [2] generated a shadow map by matching local background colors and used it to correct the image, producing the final shadow-free output.

Recent advances in deep learning have spurred extensive research on document image shadow removal using deep learning-based approaches [18]. BEDSR-Net [19] introduces a background estimation network to estimate the global background color of a document, which is then used to guide the shadow removal process. BGShadowNet [39] proposes a contextual background extraction network to generate spatially varying background maps of document colors, and further designs a background-guided shadow removal network for shadow elimination. Unlike methods [19, 39], DocRes [38] first employs dilation operations to remove textual content in documents, then applies median filtering to smooth artifacts resulting from incomplete text elimination, thereby generating document backgrounds as prior information for shadow removal. Li *et al*. [18] presented a large-scale real-world dataset comprising high-resolution shadow/shadow-free image pairs captured

(a) Input (b) BEDSR-Net [19] (c) BGShadowNet [39] (d) DocShadow [20] (e) DocRes [38]

Figure 2 The current challenges in existing document shadow elimination methods.

under diverse illumination conditions, and introduced a frequency-aware network architecture specifically designed for high-resolution document shadow removal. Liu *et al.* [21] proposed a contrast-based document shadow removal method that locates shadow regions by extracting contrast features from document images. However, these methods often suffer from incomplete shadow removal, altered document background colors, and blurred text (Fig. 2).

3. Methods

We present the method details of our methodology in this section, with the corresponding method architecture shown in Fig. 3. Our framework introduces a novel approach to document shadow removal by modeling it as a local illumination residual prediction task. The core innovation lies in our pixel-adaptive shadow-illumination residual model, which effectively separates illumination effects from document content. We develop a dual-guidance mechanism that strategically combines background image priors with learned difference maps for precise shadow localization. To capture both local and global features, we incorporate Fast Fourier Transform (FFT) operators within our network architecture. The optimization process employs a carefully designed composite loss function that enforces photometric consistency while preserving important edge and textural details.

The residual learning approach accurately decouples shadows from background content rather than directly predicts clean images, thereby avoiding excessive modification of background information. Residual estimation effectively separates shadow transparency variations (e.g., brightness attenuation coefficients in shadowed regions) while preserving the original structure of paper-based text strokes. Through cross-channel collaborative optimization, it prevents color distortion by constraining RGB channel residual correlations. The residual estimation automatically learns spatial distribution characteristics of shadows (e.g., gradient shadows, complex boundaries), maintaining high-frequency background details (such as text edges) while smoothly processing shadow regions.

3.1. Shadow-illumination residual model

Previous studies [28, 16] established a fundamental shadow removal paradigm in which a shadow-affected image I^s and its corresponding binary mask can be transformed into a shadow-free image I^f through a linear illumination adjustment model:

$$I_p^f = w \cdot I_p^s + b, \tag{1}$$

where p denotes pixel coordinates within shadow regions, while w and b represent global illumination correction parameters that remain constant across all shadowed pixels.

However, this linear model with constant relighting coefficients w and b can only handle uniformly illuminated shadows and performs poorly when processing non-uniform shadows [29]. To overcome this limitation, Self-ShadowGAN [11] proposes a pixel-adaptive shadow relighting model:

$$I_p^f = W_p \cdot I_p^s + B_p, \tag{2}$$

where W_p and B_p are spatially varying relighting coefficients for each shadow pixel p, designed to accommodate complex shadows with non-uniform illumination. To preserve shadow-free regions during the relighting process, the values of W and B for original shadow-free pixels are fixed at 1 and 0, respectively.

Accurate shadow region detection remains a challenging problem, as imperfect detection significantly impacts shadow removal performance. The conventional approach of fixing non-shadow region coefficients to $W = 1$ and $B = 0$ proves theoretically unsound and practically problematic

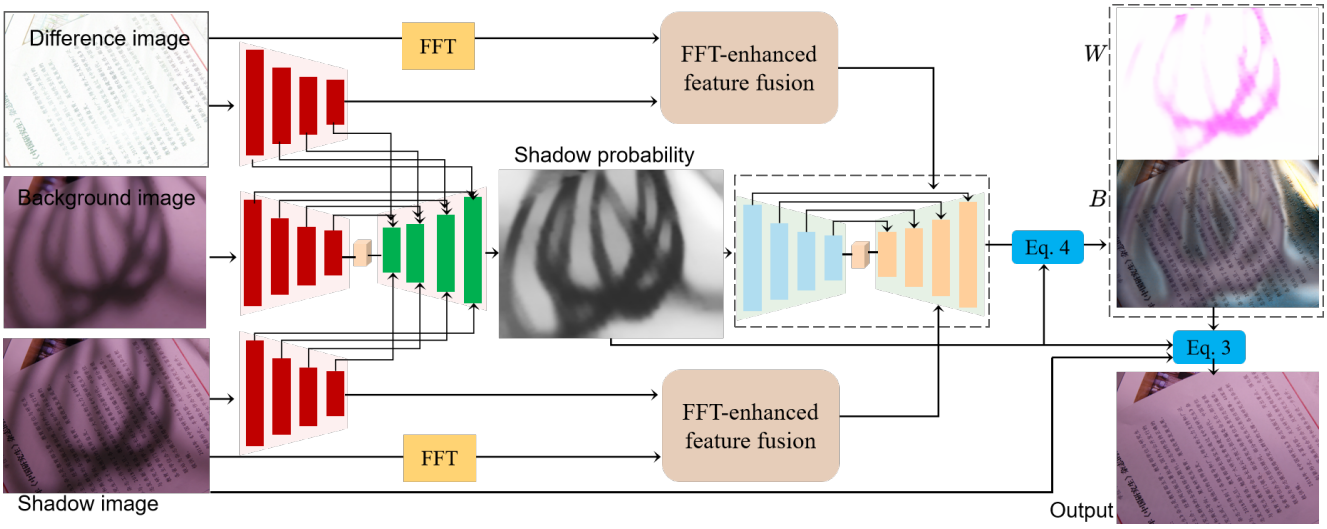


Figure 3 Overview of the proposed method.Our approach integrates shadow probability estimation and illumination residual prediction within an end-to-end framework. First, we extract multi-scale features from the input images using a shared encoder, then decode them to generate shadow probability maps, which serve as guidance for the illumination residual estimation network. Next, we predict illumination residuals by fusing the encoded features, shadow probability map features, and FFT-transformed representations. Finally, we reconstruct shadow-free images using our illumination residual model.

for three key reasons: First, real-world shadow boundaries often exhibit gradual transitions rather than sharp edges, making binary classification inherently inaccurate. Second, illumination conditions in non-shadow areas may still require subtle adjustment to achieve photometric consistency with processed shadow regions. Third, detection errors inevitably propagate through the relighting process, causing either incomplete shadow removal or over-correction of non-shadow areas.

To address these limitations, we propose a unified framework that jointly optimizes shadow detection and illumination correction. Rather than enforcing rigid constraints, we formulate the problem as:

$$I_p^f = (1 - M_p) \cdot I_p^s + M_p \cdot (W_p \cdot I_p^s + B_p), \quad (3)$$

where $M_p \in [0, 1]$ represents a continuous shadow mask probability map instead of a binary mask, as shown in Fig. 4. This soft formulation provides three advantages: (1) it naturally handles shadow penumbra regions through probabilistic weighting, (2) allows partial correction for potential false negatives in detection, and (3) maintains differentiability for end-to-end optimization. The coefficients W_p and B_p are now defined over all image pixels, with their magnitudes automatically regulated through the following mechanisms:

$$W_p = 1 + M_p \cdot \Delta W_p, \quad B_p = M_p \cdot \Delta B_p, \quad (4)$$

where ΔW_p and ΔB_p are learned residual corrections. This formulation ensures that non-shadow areas ($M_p \rightarrow 0$) maintain near-identity transformation while shadow regions ($M_p \rightarrow 1$) receive appropriate illumination adjustment.

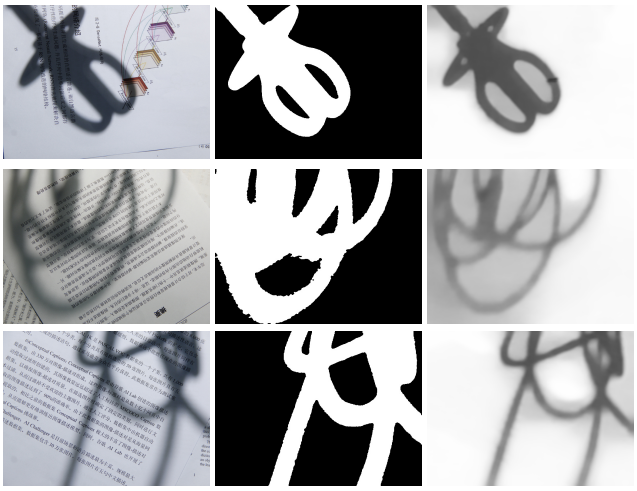
Directly predicting illumination adjustment parameters faces a severely ill-posed problem, as there could be countless parameter combinations for the same shadow region that yield similar correction effects. This makes the optimization process difficult to converge to a reasonable solution. More critically, such global operations tend to destroy high-frequency structural details in the image, such as text strokes or fine textures, resulting in blurred reconstructions or artifacts.

Additionally, the network must simultaneously accomplish two coupled tasks: shadow region localization and illumination parameter estimation. Any misjudgment in non-shadow regions will be amplified through multiplicative operations, leading to overall brightness imbalance. Since the attenuation characteristics of shadows vary across different color channels, direct parameter regression also struggles to meet physical consistency requirements.

When the objective function contains flat regions, the training process is prone to getting trapped in local optima. These factors collectively limit the performance of direct prediction methods, necessitating the introduction of shadow probability guidance and the design of an illumination residual model.

The brightness adjustment ΔB_p is bounded using an activation function:

$$\Delta B_p = \beta \cdot \tanh(\Delta \tilde{B}_p), \quad (5)$$



(a) (b) (c)

Figure 4 Comparison of (b) Binary shadow mask and (c) Probability map (derived from (a) the input shadow image). In our probability map, darker regions correspond to higher shadow likelihood values.

where $\beta = 0.5$, $\Delta\tilde{B}_p$ represents the unbounded raw prediction of brightness adjustment at pixel p before constraint application. This constrains the brightness modification to the physically plausible range $[-0.5, 0.5]$.

The shadow attenuation factor ΔW_p is governed by a composite constraint that integrates physical bounds with shadow-depth adaptation:

$$\Delta W_p = 0.8 \cdot \tanh(\Delta\tilde{W}_p) \tag{6}$$

3.2. Network detail

We adopt the shadow image background estimation method proposed by DocRes [38] to generate the background image of the shadow image. Initially, dilation operations are employed to eliminate the textual content within the document. Subsequently, a median filter is applied to smooth out artifacts introduced due to incomplete removal.

The background image inherently contains both shadows and document background colors. Directly using it as guidance information may cause the network to misinterpret shadow colors as part of the document background, leading to incomplete shadow removal or color distortion in non-shadow regions, as illustrated in Fig. 1.

As illustrated in Fig. 5, we observe that the difference map between the shadow image and background image can effectively eliminate most shadow information while preserving the majority of textual content. By incorporating both the difference map and background image as guidance inputs to the network, the system can accurately identify shadow regions and extract text features from shadowed areas, thereby achieving high-precision shadow removal with text fidelity preservation.

(a) Shadow image (b) Background image (c) Difference map

Figure 5 Visualization shadow image, background image, and their difference map.

As shown in Fig. 3, we employ a U-Net encoder-decoder network architecture to estimate residual correction coefficients. During the encoding stage, a ResNet-18 backbone is employed for image feature extraction. Considering that image downsampling during the encoding stage leads to partial information loss, we introduce Fast Fourier Transform (FFT) of both the shadow image and the difference image in the decoding process of shadow removal to supplement global features. By leveraging the global representation capability of Fourier transform, our method can simultaneously model local texture details (e.g., character strokes) and global structural features (e.g., page layout) in document images. Specifically, the magnitude spectrum primarily captures high-frequency text edge features, while the phase spectrum preserves the geometric topology of the document. This dual-spectrum collaboration mechanism effectively mitigates content distortion during shadow removal and significantly enhances the preservation of text region integrity.

To address this, we propose a Fast Fourier Transform-enhanced feature fusion model (FFT-enhanced feature fusion) . As shown in Fig. 6, the detailed implementation pipeline is as follows: First, we perform Fast Fourier Transform (FFT) on the shadow image, then downsample the transformed results to different scale levels. After extracting features through multi-layer convolutional networks, we concatenate (concat) these features with encoder features for fusion. Subsequently, we employ a Convolutional Block Attention Module (CBAM) [33] to simultaneously extract channel attention and spatial attention features. Finally, the processed features are connected to the decoder. For feature processing of the difference image, we follow the same pipeline approach.

We selected the Fourier Transform (FT) because it effectively separates shadow and text components in the

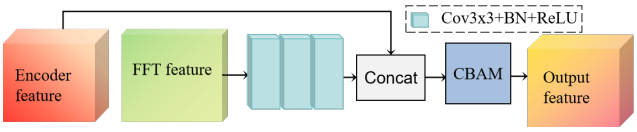


Figure 6 FFT-enhanced feature fusion model.

frequency domain, in which shadows primarily affect low-frequency regions while text features dominate high-frequency bands. This frequency-based separation enables precise processing of shadow regions while preserving text details.

3.3. Loss function

During training, we optimize the network model by imposing constraints on: (1) the shadow mask probability M, (2) the learned residual correction coefficients ΔW and ΔB, (3) the relighting coefficients W and B, and (4) the shadow-removed image. The specific constraints are as follows.

Appearance consistency loss. We measure the pixel-wise L1 distance between predicted shadow-free image I_{free} and the ground truth image I_{gt} to maintain visual fidelity:

$$\mathcal{L}_{\text{a}} = \|I_{\text{gt}} - I_{\text{free}}\|_1, \tag{7}$$

Adversarial loss. We employ a relativistic adversarial loss [12] that evaluates not only the absolute discriminator scores for real and generated images but also their relative discrepancies:

$$\begin{aligned}\mathcal{L}_{\text{adv}} = \text{BCE}\left(\sigma\left(D(I_{\text{gt}}) - D(I_{\text{free}})\right), g\right) \\ +\text{BCE}\left(\sigma\left(I_{\text{gt}}\right) - D(I_{\text{free}})\right), d)\end{aligned} \tag{8}$$

where σ denotes the sigmoid activation function, $\text{BCE}(\cdot)$ represents binary cross-entropy, and g/d are training targets ($g = 1$, $d = 0$ for generator updates; $g = 0$, $d = 1$ for discriminator training).

Gradient smoothness loss. To ensure smooth spatial transitions in the shadow probability mask M, W, B, we impose a gradient smoothness constraint:

$$\mathcal{L}_{\text{g}} = \|\nabla M\|_2^2 + \|\nabla W\|_2^2 + \|\nabla B\|_2^2, \tag{9}$$

where ∇ denotes the spatial gradient operator.

Non-shadow identity preservation loss. To maintain color fidelity in non-shadow regions, we enforce the identity constraint when the shadow probability M_p approaches zero ($M_p < \epsilon$):

$$\mathcal{L}_{\text{i}} = \sum_{M_p < \epsilon} \left(\|W_p - 1\|_2^2 + \|B_p\|_2^2\right) \tag{10}$$

where $\epsilon = 0.1$ is the non-shadow probability threshold.

Shadow consistency loss. To prevent over-enhancement of shadow regions beyond non-shadow illumination levels, we introduce a quadratic penalty term:

$$\mathcal{L}_{\text{s}} = \sum_{p} \max(0, W_p - 1)^2 \cdot M_p \tag{11}$$

Regional similarity constraint loss. To enforce spatially coherent shadow removal in homogeneous regions while preserving natural boundaries, we propose a novel color-guided similarity constraint on brightness adjustment fields:

$$\mathcal{L}_{\text{r}} = \sum_{p,q \in \mathcal{N}_p} \|\Delta B_p - \Delta B_q\| \cdot \mathbb{I}(\|I_p - I_q\| < \tau) \tag{12}$$

where $\mathcal{N}_p$ is the $k \times k$ neighborhood ($k = 3$ typically); τ is a chromatic similarity threshold (10/255 in RGB space); $\mathbb{I}(\cdot)$ is the indicator function:

$$\mathbb{I}(x) = \begin{cases} 1 & \text{if } x < \tau \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

This loss operates on the principle that neighboring pixels with similar colors should undergo analogous illumination corrections.

Cross-channel consistency loss. To enforce consistent adjustments across RGB channels and avoid color distortion, we apply a covariance-based penalty:

$$\mathcal{L}_{\text{c}} = \sum_{c_1 \neq c_2} \text{Cov}(\Delta W^{c_1}, \Delta W^{c_2}), \tag{14}$$

where $c_1, c_2 \in \{\text{R}, \text{G}, \text{B}\}$ and $c_1 \neq c_2$ represent all possible inter-channel relationships.

The overall objective function is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{a}} + \mathcal{L}_{\text{adv}} + t_1\mathcal{L}_{\text{g}} + t_2\mathcal{L}_{\text{i}} + t_3\mathcal{L}_{\text{s}} + t_4\mathcal{L}_{\text{r}} + t_5\mathcal{L}_{\text{c}}, \tag{15}$$

In our experiments, the weighting parameters are set as: $\lambda_1 = 0.5$, $\lambda_2 = \lambda_3 = 0.2$, and $\lambda_4 = \lambda_5 = 0.1$. Among all the loss terms, the appearance consistency loss, adversarial loss, and gradient smoothness loss play the most crucial roles, which is why we assign them larger weights. Extensive experimental validation confirms that this particular weight configuration yields optimal performance.

4. Experiments

Implementation detail. Our neural network was implemented using the PyTorch framework and trained over 200 epochs on an NVIDIA GeForce RTX 4080 GPU. We employed the Adam optimizer with the following configurations: momentum parameters (β_1, β_2) were set to $(0.5, 0.999)$ for both generator and discriminator, and the

Table 1 Quantitative comparison between our method and state-of-the-art approaches on RDD [39] and SD7K [18] datasets. In each set of test data, highlight the best result in **bold** and the second-best result with an underline.

Method	Venue/Year	RDD [39]				SD7K [18]			
		PSNR ↑	SSIM ↑	RMSE ↓	OCR ↑	PSNR ↑	SSIM ↑	RMSE ↓	OCR ↑
Water-Filling [14]	ACCV/2018	19.580	0.876	38.015	70.214	18.406	0.871	37.826	68.303
Liu *et al*. [20]	ICASSP/2023	21.631	0.882	29.746	72.356	19.056	0.885	31.504	71.314
BEDSR-Net [19]	CVPR/2020	22.915	0.876	25.140	71.301	20.905	0.875	27.208	72.148
BGShadowNet [39]	CVPR/2023	26.114	0.920	10.102	68.450	25.880	0.941	10.047	68.366
DocShadow [18]	ICCV/2023	24.630	0.910	10.052	73.156	26.015	0.946	9.691	71.508
DocRes[38]	CVPR/2024	27.448	0.931	8.041	79.045	26.610	0.953	8.932	**79.368**
Liu *et al*. [21]	2025	25.028	0.910	10.964	74.512	25.901	0.948	9.045	72.904
ShadowDiffusion [8]	CVPR/2023	25.047	0.926	9.905	62.567	25.803	0.940	10.460	64.025
Omnisr [35]	AAAI/2025	23.812	0.914	12.006	68.401	23.015	0.926	13.702	65.360
Des3 [13]	CVPR/2024	23.006	0.911	16.805	71.259	22.002	0.927	16.205	72.304
Diff-Shadow [24]	AAAI/2024	23.704	0.909	18.532	64.521	23.104	0.925	17.206	62.356
Le [17]	TPAMIR/2021	22.490	0.904	21.710	73.802	21.812	0.915	22.380	70.956
Self-ShadowGAN [11]	IJCV/2023	20.620	0.877	26.507	75.436	21.046	0.894	28.004	71.248
Ours	CVM/2026	**29.815**	**0.950**	**6.824**	**83.241**	**28.570**	**0.961**	**6.910**	78.950

initial learning rate was 0.0001. The input images were resized to a resolution of 512×512 pixels, and we used a batch size of 3 to balance memory efficiency with stable gradient updates. To improve model robustness and mitigate overfitting, we implemented data augmentation strategies including random rotation and cropping.

4.1. Datasets and evaluation metrics

Datasets. We adopted the SD7K dataset proposed by Li *et al*. [18] and the RDD dataset introduced by Zhang *et al*. [39] as the training and testing benchmarks for our method and comparative approaches. These two datasets are currently publicly available large-scale document shadow removal datasets, both containing paired high-resolution document images. Specifically, the SD7K dataset consists of 6,479 training samples and 760 test samples, while the RDD dataset comprises 4,371 training samples and 545 test samples. In our experiments, we randomly selected 4,000 samples from each training set for model training and evaluated the model performance on both test sets.

Evaluation metrics. In all experiments, we employed three evaluation metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Root Mean Square Error (RMSE), and word accuracy (OCR). Specifically, higher PSNR SSIM, and OCR values indicate better shadow removal performance, while a lower RMSE value suggests smaller deviations between the model output and the ground truth.

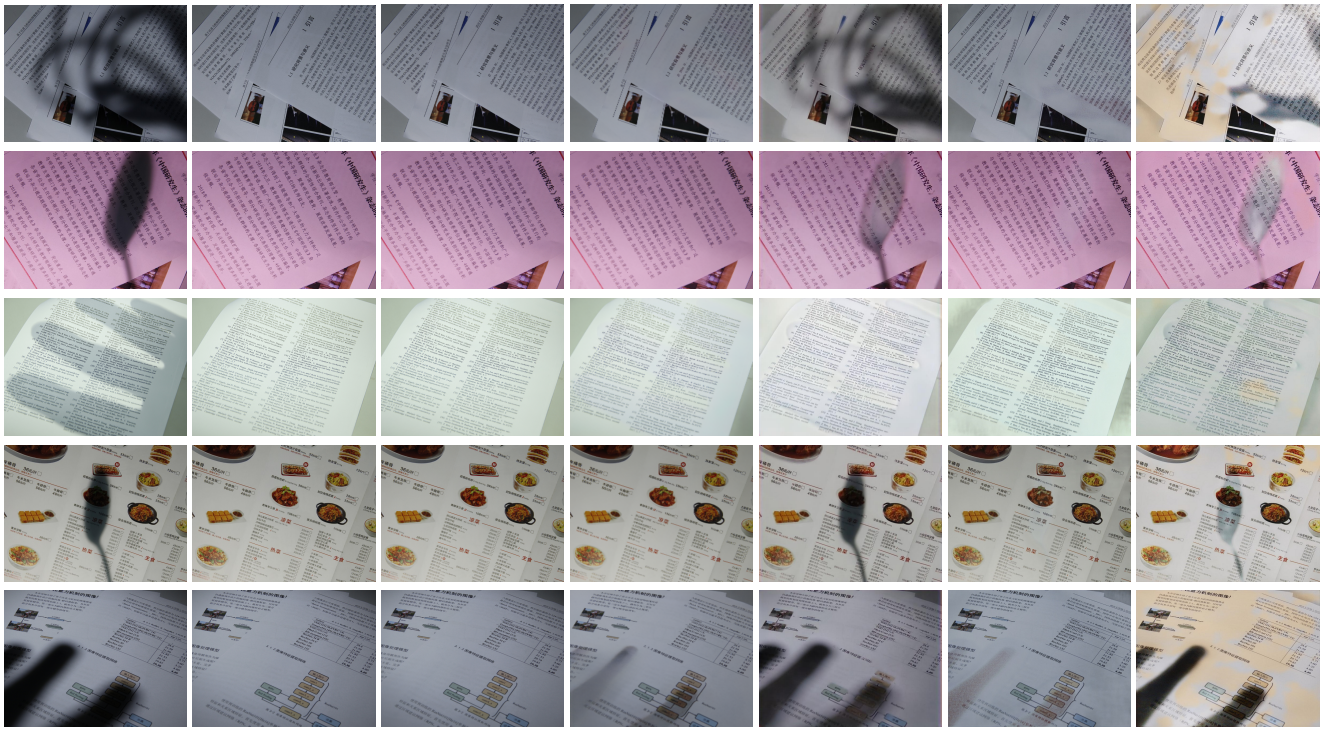
4.2. Comparison with state-of-the-art methods

To validate the effectiveness of our proposed method, we conducted comprehensive comparisons with state-of-the-art approaches spanning two domains: (1) seven document image shadow removal methods (Water-Filling [14], Liu *et al*. [20], BEDSR-Net [19], BGShadowNet [39], DocShadow [18], DocRes [38], and Liu *et al*. [21]) and (2) six natural image shadow removal techniques (ShadowDiffusion [8], Omnisr [35], Des3 [13], Diff-Shadow [24], Le *et al*. [17], and Self-ShadowGAN [11]).

To ensure a fair comparison, all learning-based methods were trained under identical hardware configurations using the RDD and SDK datasets. Evaluation metrics were uniformly computed on 512×512 resolution outputs. As demonstrated in Table 1, our method achieves superior performance across all evaluation metrics, establishing its significant advantage over existing approaches.

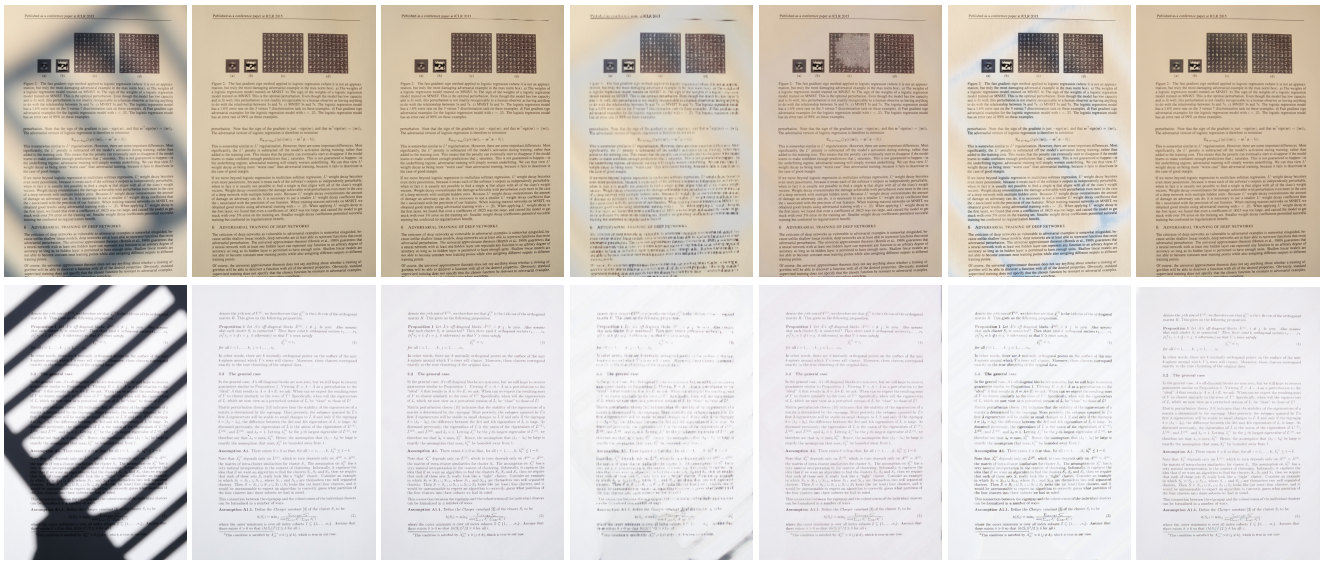
As shown in Figures 7 and 8, we present a visual comparison of our method with existing document image shadow removal approaches on datasets RDD and SD7K. The experimental results demonstrate that our method can completely eliminate shadows in document images while effectively preserving text information. In contrast, other methods exhibit the following issues: (1) shadow artifact, (2) altered background colors, and (3) blurred text. Both qualitative and quantitative analyses confirm that our method achieves significant advantages and outperforms existing approaches by a clear margin.

To validate the robustness of our method, we conducted a user evaluation using 500 shadowed document images captured under natural lighting conditions (outdoor sunlight and indoor mixed lighting) with various devices. The evaluation compared our method against state-of-the-art baselines through blind assessment by 200 image-processing students using a 1-5 scoring scale (5 being best). The Per-



(a) (b) (c) (d) (e) (f) (g)

Figure 7 Visual comparison among state-of-the-art shadow removal methods on testing images from the RDD dataset: (a) Input images, (b) Ground truth, (c) Ours, (d) BGShadowNet [39], (e) DocShadow [20], (f) DocRes [38], and (g) Liu [21].



(a) (b) (c) (d) (e) (f) (g)

Figure 8 Visual comparison among state-of-the-art shadow removal methods on testing images from the SD7K dataset: (a) Input images, (b) Ground truth, (c) Ours, (d) BGShadowNet [39], (e) DocShadow [20], (f) DocRes [38], and (g) Liu [21].

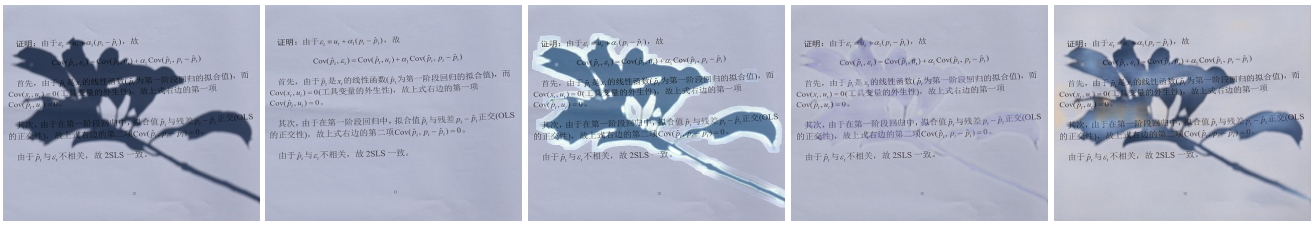
centage of Votes (PoV) metric was computed as:

$$\text{PoV} = \frac{1}{N \cdot M} \sum_{i=1}^{N} \sum_{j=1}^{M} V_{ij} \times 100\% \tag{16}$$

where $N = 500$ is the test set size, $M = 200$ is the number of evaluators, and V_{ij} denotes the score from the j-th evaluator for the i-th result.

Table 2 Comparative performance evaluation of shadow removal methods (Percentage of Votes).

Method	[14]	[20]	[19]	[39]	[18]	[38]	[21]	[8]	[35]	[13]	[24]	[17]	[11]	Ours
Naturalness	2.31	3.02	3.45	8.17	8.63	10.87	7.92	6.85	4.11	4.59	4.08	3.25	2.37	30.38
Readability	2.12	2.01	3.41	8.24	8.70	8.23	7.86	6.88	4.74	4.66	4.01	3.17	2.97	33.00
Artifacts	3.47	2.83	3.72	7.89	8.35	9.68	8.21	6.53	4.39	4.31	4.36	3.52	3.62	29.12
Overall quality	2.16	3.20	3.21	8.35	8.48	11.10	7.85	7.06	4.01	4.72	4.35	3.20	2.22	30.09



(a) Shadow image (b) Ours (c) Self-ShadowGAN [11] (d) DocRes [38] (e) Liu [21]

Figure 9 Visual comparison among state-of-the-art shadow removal methods.

Participants evaluated four criteria: naturalness, readability, absence of artifacts, and overall quality. As shown in Table 2, our method achieved the highest PoV scores across all four metrics. Notably, although it was not explicitly trained on sunlight-affected samples, our method successfully eliminates non-uniform hard shadows while preserving text edge sharpness and avoiding common artifacts. These results demonstrate its strong adaptability to lighting variations and generalization capability.

In Fig. 9, we present a comparative analysis of document images captured under sunlight using existing shadow removal methods. It is noteworthy that although our training data consists solely of indoor-captured scenes and does not include any outdoor sunlight conditions, our method still effectively removes shadows while completely preserving background document information. These results demonstrate the outstanding robustness of our algorithm.

4.3. Ablation study

Ablation of shadow-illumination residual model. To validate the effectiveness of our proposed shadow-illumination residual model, we designed the following three comparative experiments:

A0: direct RGB prediction baseline.

A1: Employs the pixel-adaptive shadow relighting model [11], formulated as:

$$I_p^f = W_p \cdot I_p^s + B_p, \tag{17}$$

where W_p and B_p are directly learned by the network.

A2: Introduces a shadow mask M_p into the model design:

$$I_p^f = (1 - M_p) \cdot I_p^s + M_p \cdot (W_p \cdot I_p^s + B_p), \tag{18}$$

where the network jointly learns M_p, W_p, and B_p.

A3: Further improves upon Experiment 2 by learning residual corrections ΔW_p and ΔB_p, while constraining the forms of W_p and B_p:

$$W_p = 1 + M_p \cdot \Delta W_p, \quad B_p = M_p \cdot \Delta B_p. \tag{19}$$

This constitutes our shadow-illumination residual model.

All three experiments were trained on the same dataset and evaluated on two test sets. The quantitative comparison results in Table 3 demonstrate that: The model in A2 outperforms the pixel-adaptive shadow relighting model (A1); The shadow-illumination residual model (A3) achieves the best performance. Additionally, Fig. 10 presents a set of visual comparisons. The visual results confirm that our proposed pixel-adaptive shadow relighting model delivers optimal shadow removal quality. The shadow removal results of Self-ShadowGAN [11] are unsatisfactory, particularly in handling hard shadows, which require multiple iterative optimizations and cannot achieve real-time processing. Both quantitative metrics and visual evaluations consistently validate the effectiveness of our model.

Table 3 Ablation of shadow illumination residual model.

	RDD [39]			SD7K [18]		
	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE
A0	26.412	0.923	7.723	26.926	0.944	8.013
A1	28.007	0.939	7.512	27.304	0.948	7.952
A2	29.474	0.947	7.102	28.401	0.956	7.261
A3	**29.815**	**0.950**	**6.824**	**28.570**	**0.961**	**6.910**

Ablation of dual guidance fusion strategy. To validate the effectiveness of our proposed joint guidance framework using both background and residual images, we con-

Table 4 Ablation study on dual guidance fusion strategy.

	RDD [39]			SD7K [18]		
	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE
B1	28.991	0.934	7.300	27.704	0.933	7.498
B2	29.610	0.942	7.090	28.312	0.949	7.240
B3	29.540	0.940	6.992	28.376	0.951	7.103
B4	**29.815**	**0.950**	**6.824**	**28.570**	**0.961**	**6.910**
C1	27.448	0.931	8.041	26.610	0.953	8.932
C2	**27.702**	**0.939**	**7.920**	**26.801**	**0.958**	**8.926**

Table 5 Ablation study on FFT-enhanced feature fusion.

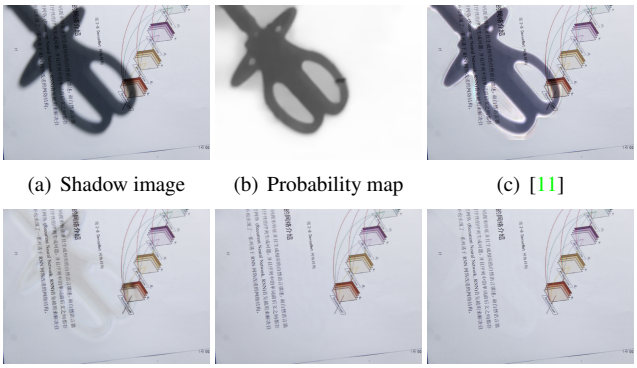
	RDD [39]			SD7K [18]		
	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE
D1	29.740	0.942	6.907	28.425	0.948	7.114
D2	**29.815**	**0.950**	**6.824**	**28.570**	**0.961**	**6.910**

ducted ablation experiments comparing our method with DocRes[38].

In our method, we established four experimental configurations to systematically evaluate model performance:
B1: Without difference map and background guidance;
B2: Background guidance only;
B3: Difference map guidance only;
B4: Combined difference map and background guidance.
All models were retrained on identical datasets and evaluated on two benchmark datasets (RDD and SD7K). As demonstrated in Table 4, both background and difference map guidance contribute to significant performance improvements, with optimal results achieved when both modalities jointly guide the shadow removal process. Visual comparisons in Fig. 11 further substantiate that our dual-guidance approach not only effectively eliminates shadows but also better preserves fine-grained texture details and textual integrity.

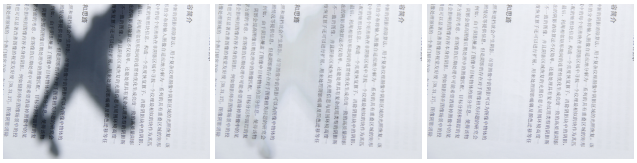
We conducted extended experiments on the DocRes [38] with two variants:
C1: Original method;
C2: Combined difference map and background guidance.
Experimental results (As shown in Table 4) confirm that incorporating difference map guidance yields statistically significant performance gains in DocRes. Collectively, these experiments provide comprehensive evidence for the efficacy and general applicability of our proposed dual-guidance strategy combining background and difference map information.

Ablation of FFT-enhanced feature fusion. To validate the effectiveness of our proposed FFT-enhanced feature fusion, we designed two experimental configurations for systematic performance evaluation:



(a) Shadow image (b) Probability map (c) [11]
(d) Experiment 1 (e) Experiment 2 (f) Experiment 3

Figure 10 Ablation of shadow-illumination residual model.



(a) Shadow image (b) w/o difference map (c) Ours

Figure 11 Ablation study on dual guidance fusion strategy.

D1: Without the FFT-enhanced feature fusion module;
D2: Model with the FFT-enhanced feature fusion module.
All models were retrained on the same dataset and evaluated on two benchmark datasets (RDD and SD7K). As shown in Table 5, the experimental results demonstrate that the FFT-enhanced feature fusion module significantly improves model performance.

Limitation. Although our method demonstrates superior performance compared to state-of-the-art approaches, subtle artifacts may occasionally appear in the shadow rendering of glossy materials, as illustrated in Fig. 12. We identify this as a key limitation and plan to investigate more accurate solutions in future work, potentially by incorporating ray-traced shadows or neural rendering techniques.

(a) Input (b) Ours

Figure 12 Limitation of our method.

5. Conclusion

In this paper, we proposed a novel shadow removal framework guided by the joint utilization of background images and difference maps, and further innovatively presented the Shadow-Illumination Residual Model. The model collaboratively optimizes the shadow removal process through a dual-branch architecture: the illumination correction explicitly decouples shadow-illumination effects, while the residual learning adaptively separates content residuals, thereby effectively preserving fine-grained textures and text details while eliminating shadows. Systematic ablation experiments demonstrated that our proposed joint guidance strategy significantly improves shadow removal performance. Quantitative and qualitative evaluations on multiple test datasets indicated that our method outperforms existing techniques in handling complex shadow boundaries, maintaining background consistency, and adapting to texture-sensitive scenarios. We will incorporate modern backbones (e.g., ConvNeXt, ViT) in future experiments to explore their potential for performance enhancement, and report comparative results in subsequent publications.

Acknowledgement

This work is partly supported by the National Natural Science Foundation of China (No.62366056); the Caiyun Postdoctoral Innovation Program of Yunnan Province; Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University; Natural Science Foundation of Yunnan Province (No.202401AU070120).

References

[1] E. Arbel and H. Hel-Or. Shadow removal using intensity surfaces and texture anchor points. *IEEE transactions on pattern analysis and machine intelligence*, 33(6):1202–1216, 2010. 3

[2] S. Bako, S. Darabi, E. Shechtman, J. Wang, K. Sunkavalli, and P. Sen. Removing shadows from images of documents. In *Asian Conference on Computer Vision*, pages 173–183, 2016. 3

[3] X. Chen, X. Cun, C.-M. Pun, and S. Wang. Shadocnet: Learning spatial-aware tokens in transformer for document shadow removal. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2023. 1

[4] Z. Chen, C. Long, L. Zhang, and C. Xiao. Canet: A context-aware network for shadow removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4743–4752, 2021. 1

[5] Y.-Y. Chuang, D. B. Goldman, B. Curless, D. H. Salesin, and R. Szeliski. Shadow matting and compositing. In *ACM SIGGRAPH 2003 Papers*, pages 494–500. 2003. 2

[6] G. D. Finlayson, M. S. Drew, and C. Lu. Entropy minimization for shadow removal. *International Journal of Computer Vision*, 85(1):35–57, 2009. 3

[7] L. Guo, S. Huang, D. Liu, H. Cheng, and B. Wen. Shadowformer: Global context helps shadow removal. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 710–718, 2023. 1

[8] L. Guo, C. Wang, W. Yang, S. Huang, Y. Wang, H. Pfister, and B. Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14049–14058, 2023. 3, 8, 10

[9] R. Guo, Q. Dai, and D. Hoiem. Single-image shadow detection and removal using paired regions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2033–2040, 2011. 3

[10] Y. Huang, X. Lu, Y. Quan, Y. Xu, and H. Ji. Image shadow removal via multi-scale deep retinex decomposition. *Pattern Recognition*, 159:111126, 2025. 1

[11] H. Jiang, Q. Zhang, Y. Nie, L. Zhu, and W.-S. Zheng. Learning to remove shadows from a single image. *International Journal of Computer Vision*, 131(9):2471–2488, 2023. 2, 4, 8, 10, 11

[12] Z. Jiang, J. Hu, L. Zhang, G. Fu, and C. Xiao. Hierarchical adaptive filtering network for text image specular highlight removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2408–2417, 2025. 7

[13] Y. Jin, W. Ye, W. Yang, Y. Yuan, and R. T. Tan. Des3: Adaptive attention-driven self and soft shadow removal using vit similarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2634–2642, 2024. 8, 10

[14] S. Jung, M. A. Hasan, and C. Kim. Water-filling: An efficient algorithm for digitized document shadow removal. In *Asian Conference on Computer Vision*, pages 398–414, 2018. 1, 3, 8, 10

[15] N. Kligler, S. Katz, and A. Tal. Document enhancement using visibility detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2374–2382, 2018. 3

[16] H. Le and D. Samaras. Shadow removal via shadow image decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8578–8587, 2019. 2, 4

[17] H. Le and D. Samaras. Physics-based shadow image decomposition for shadow removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9088–9101, 2021. 8, 10

[18] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12415–12424, 2023. 1, 3, 8, 10, 11

[19] Y.-H. Lin, W.-C. Chen, and Y.-Y. Chuang. Bedsr-net: A deep shadow removal network from a single document image. In *Proceedings of the IEEE/CVF Conference on Computer Vi-*

sion and Pattern Recognition, pages 12905–12914, 2020. 1, 2, 3, 4, 8, 10

[20] W. Liu, B. Wang, J. Zheng, and W. Wang. Shadow removal of text document images using background estimation and adaptive text enhancement. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2023. 3, 4, 8, 9, 10

[21] Y. Liu, J. Huang, N. Liu, M. Yan, Y. Huang, and S. Chen. Leveraging contrast information for efficient document shadow removal. *arXiv preprint arXiv:2504.00385*, 2025. 4, 8, 9, 10

[22] Z. Liu, H. Yin, Y. Mi, M. Pu, and S. Wang. Shadow removal by a lightness-guided network with training on unpaired data. *IEEE Transactions on Image Processing*, 30:1853–1865, 2021. 3

[23] X. Lu, Y. Zhu, X. Wang, D. Li, J. Xiao, Y. Zhang, X. Fu, and Z.-J. Zha. Hirformer: Dynamic high resolution transformer for large-scale image shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6513–6523, 2024. 3

[24] J. Luo, R. Li, C. Jiang, X. Zhang, M. Han, T. Jiang, H. Fan, and S. Liu. Diff-shadow: Global-guided diffusion model for shadow removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5856–5864, 2025. 3, 8, 10

[25] D. M. Oliveira, R. D. Lins, and G. de França Pereira e Silva. Shading removal of illustrated documents. In *Proc eedings of the International Conference Image Analysis and Recognition*, pages 308–317, 2013. 3

[26] L. Qu, J. Tian, S. He, Y. Tang, and R. W. Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4067–4075, 2017. 3

[27] V. Shah and V. Gandhi. An iterative approach for shadow removal in document images. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1892–1896, 2018. 1

[28] Y. Shor and D. Lischinski. The shadow meets the mask: Pyramid-based shadow removal. In *Computer Graphics Forum*, volume 27, pages 577–586, 2008. 4

[29] F.-A. Vasluianu, A. Romero, L. Van Gool, and R. Timofte. Shadow removal with paired and unpaired learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshop*, pages 826–835, 2021. 4

[30] B. Wang and C. P. Chen. An effective background estimation method for shadows removal of document images. In *IEEE International Conference on Image Processing*, pages 3611–3615, 2019. 1, 3

[31] J. Wang, X. Li, and J. Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1788–1797, 2018. 3

[32] J.-R. Wang and Y.-Y. Chuang. Shadow removal of text document images by estimating local and global background colors. In *EEE International Conference on Acoustics, Speech and Signal Processing*, pages 1534–1538, 2020. 1

[33] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, pages 3–19, 2018. 6

[34] C. Xiao, D. Xiao, L. Zhang, and L. Chen. Efficient shadow removal using subregion matching illumination transfer. In *Computer Graphics Forum*, volume 32, pages 421–430, 2013. 3

[35] J. Xu, Z. Li, Y. Zheng, C. Huang, R. Gu, W. Xu, and G. Xu. Omnisr: Shadow removal under direct and indirect lighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8887–8895, 2025. 3, 8, 10

[36] Q. Yang, K.-H. Tan, and N. Ahuja. Shadow removal using bilateral filtering. *IEEE Transactions on Image Processing*, 21. 3

[37] M. K. Yücel, V. Dimaridou, B. Manganelli, M. Ozay, A. Drosou, and A. Saa-Garriga. Lra&ldra: Rethinking residual predictions for efficient shadow detection and removal. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4925–4935, 2023. 3

[38] J. Zhang, D. Peng, C. Liu, P. Zhang, and L. Jin. Docres: A generalist model toward unifying document image restoration tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15654–15664, 2024. 1, 2, 3, 4, 6, 8, 9, 10, 11

[39] L. Zhang, Y. He, Q. Zhang, Z. Liu, X. Zhang, and C. Xiao. Document image shadow removal guided by color-aware background. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1818–1827, 2023. 1, 2, 3, 4, 8, 9, 10, 11

[40] L. Zhang, A. M. Yip, and C. L. Tan. Removing shading distortions in camera-based document images using inpainting and surface fitting with radial basis functions. In *Proceedings of the International Conference on Document Analysis and Recognition*, volume 2, pages 984–988, 2007. 3

[41] L. Zhang, Q. Zhang, and C. Xiao. Shadow remover: Image shadow removal based on illumination recovering optimization. *IEEE Transactions on Image Processing*, 24(11):4623–4636, 2015. 1, 3

[42] R. Zheng, Q. Zhang, Y. Nie, and W.-S. Zheng. When shadow removal meets intrinsic image decomposition: A joint learning framework using unpaired data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10599–10607, 2025. 1, 3

[43] C. Zhou, C. Xu, and B. Shi. Polarization guided mask-free shadow removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10716–10724, 2025. 1