Shift-Window Meets Dual Attention: A Multi-Model Architecture for Specular Highlight Removal

Tianci Huo, Lingfeng Qi, Yuhan Chen, Qihong Xue, Jinyuan Shao, Hai Yu, Jie Li, Zhanhua Zhang, and Guofa Li, *Senior Member, IEEE*

***Abstract*—Downstream tasks in visual multimedia and computer vision, such as image fusion and image segmentation, are severely affected by specular reflections. However, existing methods exhibit limited texture consistency and struggle to balance inference speed with highlight removal accuracy. To resolve these challenges, this paper proposes a multi-model architecture for specular highlight removal (MM-SHR). We employ convolution operations to extract local details in the shallow layers of MM-SHR, and utilize attention mechanism to capture global features in the deep layers, ensuring both operational efficiency and removal accuracy. To model long-range dependencies without compromising computational complexity, we propose Omnidirectional Attention Integration Block and Adaptive Region-Aware Hybrid-Domain Dual Attention Convolutional Network, which leverage pixel-shifting and window-dividing operations at the raw features to achieve specular highlight removal in a coarse-to-fine manner. Moreover, frequency processing at different feature scales is incorporated in the proposed modules to incrementally enhance spatial perception of highlight features. This facilitates capturing global illumination features while preserving the overall brightness of the original image by avoiding over-processing. Extensive experiments show that MM-SHR exceeds the state-of-the-art methods investigated in evaluation metrics. The implementation will be made publicly available at https://github.com/Htcicv/MM-SHR.**

***Index Terms*—Specular highlight removal, deep learning, attention mechanism**

I. Introduction

SPECULAR highlight is a localized highlight phenomenon caused by specular reflection on smooth surfaces under the action of light, and this localized highlight often reduces the visual consistency of images and videos and destroys the detailed texture information, which poses great challenges to computer vision tasks, including image segmentation, NeRF Rendering, and object detection [1]. Therefore, the efficient removal of specular highlight from images is of great importance for the development and application of multimedia vision technology. To solve this problem, researchers have proposed various methods of specular highlight removal. A classic approach is based on the HIS color space [2], where researchers achieve specular component separation by matching the saturation of specular pixels to the value of diffuse pixels of the same diffuse chromaticity. Another widely used method is the method [3], [4], [5], [6] based on the dichromatic reflection model [7], where researchers have achieved the removal of specular highlight from an image by separating the reflective component of the image into two parts: diffuse and specular reflection. In addition to methods based on HIS and dichromatic reflection model, filters [8], polarization characterization [9] [10] [11] and principal component analysis (PCA) [12] have been newly applied to the field of specular highlight removal. Despite their contributions, these traditional methods generally depend on restrictive assumptions regarding surface properties, lighting conditions, or material priors, which undermines their applicability to complex real-world scenes.

With the rapid development of deep learning [13], [14] and high-quality datasets [15], [16], [17], a variety of deep learning-based specular highlight removal methods have emerged, which can be broadly categorized into multi-image approaches [18], [19], [20] and single-image approaches [21], [22], [23], [24]. Multi-image methods often leverage highlight masks or multi-view information to improve accuracy but require auxiliary data, limiting their applicability. Single-image methods are more practical, but face inherent trade-offs: CNN-based networks lack long-range feature modeling, while Transformer-based architectures incur high computational costs [24], making real-time deployment challenging. Furthermore, many existing UNet-based designs employ homogeneous encoder–decoder architectures—constructed entirely with CNNs or Transformers across all scales—which fail to exploit their complementary strengths in capturing local details and global context. These limitations motivate our proposed method.

To solve the above challenges while ensuring both real-time performance and high precision for practical applications, we propose a novel end-to-end network for specular highlight removal. Our approach combines shallow and deep network components to optimize feature extraction and model efficiency. In particular, to enhance the shallow network's ability to model long-range dependencies, we incorporate the window partitioning mechanism from the Swin Transformer [25] into its architecture, enabling it to capture global context more effectively while preserving local feature details. Furthermore, to overcome the inherent limitation of isolated window processing in existing window-based models, we design a novel shallow-level cross-window modeling mechanism. In

Tianci Huo, Lingfeng Qi, Yuhan Chen, Qihong Xue, Jinyuan Shao, Hai Yu, Jie Li, and Guofa Li, are with the College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing, 400044, China.

Zhanhua Zhang is with the College of Computer Science and Technology, Zhejiang University, Hangzhou, 310058, China and Artificial Intelligence Center, Geely Automotive Research Institute, Ningbo, 315300, China.

this mechanism, feature maps are first divided into non-overlapping windows, which are then shifted in multiple directions. For each window, we perform joint global–local feature extraction guided by frequency-domain information, allowing the shallow network to capture long-range dependencies without sacrificing fine-grained local details. This design, for the first time in the field of specular highlight removal, eliminates the isolation problem between windows while enabling cross-window global modeling at the shallow stage of the network. Our primary contributions are summarized as follows:

1) We develop a multi-model architecture for specular highlight removal (MM-SHR). This network adopts a hierarchical structure that combines CNN for detail extraction and Transformer for global feature integration, leveraging the strengths of both to efficiently and effectively remove specular highlight. Meanwhile, we leverage the frequency-domain information, which includes global illumination characteristics, to guide the extraction of spatial highlight features across across multiple scales.
2) To enhance the CNN's ability for capturing long-range dependencies, we incorporate omni-directional pixel shifting and window dividing operations, which segments the raw features into interrelated windows, into both the Omnidirectional Attention Integration Block (OAIBlock) and the Adaptive Region-Aware Hybrid-Domain Dual Attention Convolutional network (HD-DAConv), which distinguishes itself from OAIBlock by decoupling these operations. This allows MM-SHR to more fully capture the dependencies of long-range information across multiple scales.
3) Extensive experiments conducted on the SHIQ (Specular Highlight Image Quadruples), PSD (Paired Specular-Diffuse), and NSH (Natural Specular Highlight) datasets demonstrate that MM-SHR outperforms state-of-the-art methods both quantitatively and qualitatively. Additional ablation studies confirm the individual effectiveness of each proposed module, providing a compelling and generalizable solution for specular highlight removal tasks in real-world imaging systems.

II. Related Work

A. Multi-image Methods

Multi-image-based techniques utilize auxiliary information such as multiple viewpoints or externally generated highlight masks to facilitate specular highlight removal. Among these methods, Feng et al. [18] introduced an approach based on the dichromatic reflection model, utilizing full variational optimization within a light field camera to eliminate highlight. Fu et al. [19] introduced a stage-wise approach for specular reflection removal. Jin et al. [26] proposed a two-step learning framework that first employs a constrained loss function based on prior knowledge of shaded and highlight-free images to estimate an initial reflectance layer. This is followed by an S-aware network that further refines the reflectance layer. Ultimately, their model classifies and separates the highlight component from the input image. Researchers [27], [28], [29] have explored the use of highlight masks to guide specular highlight removal. Wu et al. [20] developed a UNet-Transformer network that detects and removes highlight by first training a UNet-based highlight detection module and then using the generated mask to guide the removal process. Hu et al. [30] applied the NMF method to extract highlight region mask and separately trained generator networks for the luminance and full channels. Their approach produced high-quality results free of color distortion and highlight residues. Despite their theoretical soundness, mask-dependent methods often face limitations in real-world deployment because highlight mask annotations are often unavailable or inaccurate. Recently, Ha et al. [31] addressed the challenge of accurately separating diffuse and specular reflection in high dynamic range images by introducing the Temporal Dark Prior (TDP) as a pseudo-diffuse reflection. At the same time, they leveraged the periodic variation characteristics of AC-powered light sources as constraints, offering a novel approach to highlight removal. For facial highlight removal, Su et al. [32] employed a nonlinear optimization approach based on multi-view face images and Lambertian consistency to effectively reduce highlight. Overall, while multi-image-based specular highlight removal techniques have achieved significant success, their reliance on multiple images or auxiliary data limits their practical applicability.

B. Single-Image methods

Single-image specular highlight removal methods aim to eliminate highlights from a single image, offering broad applicability compared with multi-image approaches. Early works often combined traditional reflection models with hand-crafted strategies, such as coarse-to-fine associative learning [21], feature decomposition [22], and color expansion mechanisms [33]. With the rapid advancement of deep learning, specular highlight removal methods [20] [34] [35] based on new network architectures such as GAN [13] and Transformer [14] have emerged continuously. Among these, Anwer et al. [23] proposed SpecSeg, a segmentation-oriented network based on the UNet architecture, which serves as a strong baseline for subsequent highlight removal tasks. Guo et al. [24] proposed a highlight removal network DHAN-SHR built on the Transformer framework. However, the high computational and memory demands of Transformers result in slow inference speeds, making the model unsuitable for real-time applications in automated systems. Madessa et al. [36] focused on improving the efficiency of existing methods for removing highlight from images of transparent materials. They employed automatic semantic mask generation and partial convolution restoration techniques to effectively remove highlight regions, all while maintaining important local and semantic information in the image. Zhang et al. [15] recently introduced the NSH dataset for specular highlight removal, utilizing cross-polarization and the dichromatic reflection model to address the shortcomings of previous datasets such as SHIQ, SSHR, and PSD. Overall, single-image methods provide greater robustness and practicality than multi-image approaches. Despite advances driven by deep learning, challenges remain: convolutional networks struggle with capturing long-range dependencies, while Transformers often require excessive computation.

This trade-off between speed and removal quality motivates our work to develop more efficient architectures that better balance these factors and improve color and texture consistency.

III. Methodology

A. Overall

The general architecture of MM-SHR, an end-to-end multi-model architecture with a single image as input, is shown in Fig. 1. Its purpose is to enhance the speed of inference and consistency of color texture for tasks involving the removal of specular highlight. The network constructs a multi-stage U-shaped encoder-decoder structure by ingeniously fusing the global feature capture advantage of Transformer with the local detail extraction capability of CNN. MM-SHR effectively captures and processes the global context and detailed texture in images at various scales by utilizing the synergy of residual concatenation, upsampling, and downsampling modules. The framework offers a dependable solution for specular highlight removal tasks by accurately separating and restoring features in highlight and non-highlight regions using pixel-by-pixel feature accumulation and multipath feature fusion strategies.

The framework proposed in this paper consists of three key modules: the Omnidirectional Attention Integration Block (OAIBlock), the Adaptive Region-Aware Hybrid-Domain Dual Attention Convolutional Network (HDDAConv), and the Hybrid-Domain Cross-Attention Transformer (HDCTransformer).

B. Omnidirectional Attention Integration Block

Specular highlight often appear as localized highlight regions lacking detail, resulting in significant loss of local information. Relying solely on local detail features makes it challenging to reconstruct the original texture accurately. To address this, capturing broader contextual information becomes essential, especially at shallower network layers. While Transformer excel at capturing long-range dependencies, it introduce considerable memory overhead and slow down inference. To balance performance and efficiency, we propose an Omnidirectional Attention Integration Block (OAIBlock) based on convolutional operations. Positioned at the topmost layer of our MM-SHR network, the OAIBlock captures medium-range and long-range contextual information from high-resolution features. Simultaneously, it enhances local detail extraction and feature fusion without significantly increasing floating-point operations and inference time. This design ensures low-cost, high-quality extraction of low-level highlight features, laying a robust foundation for subsequent deep-level feature extraction and improving overall highlight removal performance. OAIBlock mainly realizes the coupling of medium-range and long-range features through Omnidirectional Pixel Shifting and Window Dividing Operations (Shift-Window). Global-Local Attention (GLA) and Context Anchor Attention (CAA) [37] are also utilized to efficiently extract multi-scale contextual information. And the highlight features of multiple attentions are adaptively fused by a learnable parameter α. The OAIBlock can be represented by Eq. (1):

Fig. 1. Overall architecture of Shift-Window Meets Dual Attention: A Multi-Model Architecture for Specular Highlight Removal

$$\begin{cases} F = SW(Input) \\ OAIBlock(F) = \alpha \times GLA(FP(F), F) \\ \qquad\qquad + (1-\alpha) \times CAA(FP(F), F) \end{cases} \tag{1}$$

where $Input \in R^{N\times C\times H\times W}$ represents the input features, $SW(\cdot)$ denotes the Omni-Directional Pixel Shifting and Window Dividing Operations (Shift-Window), $FP(\cdot)$ corresponds to the frequency processor, $GLA(\cdot)$ represents the Global-Local Attention mechanism, $CAA(\cdot)$ [37] denotes the Contextual Anchor Attention mechanism.

1) Shift-Window: Specular highlight in images often obscure or eliminate fine details within localized regions. To restore these areas, information from the areas surrounding the specular highlights is needed as a reference. Some existing approaches [20] [24] [38] use Transformer-based architectures to capture long-range dependencies between highlight areas and their surrounding features. While effective, these methods are computationally expensive when applied to high-resolution features, leading to higher inference costs. Other approaches combine pixel shifting with window partitioning, but these are often implemented within Transformers and yield limited performance gains. On the other hand, CNN-based architectures [19] [39] are efficient to train and run, but struggle with highlight removal due to their restricted receptive fields. To overcome these limitations, we introduce $SW(\cdot)$ into CNNs without altering their receptive field. This preserves the efficiency of convolution while enabling effective modeling of long-range dependencies. The implementation is illustrated in Fig. 2 and Eq. (2).

$$\begin{cases} \mathcal{S} = [Up, Down, Left, Right, \text{Left-Up}, \\ \qquad \text{Right-Up}, \text{Left-Down}, \text{Right-Down}] \\ SW(Input) = \sum_{i=1}^{C} \mathcal{D}(\mathcal{S}_{i \bmod 8}(Input)) \end{cases} \tag{2}$$

Given an input feature map $Input \in R^{N\times C\times H\times W}$, the i-th channel in the channel dimension C is cyclically shifted along the $\mathcal{S}_{i \bmod 8}(\cdot)$ direction. To ensure complete window partitioning along both H and W, the shifted feature maps are adaptively padded before division into windows of size $(H/E, W/E)$, where E denotes the window expansion factor. The resulting output $F \in R^{N\times CE^2\times \frac{H}{E}\times\frac{W}{E}}$ simultaneously encodes cross-window semantic correlations and local geometric details.

2) Frequency Processing: Analyzing high-frequency components in the frequency domain allows the special highlight removal algorithm to effectively capture edges, fine details, and the global distribution of light in highlighted regions. This approach addresses the challenge of perceiving global highlight patterns, which is difficult when working solely in the spatial domain. To enhance feature representation, we apply a frequency-domain transformation, denoted as $FP(\cdot)$, to the input features F. This process improves the model's ability to balance global highlight awareness with local detail preservation. The transformed features are then mapped back to the spatial domain for further processing. The specific implementation of $FP(\cdot)$ is presented in Eq. (3):

Fig. 2. Diagram of the Shift-Window process

$$FP(F) = Conv_{1\times1}\Big(Concat\Big(IFFT2D\Big(Conv_{1\times1}(BN(Re(FFT2D(F)))) + Re(FFT2D(F))\Big), F\Big)\Big) \tag{3}$$

where the operation $Concat(\cdot)$ combines the enhanced frequency domain features with the original spatial domain features, integrating both representations. $Re(FFT2D(\cdot))$ extracts key frequency information, while $IFFT2D(\cdot)$ transforms the processed frequency data back into the spatial domain. To ensure stability and consistency in the frequency feature distribution, $BN(\cdot)$ is applied for normalization.

3) Global-Local Attention: The $GLA(\cdot)$ mechanism is designed to integrate global context information with local detail features, addressing the loss of fine details in specular highlight removal. The specific implementation is formulated as shown in Eq. (4):

$$GLA(f, f_p) = SAB(CAB(f_p) \otimes f) \otimes f \tag{4}$$

where, $f \in R^{N\times C\cdot E^2\times\frac{H}{E}\times\frac{W}{E}}$ and $f_p \in R^{N\times C\cdot E^2\times\frac{H}{E}\times\frac{W}{E}}$ represent the input features, and f_p denote the corresponding frequency-domain features obtained through the $FP(\cdot)$ transformation. Channel Attention Block (CAB) and Spatial Attention Block (SAB) [40] are employed to capture the global distribution of highlight regions and enhance the model's ability to restore local details. The symbol $\otimes$ represents pixel-wise multiplication, enabling effective fusion of the attention-enhanced features.

4) Contextual Anchor Attention: Traditional 3×3 convolutions often struggle to capture critical pixel information around specular highlight in high-resolution images, resulting in the loss of fine texture details. This limitation arises from their inherently small receptive field, which prevents effective context modeling in these challenging regions. To overcome this, we introduce $CAA(\cdot)$ [37], designed to strengthen feature representation in the central highlight regions. CAA combines global average pooling to aggregate broader contextual information with 1D strip convolution, which preserves structural continuity along key spatial directions. This combination enhances the model's focus on crucial regions while maintaining the surrounding contextual integrity. This approach works in tandem with $SW(\cdot)$, improving the model's ability to capture contextual information over medium and long distances, ultimately supporting more effective highlight removal.

C. Adaptive Region-Aware Hybrid-Domain Dual Attention Convolutional Network

OAIBlock effectively captures medium-range and long-range contextual information through a comprehensive attention mechanism, allowing for initial localization and feature enhancement of highlight regions. However, it primarily focuses on coarse highlight identification and enhancement, lacking the ability to model degraded textures within these regions at a fine-grained level. To address this limitation, we propose the Adaptive Region-Aware Hybrid-Domain Dual Attention Convolutional Network (HDDAConv). Positioned between OAIBlock and HDCTransformer, this module enables region-adaptive feature decoupling and reconstruction through a dual-path synergetic architecture. By leveraging both spatial and frequency domain features, HDDAConv enhances detail representation in highlight regions and improves the separation of degraded textures from underlying real features. This leads to more effective highlight removal and better texture restoration.

As shown in Fig. 1, HDDAConv first applies a window division mechanism to transform the input feature $Input \in R^{N\times C\times H\times W}$ into a feature of size $N \times C \cdot E^2 \times \frac{H}{E} \times \frac{W}{E}$. Rather than employing pixel shifting as in OAIBlock, HDDAConv utilizes the Adaptive Hybrid Domain Dual Attention (AHDDA) mechanism to model spatial and frequency domain features at a fine-grained level. This enables precise capture of details across different scales and domains, improving the separation and restoration of degraded textures in highlight regions. The implementation of AHDDA is detailed in Eq. (5), where a cross-domain attention mechanism integrates spatial and frequency domain features. This approach enhances fine details while suppressing irrelevant noise, ensuring a more accurate reconstruction. As a result, HDDAConv effectively preserves the true features of highlight regions while restoring degraded textures with high precision.

$$AHDDA\left(f, f_p\right) = \beta \cdot CAB\left(f_p\right) \otimes f + (1-\beta) \cdot SAB\left(f_p\right) \otimes f \tag{5}$$

where $f \in R^{N\times C\cdot E^2\times \frac{H}{E}\times \frac{W}{E}}$ and $f_p \in R^{N\times C\cdot E^2\times \frac{H}{E}\times \frac{W}{E}}$ denote the input features, and f_p represents the frequency-domain transformation of f processed by the frequency domain processor $FP\left(\cdot\right)$. The learnable parameter β adaptively balances the contributions of channel attention $CAB\left(\cdot\right)$ and spatial attention $SAB\left(\cdot\right)$ to optimize feature.

Capturing dependencies between specular highlight regions and their surrounding features in high-resolution images remains challenging with dual attention alone. To overcome this limitation, HDDAConv develops the pixel bi-directional offset and windowing to extract fine highlight features and reconstruct textures more effectively. Additionally, it incorporates double-path contextual anchor attention (CAA) alongside AHDDA to further enhance feature representation. The detailed implementation is illustrated in Fig. 1.

D. Hybrid-Domain Cross-Attention Transformer

At the bottom of the UNet encoder, the feature map has been progressively refined, including high-level semantic representations. While existing Transformer-based methods, such as DHAN-SHR and UNet-Transformer, operate on raw-resolution or shallow features, they incur high computational costs when processing high-resolution images and are susceptible to redundant information. To enhance computational efficiency and improve global feature perception, we employ the Hybrid-Domain Cross-Attention Transformer (HDCTransformer) at downsampled scales of $H/4 \times W/4$ and $H/8 \times W/8$. This approach is grounded in two key principles:

1) Leveraging Deep Feature Semantics: As the network encodes the image, feature maps transition from low-level textures to abstract and high-level representations. These deeper features are more effective in capturing global structures and long-range dependencies. Applying attention mechanisms at this stage enhances both information exchange efficiency and accuracy.

2) Optimizing Computational Efficiency: The computational complexity of the self-attention mechanism is proportional to the square of the feature map size, i.e., $O(N^2)$. By downscaling the feature map to 1/4 or 1/8 of the original resolution, the complexity is reduced to $O(\frac{N^2}{16})$. This reduces computational resource consumption while maintaining efficient feature representation.

To improve the accuracy of highlight region identification, we integrate the $FP\left(\cdot\right)$ into the HDCTransformer. This processor refines the query matrix Q by incorporating prior information from the frequency domain. By transforming Q in this way, we establish a complementary relationship with the key K and value V matrices, which remain in the spatial domain. This integration of frequency and spatial information enhances the attention mechanism's ability to detect highlight regions more effectively. Additionally, the combination of these two domains provides cross-domain calibration for the attention weights used in Eq. (6). This leads to improved feature representation of highlight regions. The detailed implementation of the HDCTransformer is outlined below:

$$HDCTransformer\left(Q, K, V\right) \approx Softmax\left(\frac{Q \cdot K^T}{\delta}\right) \cdot V \tag{6}$$

where K and V are obtained by mapping the input feature tensor, defined as $Input \in R^{N\times 4C\times \frac{H}{4}\times \frac{W}{4}}$. The Q is generated by applying the frequency domain processor $FP\left(\cdot\right)$ to the same input. Dimensionality reduction decreases memory consumption during training and enhances inference efficiency, while preserving the HDCTransformer's ability to model features effectively. Next, the resulting features are reshaped into a tensor of shape $N \times Head \times Channel \times H \cdot W$, where N is the batch size, and $Head$ is the number of attention heads, calculated as $dim/16$. This structure enables the model to capture diverse information across multiple subspaces. Each attention head is assigned $dim/Head$ channels, where dim is either $4C$ or $8C$, depending on the configuration. To further enhance flexibility, we replace the standard fixed scaling factor with a learnable parameter δ, initialized to $dim/16$. This parameter is updated through backpropagation, allowing the model to adaptively control the distribution of attention weights. As a result, the model can dynamically adjust the

intensity of attention across different image regions, leading to more effective detail restoration in highlight areas.

E. Double Out

Our model is uniquely designed to output two complementary images in an end-to-end manner: one representing the highlight-free image Out_1, and the other the corresponding specular highlight residual Out_2. These two outputs sum pixel-wise to reconstruct the original input image $Input$ containing specular highlights, satisfying the following physical constraint Eq. (7):

$$Input = \sum_{c=1}^{3}\sum_{i=1}^{H}\sum_{j=1}^{W} Out_1(x_{c,i,j}) + Out_2(x_{c,i,j}) \tag{7}$$

A key challenge lies in the fact that datasets only contain paired highlight-free and highlight images, without explicit ground truth for the specular highlight residual. To address this, we implicitly extract the highlight residual by leveraging the physical property that the input image can be decomposed into a sum of diffuse (highlight-free) and specular (highlight residual) components. By enforcing the reconstruction constraint through the loss function Eq. (8):

$$\mathcal{L} = \mathcal{L}_{MSE}(Out_1 + Out_2, Input) \tag{8}$$

where $\mathcal{L}_{MSE}(\cdot,\cdot)$ denotes a pixel-wise loss function. This loss encourages the model to learn a meaningful separation between specular and specular components-free despite the absence of explicit supervision on the specular residue. By leveraging this end-to-end design, our method jointly models specular removal and residue extraction, enhancing restoration accuracy while maintaining end-to-end differentiability and training stability.

F. Loss Function

The goal of specular highlight removal is to generate a diffuse image from one that contains specular highlight. This process is a form of image-to-image translation, where the model learns to transform the input image into a more natural, highlight-free version. To achieve this, a loss function is used to quantify the difference between the ground truth and the inference result of specular highlight removal algorithm. It guides the model's training by encouraging adjustments that reduce this difference. An effective loss function should capture differences in pixel values, texture, edges, and high-level visual features to ensure the generated image is perceptually consistent with the ground truth. To address these requirements, we propose a composite loss function that combines three components: Mean Squared Error ($\mathcal{L}_{MSE}$) for pixel-wise accuracy, Structural Similarity Index Measure ($\mathcal{L}_{SSIM}$) for preserving image structure, and a semantic feature loss ($\mathcal{L}_{VGG}$) based on a pretrained neural network to capture higher-level visual characteristics. This combined approach is formalized in Eq. (9):

$$\begin{aligned}\mathcal{L} = \alpha \cdot \mathcal{L}_{MSE_1}(Out_1, GT) + \beta \cdot \mathcal{L}_{MSE_2}(Out_1 + Out_2, Input) \\ + \gamma \cdot \mathcal{L}_{SSIM} + \delta \cdot \mathcal{L}_{VGG}\end{aligned} \tag{9}$$

where the weights α, β, γ, and δ are set to 1, 1.5, 0.4, and 0.2, respectively. Out_1 refers to the generated image with specular highlight removal, while Out_2 represents the version with highlight retained. The term $Input$ denotes the original input image provided to the model. To evaluate high-level perceptual similarity, we use the semantic feature loss ($\mathcal{L}_{VGG}$).

This loss measures the difference between the generated image and the ground truth by comparing their representations in a feature space extracted using a pretrained VGG16 network. The specific formulation of this loss is provided in Eq. (10).

$$\mathcal{L}_{VGG} = \frac{1}{N}\sum_{l} \parallel \phi_l\left(\hat{I}\right) - \phi_l\left(I\right) \parallel_2^2,\ l \in 3, 8, 15 \tag{10}$$

where $\hat{I}$ represents the image generated by the model, I denotes the corresponding ground truth image, ϕ_l is the feature extraction function from the l layer of the VGG network, and N is the total number of layers used in computing the semantic feature loss.

IV. Experiments

A. Implementation Details

We implemented the specular highlight removal algorithm, MM-SHR, using the PyTorch framework and conducted all experiments on an Ubuntu 24.04 system. The hardware setup for these experiments included an AMD EPYC 7F72 24 core processor and an NVIDIA GeForce RTX 4090D GPU. For optimization, we used the AdamW optimizer with a batch size of 16. The model was trained for 100 epochs, with cosine annealing for learning rate scheduling. The initial learning rate was set to 1e-3, with a minimum value of 1e-5. All other parameters were set to their default values.

B. Evaluation Metrics

To assess the performance of our specular highlight removal method, we selected three widely used image quality assessment metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). PSNR is a traditional metric that measures image fidelity by calculating the ratio between the maximum possible pixel value and the mean squared error. While simple and computationally efficient, it does not always reflect perceived visual quality. SSIM improves on this by evaluating structural similarity between images, taking into account luminance, contrast, and local patterns. It better aligns with how the human visual system perceives image quality. LPIPS offers a more advanced, perceptually driven evaluation. It uses a pre-trained deep neural network to extract high-level features and compare differences at a semantic level, capturing subtle perceptual variations that traditional metrics may miss. Together, these three metrics provide a comprehensive evaluation of our method across multiple dimensions, including pixel-level accuracy, structural preservation, and perceptual consistency, offering a robust, quantitative basis for objective comparison.

TABLE I
QUANTITATIVE COMPARISON OF MM-SHR WITH SIX ADVANCED SPECULAR HIGHLIGHT REMOVAL METHODS ON SHIQ, PSD, AND NSH. THE TOP-PERFORMING RESULTS ARE HIGHLIGHTED IN BOLD RED, THE SECOND-BEST RESULTS IN BOLD BLACK, AND THE THIRD-BEST RESULTS ARE UNDERLINED.

Methods	SHIQ			PSD			NSH			Average		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
SLRR [41]	14.6890	0.6942	0.2703	13.3136	0.5329	0.2323	15.5497	0.5621	0.2137	14.5174	0.5964	0.2388
SpecularityNet [16]	32.1849	0.9581	**0.0456**	25.2063	0.8920	**0.0423**	40.5315	0.0059	0.0059	32.6409	0.6187	**0.0313**
TSHRNet [19]	32.7168	0.9681	0.0533	22.0323	0.8260	0.0782	36.3057	0.9571	0.0176	30.3516	0.9171	0.0497
HighlightRNet [15]	30.9728	0.9493	0.0695	**25.7563**	0.8916	0.0476	39.6003	0.9765	0.0067	32.1098	0.9391	0.0413
DHAN-SHR [24]	**34.6894**	**0.9774**	0.3667	25.1358	0.8918	0.0434	39.9630	0.9745	**0.0062**	33.2627	0.9479	0.1388
AHA [39]	34.1620	0.9699	0.0475	25.2260	**0.8979**	0.0477	**41.8759**	**0.9861**	0.0064	**33.7546**	**0.9513**	0.0339
MM-SHR(Ours)	**36.6312**	**0.9808**	**0.0274**	**25.7638**	**0.9034**	**0.0433**	**45.7816**	**0.9940**	**0.0020**	**36.0588**	**0.9594**	**0.0242**

C. Comparison with State-of-the-Art Methods

To comprehensively evaluate the effectiveness of the proposed MM-SHR framework, this paper conducts rigorous quantitative and qualitative comparative experiments on three benchmark datasets for specular highlight removal (SHIQ [17], PSD [16], NSH [15]), comparing it with six advanced methods (SLRR [41], SpecularityNet [16], TSHRNet [19], HighlightRNet [15], DHAN-SHR [24], AHA [39]). Importantly, the data used in these experiments were not part of the training datasets, ensuring an unbiased assessment. To maintain fairness, all comparison methods were re-trained on our team's hardware devices. We also optimized the hyperparameters based on the settings from the original papers to ensure the best possible performance. Furthermore, the SHIQ, PSD, and NSH datasets do not include ground truth intrinsic images required for the first stage of TSHRNet training. To address this, we adapted the TSHRNet architecture to meet the specific needs of this experiment.

1) Quantitative Comparison: Table I presents a quantitative comparison of MM-SHR with six state-of-the-art specular highlight removal methods. The results consistently demonstrate MM-SHR's superior performance across various datasets. In the SHIQ dataset, MM-SHR achieves a PSNR of 36.6312 dB, outperforming the next best method, DHAN-SHR, by 1.9418 dB (34.6894 dB). MM-SHR also achieves the highest SSIM (0.9808) and the lowest LPIPS (0.0274). Notably, DHAN-SHR's LPIPS on SHIQ is significantly higher at 0.3667, likely due to its model's sensitivity to complex reflectance variations, which leads to a loss of image details. This further emphasizes MM-SHR's ability to preserve image structure. In the NSH dataset, MM-SHR excels with a PSNR of 45.7816 dB, a 3.9057 dB improvement over the second-best method, AHA (41.8759 dB). MM-SHR also achieves a remarkable SSIM of 0.9940 and a minimal LPIPS of 0.0020. While MM-SHR shows strong performance in the PSD dataset as well, with a PSNR of 25.638 dB and an SSIM of 0.9034, its LPIPS (0.0433) is slightly higher than that of SpecialarityNet (0.0423). This small difference is likely due to limitations within the PSD dataset, including misalignment of image pairs, luminance discrepancies, and non-one-to-one correspondence, which can affect highlight removal. Despite this, MM-SHR's balanced performance in PSNR and SSIM, combined with its clear advantages on the SHIQ and NSH datasets, highlight its robust generalization capability. Its superior performance across all three datasets compared to the six other methods underscores MM-SHR's effectiveness and generalization capability in specular highlight removal. Furthermore, the computational cost of MM-SHR is analyzed, with the results presented in Table II.

TABLE II
A COMPARATIVE ANALYSIS OF FLOPS AND PARAMETERS WITH STATE-OF-THE-ART METHODS, WHERE THE INPUT FEATURE SIZE IS FIXED AT $F \in R^{1\times3\times256\times256}$

Methods	FLOPs(G)↓	Params(M)↓
SpecularityNet [16]	212.87	17.00
TSHRNet [19]	72.76	116.99
HighlightRNet [15]	250.84	48.37
DHAN-SHR [24]	**65.69**	**4.53**
AHA [39]	92.40	35.86
MM-SHR(Ours)	**18.03**	**16.09**

2) Qualitative Comparison: Fig. 3 presents a qualitative comparison among MM-SHR and six state-of-the-art methods for specular highlight removal. For a more detailed evaluation, we recommend enlarging the image. The figure demonstrates that, while previous methods achieve some level of success in removing specular highlight, they still exhibit notable limitations. For example, the SLRR method causes overall image darkening and leaves highlight residues. SpecularityNet struggles with significant highlight remnants, particularly on highly saturated materials (see the third and seventh rows). TSHRNet tends to distort the original image structure and texture (see the fourth and sixth rows). Both HighlightRNet and AHA show poor performance on transparent materials (see the second row), with HighlightRNet displaying more prominent highlight residues (see the first and seventh rows) and AHA suffering from severe visual artifacts (see the sixth row) and color distortion (see the seventh row). In contrast, our method, MM-SHR, provides superior visual results. Although DHAN-SHR also delivers strong visual effects, its use of a Transformer module leads to slower inference times, making it less suitable for practical applications. To further demonstrate MM-SHR's robustness and consistency in preserving visual texture, Fig. 4 shows the results of Zero-Shot tests. These re-

Fig. 3. Qualitative comparison of MM-SHR with leading state-of-the-art highlight removal methods.

Fig. 4. Zero-shot test results demonstrating the generalization capability of MM-SHR.

sults clearly highlight MM-SHR's ability to effectively remove specular highlight and its strong generalization capability.

D. Ablation Studies

To thoroughly evaluate the contribution of each key module and loss function in MM-SHR, we conducted a systematic ablation study on the SHIQ dataset. This experiment assesses how different module combinations affect three metrics: PSNR, SSIM, and LPIPS.

1) Fundamental module: To examine the effectiveness of each component, we designed two sets of comparison experiments under the same conditions of parameters and training methods. In the first set, we began with a baseline UNet architecture and progressively added key modules such as the OAIBlock and HDDAConv. After each addition, the network was retrained and evaluated. The results, shown in Table III, clearly demonstrate the performance improvements contributed by each module. In the second set of experiments, we started with the full MM-SHR architecture and selectively removed one module at a time, keeping all other components unchanged. The network was then retrained and evaluated. The results, summarized in Table IV, highlight the performance degradation caused by the absence of each individual module. Taken together, these results confirm that each module contributes positively to the overall performance. When fully integrated, MM-SHR achieves the highest performance across all three metrics, indicating that the architecture is well-structured and that the modules work in synergy to enhance highlight removal.

2) Loss Function: To assess the effectiveness of our composite loss function, we conducted a series of experiments by incrementally introducing additional loss components $\mathcal{L}_{MSE_2}$, $\mathcal{L}_{SSIM}$, and $\mathcal{L}_{VGG}$ on top of $\mathcal{L}_{MSE_1}$. Each loss term was

TABLE III
ABLATION RESULTS OF MM-SHR ON THE SHIQ DATASET.

Unet	OAIBlock	HDDAConv	HDCTransformer	Double Out	SHIQ		
					PSNR↑	SSIM↑	LPIPS↓
✓					33.4811	0.9742	0.0388
✓	✓				34.1417	0.9757	0.0368
✓		✓			33.6888	0.9743	0.0405
✓	✓	✓			34.5779	0.9764	0.0368
✓	✓	✓	✓		36.4637	0.9804	0.0283
✓	✓	✓	✓	✓	**36.6312**	**0.9808**	**0.0274**

TABLE IV
ABLATION RESULTS OF MM-SHR INDIVIDUAL COMPONENT ON THE SHIQ DATASET.

Without	SHIQ		
	PSNR↑	SSIM↑	LPIPS↓
OAIBlock	36.5231	0.9807	0.0279
HDDAConv	35.8800	0.9755	0.0372
HDCTransformer	33.3801	0.9678	0.0479
all	**36.6312**	**0.9808**	**0.0274**

TABLE V
ABLATION EXPERIMENTS FOR MM-SHR LOSS FUNCTION ON THE SHIQ DATASET.

$\mathcal{L}_{MSE_1}$	$\mathcal{L}_{MSE_2}$	$\mathcal{L}_{SSIM}$	$\mathcal{L}_{VGG}$	SHIQ		
				PSNR↑	SSIM↑	LPIPS↓
✓				34.944	0.9736	0.0400
✓	✓			35.8238	0.9758	0.0377
✓	✓	✓		36.6242	0.9806	0.0321
✓	✓	✓	✓	**36.6312**	**0.9808**	**0.0274**

assigned an appropriate weight to evaluate its individual contribution to highlight removal performance. As shown in Table V, MM-SHR achieves its best results only when all loss components are used together. This confirms the effectiveness and necessity of the composite loss function design.

V. CONCLUSION

We propose MM-SHR, a multi-model cooperative framework, to address two key challenges in specular highlight removal: the preservation of fine texture details and the balance between accuracy and efficiency. The framework adopts a hierarchical architecture that combines convolutional layers for detailed feature extraction with Transformer-based modules for global context integration, enabling both high-quality and efficient specular highlight removal. To enhance feature representation, we introduce two novel modules, OAIBlock and HDDAConv, based on frequency domain-guided, pixel shifting, and windowing mechanisms. These modules facilitate the effective coupling of high-resolution medium-range and long-range contextual features in the top layers of the network. In parallel, the HDCTransformer module is incorporated at the bottleneck layer to capture abstract global features, thereby improving the overall modeling capacity and accuracy of highlight suppression. Extensive evaluations on three benchmark datasets demonstrate that MM-SHR achieves strong performance and fast inference speed. Systematic ablation studies confirm the effectiveness and complementary roles of each module within the architecture.

REFERENCES

[1] G. Li, H. Lu, J. Li, Z. Li, Q. Li, X. Ren, and L. Zheng, "Alternating interaction fusion of image-point cloud for multi-modal 3d object detection," *Advanced Engineering Informatics*, vol. 65, p. 103370, 2025.

[2] J. Yang, L. Liu, and S. Z. Li, "Separating specular and diffuse reflection components in the hsi color space," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, June 2013.

[3] T. Yamamoto and A. Nakazawa, "General improvement method of specular component separation using high-emphasis filter and similarity function," *ITE Transactions on Media Technology and Applications*, vol. 7, no. 2, pp. 92–102, 2019.

[4] W. Feng, X. Cheng, X. Li, Q. Liu, and Z. Zhai, "Specular highlight removal based on dichromatic reflection model and priority-based adaptive direction with light field camera," *Optics and Lasers in Engineering*, vol. 172, p. 107856, 2024.

[5] W. Feng, J. Sun, Q. Liu, X. Li, D. Liu, and Z. Zhai, "Specular highlight removal of light field image combining dichromatic reflection with exemplar patch filling," *Optics and Lasers in Engineering*, vol. 178, p. 108175, 2024.

[6] W. Wu, L. Jin, B. Qi, G. Li, and J. Li, "Anti-noise light field depth measurement using specular highlight removal," *Optics and Lasers in Engineering*, vol. 176, p. 108079, 2024.

[7] S. A. Shafer, "Using color to separate reflection components," *Color Research & Application*, vol. 10, no. 4, pp. 210–218, 1985.

[8] Q. Yang, J. Tang, and N. Ahuja, "Efficient and robust specular highlight removal," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1304–1311, 2014.

[9] W. Feng, X. Cheng, J. Sun, Z. Xiong, and Z. Zhai, "Specular highlight removal and depth estimation based on polarization characteristics of light field," *Optics Communications*, vol. 537, p. 129467, 2023.

[10] S. Wen, Y. Zheng, and F. Lu, "Polarization guided specular reflection separation," *IEEE Transactions on Image Processing*, vol. 30, pp. 7280–7291, 2021.

[11] J. Hao, Y. Zhao, and Q. Peng, "A specular highlight removal algorithm for quality inspection of fresh fruits," *Remote Sensing*, vol. 14, no. 13, p. 3215, 2022.

[12] X. Wei, X. Xu, J. Zhang, and Y. Gong, "Specular highlight reduction with known surface geometry," *Computer Vision and Image Understanding*, vol. 168, pp. 132–144, 2018.

[13] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc., 2014.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[15] 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*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 10046–10054.

[16] 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*, vol. 24, pp. 3782–3793, 2021.

[17] 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)*, June 2021, pp. 7752–7761.

[18] W. Feng, X. Li, X. Cheng, H. Wang, Z. Xiong, and Z. Zhai, "Specular highlight removal of light field based on dichromatic reflection and total variation optimizations," *Optics and Lasers in Engineering*, vol. 151, p. 106939, 2022.

[19] 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)*, October 2023, pp. 12 857–12 865.

[20] Z. Wu, J. Guo, C. Zhuang, J. Xiao, D.-M. Yan, and X. Zhang, "Joint specular highlight detection and removal in single images via unet-transformer," *Computational Visual Media*, vol. 9, no. 1, pp. 141–154, 2023.

[21] J. Shen, H. Guan, S. Tao, K. Yan, F. Zhou, and F. Luo, "Specular highlight detection and removal based on dynamic association learning," in *Advances in Computer Graphics*. Springer Nature Switzerland, 2023, pp. 375–387.

[22] G. Huang, J. Yao, P. Huang, and L. Han, "A mutual enhancement framework for specular highlight detection and removal," in *Pattern Recognition and Computer Vision*. Springer Nature Singapore, 2023, pp. 457–468.

[23] A. Anwer, S. Ainouz, N. M. Saad, S. S. A. Ali, and F. Meriaudeau, "Specseg network for specular highlight detection and segmentation in real-world images," *Sensors*, vol. 22, no. 17, p. 6552, 2022.

[24] X. Guo, X. Chen, S. Luo, S. Wang, and C.-M. Pun, "Dual-hybrid attention network for specular highlight removal," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 10173–10181.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10 012–10 022.

[26] Y. Jin, R. Li, W. Yang, and R. T. Tan, "Estimating reflectance layer from a single image: Integrating reflectance guidance and shadow/specular aware learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, pp. 1069–1077, 2023.

[27] H. Chen, L. Li, and N. Yu, "Mask-guided joint single image specular highlight detection and removal," in *Pattern Recognition and Computer Vision*. Springer Nature Singapore, 2023, pp. 457–468.

[28] J. Xu, S. Liu, G. Chen, and Q. Liu, "Bifurcated convolutional network for specular highlight removal," *Optoelectronics Letters*, vol. 19, no. 12, pp. 756–761, 2023.

[29] Y. Lin, Y. Yang, Y. Jiang, X. Zhang, and P. Song, "Et-hdr: An efficient two-stage network for specular highlight detection and removal," in *Pattern Recognition*. Springer Nature Switzerland, 2023, pp. 273–287.

[30] G. Hu, Y. Zheng, H. Yan, G. Hua, and Y. Yan, "Mask-guided cycle-gan for specular highlight removal," *Pattern Recognition Letters*, vol. 161, pp. 108–114, 2022.

[31] J. W. Ha, K. K. Lee, J. S. Yoo, and J. O. Kim, "Deep highlight removal using temporal dark prior in high-speed domain," *IEEE Access*, vol. 11, pp. 20 136–20 149, 2023.

[32] T. Su, Y. Zhou, Y. Yu, and S. Du, "Highlight removal of multi-view facial images," *Sensors*, vol. 22, no. 17, p. 6656, 2022.

[33] M. Son, Y. Lee, and H. S. Chang, "Toward specular removal from natural images based on statistical reflection models," *IEEE Transactions on Image Processing*, vol. 29, pp. 4204–4218, 2020.

[34] H. Xu, Q. Li, and J. Chen, "Highlight removal from a single grayscale image using attentive gan," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 1988441, 2022.

[35] A. Anwer, S. Ainouz, N. M. Saad, S. S. A. Ali, and F. Meriaudeau, "Joint network for specular highlight detection and adversarial generation of specular-free images trained with polarimetric data," *Neurocomputing*, vol. 559, p. 126769, 2023.

[36] A. H. Madessa, J. Dong, Y. Gan, and F. Gao, "A deep learning approach for specular highlight removal from transmissive materials," *Expert Systems*, vol. 40, no. 2, p. e12598, 2023.

[37] M. M. Rahman, M. Munir, and R. Marculescu, "Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 11 769–11 779.

[38] Z. Shen, H. Xu, T. Luo, Y. Song, and Z. He, "Udaformer: Underwater image enhancement based on dual attention transformer," *Computers & Graphics*, vol. 111, pp. 77–88, 2023.

[39] K. Hu, Z. Huang, and X. Wang, "Highlight removal network based on an improved dichromatic reflection model," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 2645–2649.

[40] X. Cai, Q. Lai, Y. Wang, W. Wang, Z. Sun, and Y. Yao, "Poly kernel inception network for remote sensing detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 27 706–27 716.

[41] 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)*, September 2018.