International Journal of Advanced AI Applications
Online ISSN: 3104-9338
Print ISSN: 3104-932X

Advanced Unpaired Shadow Manipulation via Adversarial Generation and Image Partitioning

Shangan Zhou*

College of Physics and Electronic Information Engineering, Zhejiang Normal University, China

Received: July 15, 2026

Revised: July 20, 2026

Accepted: July 25, 2026

Published online: July 26, 2026

To appear in: *International Journal of Advanced AI Applications,* Vol. 2, No. 8 (Aughst 2026)

* Corresponding Author: Shangan Zhou (3217432128@qq.com)

Abstract. Accurate shadow processing is critical for perceptual realism in image synthesis, yet current methods often rely heavily on paired data. We propose two adversarial frameworks for shadow manipulation using unpaired data. For generation, a two-stage CycleGAN with attention predicts masks and synthesizes high-fidelity shadows without paired images. For removal, a weakly-supervised method uses patch extraction to construct intra-image training pairs and cycle-consistency to eliminate shadows without ground-truth unshaded images. Experiments on DESOBA and ISTD datasets demonstrate our methods achieve superior or competitive performance.

Keywords: *Deep Learning; Shadow Generation; Shadow Removal; Adversarial Networks; Weakly Supervised Learning*

1. Introduction

Image synthesis is a long-standing research domain involving the fusion of multiple images or their components to create novel scenes, leveraging principles from computer graphics and computer vision. With the paradigm shift driven by deep learning, the ability to seamlessly insert or remove objects in images has garnered significant attention. The visual success of these operations hinges critically on the accurate manipulation of shadows.

Shadows convey essential cues about object geometry, surface illumination, and light direction. Shadow generation aims to synthesize a plausible cast shadow for a newly inserted object onto a background scene, while shadow removal aims to eliminate unwanted shadows to restore underlying textures and colors for tasks like object detection and tracking. Furthermore, the acquisition of large-scale, well-annotated datasets for shadow manipulation remains a practical bottleneck. In the broader field of deep learning, supervised classification and recognition tasks have greatly benefited from massive labelled datasets [21-22]. However, obtaining perfectly paired shadow and non-shadow images under consistent illumination in

natural environments is exceptionally difficult, limiting the scalability of fully supervised models [48]. While traditional texture synthesis and image inpainting techniques [23-24] offer some foundational guidance for filling missing regions, generating accurate structural shadows requires specialized architectures that can reason about lighting and geometry [25].

Motivation and Challenges.Conventional shadow generation methods often rely on physical rendering and inverse lighting, requiring exhaustive inputs such as 3D geometry, material reflectance, and explicit light source parameters [44]. These methods are labor-intensive and difficult to automate. Deep learning approaches, particularly GANs, have made strides; however, many methods, such as ShadowGAN [6], are constrained by the necessity for strictly paired (shadow vs. non-shadow) training data, which is difficult and expensive to acquire in natural scenes. Similarly, while some semi-supervised methods like Mask-ShadowGAN [4] reduce data constraints, they still require unshaded images or suffer from inconsistency when background styles differ significantly from the training domain. This issue is exacerbated by the fact that variations in environmental factors—such as weather, season, and time of day—are often uncorrelated in paired datasets [46,49].

Furthermore, existing approaches often fail to fully exploit the rich illumination information present within a single image. For instance, the intricate relationship between a background object and its cast shadow can serve as a powerful reference point for generating an accurate shadow for a new object introduced within the same scene. Many recent advancements in the field have increasingly moved toward developing more robust context-aware models. For example, semantic segmentation networks [23,25] and recurrent attention architectures [42] have been shown to be particularly useful in effectively extracting spatial contexts necessary for shadow detection. Meanwhile, the integration of Transformer-based architectures [43] has demonstrated significant promise in enhancing global feature dependencies for various shadow detection tasks. We hypothesize that by leveraging unpaired data along with intrinsic intra-image correlations, we can notably reduce the dependency on expensive paired datasets, thus making the process more efficient and accessible.

Contributions. This work addresses these limitations through the following contributions:

(1) A Two-Stage Shadow Generation Method (SGIRE): We propose SGIRE, a framework capable of generating high-fidelity shadows for virtual objects in real-world backgrounds without requiring paired training data. This is achieved by explicitly learning the mapping between background objects and their shadows via an attention-guided SMG, followed by an adversarial CycleGAN module (CGN) and an Image Enhancement Network (IEN).

(2) A Weakly Supervised Shadow Removal Method (SRIRE): We introduce SRIRE, which eliminates the need for unshaded images during training. By utilizing a Patch Image Generator (PIG), it automatically constructs pair-wise training samples from a single shadow image. A CycleGAN structure refines the removal process, and we incorporate an ICF module for adaptive luminance compensation during inference, ensuring seamless background integration.

(3) Robust Experimental Validation: We conduct a comprehensive evaluation of our methods on both the DESOBA and ISTD datasets. The results clearly demonstrate that our approaches achieve superior or, at the very least, comparable performance to leading state-of-the-art techniques in terms of RMSE and SSIM metrics. Furthermore, we observe significant improvements in visual coherence and naturalness, which highlights the effectiveness of our methods.

2. Methodologies

Our frameworks are meticulously built upon standard deep learning backbones, specifically adopting the CycleGAN architecture [2] to facilitate unpaired domain translation. To further enhance generation quality, we integrate a sophisticated UNet-based encoder-decoder structure that employs skip connections for effective feature extraction and accurate reconstruction. This is complemented by a PatchGAN discriminator [19] and attention mechanisms (CBAM) to focus on maintaining local texture consistency and addressing critical spatial relationships within the images. Furthermore, we utilize pre-trained VGG networks [18] for perceptual loss calculations, which is a common practice employed to preserve high-frequency details in image translation tasks [3]. Distinct from these established baselines, our proposed methods explicitly decouple the shadow mask prediction from the image generation process and utilize innovative intra-image patch extraction strategies to effectively tackle the data scarcity issue. The following sections will provide a comprehensive and detailed analysis of our two novel frameworks: SGIRE for shadow generation and SRIRE for shadow removal, highlighting their unique contributions to the field.

2.1. SGIRE: Shadow Generation via Cycle-Consistent Adversarial Learning

SGIRE has been specifically designed to effectively synthesize realistic shadows for target objects that are inserted into natural scenes, all while relying solely on unpaired data. The overall architecture of this innovative system is illustrated in detail in Figure 1, showcasing its unique approach and methodology.

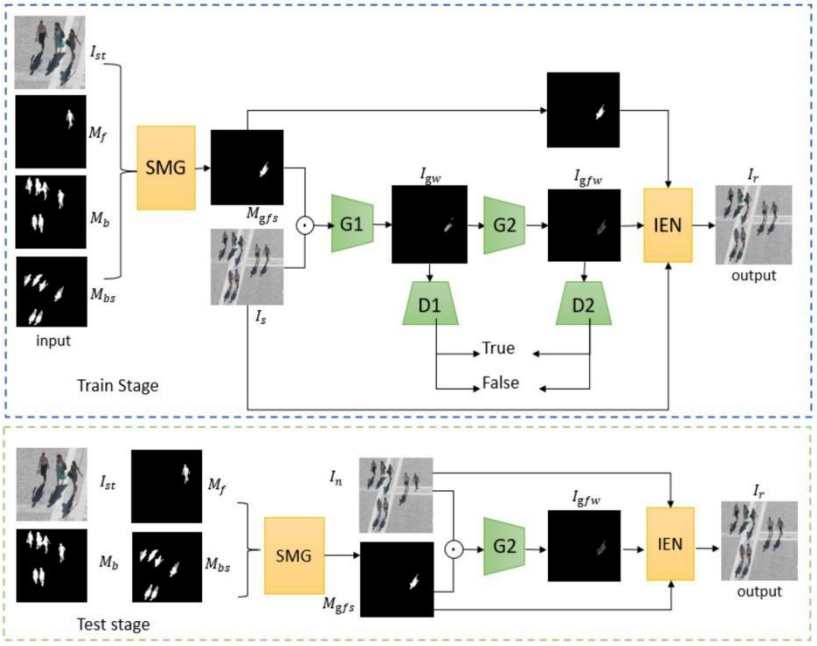

Figure 1. Architecture of the Shadow Generation In Real-world Environments (SGIRE) network.

It operates in two primary phases: Shadow Region Prediction and Shadow Generation & Fusion.

2.1.1. Shadow Mask Generator (SMG)

The SMG is the first sub-module. Unlike methods that merely infer shadow shapes, SMG explicitly learns the mapping from background lighting and object-shadow relationships. The input comprises a cropped background image I_{st}, the target object mask M_f, the background object mask M_b, and its corresponding shadow mask M_{bs}. The visual prediction results of SMG are presented in Figure 2.

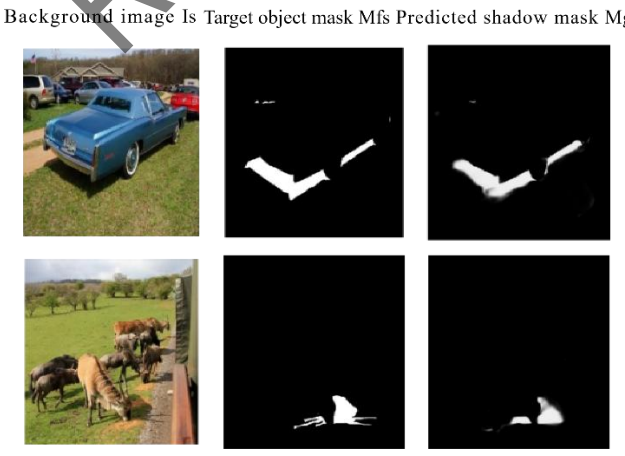

Figure 2. Visual prediction results of the Shadow Mask Generator (SMG).

To leverage semantic cues, we utilize a dual-stream Encoder-Decoder architecture integrated with attention, as illustrated in Figure 3.

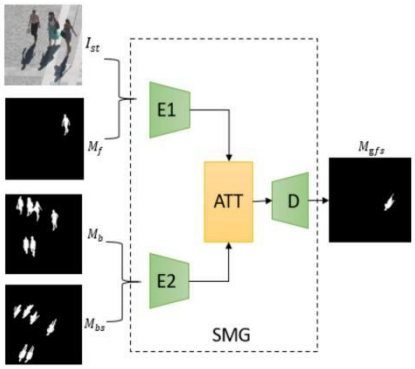

Figure 3. Input structure of the Shadow Mask Generator (SMG).

Encoder E1 captures the lighting features from I_{st} and the target mask. Encoder E2 extracts the mapping features from the background masks M_b and M_{bs}. The bottleneck consists of three Attention Blocks that fuse these feature maps, allowing the network to reason about lighting direction and shadow shape. The detailed network structure is illustrated in Figure 4.

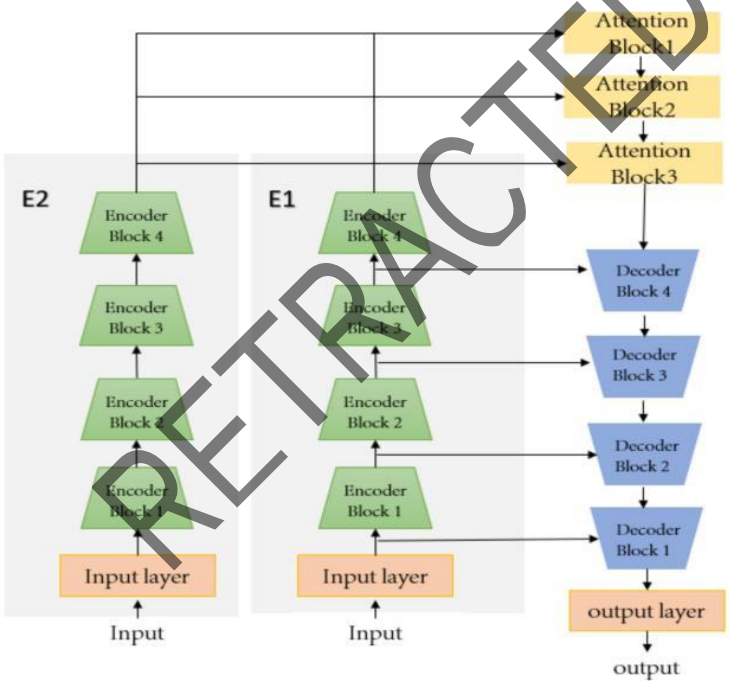

Figure 4. Network architecture of the Shadow Mask Generator (SMG).

The decoder reconstructs the predicted target shadow mask M_{gfs} using skip connections from the encoders.

To further boost the prediction accuracy, our architecture incorporates residual learning blocks, similar to those in ResNet [17]. These residual connections facilitate the transmission of high-frequency gradients during backpropagation, helping the model converge efficiently even when detecting subtle shadow contours in complex backgrounds. In the context of shadow

detection, several foundational studies have utilized conditional GANs to differentiate shadow regions [40], while others have introduced direction-aware spatial context features [41]. Additionally, methods like SegNet [24] and Pyramid Scene Parsing Networks [25] have revolutionized semantic segmentation, and we adapt these dense feature extraction strategies to the specific task of predicting shadow masks. The specific architectures of the Encoder Block and Decoder Block are depicted in Figure 5 and Figure 6 respectively.

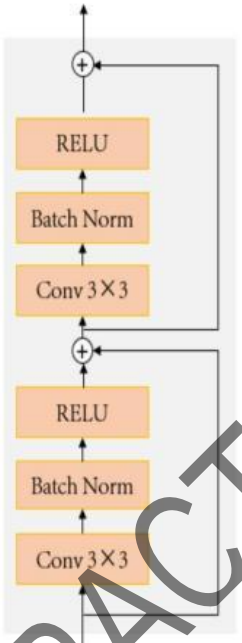

Figure 5. Architecture of the SMG Encoder Block.

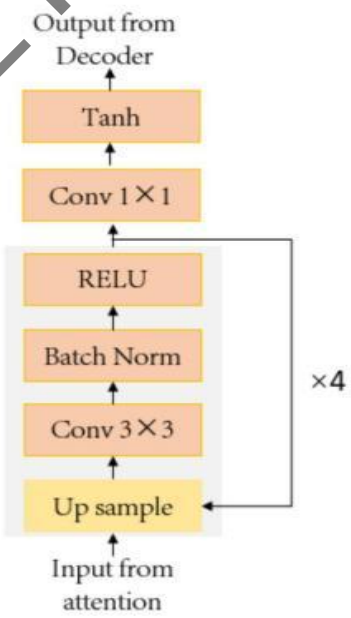

Figure 6. Architecture of the SMG Decoder Block.

These architectural influences allow SMG to maintain fidelity even when background objects are highly irregular.

SMG Loss: To optimize the SMG, we minimize the L_1distance between the predicted mask $M_{gfs}$and the ground-truth mask M_{fs}:

$$L_s = \mathbb{E}_{x\sim p(x)}\left[\left\|M_{gfs} - M_{fs}\right\|_1\right]$$

The explicit separation of lighting cues and object-shadow mappings within the Shadow Mask Generator (SMG) ensures that it predicts the shadow mask based on concrete physical evidence, rather than merely relying on implicit feature correlations that could lead to inaccuracies. This thoughtful design effectively avoids the erroneous attention shifts that are commonly observed in methods like ARShadowGAN, especially under complex dark backgrounds, where the network might mistakenly interpret intricate background textures as shadow regions. By providing clear and explicit structural guidance, the SMG establishes a robust geometric foundation for the subsequent generation network, thus enhancing the overall reliability of the shadow prediction process.

2.1.2. Cycle-Consistent Generation Network (CGN)

After obtaining M_{gfs}, the CGN—consisting of two generators (G1, G2) and two discriminators (D1, D2)—performs the image translation.

- Generator G1: Transforms the background image masked with the predicted shadow I_{fw}into a pseudo-shadow-free patch I_{gw}.
- Generator G2: Transforms the pseudo-shadow-free patch I_{gw} back into a synthetic shadow patch I_{gfw}.
- Discriminators: D1 and D2 (PatchGAN architecture [19]) distinguish between real shadow patches and generated patches, enforcing local realism.

The intermediate generation results of the CGN are illustrated in Figure 7.

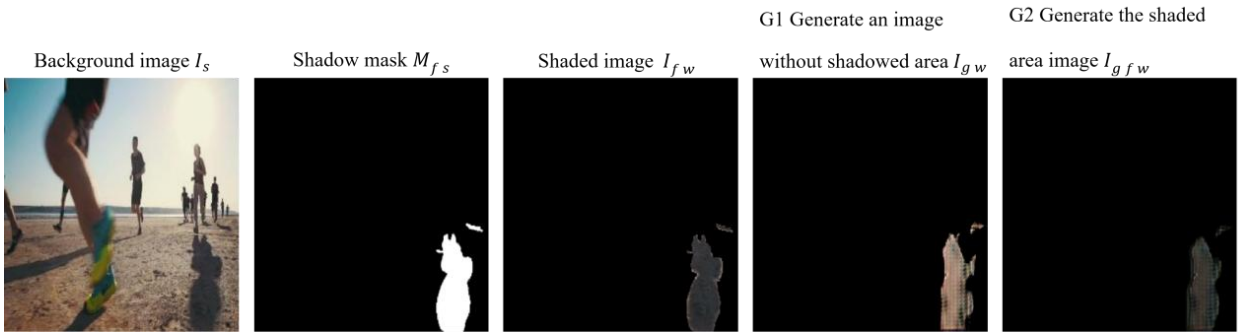

Figure 7. Intermediate artifact generation results of the Cycle-Generation Network (CGN).

The generators follow an Encoder-ResNet-Decoder architecture, as shown in Figure 8.

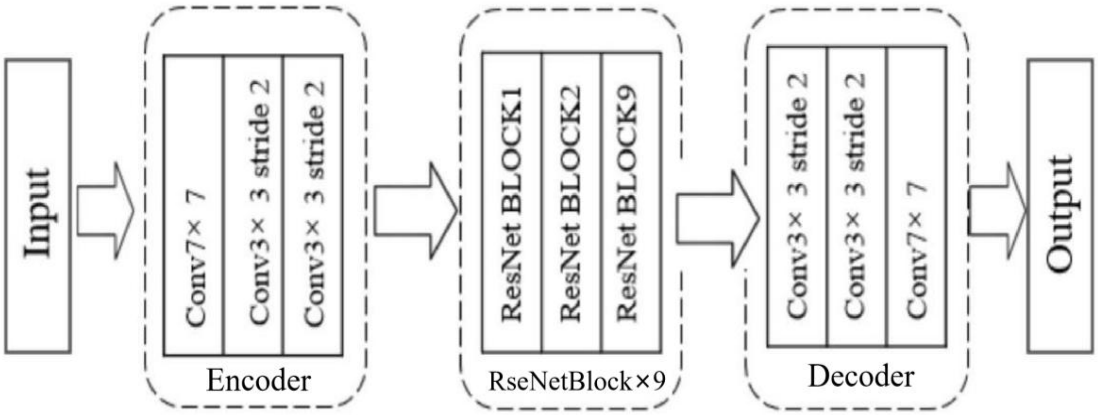

Figure 8. Network architecture of the CGN generator.

We utilize the cycle-consistency loss to ensure the bijection property:

$$L_c = \mathbb{E}_{x\sim p(x)}\left[\left|\left|G2(G1(x)) - x\right|\right|_1\right] + \mathbb{E}_{y\sim p(y)}\left[\left|\left|G1(G2(y)) - y\right|\right|_1\right]$$

Additionally, an identity loss L_{id} is applied to encourage generators to keep already correct domain characteristics unchanged (e.g., G1 should not alter an already shadow-free patch). The adversarial loss L_{gan} is defined as:

$$L_{gan} = \mathbb{E}_{x\sim p_{data}(I_h)}[log\, D\,(x)] + \mathbb{E}_{z\sim p_z(M_{gfs},I_{fs})}\left[log\left(1 - D(G(z))\right)\right]$$

The discriminator structure is presented in Figure 9.

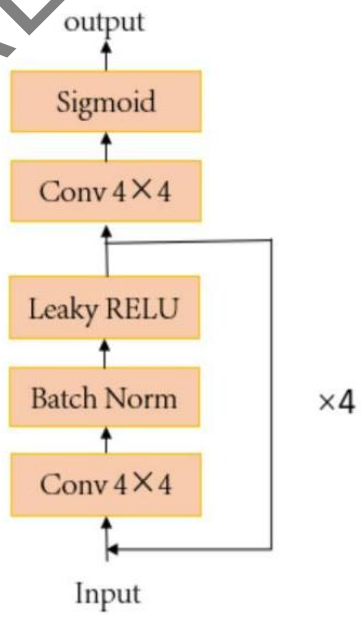

Figure 9. Architecture of the SGIRE discriminator.

Unlike end-to-end generative models that jointly optimize shadow shape and texture, our two-stage design explicitly imposes geometric constraints prior to texture synthesis. The SMG provides a rigid structural prior, which effectively prevents the subsequent CGN from producing geometrically distorted or completely missing shadows—a common failure mode in pure GAN-based generation. This decoupling allows the CGN to focus entirely on texture reconstruction and illumination consistency, thereby improving the overall stability and quality of the synthesized shadows.

2.1.3. Image Enhancement Network (IEN)

The final synthetic shadow patch I_{gfw} is combined with the original background I_s (excluding the predicted shadow region), followed by the IEN. The IEN refines the combined image to ensure seamless blending, smoothing boundary artifacts and matching color and lighting with the original background.

The IEN uses a similar UNet architecture and is trained with a pixel-level reconstruction loss and a VGG-16 perceptual loss to refine edge details and color consistency:

$$L_r = \|I_s - I_r\|_2^2$$

$$L_v = \sum_j \frac{1}{C_j H_j W_j} \left\| \phi_j(I_{gfs}) - \phi_j(I_{fs}) \right\|_2^2$$

The final training objective combines these losses:

$$L_{total} = \alpha_1 L_s + \alpha_2 L_{id} + \alpha_3 L_{gan} + \alpha_4 L_c + \alpha_5 L_r + \alpha_6 L_v$$

The IEN is specifically designed to effectively address the high-frequency discontinuities that often occur at the boundaries between the generated shadow patch and the original background. The thoughtful combination of pixel-level L2 loss and VGG perceptual loss serves a dual purpose: the L2 loss enforces global color consistency across the entire image, while the perceptual loss aligns high-level texture patterns more accurately. This synergistic approach enables the IEN to adaptively smooth transition edges and match local contrast seamlessly, ensuring that the final output exhibits no visible seams or unnatural color shifts, resulting in a visually harmonious integration.

2.2. SRIRE: Weakly Supervised Shadow Removal via Image Partitioning

SRIRE is specifically designed to effectively remove shadows from images without needing ground-truth clean images during the training process. The framework for this innovative approach is illustrated in detail in Figure 10, highlighting its unique methodology and overall functionality.

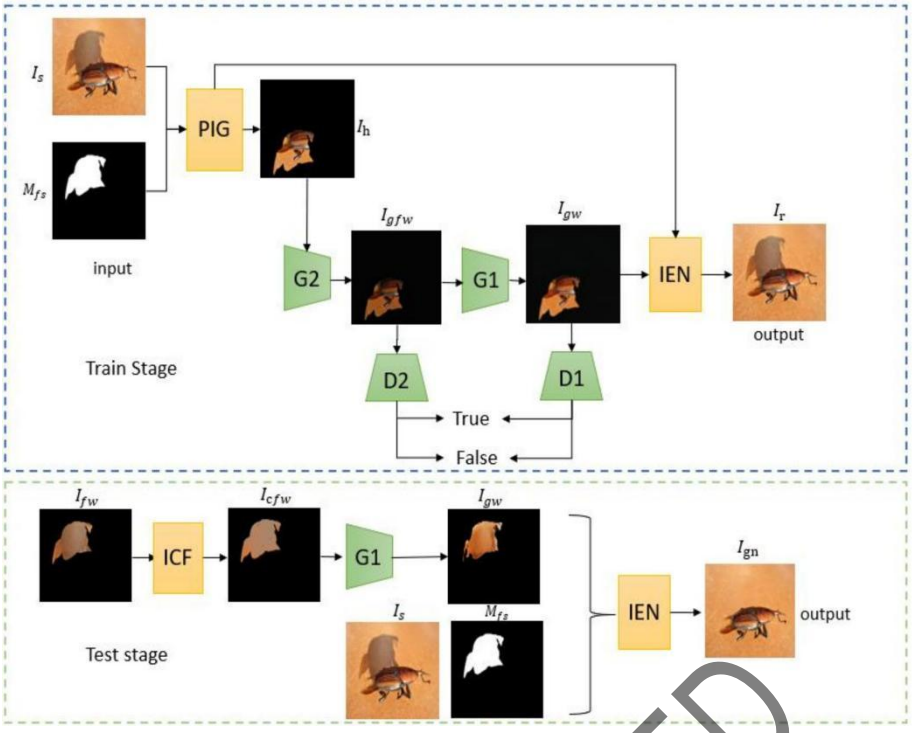

Figure 10. Architecture of the Shadow Removal In Real-world Environments (SRIRE) model.

It consists of three key sub-networks: Patch Image Generator (PIG), CGN, and IEN.

2.2.1. Patch Image Generator (PIG)

The core innovation of SRIRE is the PIG, which creates "paired" training data from a single shadow image. The PIG works by applying a sliding mask over the non-shadow background regions of the original image. It searches for a spatially proximate region that shares the same lighting context as the target shadow patch. We impose strict criteria for patch selection: the sampled patch must not intersect with the original shadow area, must possess a sufficient number of pixels to be meaningful for training, and should be geometrically adjacent to the shadow to minimize environmental variations. By extracting a shadow-free patch I_h directly from the source image, the PIG effectively provides a proxy ground-truth that retains the precise illumination and texture characteristics of the specific scene, which is a significant advantage over using external unshaded images.

It is worth noting that similar patch-based approaches have been explored in interactive shadow removal tasks [48] to significantly reduce the need for extensive manual annotation. However, our proposed Patch Integration Generative (PIG) method is entirely automated and integrates seamlessly with the downstream generative pipeline. Furthermore, this innovative method aligns closely with the overarching goals of weakly supervised learning, particularly in extreme environments, such as those found in weakly labeled astronomical or ecological

datasets [50], where obtaining perfect ground-truth data is simply unattainable. By relying on patches extracted from the same image, we ensure that the statistical properties—such as reflectance and ambient occlusion—remain consistent, which presents a distinct advantage over traditional cross-image unpaired training methods. This consistency ultimately enhances the robustness and effectiveness of our approach.

The intra-image extraction strategy offers a significant physical advantage compared to cross-image approaches. By sampling the reference patch from the exact same scene, the PIG ensures that the reference patch shares identical ambient illumination, surface reflectivity, and color temperature with the target shadow area. This intrinsic consistency fundamentally eliminates the severe domain gaps that are often encountered when pairing images from disjoint datasets. As a result, this approach allows the network to learn the shadow removal mapping from a distribution that is not only highly realistic but also deeply rooted in physical principles. Consequently, this leads to improved performance and more reliable outcomes in shadow removal tasks.

2.2.2. Cycle Generation and Enhancement for Removal

Using the patch pair (I_{fw} as the shadow patch, I_h as the reference light patch), we train the CycleGAN module. Specifically, Generator G2 learns to synthesize a realistic shadow patch I_{gfw} from I_h , allowing G1 to be trained on the complementary task of removing shadows (from I_{gfw} back to I_h). This ensures G1 is robust to various shadow intensities.

To handle luminance discrepancies between the shadow and reference regions during inference, we implement an Image Conversion with Format (ICF) technique. Before feeding the shadow patch into G1, we convert the image to the LAB color space. We adjust the L-channel (luminance) of the shadow patch to match the average luminance level of the reference patch I_h . The mathematical formulation of this adjustment is expressed as:

$$L_{nm}(I_{cfw}) = \frac{L_{nm}(I_{fw})}{L_{mean}(I_{fw})} \times L_{mean}(I_h)$$

This dynamic adjustment effectively "pre-illuminates" the shadow region, significantly reducing the burden on G1 while simultaneously preventing undesirable color artifacts. Such an operation is essentially derived from well-established physical illumination theories [34-35], which emphasize that utilizing brightness priors can greatly enhance the final rendering realism. By leveraging the LAB color space, we can independently control luminance without sacrificing any crucial color information, making the entire process more robust against unusual

and challenging lighting conditions. Finally, the IEN skillfully blends the generated shadow-free patch into the overall background, ensuring a seamless integration that enhances visual coherence.

The ICF operation is motivated by the inherent statistical independence of luminance and chrominance in the LAB color space. Unlike RGB channels, where brightness and color information are highly entangled and often lead to undesirable artifacts, scaling the L-channel in LAB allows us to robustly adjust the overall brightness without introducing any hue distortion or unwanted color shifts. This physically plausible luminance prior effectively performs a "pre-illumination" step, which enables the subsequent G1 process to focus primarily on restoring fine texture details. As a result, it minimizes the need for heavy compensation for luminance differences, thereby preventing over-smoothing and simultaneously reducing the risk of color artifacts that can detract from image quality. This careful approach ensures more accurate color representation throughout the processing.

2.2.3. Loss Functions for SRIRE

The training loss for SRIRE incorporates adversarial, cycle-consistency, and identity losses, similar to SGIRE. Additionally, we enforce a shadow-specific loss L_s in the final reconstruction to ensure the shadow-removed area aligns with the original background:

$$L_s = \|I_r - I_s\|_2^2 + \|I_r * M_{fs} - I_s * M_{fs}\|_2^2$$

The total objective is:

$$L_{total} = \alpha_1 L_{id} + \alpha_2 L_{gan} + \alpha_3 L_c + \alpha_4 L_s$$

3. Experiments and Results

3.1. Experimental Settings

All experiments were meticulously conducted on a robust Ubuntu 22.04 system equipped with an NVIDIA RTX 4090 GPU, utilizing the powerful PyTorch 1.13.1 framework for efficient computational performance and deep learning tasks.

- For SGIRE: We utilized the DESOBA dataset [14-15]. After filtering invalid instances, 4,246 images were used (8:1:1 split). Data augmentation included scaling, flipping, and translation. The SMG was pre-trained for 50 epochs, followed by joint training of all modules for 100 epochs. Learning rates were set to 1×10^{-4} for generators and 1×10^{-5} for discriminators, with a linear decay schedule. Such optimization schedules are standard in deep learning training and align with research on accelerating training

through Batch Normalization [27] and adaptive learning rates. Example 6-tuple images from the DESOBA dataset are shown in Figure 11.

Figure 11. Example 6-tuple images from the DESOBA dataset.

- For SRIRE: We used the ISTD dataset [9-10], comprising 1,870 triplets (shadow image, shadow-free image, mask). We split 200 images for testing and used the remaining for training. The training protocol followed 100 epochs with similar learning rate schedules. These datasets have played a pivotal role in developing shadow removal methods, providing a standardized benchmark for comparison [9].

3.2. Evaluation Metrics

We measured performance using Root Mean Square Error (RMSE) and Structural Similarity Index Measure (SSIM). Metrics were calculated both globally (GRMSE, GSSIM) and locally within the shadow mask area (LRMSE, LSSIM) to emphasize precision in the target regions.

To elaborate, RMSE directly penalizes large pixel deviations and is particularly sensitive to outliers in reconstructed images. SSIM, on the other hand, incorporates luminance, contrast, and structural similarity terms. In many vision tasks, SSIM is considered superior for evaluating human perception. However, given that our primary goal is to recover realistic pixel intensities in occluded shadowed regions—rather than merely preserving global structure—RMSE serves as a critical metric for evaluating the physical accuracy of the reconstruction. Prior works in shadow removal [4,7,11] have also utilized this combined evaluation framework.

3.3. Results and Analysis

3.3.1. Model Architecture

We compared SGIRE against Pix2Pix [3], Pix2Pix-Res, Mask-ShadowGAN [4], and ARShadowGAN [5] on the DESOBA test set.

- Qualitative Evaluation: As shown in Figure 12, Pix2Pix methods often produced unrealistic, diffused shadows that disregarded lighting directions. Mask-ShadowGAN exhibited limited structural accuracy, generating patchy artifacts. ARShadowGAN performed better but often produced "soft" edges lacking sharpness. In contrast, SGIRE generated distinct, realistic shadows that correctly followed the geometry and lighting cues of the background, seamlessly integrating the target object. The visual comparison verifies the superior performance of our proposed method.

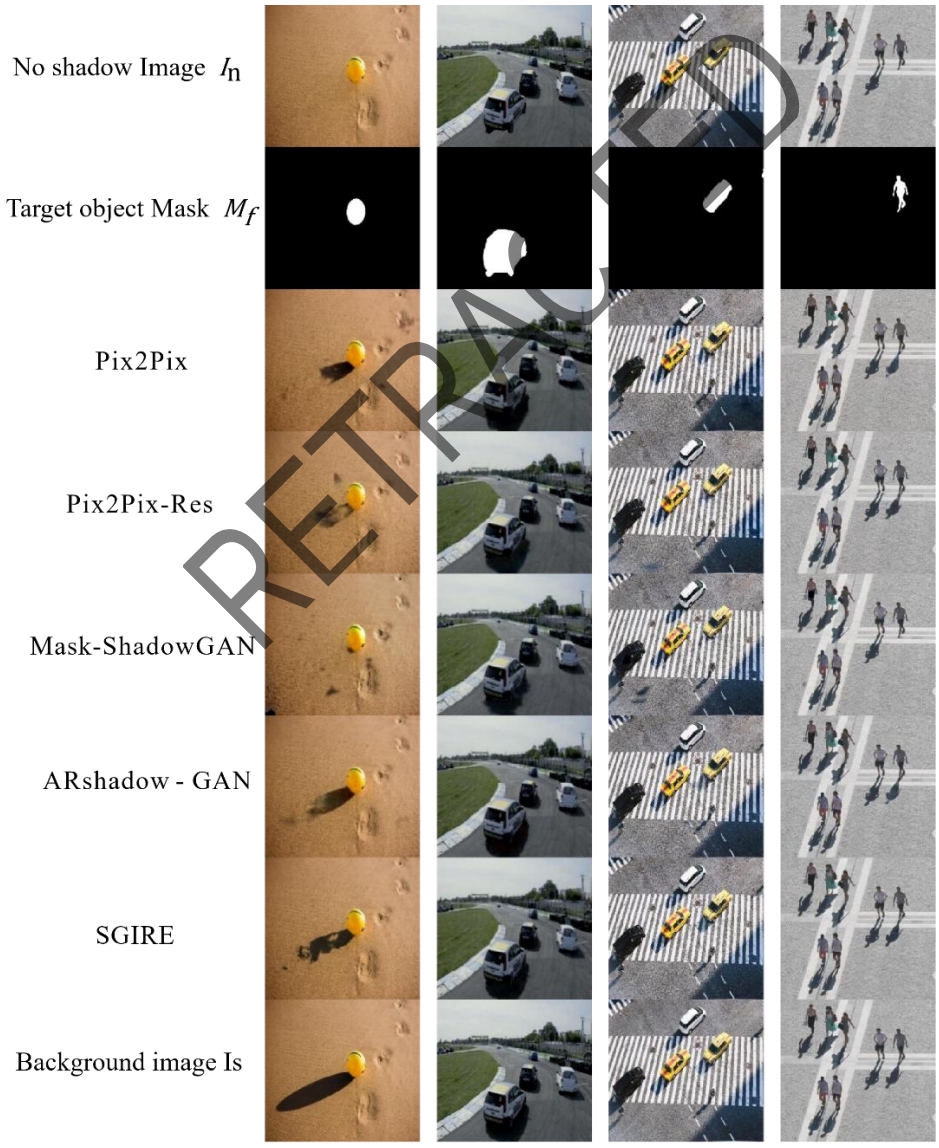

Figure 12. Visual comparison of different shadow generation methods.

- Quantitative Evaluation: The numerical comparisons are presented in Table 1. SGIRE achieved the best GRMSE (6.021) and GSSIM (0.982) and consistently outperformed others in local metrics, demonstrating superior fidelity in shadow generation.

Table 1. Quantitative comparison of shadow generation methods on DESOBA.

Method	GRMSE	GSSIM	LRMSE	LSSIM
Pix2Pix	7.822	0.909	76.338	0.62
Pix2Pix-Res	6.195	0.959	76.965	0.645
Mask-ShadowGAN	8.421	0.93	79.808	0.619
ARShadowGAN	6.629	0.971	76.006	0.64
SGIRE (Ours)	6.021	0.982	70.242	0.669

3.3.2. Shadow Removal (SRIRE)

We compared SRIRE against ShadowGAN [6], Mask-ShadowGAN [4], and LG-ShadowNet [7] on the ISTD dataset.

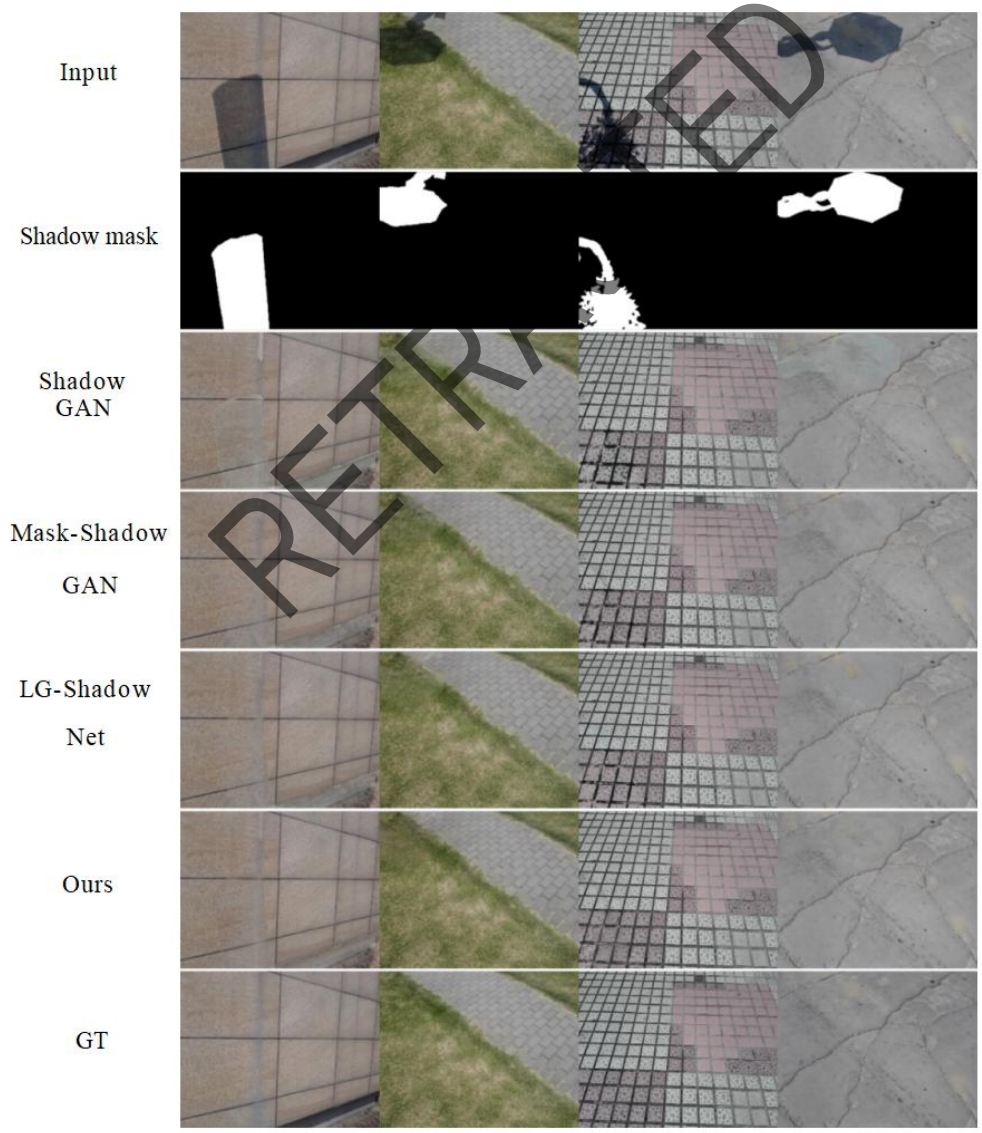

Figure 13. Visual comparison of different shadow removal methods.

- Qualitative Evaluation: As illustrated in Figure 13, ShadowGAN left obvious ghosting and color distortions. Mask-ShadowGAN and LG-ShadowNet performed better but occasionally introduced unwanted artifacts or failed to fully recover textures. SRIRE effectively eliminated shadows while preserving the underlying texture and color, outperforming the baselines in the visual fusion of shadow-removed areas with their surroundings.
- Quantitative Evaluation: Table 2 highlights the superior performance of SRIRE. We achieved the lowest GRMSE (3.92) and LRMSE (8.77). While LG-ShadowNet slightly edged us out in GSSIM (0.985 vs. 0.984), our method demonstrated marked improvement in pixel-level accuracy, which is critical for subsequent computer vision tasks.

Table 2. Quantitative Comparison of Shadow Removal Methods on ISTD.

Method	GRMSE	GSSIM	LRMSE	LSSIM
ShadowGAN	5.15	0.98	10.19	0.98
Mask-ShadowGAN	4.7	0.989	9.83	0.989
LG-ShadowNet	4.51	0.985	9.67	0.985
SRIRE (Ours)	3.92	0.984	8.77	0.984

3.4. Ablation Studies

We conducted detailed ablation experiments to verify the necessity of each module in SGIRE and SRIRE.

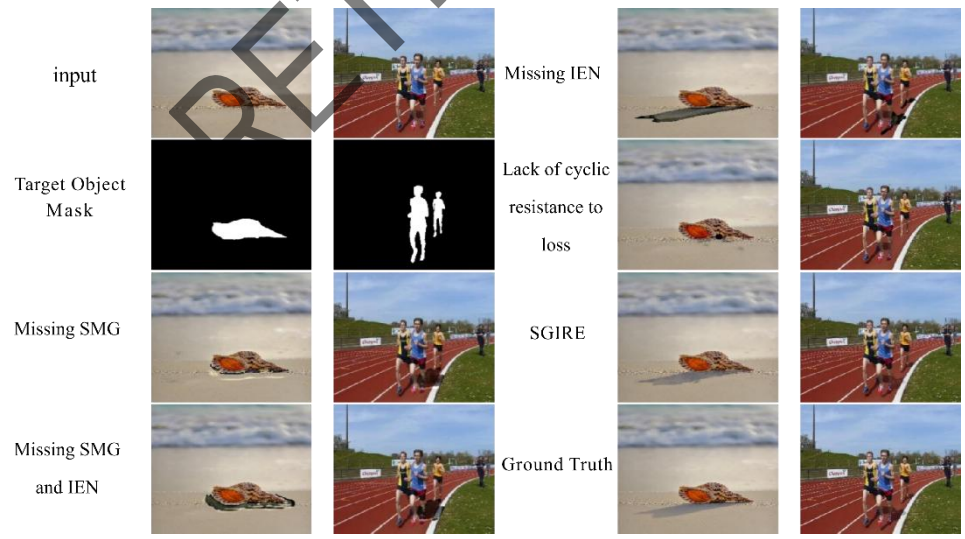

Figure 14. Visualization of the ablation study results for SGIRE.

- For SGIRE: Removing the SMG module led to severely distorted shadow shapes, as the generator lacked structural guidance. Removing the IEN resulted in visible boundaries between the generated shadow and the background. Removing the cycle-consistency

loss prevented the convergence of the domain translation, resulting in unstable outputs. The ablation study results for SGIRE are shown in Figure 14.

- For SRIRE: Eliminating the shadow-specific loss (l_s) caused distinct residual boundary artifacts, particularly affecting grassy textures. Omitting the identity and cycle consistency losses caused excessive luminance shifts, turning shadow-removed regions unnaturally bright. The full model successfully eradicated these artifacts, confirming the effectiveness of our proposed modules. The ablation study results for SRIRE are shown in Figure 15.

Figure 15. Visualization of the ablation study results for SRIRE.

These ablation results empirically validate our core design rationale. The severe shape distortions caused by missing SMG confirm that explicit geometric guidance is indispensable for structural correctness. The noticeable boundary artifacts observed when removing IEN

prove the necessity of a dedicated fusion network for seamless integration. Furthermore, the improved luminance stability in the full SRIRE model validates the effectiveness of the ICF module in bridging the brightness gap, confirming that each proposed sub-network contributes critically to the overall performance and generalization capability of our frameworks.

4. Conclusion

In this work, we successfully addressed the long-standing challenge of robust shadow manipulation in unconstrained real-world environments by proposing two complementary adversarial frameworks: SGIRE for shadow generation and SRIRE for shadow removal. Both frameworks are specifically designed to overcome the fundamental bottleneck of data acquisition in this domain—the reliance on strictly paired (shadow vs. non-shadow) images—by leveraging the power of cycle-consistency and carefully designed intra-image learning strategies.

For the shadow generation task, SGIRE demonstrates that high-fidelity shadow synthesis can be achieved without paired ground-truth data by decoupling the task into two distinct phases. The first phase, anchored by the attention-driven Shadow Mask Generator (SMG), effectively leverages semantic cues and geometric relationships from the background image itself to predict the target shadow region. This explicit structural guidance reduces the burden on the subsequent generation network, allowing the CycleGAN-based CGN to focus primarily on texture synthesis and illumination consistency. The introduction of the Image Enhancement Network (IEN) further bridges the gap between synthetic patches and the original background, eliminating boundary artifacts and ensuring color fidelity. Our experiments confirm that SGIRE not only generates structurally accurate shadows that respect physical lighting laws but also achieves superior pixel-level accuracy compared to existing methods like ARShadowGAN and Mask-ShadowGAN, which often struggle with unpaired data or produce overly soft boundaries.

For the shadow removal task, the SRIRE framework presents a paradigm shift in how weakly supervised learning can be applied to image restoration. By introducing the Patch Image Generator (PIG), we completely eliminate the requirement for external ground-truth clean images during training. The PIG's ability to automatically construct paired training samples from a single image is a pragmatic and powerful solution to the data scarcity problem. Moreover, the incorporation of the ICF module—a physically inspired luminance adjustment mechanism—provides a crucial prior that significantly enhances the network's generalization capability across diverse lighting conditions. The combination of cycle-consistency and identity loss within the CGN ensures that the removal process remains stable and content-preserving,

effectively avoiding the severe ghosting or over-smoothing artifacts often observed in unpaired adversarial networks.

Although our current frameworks achieve competitive performance on standard benchmarks (DESOBA and ISTD), it is important to acknowledge their inherent limitations. Both SGIRE and SRIRE are currently designed to operate under single-source or dominant lighting assumptions. In scenes with complex, multi-source lighting, or significant inter-reflections, the models may struggle to accurately infer the correct shadow shape or color. Additionally, the current patch-based strategy, while effective, may fail when the target shadow area is unusually large or occupies a dominant portion of the image, as finding a spatially proximate, non-overlapping patch becomes statistically difficult. These limitations highlight the need for more adaptive context-aware mechanisms in future iterations.

Looking forward, we envision several promising directions to extend this work. Firstly, integrating semantic metadata—such as estimated light source direction, time of day, or weather conditions—into the latent space of the generator could provide stronger conditional priors, allowing the model to handle dynamic environmental changes more effectively. Secondly, we plan to explore interactive frameworks that allow users to manually guide the shadow shape or intensity via a simple brush or slider input, greatly enhancing practical usability in creative industries like film production and game design. Thirdly, the recent successes of Transformer architectures [43] and diffusion models in computer vision suggest that replacing the convolution-based attention mechanisms in SMG with Transformer-based backbones could further improve global feature dependencies. Lastly, while our work focuses on 2D image manipulation, emerging paradigms like Neural Radiance Fields (NeRF) [47] hold tremendous potential for extending our methods to 3D-consistent scene editing, allowing shadows to dynamically adapt to novel viewpoints. By bridging the gap between unpaired adversarial learning and these advanced generative frontiers, we believe our proposed frameworks lay a solid foundation for the next generation of intelligent, data-efficient image editing tools.

In summary, this paper contributes two novel, resource-efficient deep learning architectures that push the boundaries of shadow manipulation in unconstrained settings. By minimizing reliance on costly paired datasets and introducing practical innovations such as semantic mask guidance and intra-image patch extraction, we provide a scalable solution ready for deployment in real-world computer vision pipelines. We anticipate that our findings will inspire further research into weakly supervised and self-supervised techniques for complex image synthesis tasks.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative adversarial nets," *in Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[2] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *in Proc. IEEE Int. Conf. Comput*. Vis. (ICCV), 2017, pp. 2242–2251.

[3] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit*. (CVPR), 2017, pp. 1125–1134.

[4] X. Hu, Y. Jiang, C. W. Fu, and P. A. Heng, "Mask-ShadowGAN: Learning to remove shadows from unpaired data," *in Proc. IEEE/CVF Int. Conf. Comput*. Vis. (ICCV), 2019, pp. 2472–2481.

[5] D. Liu, C. Long, H. Zhang, et al., "ARShadowGAN: Shadow generative adversarial network for augmented reality in single light scenes," *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 8139–8148.

[6] S. Zhang, R. Liang, and M. Wang, "ShadowGAN: Shadow synthesis for virtual objects with conditional adversarial networks," *Comput. Visual Media*, vol. 5, no. 1, pp. 105–115, 2019.

[7] Z. Liu, H. Yin, Y. Mi, et al., "Shadow removal by a lightness-guided network with training on unpaired data," *IEEE Trans. Image Process*., vol. 30, pp. 1853–1865, 2021.

[8] Y. Jin, A. Sharma, and R. T. Tan, "DC-ShadowNet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network," *in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 5027–5036.

[9] H. Le and D. Samaras, "Shadow removal via shadow image decomposition," *in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 8578–8587.

[10] J. Wang, X. Li, and J. Yang, "Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1788–1797.

[11] L. Qu, J. Tian, S. He, et al., "Deshadownet: A multi-context embedding deep network for shadow removal," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4067–4075.

[12] H. Le and D. Samaras, "From shadow segmentation to shadow removal," *in Computer Vision – ECCV 2020*, 2020, pp. 264–281.

[13] Z. Liu, H. Yin, X. Wu, et al., "From shadow generation to shadow removal," *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 4927–4936.

[14] Y. Hong, L. Niu, and J. Zhang, "Shadow generation for composite image in real-world scenes," *in Proc. AAAI Conf. Artif. Intell*., vol. 36, no. 1, 2022, pp. 914–922.

[15] Q. Liu, J. Wang, and L. Niu, "DESOBAv2: Towards large-scale real-world dataset for shadow generation," *arXiv preprint arXiv*:2308.09972, 2023.

[16] H. Le, T. F. Y. Vicente, V. Nguyen, et al., "A+D Net: Training a shadow detector with adversarial shadow attenuation," *in Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 662–

678.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv*:1409.1556, 2014.

[19] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," *in Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 702–716.

[20] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," *in Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 214–223.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *in Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[22] G. Huang, Z. Liu, L. Van Der Maaten, et al., "Densely connected convolutional networks," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4700–4708.

[23] L. C. Chen, G. Papandreou, I. Kokkinos, et al., "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2017.

[24] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[25] H. Zhao, J. Shi, X. Qi, et al., "Pyramid scene parsing network," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2881–2890.

[26] C. Szegedy, W. Liu, Y. Jia, et al., "Going deeper with convolutions," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 1–9.

[27] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *in Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 448–456.

[28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[29] R. Guo, Q. Dai, and D. Hoiem, "Paired regions for shadow detection and removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2956–2967, 2012.

[30] M. Gryka, M. Terry, and G. J. Brostow, "Learning to remove soft shadows," *ACM Trans. Graph. (TOG)*, vol. 34, no. 5, pp. 1–15, 2015.

[31] C. Xiao, R. She, D. Xiao, and K. L. Ma, "Fast shadow removal using adaptive multiscale illumination transfer," *Comput. Graph. Forum*, vol. 32, no. 8, pp. 207–218, 2013.

[32] S. H. Khan, M. Bennamoun, F. Sohel, and R. Togneri, "Automatic shadow detection and removal from a single image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 431–446, 2015.

[33] G. D. Finlayson, M. S. Drew, and C. Lu, "Entropy minimization for shadow removal," *Int. J. Comput. Vis.*, vol. 85, no. 1, pp. 35–57, 2009.

[34] I. Sato, Y. Sato, and K. Ikeuchi, "Illumination from shadows," IEEE Trans. Pattern Anal.

Mach. Intell., *vol. 25, no. 3*, pp. 290–300, 2003.

[35] A. Panagopoulos, C. Wang, D. Samaras, and N. Paragios, "Illumination estimation and cast shadow detection through a higher-order graphical model," i*n Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2011, pp. 673–680.

[36] T. Kim, M. Cha, H. Kim, et al., "Learning to discover cross-domain relations with generative adversarial networks," *in Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1857–1865.

[37] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in Proc. IEEE Int. *Conf. Comput. Vis. (ICCV)*, 2017, pp. 2849–2857.

[38] Y. Choi, M. Choi, M. Kim, et al., "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in Proc. IEEE Conf. Comput. Vis. *Pattern Recognit*. (CVPR), 2018, pp. 8789–8797.

[39] S. Tulyakov, M. Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," in Proc. IEEE Conf. Comput. Vis. *Pattern Recognit*. (CVPR), 2018, pp. 1526–1535.

[40] V. Nguyen, T. F. Yago Vicente, M. Zhao, et al., "Shadow detection with conditional generative adversarial networks," in Proc. IEEE Int. Conf. *Comput. Vis. (ICCV)*, 2017, pp. 4510–4518.

[41] X. Hu, L. Zhu, C. W. Fu, et al., "Direction-aware spatial context features for shadow detection," in Proc. IEEE Conf. Comput. *Vis. Pattern Recognit*. (CVPR), 2018, pp. 7454–7462.

[42] L. Zhu, Z. Deng, X. Hu, et al., "Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection," in Proc. Eur. Conf. *Comput. Vis*. (ECCV), 2018, pp. 121–136.

[43] K. Zhou, J. L. Fang, W. Wu, et al., "Semantic-aware Transformer for shadow detection," Comput. *Vis. Image Underst*., vol. 240, art. no. 103941, 2024.

[44] I. Arief, S. McCallum, and J. Y. Hardeberg, "Realtime estimation of illumination direction for augmented reality on mobile devices," *in Color and Imaging Conf*., 2012, pp. 111–116.

[45] K. Karsch, V. Hedau, D. Forsyth, and D. Hoiem, "Rendering synthetic objects into legacy photographs," *ACM Trans. Graph*. (TOG), vol. 30, no. 6, pp. 1–12, 2011.

[46] B. Liao, Y. Zhu, C. Liang, et al., "Illumination animating and editing in a single picture using scene structure estimation," *Comput. Graph*., vol. 82, pp. 53–64, 2019.

[47] P. P. Srinivasan, B. Deng, X. Zhang, et al., "Nerv: Neural reflectance and visibility fields for relighting and view synthesis," in Proc. IEEE/CVF Conf. Comput. *Vis. Pattern Recognit*. (CVPR), 2021, pp. 7495–7504.

[48] H. Gong and D. Cosker, "Interactive removal and ground truth for difficult shadow scenes," *J. Opt. Soc. Am. A, vol. 33*, no. 9, pp. 1798–1811, 2016.

[49] N. Su, Y. Zhang, S. Tian, et al., "Shadow detection and removal for occluded object information recovery in urban high-resolution panchromatic satellite images," IEEE J. Sel. Topics Appl. *Earth Observ. Remote Sens.*, vol. 9, no. 6, pp. 2568–2582, 2016.

[50] H. M. Le, B. C. Goncalves, D. Samaras, et al., "Weakly labeling the Antarctic: The penguin colony case," *in CVPR Workshops*, 2019, pp. 18–25.