Cross-Scale Collaborative Dual-Resolution Network for Shadow Removal

1st Wanzhi Hong
School of Cyber Security and Computer
Hebei University
Baoding, China
2023276501@qq.com

2nd Qingxuan Shi*
School of Cyber Security and Computer
Hebei University
Baoding, China
shiqingxuan@hbu.edu.cn

3rd Fang Yang
School of Cyber Security and Computer
Hebei University
Baoding, China
yangfang@hbu.edu.cn

4th Tong Wang
School of Cyber Security and Computer
Hebei University
Baoding, China
19971058644@163.com

5th Jing Zhang
School of Cyber Security and Computer
Hebei University
Baoding, China
1105295591@qq.com

6th Ming Cheng
School of Cyber Security and Computer
Hebei University
Baoding, China
549552087@qq.com

***Abstract*—Image shadow removal is inherently challenging due to its spatially non-uniform and context-dependent nature: degradation is local, while accurate correction requires reliable guidance from distant non-shadowed regions. Existing approaches struggle to balance this duality. Global or attention-based models often introduce irrelevant information that contaminates shadow regions, whereas purely local methods fail to fully exploit long-range contextual cues. To address this issue, we propose a Dual-Resolution Collaborative Encoding (DRCE) framework that explicitly models fine-grained structural details and global context through two interacting resolutions. DRCE employs a high-resolution pathway to preserve detailed spatial structures and a low-resolution pathway to capture stable global illumination context. Their interaction is enabled by a Cross-Scale Collaborative Fusion (CSCF) module, which performs content-adaptive cross-scale token interaction to selectively integrate complementary information. Extensive experiments on multiple public benchmarks demonstrate that the proposed method achieves strong quantitative performance and high visual quality.**

***Index Terms*—shadow removal, dual-resolution modeling, cross-scale feature fusion**

I. Introduction

Shadows, caused by the occlusion of direct light, lead to non-uniform illumination and color distortion across scenes. This pervasive degradation not only compromises visual quality but also severely hinders downstream vision tasks such as object detection [1], tracking [2], and scene relighting [3]. The core challenge of shadow removal lies in its inherent tension between localized restoration and global contextual reasoning. While the degradation is confined to shadow regions, the essential cues for accurate correction—true illumination and material color—must be inferred from non-shadowed areas that may be both distant and contextually heterogeneous. Existing learning-based shadow removal methods struggle to balance fine structural detail preservation with global illumination consistency.

Convolutional Neural Networks are effective at modeling local textures and boundaries, but their inherently limited receptive fields hinder long-range contextual reasoning, even when multi-scale aggregation is employed [11], [12]. Vision Transformers alleviate this limitation by enabling global information exchange through self-attention [15]. However, their high computational cost often necessitates windowed attention, which reintroduces locality constraints. More critically, standard self-attention performs indiscriminate feature mixing across all spatial positions, making it difficult to selectively control information flow between shadowed and non-shadowed regions. This often leads to feature contamination and artifacts such as blurred boundaries or residual color casts.

Recent works attempt to address these issues via structured or region-aware attention mechanisms. For example, HomoFormer [4] employs spatial shuffle and unshuffle operations to extend the effective receptive field of window-based attention, while DES3 [19] and regional decay attention [20] introduce adaptive weighting strategies to modulate cross-region interactions. Despite these advances, such methods still operate on single-scale feature representations, where fine-grained structural details and global illumination cues are mixed within the same attention space. While this overlooks the intrinsic multi-scale nature of shadow removal, it underscores a more fundamental insight: reliable shadow removal requires information from distinct spatial scales. Fine-grained details are crucial for preserving local structures like edges and textures, whereas broader, low-frequency cues govern global illumination and color consistency.

Motivated by this observation, we propose a Dual-Resolution Collaborative Encoding (DRCE) framework. Instead of relying on single-resolution attention, DRCE explicitly separates and coordinates detail-sensitive and context-stable representations through two parallel yet interacting pathways: a high-resolution branch that preserves spatial detail, and a complementary low-resolution branch that captures

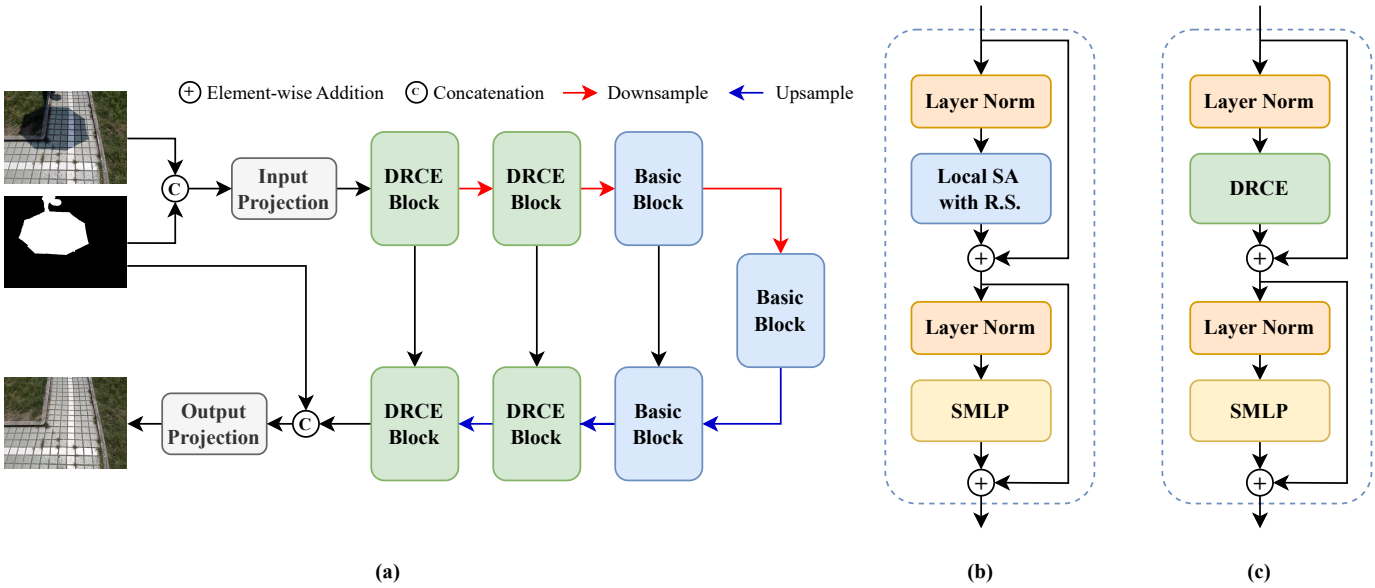


Fig. 1: An illustration of our proposed framework. (a) Overall network architecture. (b) Basic Block. (c) DRCE Block. The framework adopts a dual-resolution encoder–decoder architecture, where Basic Blocks and DRCE Blocks are stacked hierarchically to model local structural details and global illumination context.

stable, global contextual information.

The interaction between these two pathways is orchestrated by a Cross-Scale Collaborative Fusion (CSCF) module, which operates at the token level. CSCF groups high-resolution tokens with their corresponding low-resolution counterparts and employs a cross-scale token interaction mechanism to enable selective, semantically-aware information exchange. Fusion weights are dynamically predicted from aggregated token statistics, ensuring that contextual information guides local restoration only when relevant. This design effectively mitigates the feature contamination prevalent in conventional attention mechanisms, allowing for precise local correction to be informed by reliable, distant context.

Extensive experiments on multiple benchmark datasets demonstrate that our approach achieves state-of-the-art performance. The results confirm that by explicitly orchestrating cross-scale collaboration through CSCF, our model successfully resolves the core challenge of shadow removal: leveraging distant contextual guidance to perform precise, localized restoration without introducing structural or chromatic artifacts.

In summary, the main contributions of this paper are as follows:

- We propose a novel Dual-Resolution Collaborative Encoding (DRCE) framework that explicitly models high-resolution structural details and low-resolution global context through two interacting, functionally complementary pathways.
- We introduce a Cross-Scale Collaborative Fusion (CSCF) module that enables content-adaptive and selective interaction between cross-scale tokens, achieving a principled balance between local structural precision and global illumination consistency.
- Extensive experiments on multiple public benchmarks demonstrate that the proposed method achieves state-of-the-art or highly competitive performance in both quantitative metrics and visual quality, producing shadow-free images with sharper boundaries, improved structural integrity, and more natural color consistency.

II. Method

In this section, we present the proposed Dual-Resolution Collaborative Encoding (DRCE) framework for image shadow removal. The framework is designed to jointly capture fine-grained spatial details and global contextual cues through collaborative multi-scale encoding. We first introduce the overall encoder–decoder architecture, followed by the dual-resolution encoding mechanism that leverages the complementary roles of different spatial resolutions. Finally, we describe the cross-scale token interaction strategy used for effective multi-scale feature fusion.

A. Overall Framework

As illustrated in Fig. 1, our network adopts a U-shaped encoder–decoder architecture [5], [6] built upon the proposed Dual-Resolution Collaborative Encoding (DRCE) paradigm. Following common practice in shadow removal, the input shadow image and its corresponding shadow mask are concatenated along the channel dimension and projected into an initial feature space through a convolutional input projection layer. This fused representation serves as the basis for all subsequent hierarchical feature learning.

The encoder comprises multiple stages with progressively reduced spatial resolution. In the shallow stages, where shadow boundaries and local contrast variations are most prominent,

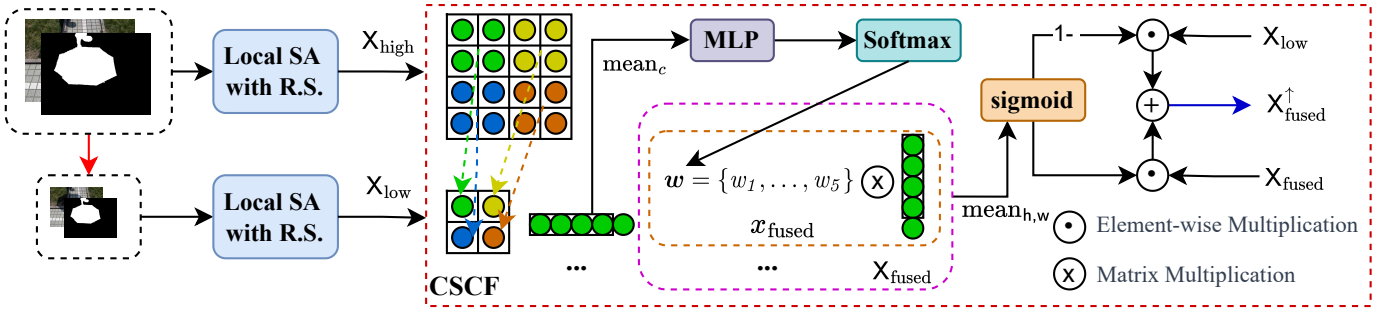


Fig. 2: Illustration of the Dual-Resolution Collaborative Encoding (DRCE) with the Cross-Scale Collaborative Fusion (CSCF). Within each DRCE, high-resolution and low-resolution branches interact through the proposed Cross-Scale Collaborative Fusion (CSCF) mechanism, enabling effective cross-scale information collaboration for shadow removal.

we employ DRCE Blocks to enhance feature representations with explicit cross-scale awareness. Each DRCE Block is a residual encoding unit that integrates a DRCE module and a Structure-enhanced MLP (SMLP). The DRCE module is responsible for modeling cross-resolution interactions by selectively incorporating complementary information from the other resolution stream, while the SMLP further refines the fused features to strengthen local structural representations. By stacking DRCE Blocks, the network repeatedly performs cross-resolution interaction within each DRCE module, enabling effective coordination between fine-grained structural details and global illumination context from the early stages of feature extraction.

As the encoder proceeds to deeper stages with lower spatial resolution, the features naturally exhibit stronger semantic abstraction. At this stage, we switch to Basic Blocks to efficiently refine representations without the overhead of explicit dual-resolution modeling. Each Basic Block follows a residual design and consists of a normalization layer, a local modeling module implemented via shuffled window-based self-attention [7], and a Structure-enhanced MLP (SMLP) [4]. This design stabilizes optimization and consolidates local representations while avoiding unnecessary cross-scale complexity in deeper feature spaces. Downsampling between adjacent encoder stages is performed using strided convolutions, progressively expanding the receptive field.

The decoder mirrors the encoder in a symmetric manner, progressively restoring spatial resolution through a series of upsampling stages. At each stage, decoder features are concatenated with their corresponding encoder features via skip connections, which is crucial for recovering high-frequency details, particularly around complex shadow boundaries. The fused features are then refined using a combination of DRCE Blocks and Basic Blocks, allowing cross-scale collaboration to guide the reconstruction process while preserving local structural fidelity and global illumination consistency.

Finally, an output projection layer maps the decoded features back to the RGB space. A global residual connection adds the network output to the original shadow image, encouraging the model to focus on learning residual shadow corrections rather than reconstructing the entire image. Through this hierarchical encoder–decoder design, empowered by the complementary use of DRCE Blocks and Basic Blocks, the proposed framework effectively balances local structural preservation and global illumination correction, leading to high-quality shadow removal results.

B. Dual-Resolution Collaborative Encoding

Given an input feature map $\mathsf{X} \in \mathbb{R}^{H\times W\times C}$, our DRCE module internally derives and processes complementary representations at two distinct spatial resolutions to facilitate subsequent cross-scale collaboration. Rather than employing completely disparate operators for each branch, we adopt a unified architectural template, allowing functional specialization to emerge naturally from the resolution difference itself. This design promotes both modularity and learning efficiency.

a) High-Resolution Branch: This branch operates directly on the original feature map X at full resolution. To break the locality constraints of standard window-based attention and enable long-range interactions while maintaining computational efficiency, we apply a global spatial shuffle operation $\mathcal{S}(\cdot)$, followed by window-based multi-head self-attention (WinAttn), and finally the inverse shuffle $\mathcal{S}^{-1}(\cdot)$:

$$\mathsf{X}_{\text{high}} = \mathcal{S}^{-1}\big(\text{WinAttn}\big(\mathcal{S}(\mathsf{X})\big)\big)\,. \tag{1}$$

By randomly permuting token positions before window partitioning, the shuffle operation allows distant tokens to be reassigned into the same local window, thereby enabling the high-resolution branch to capture fine-grained spatial details while also modeling long-range dependencies. This capability is particularly important for accurately representing shadow boundaries and local texture variations.

b) Low-Resolution Branch: The low-resolution branch aims to derive a compact and semantically stable representation by abstracting away fine spatial details. It first downsamples the input feature map X by a factor of two using average pooling:

$$\tilde{\mathsf{X}}_{\text{low}} = \text{AvgPool}(\mathsf{X}). \tag{2}$$

The same shuffle-based window attention pipeline is then applied to the downsampled feature map:

$$\mathsf{X}_{\text{low}} = \mathcal{S}^{-1}\Big(\text{WinAttn}\big(\mathcal{S}(\tilde{\mathsf{X}}_{\text{low}})\big)\Big)\,. \tag{3}$$

Due to the reduced spatial resolution, each token in X_{low} corresponds to a larger receptive field in the original image. As a result, this branch is naturally suited to modeling global, low-frequency contextual information, such as overall illumination trends and coherent color distributions, which provides stable semantic guidance for shadow region restoration.

c) Resolution-Induced Functional Complementarity: Although both resolution streams share an identical sequence of operations—namely shuffle, window-based attention, and inverse shuffle—their functional behaviors diverge significantly due to the difference in input resolution. The high-resolution branch becomes specialized in capturing local, high-frequency signals such as textures, edges, and precise shadow boundaries. In contrast, the low-resolution branch focuses on integrating broader, smoother contextual cues. This resolution-induced functional complementarity, yielding detail-sensitive and context-stable representations, forms the foundation of our collaborative encoding strategy. Effectively fusing these complementary representations is the key challenge addressed by the proposed Cross-Scale Collaborative Fusion (CSCF) module, which is described in the following section.

C. Cross-Scale Collaborative Fusion

To effectively integrate complementary information from high-resolution and low-resolution representations, we propose the Cross-Scale Collaborative Fusion (CSCF) module. As illustrated in Fig. 2, CSCF performs dynamic, content-aware fusion by emphasizing fine-grained structural cues from high-resolution features in detail-rich regions, while leveraging stable global context from low-resolution features in homogeneous or heavily shadowed areas. This design enables accurate reconstruction across shadow boundaries while maintaining both local precision and global consistency.

The core strategy of CSCF consists of cross-scale token grouping, token-level adaptive weighting mechanism, and channel-level gating. For each position in the low-resolution feature map, four high-resolution features with different spatial offsets corresponding to the same low-resolution position are collected:

$$X^*_{\text{high}} = \left\{ \boldsymbol{x}^{0,0}_{\text{high}},\, \boldsymbol{x}^{1,0}_{\text{high}},\, \boldsymbol{x}^{0,1}_{\text{high}},\, \boldsymbol{x}^{1,1}_{\text{high}} \right\}, \tag{4}$$

where the superscript $(\cdot,\cdot)$ indexes spatial offsets within the corresponding high-resolution neighborhood. The corresponding low-resolution feature at the same spatial location is denoted as $\boldsymbol{x}^*_{\text{low}}$. These features are concatenated to form a five-token group:

$$T = \text{Concat}\big(X^*_{\text{high}},\, \boldsymbol{x}^*_{\text{low}}\big), \tag{5}$$

which jointly represents fine-grained local details and coarse global context.

To adaptively fuse these tokens, CSCF estimates token importance based on channel-wise statistics by first averaging each token along the channel dimension and then applying a lightweight MLP followed by a softmax operation:

$$\boldsymbol{w} = \text{Softmax}(\text{MLP}(\text{mean}_c(T))). \tag{6}$$

For each spatial location, let $T = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_3, \boldsymbol{t}_4, \boldsymbol{t}_5\}$ denote the five tokens, where $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_4$ correspond to the four high-resolution tokens sampled at different spatial offsets, and $\mathbf{t}_5$ denotes the associated low-resolution token. Each token $\boldsymbol{t}_i \in \mathbb{R}^C$ represents a C-channel feature vector, and the predicted weights $\boldsymbol{w} = \{w_1, \ldots, w_5\}$ satisfy $w_i \geq 0$ and $\sum_{i=1}^{5} w_i = 1$. The token-level fused feature at each spatial location (h, w) is obtained via weighted aggregation:

$$\boldsymbol{x}_{\text{fused}} = \sum_{i=1}^{5} w_i \cdot \boldsymbol{t}_i. \tag{7}$$

Stacking $\boldsymbol{x}_{\text{fused}}$ over all spatial locations yields the fused feature map $\mathsf{X}_{\text{fused}} \in \mathbb{R}^{H \times W \times C}$. To further suppress potentially noisy channels and stabilize cross-scale fusion, CSCF employs a lightweight channel-level gating mechanism. The gate is computed by applying a sigmoid function to the spatially averaged fused feature map:

$$\boldsymbol{g} = \sigma(\text{mean}_{\mathsf{h},\mathsf{w}}(\mathsf{X}_{\text{fused}})), \tag{8}$$

where $\text{mean}_{\mathsf{h},\mathsf{w}}(\cdot)$ denotes averaging over spatial dimensions. The gated fusion output is then obtained by adaptively combining the fused feature with the original low-resolution representation:

$$\tilde{\mathsf{X}}_{\text{fused}} = \boldsymbol{g} \odot \mathsf{X}_{\text{fused}} + (1 - \boldsymbol{g}) \odot \mathsf{X}_{\text{low}}, \tag{9}$$

where $\odot$ denotes element-wise multiplication. Since the fused representation resides at the low-resolution scale, CSCF upsamples it to match the spatial resolution of the block input:

$$\mathsf{X}^{\uparrow}_{\text{fused}} = \text{Upsample}(\tilde{\mathsf{X}}_{\text{fused}}). \tag{10}$$

Through this design, CSCF dynamically balances local precision and global consistency by performing content-aware, token-level fusion across scales, while the channel-level gate further mitigates cross-scale interference. By selectively enhancing shadow boundary details using high-resolution cues and relying on low-resolution features for stable illumination and semantic guidance, CSCF helps reduce boundary artifacts and color inconsistency in shadowed regions.

III. Experiments

A. Experimental Settings

All experiments are implemented using PyTorch and conducted on a workstation equipped with a single NVIDIA GeForce RTX 4090 GPU. The network is optimized using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We initialize the learning rate to 2×10^{-4} and gradually decay it to zero using a cosine annealing schedule. Weight decay is applied for regularization to alleviate overfitting. During training, input shadow images and their corresponding masks are resized to a fixed resolution and randomly augmented using rotations as well as horizontal and vertical flips. Unless otherwise specified, the model is trained for *600* epochs with a batch size of *4*.

TABLE I: Comparison with SOTA methods on the SRD dataset. The best result is in **bold**, and the second-best is underlined.

Method	Shadow			Non-Shadow			All		
	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑
DHAN(Cun et al. 2020) [11]	8.94	33.67	0.978	4.80	34.79	0.979	5.67	30.51	0.949
AEF(Fu et al. 2021) [12]	7.97	32.05	0.955	5.30	31.75	0.939	6.14	28.26	0.866
BMNet(Zhu et al. 2022) [13]	6.61	35.05	0.981	3.61	36.02	0.982	4.46	31.69	0.956
SGNet(Wan et al. 2022) [14]	6.52	33.44	0.968	3.14	37.18	0.982	4.24	31.35	0.934
ShadowFormer(Guo et al. 2023a) [15]	5.90	36.91	0.982	3.44	36.22	0.983	4.04	32.90	0.957
ShadowDiffusion(Guo et al. 2023b) [16]	4.98	38.72	0.987	3.44	37.78	0.985	3.63	34.73	0.970
Inpaint4(Li et al. 2023) [17]	5.39	35.70	0.974	3.14	37.40	0.983	3.89	32.90	0.943
RRLNet(Liu et al. 2024) [18]	5.49	36.51	0.983	3.00	37.71	0.986	3.66	33.48	0.967
DeS3(Jin et al. 2024) [19]	5.88	37.45	0.984	2.83	38.12	0.988	3.72	34.11	0.968
HomoFormer(Xiao et al. 2024a) [4]	4.25	38.81	0.987	2.85	39.45	0.988	3.33	35.37	0.972
FW-Former(Zhu et al. 2025) [20]	4.48	38.61	**0.989**	2.91	38.35	**0.992**	3.29	34.88	**0.975**
Ours	**4.07**	**38.96**	0.985	**2.45**	**40.48**	**0.992**	**3.00**	**35.90**	0.972

TABLE II: Comparison with SOTA methods on the ISTD+ dataset. The best result is in **bold**, and the second-best is underlined.

Method	Shadow			Non-Shadow			All		
	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑
DHAN(Cun et al. 2020) [11]	9.60	32.98	0.987	7.44	27.12	0.973	7.80	25.65	0.955
AEF(Fu et al. 2021) [12]	6.64	36.18	0.977	3.75	31.31	0.884	4.20	29.59	0.849
BMNet(Zhu et al. 2022) [13]	6.24	37.37	0.990	2.46	37.85	0.984	3.02	33.95	0.967
SGNet(Wan et al. 2022) [14]	6.46	36.91	0.989	2.95	35.47	0.976	3.45	32.46	0.956
ShadowFormer(Guo et al. 2023a) [15]	5.34	39.54	0.992	2.34	38.72	0.984	2.81	35.44	0.972
ShadowDiffusion(Guo et al. 2023b) [16]	4.97	39.69	0.992	2.28	38.89	**0.987**	2.72	35.67	**0.975**
Inpaint4(Li et al 2023) [17]	6.12	38.09	0.989	2.92	36.95	0.977	3.43	33.81	0.960
RRLNet(Liu et al. 2024) [18]	5.69	38.04	0.990	2.31	39.15	0.984	2.87	34.96	0.968
DeS3(Jin et al. 2024) [19]	6.57	36.38	0.988	3.45	34.00	0.966	3.98	30.97	0.946
HomoFormer(Xiao et al. 2024a) [4]	4.92	39.51	0.991	2.27	38.65	0.982	2.68	35.32	0.970
FW-Former(Zhu et al. 2025) [20]	4.71	40.11	**0.993**	**2.15**	**39.21**	**0.987**	2.58	35.84	**0.975**
Ours	**4.63**	**40.22**	0.992	2.19	39.08	0.985	**2.56**	**35.92**	0.973

B. Datasets

We evaluate our method on two standard single-image shadow removal benchmarks: SRD [8] and ISTD+ [9]. The SRD dataset [8] contains 3,088 paired shadow and shadow-free images, with 2,680 images for training and 408 for testing. Since SRD does not provide ground-truth shadow masks, we follow prior works and use the DHAN-predicted masks [11] for both training and evaluation. The ISTD+ dataset [9] is an improved version of ISTD [10], in which illumination inconsistencies in the original annotations are corrected. It consists of 1,870 triplets (shadow image, shadow-free image, and shadow mask), split into 1,330 training and 540 testing samples. The availability of accurate shadow masks in ISTD+ provides reliable supervision for both shadow boundary modeling and appearance restoration.

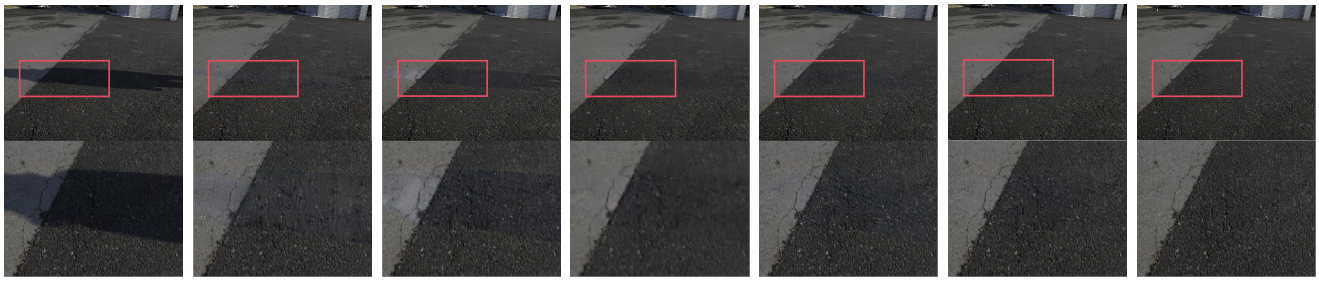
C. Evaluation Metrics

To quantitatively evaluate different methods, following the previous methods [12], [21], [22], all predicted shadow-free images and their corresponding ground-truth images are resized to 256×256 for fair comparison. We adopt three metrics: root mean square error (RMSE) in the LAB color space, Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) in the RGB space. Lower RMSE and higher PSNR/SSIM indicate better restoration. These metrics jointly reflect color accuracy, signal fidelity, and structural consistency.

D. Comparison with State-of-the-Art

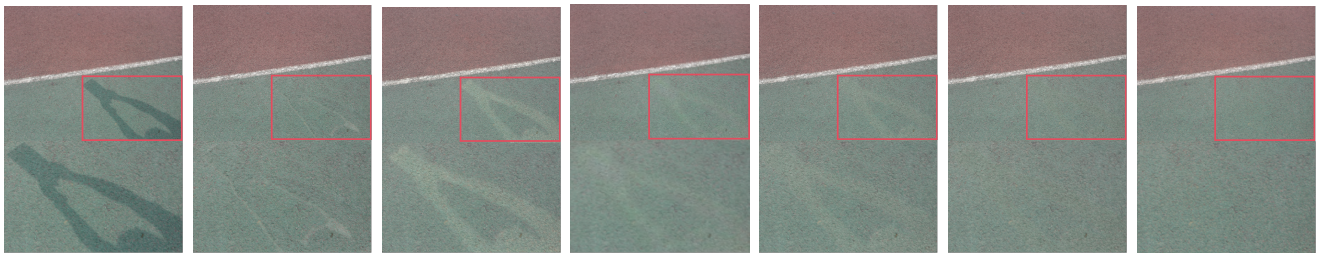
We compare our proposed method with representative state-of-the-art shadow removal approaches, including DHAN [11], AEF [12], BMNet [13], SGNet [14], ShadowFormer [15], ShadowDiffusion [16], Inpaint4 [17], RRLNet [18], DES3 [19], HomoFormer [4], and FW-Former [20]. All results are obtained from official publications or publicly released code.

*a) **Quantitative Evaluation**:* The quantitative comparisons on the SRD [8] and ISTD+ [9] datasets are summarized



(a) Input (b) SGNet (c) BMNet (d) Inpaint4 (e) HomoFormer (f) Ours (g) GT

Fig. 3: Visual comparison on the SRD dataset. First row: full-resolution images. Second row: zoomed-in regions.



(a) Input (b) SGNet (c) BMNet (d) Inpaint4 (e) HomoFormer (f) Ours (g) GT

Fig. 4: Visual comparison on the ISTD+ dataset. First row: full-resolution images. Second row: zoomed-in regions.

in Tab. I and Tab. II, respectively. Our method consistently achieves the best or highly competitive performance across all evaluation metrics.

On the SRD dataset [8], our method achieves the best overall performance in the *All* region, with an RMSE of 3.00 and a PSNR of 35.90 dB. Compared with the second-best method, our approach reduces RMSE by 0.29 and improves PSNR by 0.53 dB, demonstrating superior global restoration quality. In shadow regions, our method also achieves the lowest RMSE (4.07) and the highest PSNR (38.96 dB), indicating more accurate correction in challenging shadow areas.

On the ISTD+ dataset [9], our method again delivers the best overall performance. For the *All* region, it achieves an RMSE of 2.56 and a PSNR of 35.92 dB, surpassing the second-best FW-Former [20] by 0.02 in RMSE and 0.08 dB in PSNR. Although the numerical margins are relatively small, they consistently favor our method, reflecting its stable advantage across different datasets and evaluation settings. These improvements can be attributed to the dual-resolution collaborative encoding and the content-adaptive cross-scale fusion mechanism, which effectively integrates fine-grained structural details and global illumination cues.

*b) **Qualitative Evaluation**:* Qualitative comparisons on the SRD [8] and ISTD+ [9] datasets are shown in Fig. 3 and Fig. 4. From the visual results, our method produces cleaner shadow removal and smoother transitions across shadow boundaries compared with existing approaches. Methods such as SGNet [14] and BMNet [13] often leave residual dark traces or over-smooth local structures, while diffusion-based or transformer-based models may introduce slight color shifts near shadow edges. In contrast, our results exhibit more accurate illumination recovery and better color consistency, especially in regions with complex textures or sharp shadow boundaries. The zoomed-in regions highlight that our method preserves fine details while effectively removing shadow artifacts, leading to visually more natural and coherent shadow-free images.

E. Ablation Study

We conduct ablation studies on the ISTD+ dataset [9] to evaluate the effectiveness of the proposed Dual-Resolution Collaborative Encoding (DRCE) and Cross-Scale Collaborative Fusion (CSCF). As summarized in Tab. III, we progressively analyze the impact of introducing dual-resolution encoding and advanced cross-scale fusion strategies.

The first variant (w/o DRCE) removes the low-resolution branch and retains only the high-resolution pathway, thereby disabling DRCE and relying solely on single-scale feature encoding. This setting leads to noticeable performance degradation, particularly in shadow regions, where RMSE increases from 4.63 to 4.92 and PSNR drops from 40.22 dB to 39.51 dB. The results indicate that modeling global illumination and semantic context through a complementary low-resolution branch is essential for accurate shadow restoration.

To further assess the role of cross-scale interaction, the second variant (w/o CSCF) employs both high-resolution and low-resolution branches but replaces CSCF with a simple linear fusion strategy that directly averages multi-scale features. As shown in Tab. III, linear fusion causes the PSNR on the full image to decrease from 35.92 dB to 35.78 dB and RMSE to increase from 2.56 to 2.60, with more pronounced degradation observed in shadow regions. This suggests that naive linear

TABLE III: Ablation study on the ISTD+ dataset.

Method \ Region	Shadow (S)			Non-Shadow (NS)			All (ALL)		
	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑
w/o DRCE	4.92	39.51	0.991	2.27	38.65	0.982	2.68	35.32	0.970
w/o CSCF	4.75	40.03	0.992	2.21	39.01	0.984	2.60	35.78	0.972
Ours (default)	4.63	40.22	0.992	2.19	39.08	0.985	2.56	35.92	0.973

fusion is insufficient to fully exploit the complementary nature of cross-scale representations.

By contrast, the full model integrates dual-resolution features through the proposed CSCF module, which enables content-aware and selective cross-scale collaboration. This design consistently achieves the best performance across all regions.

IV. Conclusion

In this paper, we address the core challenge in shadow removal: the conflict between precise local correction within shadows and the need for reliable global context from non-shadow regions. We propose a novel Dual-Resolution Collaborative Encoding (DRCE) framework that explicitly models high-resolution spatial details and low-resolution contextual information in separate yet interacting pathways. The interaction is orchestrated by a Cross-Scale Collaborative Fusion (CSCF) module, which enables semantically-aware, artifact-free integration through cross-scale token attention and adaptive fusion. Extensive experiments on multiple benchmarks demonstrate that our approach achieves state-of-the-art performance, delivering visually superior results with enhanced structural integrity and color fidelity. This work validates the efficacy of explicit cross-scale modeling and collaborative fusion for shadow removal.

References

[1] J. Li, Q. Shi, F. Yang, S. Fu, and F. Hou, "MCRO-YOLO: An evolved version of YOLO for small object detection," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2025.

[2] C. Li, N. Zhao, Y. Lu, C. Zhu, and J. Tang, "Weighted sparse representation regularized graph learning for RGB-T object tracking," in *Proceedings of the 25th ACM International Conference on Multimedia (MM)*, pp. 1856–1864, 2017.

[3] J. Wang, J. Liu, X. Sun, K. K. Singh, Z. Shu, H. Zhang, J. Yang, N. Zhao, T. Y. Wang, S. S. Chen, et al., "Comprehensive relighting: Generalizable and consistent monocular human relighting and harmonization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 380–390, 2025.

[4] J. Xiao, X. Fu, Y. Zhu, D. Li, J. Huang, K. Zhu, and Z.-J. Zha, "HomoFormer: Homogenized transformer for image shadow removal," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25617–25626, 2024.

[5] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1125–1134, 2017.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

[7] J. Xiao, X. Fu, M. Zhou, H. Liu, and Z.-J. Zha, "Random shuffle transformer for image restoration," in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 38039–38058, 2023.

[8] L. Qu, J. Tian, S. He, Y. Tang, and R. W. H. Lau, "DeshadowNet: A multi-context embedding deep network for shadow removal," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4067–4075, 2017.

[9] H. Le and D. Samaras, "Shadow removal via shadow image decomposition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8578–8587, 2019.

[10] J. Wang, X. Li, and J. Yang, "Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1788–1797, 2018.

[11] X. Cun, C.-M. Pun, and C. Shi, "Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting GAN," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 34, pp. 10680–10687, 2020.

[12] L. Fu, C. Zhou, Q. Guo, F. Juefei-Xu, H. Yu, W. Feng, Y. Liu, and S. Wang, "Auto-exposure fusion for single image shadow removal," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10571–10580, 2021.

[13] Y. Zhu, J. Huang, X. Fu, F. Zhao, Q. Sun, and Z. J. Zha, "Bijective mapping network for shadow removal," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5627–5636, 2022.

[14] J. Wan, H. Yin, Z. Wu, X. Wu, Y. Liu, and S. Wang, "Style-guided shadow removal," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 361–378, 2022.

[15] L. Guo, S. Huang, D. Liu, H. Cheng, and B. Wen, "ShadowFormer: Global context helps shadow removal," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 37, pp. 710–718, 2023.

[16] L. Guo, C. Wang, W. Yang, S. Huang, Y. Wang, H. Pfister, and B. Wen, "ShadowDiffusion: When degradation prior meets diffusion model for shadow removal," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14049–14058, 2023.

[17] X. Li, Q. Guo, R. Abdelfattah, D. Lin, W. Feng, I. Tsang, and S. Wang, "Leveraging inpainting for single-image shadow removal," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13055–13064, 2023.

[18] Y. Liu, Z. Ke, K. Xu, F. Liu, Z. Wang, and R. W. H. Lau, "Recasting regional lighting for shadow removal," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 38, pp. 3810–3818, 2024.

[19] Y. Jin, W. Ye, W. Yang, Y. Yuan, and R. T. Tan, "DES3: Adaptive attention-driven self and soft shadow removal using ViT similarity," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 38, pp. 2634–2642, 2024.

[20] X. Zhu, C.-O. Chow, and J. H. Chuah, "Regional decay attention for image shadow removal," *Journal of Visual Communication and Image Representation*, p. 104694, 2025.

[21] R. Guo, Q. Dai, and D. Hoiem, "Paired regions for shadow detection and removal," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 35, no. 12, pp. 2956–2967, 2013.

[22] Y. Jin, A. Sharma, and R. T. Tan, "DC-ShadowNet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5027–5036, 2021.