Dark Channel Prior Guided Semi-Supervised for Text Image Restoration Under Specular Highlights and Blur

Yuanzhen Li[1], Fubao Yang[1], Hongzhi Liu[1], Fengli Yang[2], Leran Ye[1], Yue Zhao[1]*, Xiaoxiao Wang[3]
[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, zhao6685@yeah.net, xiaoxiaowang6@126.com

Abstract

Current research highlights that text image deblurring and highlight removal techniques predominantly rely on supervised learning approaches, which suffer from limited generalization. Existing methods often perform inadequately when faced with the challenging scenario of simultaneous highlights and blurring in text images. To address these limitations, we make two key contributions: (1) the creation of a real-world dataset of text images with highlight-blur degradation, and (2) the development of a novel semi-supervised framework for joint text image deblurring and highlight removal. Our approach is based on two key observations: (1) images degraded by varying levels of highlight and blur share the same underlying clean image, and (2) the dark channel components of specular highlight and blurred text images are generally non-zero. Building on these insights, we design a Siamese network training mechanism that processes multiple degraded versions of the same image (with different blur kernels and highlight intensities) through a weight-shared network, optimized using output consistency constraints. In addition, we employ dark channel priors to guide both deblurring and highlight removal. The resulting multi-scale processing network achieves high efficiency with a lightweight architecture. Experimental results show that our method effectively handles highlight-only, blur-only, and combined highlight-blur degradations, while demonstrating superior generalization compared to existing methods.

Keywords: ***Text image debuling, Semi-Supervised, Specular highlights, Dark channels prior.***

1. Introduction

Improving text image deblurring is crucial for enhancing the readability and accessibility of text in various applications, particularly in fields like document analysis, visual recognition, and assistive technologies for individuals with visual impairments. Text images captured in natural scenes often suffer from localized highlight and quality degradation due to a combination of factors, including uneven illumination, specular reflections, motion blur, noise, compression artifacts, device limitations, and surface characteristics of materials. These impairments not only reduce human readability but also significantly degrade the performance of downstream applications such as Optical Character Recognition (OCR) [8]. Consequently, it is essential to develop effective methods for text image highlight removal and restoration.

(a) Input (b) Output

Figure 1 Examples of our method for restoring low-quality specular highlight-blurred text images: The first input image is from real scene images captured by mobile phones, while the second input image is synthetic data.

Previous studies [16, 14] introduced specialized datasets for text images with specular highlight and proposed distinct highlight removal strategies. For example, Hou *et al.* [14] presented a novel two-stage framework that sequentially detected and removed highlighted regions using dedicated sub-networks, while Jiang *et al.* [16] developed a hi-

*Corresponding author: zhao6685@yeah.net

erarchical adaptive filtering network for handling large-area highlight in text images.

Research on image deblurring has made progress along two main paradigms: (1) optimization-based approaches [25, 23] that employ mathematical priors and iterative refinement, and (2) learning-based methods [32, 22, 37] that leverage deep neural networks to learn deblurring directly from data. However, most existing approaches treat deblurring and specular highlight removal as separate tasks and are trained exclusively on synthetic datasets within a supervised learning framework. This limits their generalization capability when applied to real-world data, often resulting in suboptimal restoration performance.

In this paper, we propose a unified framework for restoring degraded text images with both specular highlight and blurring (Figure 1). Specifically, we present a semi-supervised dark channel-guided framework that simultaneously addresses text image deblurring and highlight removal. Our method is built on two key observations: (1) images degraded by varying levels of highlight and blur share the same underlying clean image, which allows us to design a Siamese network training mechanism that enforces output consistency across differently degraded versions of the same text content, thereby introducing an unsupervised constraint; and (2) the dark channel elements of specular highlight and blurred text images are predominantly non-zero, with highlighted regions showing elevated values in dark channel maps. These insights motivate our proposed dark channel-guided joint framework for simultaneous deblurring and highlight removal. Furthermore, we introduce a dark channel loss function that precisely localizes degraded regions, significantly enhancing structural integrity and detail preservation under challenging conditions such as highlight and low-light environments.

To support this framework, we construct a real-world dataset of highlight-blurred text images, comprising both labeled and unlabeled samples. The labeled dataset includes clear original images paired with their highlight-blurred counterparts, while the unlabeled set contains a variety of highlight-only images without corresponding reference images. The dataset includes images captured in both indoor and outdoor scenarios using different smartphones and cameras. For indoor data, cameras were mounted on tripods, and lighting conditions were adjusted to obtain highlight-blurred and clear images. For outdoor data, highlight-blurred images were collected under direct sunlight, while umbrellas were used to capture clear images without specular highlight. This diverse dataset provides comprehensive coverage of real-world text degradation scenarios.

As shown in Figure 2, our framework employs a weight-sharing Siamese network that processes multiple degraded versions (with varying blur and highlight levels) of the same high-definition image through parallel branches, while enforcing output consistency. The encoder-decoder-based multi-scale restoration network integrates two main innovations: (1) dark channel prior integration during encoding, leveraging the sparsity of clean image dark channels compared to degraded ones; and (2) CBAMamba, a lightweight feature extraction module that improves upon CBAM by combining channel attention with wavelet transforms while maintaining efficiency. Extensive comparative experiments and ablation studies confirm the effectiveness of the proposed method.

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

- We develop a dark channel prior-guided semi-supervised framework for joint highlight and blur removal, achieving strong generalization performance.
- We construct a diverse real-world dataset of highlight-blurred text images, sourced from various materials (e.g., cards, billboards). The dataset includes samples with varying highlight intensities and multiple types of highlight-blur effects.
- We introduce a novel dark channel consistency loss function that explicitly optimizes the restoration of degradation-prone regions, guided by physical priors, improving structure and detail preservation.

2. Related Work

2.1. Specular highlight removal

Early highlight removal methods primarily relied on optimization models or chromaticity propagation [20, 11, 40]. For instance, Kim *et al*. [17] proposed an optimization framework based on the observation that dark channels in natural images approximate specular-free images. Fu *et al*. [5] built on this by modeling diffuse components as sparse linear combinations of basis colors, proposing a joint specular-diffuse separation framework. However, such methods often fail to reconstruct missing details in highlighted regions, limiting practical applicability.

Deep learning has enabled more advanced highlight removal techniques [36, 42, 43, 6]. Shi *et al*. [30] developed a unified framework for joint estimation of albedo, shading, and specular residues. Fu *et al*. [7] proposed a three-stage network that progressively removed visual artifacts by decomposing inputs into albedo, shading, and specular components, refining outputs, and adjusting tones for consistency. Zheng *et al*. [43] advanced the field using a patch-based diffusion model, incorporating residual-based training to reduce computation and patch-aware highlight removal to leverage non-highlighted regions. However, most of these methods target small-area highlight in natural scenes and are less effective for large-area highlight in text images.

For text-specific highlight removal, Hou *et al*. [14] and Jiang *et al*. [16] developed methods based on synthetic datasets. Hou *et al*. introduced a two-stage framework for sequential highlight detection and removal, while Jiang *et al*. proposed HAFNet, a hierarchical adaptive filtering network for large-area highlight removal. Nonetheless, because these models rely on synthetic training data and fully supervised learning, they struggle to generalize effectively to real-world cases.

To overcome these limitation, we constructed a comprehensive real-world training dataset and proposed a semi-supervised consistency learning framework that exploits unlabeled real data through perturbation-invariant constraints. Initially, the model is trained on label data in a supervised manner. Then, multi-intensity highlight variations are generated from real images and passed through the highlight removal network, where color-space consistency is enforced by minimizing output discrepancies between differently perturbed versions of the same image.

2.2. Image deblurring

Traditional methods rely on scene-specific assumptions, such as gradient sparsity in natural images, making them difficult to generalize to specialized scenarios like text and facial images. Ren *et al*. [28] utilized the low-rank property of similar local image patch groups. However, when the blurred image contains rich textures, this low-rank property has limitations. Pan *et al*. [26] discovered that the blurring process reduces the sparsity of dark channels, and proposed an L0-regularized dark channel sparsity constraint. However, the dark channel prior performance degrades when the image is dominated by bright pixels. Subsequently, Cheng *et al*. [1] proposed a new local maximum gradient prior.

Image deblurring has also been widely studied through both optimization-based [29, 28, 26] and learning-based [32, 39, 22, 37, 24] methods. Traditional approaches often rely on scene-specific priors, such as gradient sparsity in natural images. Ren *et al*. [28] exploited low-rank similarity among local image patches, but their method struggles with rich textures.

Traditional natural image deblurring methods perform poorly on text images, as they fail to model document-specific degradations and preserve fine textual details critical for readability. To address this, several works [2, 3, 25, 23] have proposed text-specific approaches. Chen *et al*. [2] derived the intensity probability density function (PDF) of sharp images from blurred histograms and used it as a prior. Pan *et al*. [25] introduced an L0-regularized prior based on intensity and gradient features. Yet, these traditional models often oversimplify degradation processes, making them ineffective for real-world mixed noise and unsuitable for real-time applications.

Deep learning has significantly advanced text image deblurring. Recurrent networks, including scale-recurrence [4], improved performance by leveraging multi-scale dependencies. More recently, transformer-based approaches [19, 33, 41] have introduced novel attention mechanisms (channel-wise, strip-based, and frequency-based) for robust blur feature extraction. ID-Blau [35] proposed a data augmentation technique that simulates diverse blurred images using continuous motion trajectory models, further boosting performance.

Text-specific deep learning approaches have also emerged. Hradiš *et al*. [15] presented an early CNN-based method for text deblurring. Sun *et al*. [32] introduced a CNN framework to estimate non-uniform motion blur by predicting patch-level kernels, refined through MRF fusion and rotation augmentation, enabling more accurate restoration.

3. Method

We present a semi-supervised learning approach for restoring single-view text images degraded by both specular highlight and blurring, alongside the first real-world dataset of such images. The key innovation of our method lies in exploiting the inherent property that multiple images with varying highlight positions/intensities and blur patterns correspond to the same high-definition content. This enables the formulation of a self-supervised optimization objective, where network output consistency is enforced across different highlight-blurred versions of the same image, thereby facilitating effective feature learning.

As illustrated in Figure 2, we design a dark channel-guided Siamese network training framework with two principal components: (1) Unsupervised phase—paired highlight-blurred images and their dark channel maps are fed into a weight-shared dual-branch network, with feature learning driven by output color consistency constraints; and (2) Supervised phase—network outputs are aligned with ground truth labels on annotated data. The following sections describe the dataset construction process and technical implementation of our method.

3.1. Dataset Construction

Hou *et al*. [14] constructed a controlled-illumination text-image highlight dataset (RD) by overlaying plastic films to simulate surface highlight (non-inherent material properties) and capturing images under multi-light switching. However, this approach has inherent limitations: the generated highlight deviates from real-world conditions, and noticeable tone inconsistencies exist between image pairs. They also rendered the synthetic highlight datasets SD1 and SD2 using Blender. Later, Jiang *et al*. [16] employed Unity3D to render large-area highlight images and constructed the synthetic LST dataset. Nonetheless, a significant gap remains between synthetic and real-world

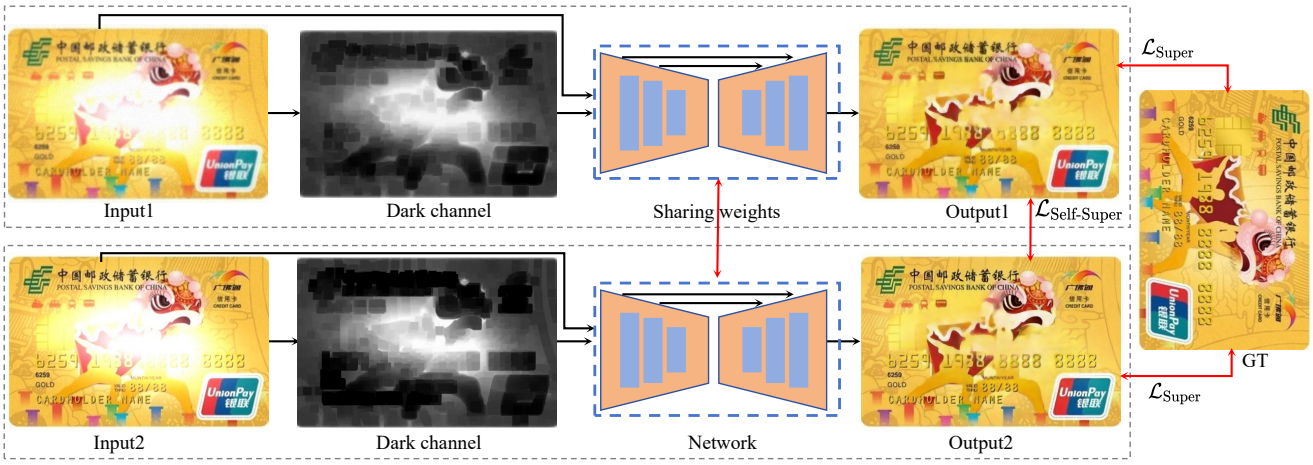


Figure 2 Overview of our method. The low-quality images with varying degrees of blurring or highlight positions, along with their corresponding dark channel images, all derived from the same high-definition reference image, are fed into a weight-sharing siamese network, where consistency constraints are applied to ensure output image alignment.

data, often causing performance degradation when models trained on synthetic datasets are applied in practice.

To address this, we built a comprehensive dataset of specular highlight-blurred text, combining both real and synthetic images. The synthetic component was generated by applying a Gaussian blur to the virtual highlight datasets of [16, 14], thereby simulating highlight-blurred conditions.

Figure 3 presents the fixed tripod and samples from our dataset. Our real dataset consists of three categories: labeled highlight images, low-quality images, and low-quality highlight images, each with corresponding quality assessment labels. We additionally collected unlabeled real-world scene images spanning these categories. The dataset contains both indoor and outdoor samples, captured using consumer-grade smartphones and cameras mounted on fixed tripods.

For indoor scenes, high-definition text images under natural lighting served as ground truth, while highlight-blurred images were captured by illuminating targets from different angles, producing varied intensity and position highlight as network input. For outdoor scenes, highlight-blurred images under direct sunlight served as input, while clear reference images were obtained using sunlight-blocking umbrellas. Additional samples are collected under varying sunlight intensities to diversify highlight characteristics.

During data preparation, we implemented a rigorous quality control protocol by eliminating invalid images (camera shake or text motion artifacts). All valid samples were then standardized to a uniform 512×352 resolution to optimize GPU memory utilization during batch processing. To improve model robustness, we employed comprehensive data augmentation strategies including random rotation within $\pm 30°$, horizontal flipping, and brightness adjustment with $\pm 20\%$ variation.

3.2. Dark Channel Prior

The dark channel prior is a statistical observation in natural haze-free images. In most local regions, at least one color channel (RGB) exhibits very low intensity. Mathematically, the dark channel J^{dark} of an image J is defined as follows [13]:

$$J^{\text{dark}}(\mathbf{x}) = \min_{c \in \{R,G,B\}} \left(\min_{\mathbf{y} \in \Omega(\mathbf{x})} J^c(\mathbf{y}) \right) \tag{1}$$

where J^c is a color channel of J, $\Omega(\mathbf{x})$ is a local patch centered at pixel $\mathbf{x}$. The dark channel prior was primarily used to describe the minimum values within the image patches.

He [13] observed that the dark channel of haze-free outdoor images is almost zero. Pan [26] discovered that most dark channel elements in natural images are zero, whereas in blurred images, most dark channel elements are nonzero.

We observed that the dark channel elements of specular highlight text images are predominantly nonzero, with values significantly higher in highlighted regions. Figure 4 compares dark channel representations of highlight-free and highlight-affected images, clearly showing the elevated values in the latter. Leveraging this distinctive property, we introduce dark channel priors to guide our network, enabling accurate highlight localization while improving the learning of textural and chromatic attributes. Additionally, we design a channel-wise loss function to further optimize performance.

Building upon the theoretical framework in [13], we systematically validated the Dark Channel Prior (DCP) using a dataset of 500 images from our proprietary collection. As presented in Table 1, highlight regions (defined by intensity

Figure 3 Fixed tripod and sample images from our dataset include: (1) outdoor blurred images under highlight conditions and their corresponding images with varying highlight intensities; (2) blurred images and their label data; (3) indoor images with highlight of different intensities and positions; and (4) outdoor highlight images and their corresponding label data.

$I > 0.9$) exhibit significantly higher dark channel values ($\mu = 0.82$, $\sigma^2 = 0.11$). compared with non-highlight regions ($\mu = 0.08$, $\sigma^2 = 0.03$), with this difference being statistically significant ($p < 0.001$ in paired t-test).

Table 1 Statistical comparison of dark channel values.

Region Type	Mean (μ)	Variance (σ^2)	p-value
Highlight	0.82	0.11	< 0.001
Non-Highlight	0.08	0.03	–

3.3. Mamba module

The Mamba module enhances state-space models (SSMs) through *input-dependent parameterization*. Its selective SSM dynamically adjusts parameters via:

$$\mathbf{B}_t = W_B x_t, \quad \Delta_t = \text{softplus}(W_\Delta x_t) \tag{2}$$

where W_B, W_Δ are learned projections. The continuous-time SSM:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}_t x(t) \tag{3}$$

undergoes zero-order hold discretization:

$$\bar{\mathbf{A}}_t = e^{\Delta_t \mathbf{A}}, \quad \bar{\mathbf{B}}_t = (\Delta_t \mathbf{A})^{-1}(e^{\Delta_t \mathbf{A}} - I)\Delta_t \mathbf{B}_t \tag{4}$$

yielding the discrete recurrence:

$$h_t = \bar{\mathbf{A}}_t h_{t-1} + \bar{\mathbf{B}}_t x_t \tag{5}$$

This achieves linear-time sequence modeling with parallel scans while maintaining global receptive fields.

Mamba leverages its linear computational complexity and dynamic selection mechanism to efficiently process long-sequence images (e.g., video frames or high-resolution images) in highlight removal tasks, while achieving precise local enhancement by selectively focusing on abrupt illumination changes. Compare with traditional CNN and Transformer architectures, Mamba significantly reduces computational overhead while maintaining a global receptive field thus making it highly suitable for real-time highlight removal tasks.

(a) (b) (c) (d)

Figure 4 Visualization of dark channel prior under varying image quality conditions from our dataset: (a) Original high-quality image; (b) Corresponding dark channel map demonstrating characteristic low-intensity values; (c) Degraded image with specular highlight and motion blur; (d) Dark channel of (c) exhibiting substantially elevated intensity values.

3.4. Method detail

Because multiple highlight-blurred variants can originate from the same clear text image, we propose a semi-supervised framework that enforces output consistency across differently degraded versions (with varying blur kernels and highlight intensities). This is realized via a weighted Siamese training strategy, where pairs of degraded text inputs are processed by a weight-sharing network to generate high-definition reconstructions. Consistency constraints ensure restored images preserve texture and color fidelity.

Finally, building on the dark channel prior, we develop a dedicated dark-channel loss to enhance color restoration in highlight regions. Integrated into the complete training

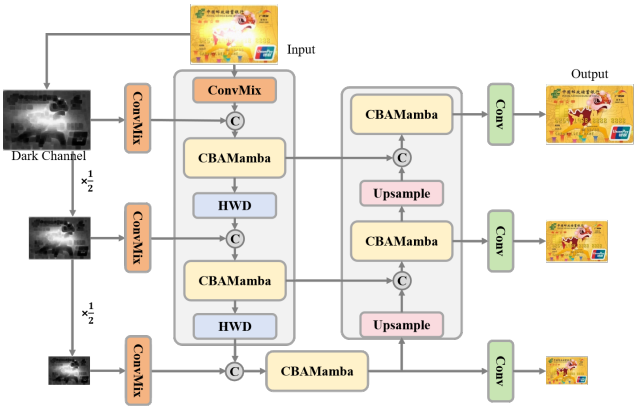


Figure 5 Network details for text image restoration with specular highlight and motion blur degradation. It follows an encoder-decoder architecture that leverages dark channel priors as guidance to progressively restore multi-scale textual content through hierarchical feature processing.

framework (2), this design leverages intrinsic dark channel distributions to effectively mitigate degradation caused by both blurring and specular highlight.

Network detail. As shown in Figure 5, we adopt an encoder–decoder deep learning architecture for restoring text images degraded by specular highlight and blurring. In the encoding stage, the network extracts multi-scale features using a Haar wavelet downsampling (HWD) module [38], replacing conventional methods such as max pooling and average pooling, thereby preserving high-frequency detail information. Wavelet downsampling preserves high-frequency structural details during resolution reduction by decomposing features into multi-scale subbands, effectively addressing the problem of information loss in conventional pooling operations. Additionally, we introduce a dark channel prior-guided feature extraction mechanism that employs Mamba modules [10]for long-range dependency modeling across multiple scales. In the decoding stage, resolution is progressively restored through upsampling, with skip connections fusing features from corresponding encoder layers, ultimately generating multi-scale outputs.

We construct a Convolutional Feature Mixer (ConvMix) model that encodes input images to extract multi-scale features through parallel hierarchical processing. The architecture employs three parallel convolutional streams using 1×1, 2×2 (with 1-pixel padding), and 3×3 (standard 1-pixel padding) kernels to simultaneously capture point-wise cross-channel interactions, small-region features, and neighborhood characteristics respectively. These multi-scale features are fused via element-wise summation while maintaining the original spatial resolution, preserving high-frequency details through identity mappings and enabling diverse receptive field coverage from 1×1 to 3×3 scales. This resolution-invariant integration approach eliminates information loss from strided operations, explicitly models scale interdependencies, and facilitates gradient flow across all scales, ultimately enhancing the network's ability to model complex multi-scale patterns more effectively than sequential pyramid approaches for dense prediction tasks.

We propose CBAMamba as the central network component—an optimized variant of MobileMamba [12], enhanced with the CBAM attention module [34] for stronger feature representation (Figure 6). Mamba [10] is a recently introduced state-space model (SSM) that effectively captures long-range dependencies with linear computational complexity. Unlike Transformer architectures, it achieves global receptive fields with significantly lower computational cost. To adapt Mamba for visual tasks, MobileMamba [12] incorporates the Multi-Receptive Field Feature Interaction (MRFFI) module, which processes input features through three branches:

- WTEMamba (Wavelet Transform-Enhanced Mamba) – extracts global features while preserving edge details via wavelet-based enhancement.
- MK-DeConv (Multi-Kernel Depthwise Convolution) – captures multi-scale local features.
- Efficient Channel Processing Branch – reduces redundancy in high-dimensional feature spaces by removing unnecessary identity mappings.

The MRFFI module synergistically integrates these representations, enabling simultaneous modeling of global contexts and local multi-scale patterns. This significantly enhances edge-detail preservation while maintaining computational efficiency.

The CBAM module enhances feature representation by sequentially applying channel attention and spatial attention mechanisms. The channel attention recalibrates feature importance across channels through global average pooling and multilayer perceptrons, while the spatial attention emphasizes salient regions via spatial pooling operations. This dual-attention design enables adaptive refinement of both channel-wise and spatial-wise feature responses, effectively suppressing irrelevant background interference while preserving critical structural details.

To further boost performance, we introduce parallel CBAM blocks after both the MK-DeConv and MRFFI branches. This yields two key improvements: (1) The dual channel–spatial attention mechanism of CBAM overcomes MobileMamba's limitation in explicit attention modeling, improving discriminative feature selection. (2) CBAM optimally coordinates the integration of WTEMamba's global representations with MK-DeConv's local features, ensuring comprehensive long-range dependencies while strengthening sensitivity to fine-grained textural and edge details.

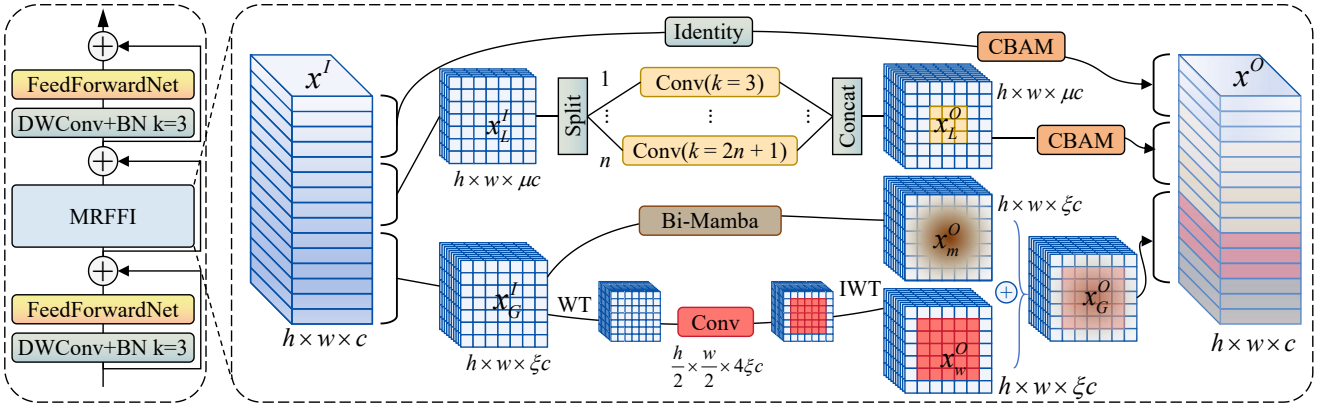


Figure 6 The detail of the CBAMamba network model.For more details, please refer to MobileMamba [12].

3.5. Loss function

Our loss function incorporates a supervised loss term for labeled data and a self-supervised regularization for unlabeled data. Specifically, the supervised objective $\mathcal{L}_{Sup}$ evaluating the predicted image I^{out} against ground truth I^{gt} consists of four fundamental components: color loss $\mathcal{L}_c$, adversarial loss $\mathcal{L}_{adv}$, feature space loss $\mathcal{L}_s$, and dark channel loss $\mathcal{L}_d$. These are combined into the total loss function:

$$\mathcal{L}_{Super} = \lambda_c \mathcal{L}_c + \lambda_{adv} \mathcal{L}_{adv} + \lambda_s \mathcal{L}_s + \lambda_d \mathcal{L}_d \tag{6}$$

where λ_c, λ_{adv}, λ_s, and λ_d are hyperparameters that balance the contributions of each term. In the following, we describe the formulation and rationale for each component.

Color loss. Color loss measures the L1 distance between the multi-scale ground-truth and output images:

$$\mathcal{L}_c = \sum_{k=1}^{K} \frac{1}{t_k} \left\| I_k^{gt} - I_k^{out} \right\|_1 \tag{7}$$

where K is the scale levels (in our method, $K = 3$), I_k^{gt} is the ground truth image at scale level k, I_k^{out} is the model output at scale level k, t_k is the total number of elements in the image at scale level k, used for the normalization of the loss.

Adversarial loss. We employ a relativistic adversarial loss function [16] that evaluates not only the absolute discriminator scores for real and generated images but also their relative discrepancies:

$$\begin{aligned} \mathcal{L}_{adv} = &\sum_{k=1}^{K} \mathrm{BCE}\left(\sigma\left(D(I_k^{gt}) - D(I_k^{out})\right), g\right) \\ &+ \sum_{k=1}^{K} \mathrm{BCE}\left(\sigma\left(D(I_k^{gt}) - D(I_k^{out})\right), d\right) \end{aligned} \tag{8}$$

where σ denotes the sigmoid activation function, $\mathrm{BCE}(\cdot)$ represents binary cross-entropy, and g/d are training targets ($g = 1$, $d = 0$ for generator updates; $g = 0$, $d = 1$ for discriminator training). This relative assessment mechanism enhances the ability of the model to discern fine-grained quality distinctions, thereby facilitating the synthesis of more realistic images.

Feature space loss. We utilize the feature space consistency constraint [9] between the predicted image and the ground truth image:

$$\begin{aligned} \mathcal{L}_s = &\sum_{k=1}^{K} \frac{1}{t_k} \left\| \Phi(I_k^{gt}) - \Phi(I_k^{out}) \right\|_1 \\ &+ \sum_{k=1}^{K} \frac{1}{t_k} \left\| G(I_k^{gt}) - G(I_k^{out}) \right\|_1 \end{aligned} \tag{9}$$

where Φ represents the latent space feature extracted from the first layer of the pre-trained VGG19 [31], which is particularly effective at capturing low-level color characteristics of the reconstructed texture; $G(.) = \Phi\Phi^T$ denotes the Gram matrix. The first term represents content loss, and the second term denotes style loss. Content loss enforces invariance in human-perceived abstract features, and style loss measures stylistic differences.

Dark channel loss. The color loss $\mathcal{L}_c$ only constrains pixel-level brightness errors, whereas dark channel loss explicitly enforces the physical sparsity of highlighted regions through local minimum filtering (min(R,G,B)), ensuring the restoration results align better with the prior knowledge that highlights occupy only small localized areas in natural scenes (as described in dark channel prior [13]). By operating on local extremum values rather than pixel intensities, the dark channel loss avoids the blurring effect of the color loss on the text image:

$$\mathcal{L}_d = \sum_{k=1}^{K} \frac{1}{t_k} \left\| R_k - C_k \right\|_1 \tag{10}$$

Table 2 Quantitative comparison between our method and state-of-the-art approaches including LSH, RD, and SD1 on both benchmark and our proposed dataset. The best results in each test set are highlighted in bold, with abbreviations w/o for without and DII for Dark Input Image. Recall is reported in percentages.

Method	Time.(S)	Recall	LSH [16]		RD [14]		SD1 [14]		Ours	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Kim *et al*. [17]	5.062	76.801	11.472	0.421	11.634	0.435	11.692	0.450	12.294	0.479
Fu *et al*. [5]	4.639	78.106	12.312	0.462	12.743	0.450	12.549	0.470	13.805	0.519
Wu *et al*. [36]	0.570	83.501	25.013	0.858	25.348	0.859	25.548	0.873	24.358	0.841
HighlightRNet [42]	0.358	85.023	25.236	0.872	25.741	0.886	25.084	0.882	24.693	0.864
TASHR [14]	0.410	81.530	23.754	0.866	23.003	0.853	23.419	0.870	22.987	0.854
HAFNet [16]	0.461	86.421	26.428	0.906	26.122	0.914	26.093	0.907	25.726	0.892
Shan *et al*. [29]	3.852	80.024	12.421	0.472	12.326	0.447	12.198	0.436	13.420	0.527
Pan *et al*. [26]	2.449	81.659	13.924	0.506	13.736	0.482	13.684	0.495	14.308	0.524
Wu *et al*. [35]	43.147	85.350	25.135	0.874	25.941	0.885	25.529	0.866	24.023	0.821
Min *et al*. [23]	2.387	82.046	24.110	0.856	24.054	0.886	24.801	0.881	23.113	0.824
Adarevd [22]	1.362	85.962	24.412	0.885	24.170	0.898	24.209	0.890	23.540	0.872
EGDeblurring [37]	0.489	87.642	25.581	0.908	25.747	0.911	25.507	0.899	24.312	0.863
Ours w/o $\mathcal{L}_{Self\text{-}Super}$	0.324	90.132	26.752	0.914	26.986	0.918	26.367	0.918	26.826	0.914
Ours MobileMamba	0.318	91.604	27.384	0.937	27.678	0.935	27.207	0.935	27.284	0.929
Ours w/o $\mathcal{L}_d$, w/o DII	0.307	93.204	27.145	0.929	27.437	0.928	27.114	0.928	27.044	0.920
Ours w/o $\mathcal{L}_d$	0.324	91.648	27.408	0.928	27.606	0.919	27.300	0.929	27.230	0.926
Ours w/o DII	0.307	92.036	27.401	0.931	27.605	0.921	27.308	0.933	27.220	0.923
Ours $K = 1$	0.319	92.212	27.412	0.939	27.658	0.929	27.312	0.932	27.215	0.924
Ours $K = 2$	0.324	92.483	27.506	0.930	27.810	0.938	27.417	0.935	27.392	0.936
Ours	0.324	**92.527**	**27.547**	**0.945**	**27.842**	**0.941**	**27.419**	**0.937**	**27.410**	**0.938**

where R_k denotes the dark channel image of the model output I_k^{out} at scale level k and C_k denotes the dark channel image of the ground truth I_k^{gt} at scale level k.

The color loss $\mathcal{L}_c$ controls global brightness convergence, while the dark channel loss focuses on suppressing local outliers, forming a global-local dual-granularity supervision framework.

The self-supervised objective functions for the two predicted images, I^{out1} and I^{out2}, consist of three key components: color loss $\mathcal{L}_c$, feature space loss $\mathcal{L}_s$, and dark channel loss $\mathcal{L}_d$. These are combined into the total loss function:

$$\mathcal{L}_{Self\text{-}Super} = \lambda_c \mathcal{L}_c + \lambda_s \mathcal{L}_s + \lambda_d \mathcal{L}_d \tag{11}$$

We design the training protocol as a two-stage progressive framework: first pre-training the model solely with $\mathcal{L}Super$ to establish fundamental restoration capabilities, then jointly optimizing $\mathcal{L}$ and $\mathcal{L}_{Self\text{-}Super}$ on mixed clean-degraded data until validation loss plateaus.

4. Experiments

To validate the effectiveness of the proposed method, we conducted systematic comparative experiments and ablation studies, followed by an in-depth analysis of the experimental results. Additionally, we discussed the limitations of the current approach and suggested potential directions for future improvement.

Implementation details. We implemented our method in PyTorch [27] and trained it for 100 epochs with a batch size of four on a PC equipped with an NVIDIA GeForce RTX 4080 Ti GPU. Optimization was performed using the Adam optimizer [18] with a fixed learning rate of 10^{-4}. Through extensive parameter tuning, we found that the model performs optimally with hyperparameters $\lambda_c = 1$, $\lambda_{adv} = 0.5$, $\lambda_s = 0.2$, and $\lambda_d = 0.5$.

4.1. Datasets and evaluation metrics

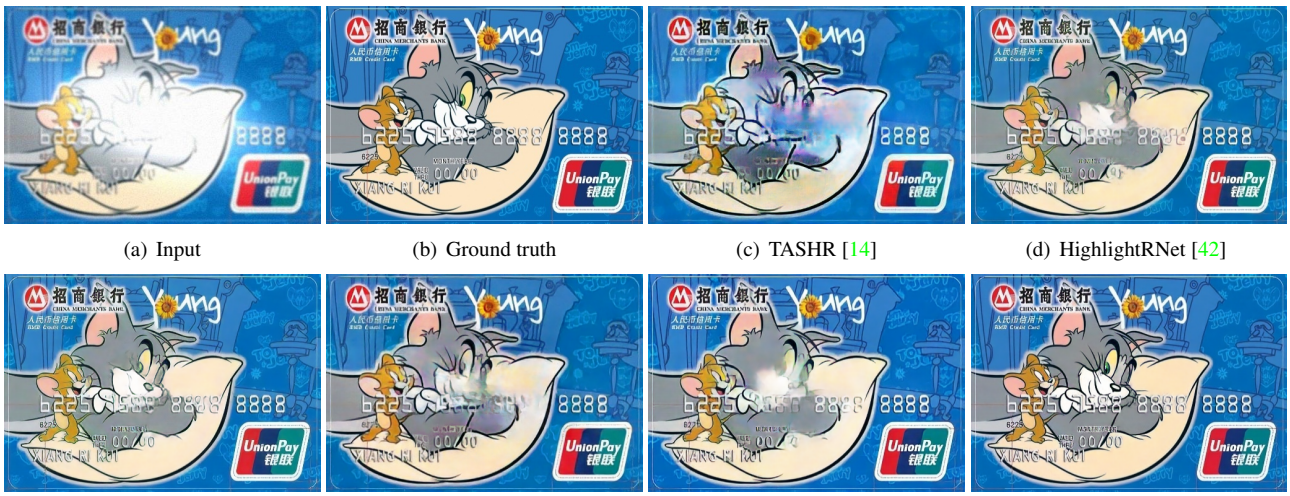
Datasets. We performed blur degradation processing on the LSH [16], RD [14], and SD1 [14] datasets to generate blurred image pairs, specular highlight image pairs, and specular highlight-blurred pairs as our virtual dataset. The training data consists of 500 labeled pairs from LSH, 100 labeled pairs from RD, as well as 1000 labeled pairs, and 500 unlabeled pairs from our collected real-world data. For evaluation, we used four datasets: LSH (100 labeled test pairs), RD (100 labeled test pairs), SD1 (100 labeled test pairs), and our collected data (200 labeled test pairs).

Evaluation metrics. We evaluated network performance using two standard metrics—Peak Signal-to-Noise

(a) Input (b) Ground truth (c) TASHR [14] (d) HighlightRNet [42]

(e) HAFNet [16] (f) Adarevd [22] (g) EGDeblurring [37] (h) Ours

Figure 7 Qualitative comparison with state-of-the-art highlight removal methods on LSH dataset.



(a) Input (b) Ground truth (c) TASHR [14] (d) HighlightRNet [42]

(e) HAFNet [16] (f) Adarevd [22] (g) EGDeblurring [37] (h) Ours

Figure 8 Qualitative comparison with state-of-the-art highlight removal methods on LSH dataset.

Ratio (PSNR) and Structural Similarity Index (SSIM)—to measure color fidelity and structural similarity between the predicted outputs and expert-retouched reference images. Higher values indicate better quality.

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

We compared our method with state-of-the-art approaches in specular highlight removal and image reblurring.

- Highlight removal: Optimization-based methods (Kim *et al*. [17], Fu *et al*. [5]), and deep learning-based approaches (Wu *et al*. [36], HighlightRNet [42], TASHR [14], and HAFNet [16]).
- Reblurring: Optimization-based works (Shan *et al*. [29], Pan *et al*. [26]), and deep learning-based methods (Wu *et al*. [35], Min *et al*. [23], Adarevd [22], and EGDeblurring [37]).

To ensure fair comparisons, we: (1) ran baseline methods with their official source codes, (2) strictly followed parameter configurations recommended in the original publications, and (3) retrained all learning-based methods on identical datasets. Our analysis focused on three aspects.

Quantitative comparison. To evaluate the learning effectiveness and generalization capability of our network, we conducted comprehensive quantitative comparisons with state-of-the-art methods. Table 2 lists the experimental re-

(a) Input (b) Ground truth (c) TASHR [14] (d) HighlightRNet [42]

(e) HAFNet [16] (f) Adarevd [22] (g) EGDeblurring [37] (h) Ours

Figure 9 Qualitative comparison with state-of-the-art highlight removal methods on our dataset.

(a) Input (b) Ground truth (c) TASHR [14] (d) HighlightRNet [42]

(e) HAFNet [16] (f) Adarevd [22] (g) EGDeblurring [37] (h) Ours

Figure 10 Qualitative comparison with state-of-the-art highlight removal methods on our dataset.

sults. Our method consistently outperforms competing approaches across all four test datasets. While other methods exhibit significant performance degradation on our test set compared to standard benchmarks, our approach maintains stable performance, demonstrating robust generalization. This strength primarily stems from our self-supervised training strategy, which reduces reliance on labeled data while maintaining high performance.

As shown in Table 2, we further compare the runtime of different methods. The experimental results show that the proposed approach requires significantly less computation time than most competing methods, demonstrating its efficiency advantage. Additionally, we evaluate OCR word recognition accuracy (measured by Recall [21]) as a key performance metric. Quantitative analysis confirms that our method achieves superior accuracy compared to existing approaches, further validating its effectiveness and robustness.

Visual comparison. Qualitative results from challenging cases in the LSH dataset and our dataset (Figures 7-10) demonstrate that our method better restores both text and background colors in highlight-occluded and blurred regions. Particularly on real-world images, our model significantly outperformes competing methods, underscoring its strong generalization capability.

User study. Moreover, to validate the effectiveness of

(a) Input (b) TASHR (c) HighlightRNet (d) HAFNet (e) Adarevd (f) EGDeblurring (g) Ours

Figure 11 Qualitative comparison with state-of-the-art highlight removal methods on our dataset.

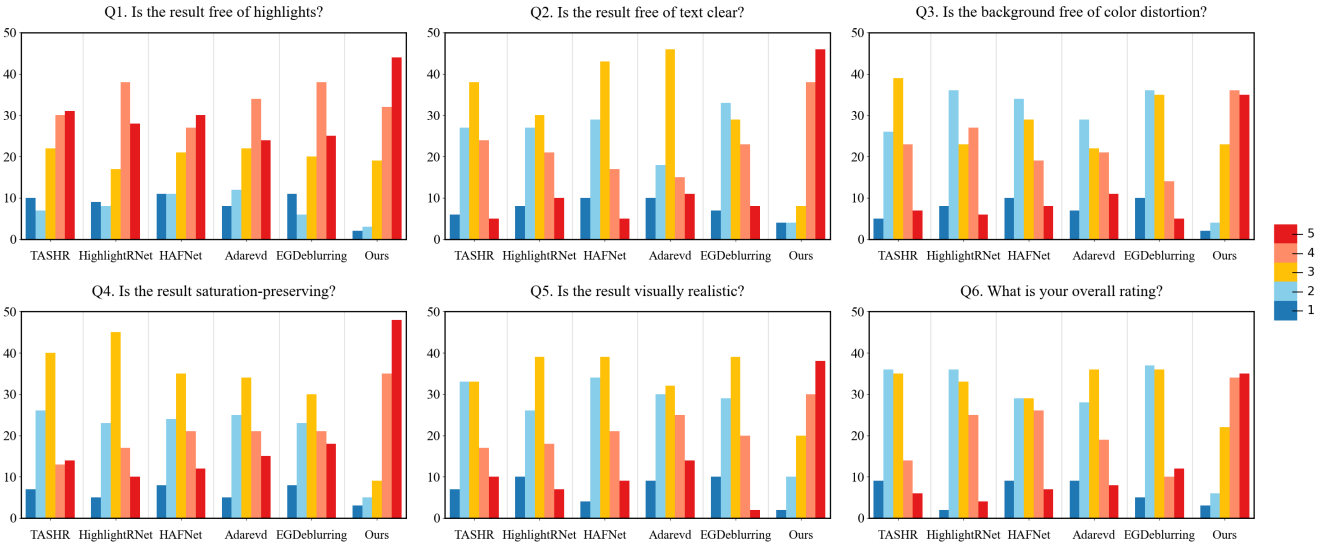


Figure 12 Rating distributions of different methods across the six questions in the user study. The y-axis indicates the percentage of ratings that participants assigned to each method.

our method on real-world data, we conducted a user study. We collected 100 test images—50 crawled from the Internet and 50 contributed by volunteers—containing text affected by specular highlight, blurring, or both, captured under diverse scenarios using different cameras (see Figure 11). Each image was processed using our method along with competing approaches for highlight and blur removal. To ensure an objective evaluation, we recruited 500 campus volunteers to rate the results, with all processed outputs presented in a randomized order to minimize subjective bias.

Participants were instructed to evaluate each processed result using a 5-point Likert scale (1 = worst, 5 = best) across the six assessment dimensions, as shown in Figure 11. Figure 12 shows the aggregated rating distributions, where each subplot displays the score distribution of the different methods for each evaluation criterion. Statistical analysis reveals that our method achieves a significantly higher density in the high-quality range (4-5 points) compared to baseline approaches, while showing substantially fewer low-score evaluations. This distinct distribution pattern clearly demonstrates that our method not only receives stronger preference from evaluators but also exhibits better generalization capability across diverse scenarios when compared to existing methods.

4.3. Ablation study and limitation

Ablation study on self-supervised framework. We conducted an ablation study to evaluate the impact of our self-supervised training strategy. Specifically, we compared two loss configurations: (1) using only $\mathcal{L}_{Super}$ (without $\mathcal{L}_{Self\text{-}Super}$) and (2) our proposed combination, $\mathcal{L}_{Super} + \mathcal{L}_{Self\text{-}Super}$ (Ours). Quantitative comparisons (Table 2 shows that incorporating the self-supervised loss yields improvements across both evaluation metrics, demonstrating the effectiveness of the proposed approach.

Figure 13 presents a visual result on the input image randomly captured with mobile phones, which is excluded from the training set. The comparison illustrates that in-

(a) Input (b) w/o $\mathcal{L}_{Self\text{-}Super}$ (c) Ours

Figure 13 Ablation study on self-supervised loss.

corporating the self-supervised loss significantly improves restoration quality, enabling accurate recovery of text previously occluded by specular highlight.

Ablation study on CBAMamba model. To highlight the advantages of our CBAMamba, built upon MobileMamba [12], we replaced CBAMamba with the baseline MobileMamba and trained it on the same dataset. Quantitative results (Table 2) confirm the superior performance of CBAMamba. By combining dual attention through CBAM, our model enhances feature selection and achieves refined fusion of global-local features, resulting in stronger performance compared to the baseline.

Ablation study on multi-scale estimation. We further verified the effectiveness of our proposed multi-scale restoration design by varying the output scales ($K = 1, 2, 3$). A single-scale model ($K = 1$) outputs only the restored image at the original resolution, while the two-scale model ($K = 2$) outputs both full and half-resolution results. Training and evaluation show that our three-scale approach ($K = 3$) consistently achieves the best quantitative performance across all test sets (Table 2). Visual comparisons in Figure 14 further confirm that the three-scale design produces more realistic results with superior detail preservation and structural integrity, significantly outperforming single-scale alternatives.

Ablation study on dark channel prior. To validate the role of dark channel guidance, we conducted four comparative configurations: (1) the complete method (baseline), (2) removal of the dark channel image input from the network (without DII), (3) elimination of the dark channel loss function (without $\mathcal{L}_d$), and (4) simultaneous removal of both components.

All configurations were trained on the same dataset and evaluated on our test set (Table 2). Results show that dark channel guidance substantially improves performance, both as an input feature and as a loss component. The complete method achieves the best performance, confirming the effectiveness of dark channel priors in our framework.

Limitations. Despite its superior performance on both synthetic and real-world datasets, our method still faces two challenges: (1) difficulty in restoring images with extremely strong highlight and severe blurring, and (2) color distortion

(a) Input (b) $K = 1$ (c) $K = 3$

Figure 14 Ablation study on multi-scale estimation.

in non-highlight areas, which arises from inconsistent color matching in supervised training pairs. Future work will focus on more robust self-supervised learning strategies and improved loss functions to address color inconsistency.

5. Conclusion

In this paper, we proposed a semi-supervised restoration method for text images affected by specular highlight and blurring, leveraging dark channel priors. Our approach demonstrated strong generalization capability, effectively removing specular highlight while improving overall image quality. Unlike prior methods that addressed highlight or blurring in isolation and exhibited limited robustness, our framework tackled both simultaneously. Extensive ablation experiments confirmed the effectiveness of the proposed modules, and quantitative results established state-of-the-art performance.

Acknowledgement

This work is 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] L. Chen, F. Fang, T. Wang, and G. Zhang. Blind image deblurring with local maximum gradient prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1742–1750, 2019. 3

[2] X. Chen, X. He, J. Yang, and Q. Wu. An effective document image deblurring algorithm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 369–376, 2011. 3

[3] H. Cho, J. Wang, and S. Lee. Text image deblurring using text-specific properties. In *Proceedings of the European*

Conference on Computer Vision (ECCV), pages 524–537, 2012. 3

[4] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision (ICCV)*, pages 4641–4650, 2021. 3

[5] G. Fu, Q. Zhang, C. Song, Q. Lin, and C. Xiao. Specular highlight removal for real-world images. *Computer Graphics Forum (CGF)*, 38(7):253–263, 2019. 2, 8, 9

[6] G. Fu, Q. Zhang, L. Zhu, P. Li, and C. Xiao. A multi-task network for joint specular highlight detection and removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7752–7761, 2021. 2

[7] G. Fu, Q. Zhang, L. Zhu, C. Xiao, and P. Li. Towards high-quality specular highlight removal by leveraging large-scale synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12857–12865, 2023. 2

[8] M. Fujitake. Dtrocr: Decoder-only transformer for optical character recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8025–8035, 2024. 1

[9] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016. 7

[10] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 6

[11] J. Guo, Z. Zhou, and L. Wang. Single image highlight removal with a sparse and low-rank reflection model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 268–283, 2018. 2

[12] H. He, J. Zhang, Y. Cai, H. Chen, X. Hu, Z. Gan, Y. Wang, C. Wang, Y. Wu, and L. Xie. Mobilemamba: Lightweight multi-receptive visual mamba network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4497–4507, 2025. 6, 7, 12

[13] K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(12):2341–2353, 2010. 4, 7

[14] S. Hou, C. Wang, W. Quan, J. Jiang, and D.-M. Yan. Text-aware single image specular highlight removal. In *Proceedings of the Pattern Recognition and Computer Vision (PRCV)*, pages 115–127, 2021. 1, 3, 4, 8, 9, 10

[15] M. Hradiš, J. Kotera, P. Zemcık, and F. Šroubek. Convolutional neural networks for direct text deblurring. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 10, 2015. 3

[16] 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 (CVPR)*, pages 2408–2417, 2025. 1, 3, 4, 7, 8, 9, 10

[17] H. Kim, H. Jin, S. Hadap, and I. Kweon. Specular reflection separation using dark channel prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1460–1467, 2013. 2, 8, 9

[18] D. P. Kingma. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2014. 8

[19] L. Kong, J. Dong, J. Ge, M. Li, and J. Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5886–5895, 2023. 3

[20] Y. Liu, Z. Yuan, N. Zheng, and Y. Wu. Saturation-preserving specular reflection separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3725–3733, 2015. 2

[21] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto, et al. Icdar 2003 robust reading competitions: entries, results, and future directions. *International Journal of Document Analysis and Recognition (IJDAR)*, 7(2):105–122, 2005. 10

[22] X. Mao, Q. Li, and Y. Wang. Adarevd: Adaptive patch exiting reversible decoder pushes the limit of image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25681–25690, 2024. 2, 3, 8, 9, 10

[23] Z. Min, G. M. Hassan, and G.-S. Jo. Robust blind text image deblurring via maximum consensus framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4242–4250, 2024. 2, 3, 8, 9

[24] J. Pan, J. Dong, Y. Liu, J. Zhang, J. Ren, J. Tang, Y.-W. Tai, and M.-H. Yang. Physics-based generative adversarial models for image restoration and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(7):2449–2462, 2021. 3

[25] J. Pan, Z. Hu, Z. Su, and M.-H. Yang. Deblurring text images via L0-regularized intensity and gradient prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2908, 2014. 2, 3

[26] J. Pan, D. Sun, H. Pfister, and M.-H. Yang. Deblurring images via dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(10):2315–2328, 2017. 3, 4, 8, 9

[27] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2017. 8

[28] W. Ren, X. Cao, J. Pan, X. Guo, W. Zuo, and M.-H. Yang. Image deblurring via enhanced low-rank prior. *IEEE Transactions on Image Processing (TIP)*, 25(7):3426–3437, 2016. 3

[29] Q. Shan, J. Jia, and A. Agarwala. High-quality motion deblurring from a single image. *ACM Transactions on Graphics (TOG)*, 27(3):1–10, 2008. 3, 8, 9

[30] J. Shi, Y. Dong, H. Su, and S. X. Yu. Learning non-lambertian object intrinsics across shapenet categories. In *Proceedings of the IEEE/CVF Conference on Computer Vi-*

sion and Pattern Recognition (CVPR), pages 1685–1694, 2017. 2

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 7

[32] J. Sun, W. Cao, Z. Xu, and J. Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 769–777, 2015. 2, 3

[33] F.-J. Tsai, Y.-T. Peng, Y.-Y. Lin, C.-C. Tsai, and C.-W. Lin. Stripformer: Strip transformer for fast image deblurring. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 146–162. Springer, 2022. 3

[34] 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 (ECCV)*, pages 3–19, 2018. 6

[35] J.-H. Wu, F.-J. Tsai, Y.-T. Peng, C.-C. Tsai, C.-W. Lin, and Y.-Y. Lin. Id-blau: Image deblurring by implicit diffusion-based reblurring augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25847–25856, 2024. 3, 8, 9

[36] Z. Wu, C. Zhuang, J. Shi, J. Guo, J. Xiao, X. Zhang, and D.-M. Yan. Single-image specular highlight removal via real-world dataset construction. *IEEE Transactions on Multimedia (TMM)*, 24:3782–3793, 2021. 2, 8, 9

[37] X. Xie, Q. Zhang, and W.-S. Zheng. Diffusion-based event generation for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2194–2203, 2025. 2, 3, 8, 9, 10

[38] G. Xu, W. Liao, X. Zhang, C. Li, X. He, and X. Wu. A simple but effective downsampling module for semantic segmentation. *Pattern Recognition (PR)*, 143:109819, 2023. 6

[39] S. Xu, Z. Sun, M. Zhong, C. Cao, Y. Liu, X. Fu, and Y. Chen. Motion-adaptive transformer for event-based image deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 8942–8950, 2025. 3

[40] Q. Yang, S. Wang, and N. Ahuja. Real-time specular highlight removal using bilateral filtering. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision (ECCV)*, pages 87–100, 2010. 2

[41] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5728–5739, 2022. 3

[42] L. Zhang, Y. Ma, Z. Jiang, W. He, Z. Bao, G. Fu, W. Xu, and C. Xiao. Highlightremover: Spatially valid pixel learning for image specular highlight removal. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, pages 10046–10054, 2024. 2, 8, 9, 10

[43] H. Zheng, Z. Bao, G. Fu, X. Jiao, and C. Xiao. Phr-diff: Portrait highlights removal via patch-aware diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 10555–10563, 2025. 2