Deep Learning Approach for Shadow Removal Using Semantic Segmentation and Attention Mechanism

Po-Chin Chang
Dept. of Computer Science & Info. Eng.,
National Taiwan University of Science and Technology,
Taipei 106, Taiwan
e-mail: pp0956pp@gmail.com

Shi-Jinn Horng
Dept. of Computer Science & Info. Eng., Asia University,
Dept. of Med. Res., CMUH, China Medical University,
Taichung 413305, Taiwan
e-mail: horngsj@yahoo.com.tw

***Abstract*—This paper presents a novel architecture for shadow removal that leverages semantic segmentation to divide the image into distinct regions: shadow areas, foreground areas, and shadow boundaries. To capture the intricate interactions among these regions, the model incorporates a self-attention mechanism. To tackle the persistent issue of shadow boundary residues found in existing models, this approach introduces a shadow feature fusion mechanism. This mechanism employs area attention to accurately blend features across different regions, enhancing the natural transition at shadow edges and improving shadow region restoration quality. Experimental results on public datasets validate the model's effectiveness in shadow recovery and detail preservation, as evidenced by metrics such as Structural Similarity Index Measure (SSIM) and Root Mean Square Error (RMSE). Additionally, the model demonstrates strong generalization across various test settings, highlighting its practical applicability for shadow removal tasks.**

Keywords- Shadow removal; Area attention; Shadow region restoration; SSIM; RMSE.

I. INTRODUCTION

In the past decade, deep learning has driven major advances in image processing, with models like Convolutional Neural Networks [1] and Vision Transformers [2] greatly improving the accuracy and efficiency of image analysis.

Shadow removal remains a key challenge in image processing due to its impact on visual quality and algorithm performance. Shadows can distort object boundaries and colors, hindering tasks like object detection, face recognition, and scene parsing, especially in outdoor settings. Effective shadow removal is essential for improving both image clarity and recognition accuracy [3].

Shadow removal research faces key challenges, including limited and less diverse datasets like Image Shadow Triplets Dataset (ISTD) [4], SRD [5], and SBU [6], which hinder model robustness. Shadow variations caused by lighting and object interactions further complicate detection, especially when shadow and object colors are similar. This study aims to develop more accurate and adaptable deep learning-based shadow removal models to advance image processing applications.

This paper integrates Segment Anything Model (SAM) [7], Swin Transformer [8], and U-Net [9] with a selective shadow fusion mechanism to build an efficient shadow removal model. SAM provides zero-shot segmentation using pre-trained masks, the Swin Transformer captures global shadow context through window attention, and U-Net excels at restoring fine image details. Together, they enable accurate shadow detection and natural, high-quality shadow-free image reconstruction.

The paper is organized as follows: Section 1 outlines the background and motivation for shadow removal. Section 2 reviews related work and deep learning approaches. Section 3 details the methodology, including experimental setup and model design. Section 4 presents and analyzes the results. The final section summarizes key findings and future directions.

II. RELATED WORK

A. Related Reseach

In recent years, shadow removal has advanced through both physical modeling and deep learning techniques. STacked Conditional Generative Adversarial Network (ST-CGAN) [4] uses dual Conditional-GANs to jointly learn shadow detection and removal in a unified framework. SP+M-Net [10] combines deep networks with a linear illumination model to simulate shadows and predict lighting parameters. Mask-ShadowGAN [11] applies Cycle-GANs [15] to unpaired data, removing shadows without needing paired training samples. Auto-Exposure [12] enhances lighting consistency by automatically adjusting exposure across shadowed regions. CRFormer [13] leverages a Transformer with unidirectional attention for efficient pixel restoration. SpA-Former [14] merges Transformer and Convolutional Neural Network (CNN) architectures with spatial attention for fast, accurate single-stage shadow removal.

B. Transformer

(1) Attention is All You Need

The Transformer [16] replaced Recurrent Neural Networks (RNNs) [17] and Long Short-Term Memory networks (LSTMs) [18] by enabling parallel processing of sequence data, greatly improving training efficiency. It uses multi-head attention in layered encoders and decoders to learn complex patterns. Positional encoding with sine and cosine functions helps retain token order despite parallel input processing.

The self-attention mechanism in Transformers computes relationships between tokens by converting inputs into queries, keys, and values. Attention scores from query-key dot products are normalized with softmax and used to weight the value vectors, enabling the model to capture long-range dependencies effectively.

(2) Swin Transformer

The Swin Transformer [8], or Shifted Window Transformer, improves visual task performance and efficiency using a hierarchical structure and window shifting, effectively handling scale variation and high-resolution images.

Prior Vision Transformer [2] used a global self-attention mechanism, and the Swin Transformer introduces a shifted window mechanism that limits self-attention to local windows, reducing computational complexity from quadratic to linear. By shifting windows, it enables cross-window interactions. Its hierarchical structure merges patches progressively, boosting multi-scale representation learning and making it a strong alternative to CNN backbones in visual tasks.

(3) DehazeFormer

DehazeFormer [19], based on the Swin Transformer, addresses dehazing by improving edge handling in window-based self-attention. Unlike cyclic shifting, which reduces patch use at image edges, DehazeFormer uses reflection padding to extend boundaries, ensuring consistent patch numbers and better feature continuity. After attention, center cropping restores the original size. This method is also effective for shadow removal, where edge detail is crucial.

C. Semantic Segmentation

(1) Segment Anything Model

The Segment Anything Model (SAM) [7] performs zero-shot image segmentation using minimal cues like points or rough selections. Pre-trained on large datasets, SAM delivers high-quality masks instantly and serves as a versatile backbone for tasks like segmentation, data annotation, and real-time image analysis.

SAM also introduces the SA-1B dataset, with over 10 million images and 1 billion masks, enriching model training. Its architecture includes an Image Encoder (based on Vision Transformer [2]), a Prompt Encoder, and a Mask Decoder. The encoders extract features from images and prompts, while the decoder combines them to generate accurate masks and confidence scores, even under ambiguous input.

(2) SAM-Adapter

The SAM-Adapter [20] enhances the original SAM by adding adapters to improve performance on specific tasks. This boosts generalizability and makes it more effective for shadow detection.

SAM-Adapter retains SAM's original image encoder but adds adapters between Transformer layers. Each adapter uses two Multilayer Perceptron (MLPs) and a Gaussian Error Linear Units (GELU) activation: one creates task-specific hints, the other adjusts them to fit the encoder. These refined features improve mask accuracy, making SAM-Adapter effective for shadow detection, which this study adopts.

D. U-Net

U-Net [9] is a convolutional neural network widely used for image restoration and segmentation due to its symmetric U-shaped design. It consists of a contracting path for feature extraction using 3x3 convolutions and 2x2 max pooling, and an expansive path for upsampling and feature fusion. Skip connections between corresponding layers help preserve spatial details, making U-Net effective for high-precision image restoration and the backbone of our model.

E. Selective Kernel Networks

Selective Kernel Networks (SK-Net) [21] overcome the fixed receptive field limitation in CNNs by enabling neurons to adaptively adjust their receptive field size for better multi-scale processing. SK-Net operates in three phases: Split, where input features are processed through multiple convolution paths with different kernel sizes; Fuse, where these are combined and condensed into a global feature vector; and Select, where attention-based weights determine the contribution of each path, dynamically adjusting the receptive fields. This efficient, flexible mechanism makes SK-Net ideal for feature fusion, which is critical in achieving precise shadow removal in our model.

III. PROPOSED METHOD

A. Dataset

(1) Image Shadow Triplets Dataset

The ISTD dataset [4] is a widely used benchmark for shadow detection and removal, containing 1,870 triplets—each with a shadow image, shadow mask, and shadow-free image—across 135 diverse scenes. It includes 1,330 training and 540 testing triplets, featuring varied lighting and shadow types, making it essential for evaluating shadow removal methods.

Shadow Image Shadow Mask Ground Truth

Figure 1. ISTD triplet [4].

(2) Adjusted Image Shadow Triplets Dataset

The Adjusted Image Shadow Triplets Dataset (AISTD) [10] addresses color inconsistencies in the ISTD dataset [4], caused by varying lighting when shadow and non-shadow images were taken. These differences result in notable RMSE values—up to 12.9 in non-shadow areas (Figure 2) and 6.83 on average in the test set. To correct this, the authors applied linear regression to align pixel values in non-shadow regions, using shadow masks and adjusting each Red Green and Blue (RGB) channel separately.

Initially, a shadow mask was used to select the non-shadow regions from each pair of shadow and non-shadow images. Subsequently, a separate linear regression model was applied independently to each color channel (red, green, blue).

$$I_{corrected}(x) = a \cdot I_{shadow-free}(x) + b \qquad (1)$$

The linear regression model, as per Equation (1), uses $I_{corrected}(x)$ to denote the color-corrected pixel values of the shadow-free image, and $I_{shadow-free}(x)$ to represent the original pixel values of the shadow-free image. The parameters a and b of the linear regression model are obtained through the Least Squares Method, fitted within the non-shadow regions. This method significantly reduces the color discrepancy between shadowed and shadow-free images, thereby enabling more accurate performance evaluations during the training of shadow removal models.
As shown in the Corrected GT in Figure 2, color correction reduced RMSE in non-shadow regions from 12.9 to 2.9, and from 6.83 to 2.6 across the test set, improving dataset quality. This corrected version is widely adopted in shadow removal research, including this study.

Shadow Image Original GT Corrected GT

Figure 2. Corrected AISTD dataset [10].

B. Network Architecture

(1) System architecture and process

This paper proposes a novel shadow removal model (Figure 3) that generates shadow-free images from shadow inputs. It combines SAM-Adapter, U-Net, and Swin Transformer, with a Selective Shadow Fusion module to improve integration in shadowed areas.

Experiments showed that end-to-end training with shadow and shadow-free images often leaves residual shadows due to large pixel value differences between non-shadow and shadow areas, with the former averaging 2.3 times brighter than the latter, as shown in Figure 2. To address this, our model uses a pre-trained SAM-Adapter to extract shadow masks, then applies mask inversion and morphological gradients operations [22] to segment the image into shadow, foreground, and boundary regions. Each is processed separately: preserving content in the foreground, smoothing transitions at boundaries, and restoring brightness in shadows. These weighted feature maps are then fused for more accurate shadow removal.

Figure 3 illustrates the architecture of the shadow restoration model, consisting of a contracting path and an expanding path. The sequence of feature map depths transitions from input to output as (3, 24, 48, 96, 48, 24, 3). The Transformer modules are stacked in sequences of (16, 16, 16, 8, 8), with pairs executing computations of window attention and shifted window attention, respectively.

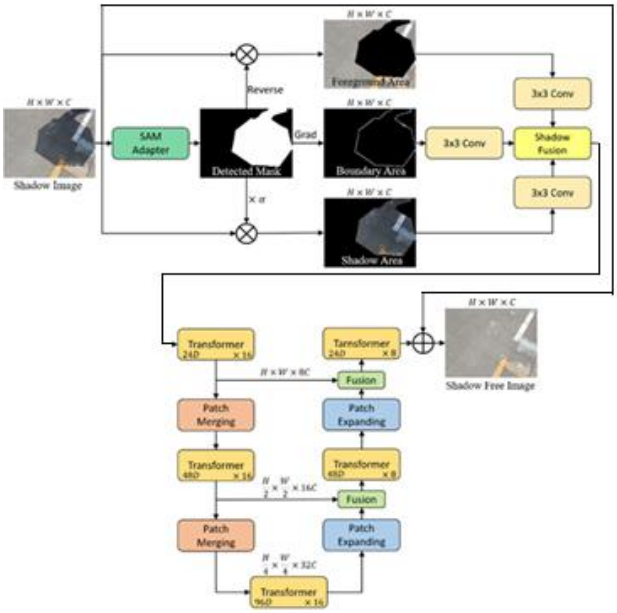

Figure 3. Network architecture.

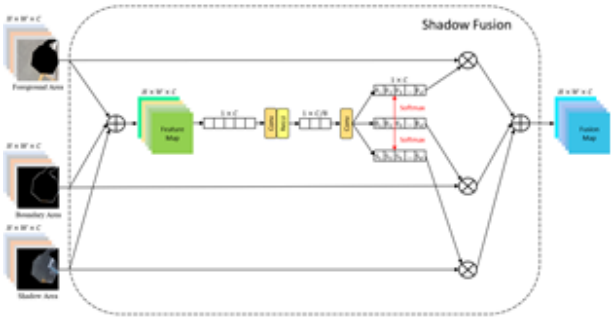

Figure 4. Shadow fusion block.

In the contracting pathway, downsampling and feature extraction are primarily conducted through two Patch Merging modules and three Transformer modules. Each passage through a Patch Merging module halves the dimensions of the feature map while doubling the number of channels. Conversely, the expanding pathway is responsible for upsampling and feature fusion, comprising two Patch Expanding modules, two Transformer modules, and two Fusion modules. The Patch Expanding modules increase the dimensions of the feature map by a factor of two while halving the channel count. The Fusion modules utilize a Selective Kernel Network (SK-Net) and integrate features from both the contracting and expanding pathways via Skip Connections. Ultimately, the output is merged with the original shadow image through a Residual Connection to produce the corresponding shadow-free image.

(2) Shadow Fusion Block

In the task of shadow removal, to enhance the detail at the shadow boundaries and the recovery of shadowed areas, this study introduces a Shadow Fusion module. This module employs an Area Attention mechanism to boost the model's capability to discern features in different shadow regions. The architecture is illustrated in Figure 4. During the fusion phase, element-wise addition is initially applied to the feature maps of the shadow region, foreground area, and shadow boundary area. Subsequently, these combined feature maps undergo Global Average Pooling. Following this, a convolution layer and Rectified Linear Unit (ReLU) activation function reduce the channel dimension of the feature maps to one-eighth of its original size, after which convolution operations expand the channel count back to its initial dimension. Throughout this process, three vectors of the same dimensions, F, E, and S, are generated. During the selection phase, corresponding elements of these three vectors undergo Softmax computation, producing regional attention vectors for the three paths. These vectors are then multiplied by their corresponding regional feature maps. The resulting products are cumulatively added to obtain a channel-fused shadow feature map, referred to as the Fusion Map.

The proposed shadow fusion module uses regional attention to better recognize and process shadow features, offering three key advantages:

a) Regional differentiation processing: The attention mechanism adjusts channel weights by region, enabling flexible handling of varying shadows and enhancing feature extraction in dark, dense areas to prevent detail loss.

b) Smooth shadow boundary transitions: Rather than simply adding features, our model uses channel fusion to integrate shadow, foreground, and boundary regions, enabling smoother, more natural transitions at shadow edges for improved restoration detail.

c) Enhanced model generalization: Channel attention allows the model to adaptively adjust weights, enabling effective shadow removal and strong generalization under complex lighting and irregular shadow conditions.

(3) Transformer Block

Figure 5 illustrates the modified architecture of the Transformer module. Initially, the feature map undergoes normalization through the Layer Normalization module, which normalizes across channels. Subsequently, the feature map is directed to the Reflection Padding module for expansion. This process involves mirror padding on the right and bottom sides of the feature map, ensuring that the patches within the window are expanded to multiples of the window size. Following this, the feature map enters the Window-based Multi-head Self-Attention (W-MSA) module, where window attention computations are performed.

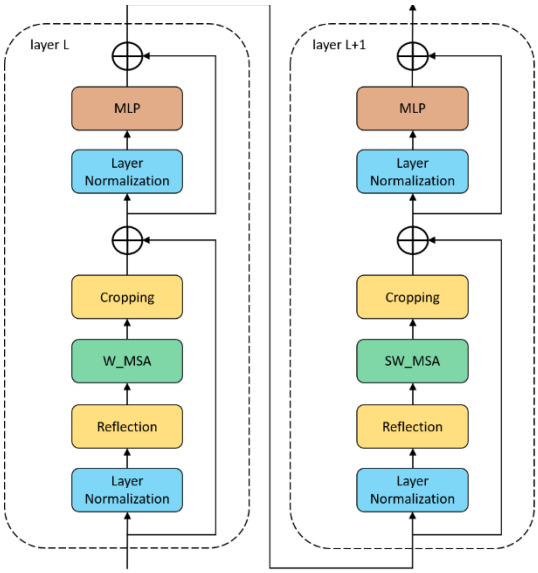

Figure 5. Revised transformer block.

Additionally, relative position embedding is incorporated to enhance the model's spatial awareness. The overall computation is described in (2). Here, Q, K, V represent the Query, Key, and Value vectors, respectively, while B denotes the Relative Position Bias.

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d}} + B\right) \times V \quad (2)$$

Unlike the Vision Transformer, the window attention mechanism restricts computations within each window, eliminating the need for absolute position embedding of feature maps at the input stage. Instead, relative positional biases are incorporated during the self-attention computations within each window. The relative position vectors are combined with the results of the dot product between the query and key vectors, influencing the final attention scores. This approach allows the attention mechanism in each window not only to consider the similarity of features but also to dynamically adjust for positional relationships, enhancing the adaptability and generalization of feature representation. Additionally, this mechanism enables the model to effectively handle feature maps of varying sizes.

To better restore the pixel quality in the edge regions of images, this study employs a mechanism distinct from the Swin Transformer. In the calculation of shifted window attention within the SW-MSA module, Window Masks are not utilized to cover the windows. Instead, the feature maps are mirrored and padded on all four sides to integer multiples of the window size before calculating the window attention.

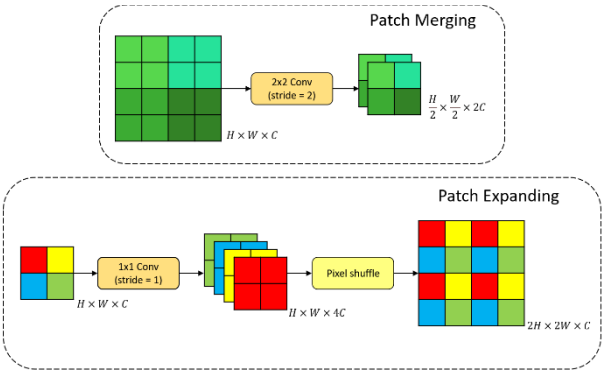

Figure 6. Patch Merging (top) and patch expanding (bottom) block.

This mechanism offers three advantages for shadow recovery models:

(a) Enhanced processing of image edge regions: Traditional padding methods, such as zero padding or cyclic shift mechanisms, may introduce irrelevant information or cause unreasonable element arrangements at image edges, which are detrimental to image restoration models. By reflecting edge pixels, our model effectively maintains consistency in window size at the edges, avoiding biases in model training due to insufficient patch numbers within the windows, thereby improving the processing quality of image edge regions.

(b) Improved quality of feature representation: By using reflection padding to extend the image content naturally at the edges, this method maintains contextual information more effectively compared to other padding techniques, thus enhancing the quality of feature representation.

(c) Reduced additional computational costs: Reflection Padding simplifies computation by removing the need for Window Mask operations. Though slightly more costly than cyclic shift, its impact is minimal on large images, where edge regions are less significant.

(4) Patch Merging & Patch Expanding Block

The architectural framework of the Patch Merging and Patch Expanding modules used in this study is shown in Figure 6. The Patch Merging module employs a 2x2 convolutional kernel with a stride of 2, merging four adjacent patches into one. This approach effectively reduces the feature map by half while doubling the number of channels. Conversely, the Patch Expanding module utilizes a 1x1 convolution to quadruple the channel count of the input feature map. Subsequently, a Pixel Shuffle Layer [23] transforms channel information into pixel information necessary for upsampling, ultimately achieving a twofold increase in the feature map resolution.

C. Loss Function

In this paper, the L1 Loss is employed as the loss function for the shadow removal model. The principal mechanism of the L1 Loss involves measuring the model's error by calculating the absolute differences between the predicted values and the true values. For a given set of input image pairs x_i, y_i where $i = 1$ to n, the L1 Loss function is defined in (3).

$$L1(x_i, y_i) = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \quad (3)$$

Here, x_i represents the true values, y_i denotes the model's predicted values, and n indicates the number of pixels. The L1 Loss function computes the overall loss by summing the absolute differences between the predicted and true values of each pixel and then averaging these sums. Employing the L1 Loss in shadow removal models assists in more accurately restoring details in areas obscured by shadows. Given its involvement in the restoration of image brightness and color, the L1 Loss effectively handles subtle variations in brightness, thereby maintaining the naturalness and visual continuity of the image while removing shadows.

IV. EXPERIMENTAL RESULTS

A. Evaluation Metrics

(1) Structural Similarity Index Measure

The Structural Similarity Index Measure (SSIM) [24] is used to measure visual similarity between images based on luminance, contrast, and structure, which better reflects human perception than pixel-level metrics. SSIM is calculated based on luminance, contrast, and structure:

(a) Luminance Function: Luminance influences human perception. In (4), $L(x, y)$ represents luminance similarity, with μ_x and μ_y as the average luminance of the images, and $C_1 = 6.5025$ to prevent division by zero.

(b) Contrast Function: Contrast refers to the difference between the brightest and darkest parts of an image. In (5), $C(x, y)$ measures contrast by calculating the standard deviations of the images, ensuring similar luminance distribution and range. σ_x and σ_y the images' standard deviations, with $C_2 = 58.5225$ to prevent division by zero.

(c) Structure Function: The structure function evaluates the preservation of details and textures. In (6), $S(x, y)$ calculates the covariance σ_{xy} between images, with $C_3 = 29.26125$ ensuring calculation stability.

$$L(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (4)$$

$$C(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (5)$$

$$S(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \quad (6)$$

$$SSIM(x,y) = \left[L(x,y)^\alpha \cdot C(x,y)^\beta \cdot S(x,y)^\gamma\right] \quad (7)$$

SSIM in (7) uses parameters α, β, and γ to weight its three components, typically set to 1 for balanced evaluation.

(2) Root Mean Square Error

Root Mean Square Error (RMSE) intuitively reflects the magnitude of error by calculating the root mean square difference between predicted and actual values. It is shown in (8), where n represents the number of samples, x_i represents the shadow-free image, and y_i represents the image after shadow removal.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \quad (8)$$

B. Metrics Used to Evaluation

To evaluate the model, SSIM and RMSE were measured on the AISTD test set, analyzing both entire images and shadow/non-shadow regions to assess restoration and preservation. The proposed architecture performs shadow detection followed by removal without requiring shadow masks. Images were resized to 256×256 and trained for 300 epochs using the AdamW optimizer [25], with a learning rate of 2×10^{-4} and batch size of 4.

TABLE I shows SSIM scores above 0.98 in both shadowed and non-shadowed regions, confirming the model's effectiveness in shadow removal and preserving structural integrity without using shadow masks.

TABLE I. COMPARISON OF SSIM ON AISTD

Scheme	Method	Shadow	Non-shadow	All
Mask-Based	Input image	0.926	0.984	0.894
	SP+M-Net [10]	0.987	0.972	0.947
	Auto-Exposure [12]	0.976	0.875	0.840
	Inpaint4Shadow [26]	0.989	0.977	0.960
	ShadowFormer [27]	0.990	0.979	0.966
	RRL-Net [28]	0.990	**0.984**	0.968
Mask-Free	DC-ShadowNet [29]	0.975	0.963	0.921
	G2RShadowNet [30]	0.988	0.975	0.953
	BMNet [31]	0.990	0.977	0.962
	Ours	**0.991**	0.982	**0.969**

TABLE II shows that the proposed model achieves lower RMSE in shadowed areas compared to other mask-free methods, highlighting its accuracy in shadow restoration.

TABLE II. COMPARISON OF RMSE ON AISTD

Scheme	Method	Shadow	Non-shadow	All
Mask-Based	Input image	40.2	2.6	8.5
	CRFormer [13]	5.9	2.9	3.4
	Inpaint4Shadow [26]	5.9	2.9	3.3
	ShadowFormer [27]	5.4	2.4	2.8
	RRL-Net [28]	5.6	**2.3**	**2.8**
Mask-Free	G2RShadowNet [30]	7.3	3.0	3.6
	BMNet [31]	6.1	2.9	3.5
	Ours	**5.2**	2.5	3.0

C. Comparative Results on the AISTD Dataset

Figure 7 illustrates the comparative results on the AISTD dataset against other studies. In the first four columns of tested images, the method proposed in this paper effectively removes shadows while minimizing residual shadows at the shadow boundaries. In the fifth column of images, the colors within the shadow regions are accurately transformed, and the transitions at the shadow edges appear more natural.

Figure 7. Visualization results on AISTD dataset.

D. Ablation Study

To evaluate the proposed methods, ablation experiments were conducted to assess the impact of the SAM-Adapter and Shadow Fusion modules on shadow removal. TABLE III compares the results using the RMSE metric.

Removing the semantic segmentation module hinders accurate shadow region restoration and increases RMSE by affecting non-shadow areas. Without the shadow fusion module, segmentation alone restores structure but causes unnatural transitions between shadow and non-shadow regions.

TABLE III. COMPARISON OF RMSE IN ABLATION STUDIES

Setting	Shadow	Non-shadow	All
Input image	40.2	2.6	8.5
w/o SAM-Adapter	6.4	3.2	3.8
w/o Shadow Fusion	5.8	2.6	3.2
Ours	**5.2**	**2.5**	**3.0**

w/o SAM-Adapter w/o Shadow Fusion Ours

Figure 8. Visualization results on ablation studies.

Figure 8 shows how the SAM-Adapter and Shadow Fusion modules enhance shadow removal. Without semantic segmentation, residual shadows remain. With only semantic segmentation, shadows are removed but without optimal refinement.

V. CONCLUSION

This paper proposes an architecture that combines semantic segmentation and attention mechanisms for shadow removal, enhanced by a shadow fusion module to restore image details. The Semantic Attention Module (SAM) segments shadow and non-shadow regions across diverse scenes, while attention mechanisms capture their relationships. The fusion module refines shadow boundaries, producing high-quality shadow-free images. Experiments on the AISTD dataset show strong performance, with high SSIM and low RMSE scores. The model also generalizes well to various scenes, including game environments, outdoor landscapes, and facial images, demonstrating its strong generalization capability.

ACKNOWLEDGEMENT

This work was supported in part by the NSTC under contract 111-2221-E-011 -134 -, 112-2221-E-468 -023 -.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol. 25, 2012.

[2] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[3] Y. Zhong, X. Liu, D. Zhai, J. Jiang, and X. Ji, "Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon," IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15345–15354.

[4] J. Wang, X. Li, and J. Yang, "Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal," IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1788–1797.

[5] L. Qu, J. Tian, S. He, Y. Tang, and R. W. Lau, "Deshadownet: A multi-context embedding deep network for shadow removal," IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4067–4075.

[6] T. F. Y. Vicente, M. Hoai, and D. Samaras, "Noisy label recovery for shadow detection in unfamiliar domains," IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3783–3792.

[7] A. Kirillov et al., "Segment anything," IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.

[8] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," Medical image computing and computer-assisted intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, 2015, pp. 234–241.

[10] H. Le and D. Samaras, "Shadow removal via shadow image decomposition," IEEE/CVF International Conference on Computer Vision, 2019, pp. 8578–8587.

[11] X. Hu, Y. Jiang, C.-W. Fu, and P.-A. Heng, "Mask-shadowgan: Learning to remove shadows from unpaired data," IEEE/CVF International Conference on Computer Vision, 2019, pp. 2472–2481.

[12] L. Fu et al., "Auto-exposure fusion for single-image shadow removal," IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10571–10580.

[13] J. Wan, H. Yin, Z. Wu, X. Wu, Z. Liu, and S. Wang, "Crformer: A cross-region transformer for shadow removal," arXiv preprint arXiv:2207.01600, 2022.

[14] X. F. Zhang, C. C. Gu, and S. Y. Zhu, "Spa-former: Transformer image shadow detection and removal via spatial attention," arXiv preprint arXiv:2206.10910, 2022.

[15] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[16] A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[17] M. I. Jordan, "Serial order: A parallel distributed processing approach," Advances in psychology, vol. 121, Elsevier, 1997, pp. 471–495.

[18] A. Graves and A. Graves, "Long short-term memory," Supervised sequence labelling with recurrent neural networks, pp. 37–45, 2012.

[19] Y. Song, Z. He, H. Qian, and X. Du, "Vision transformers for single image dehazing," IEEE Transactions on Image Processing, vol. 32, pp. 1927–1941, 2023.

[20] T. Chen et al., "SAM Fails to Segment Anything?–SAM-Adapter: Adapting SAM in Underperformed Scenes: Camouflage, Shadow, Medical Image Segmentation, and More," arXiv preprint arXiv:2304.09148, 2023.

[21] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 510–519.

[22] R. M. Haralick, S. R. Sternberg, and X. Zhuang, "Image analysis using mathematical morphology," IEEE Transactions on Pattern Analysis and Machine Intelligence, no. 4, 1987, pp. 532–550.

[23] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1874–1883.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, 2004, pp. 600–612.

[25] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," arXiv preprint arXiv:1711.05101, 2017.

[26] X. Li et al., "Leveraging inpainting for single-image shadow removal," IEEE/CVF International Conference on Computer Vision, 2023, pp. 13055–13064.

[27] L. Guo, S. Huang, D. Liu, H. Cheng, and B. Wen, "Shadowformer: Global context helps image shadow removal," arXiv preprint arXiv:2302.01650, 2023.

[28] Y. Liu, Z. Ke, K. Xu, F. Liu, Z. Wang, and R. W. Lau, "Recasting regional lighting for shadow removal," AAAI Conference on Artificial Intelligence, 2024, vol. 38, no. 4, pp. 3810–3818.

[29] Y. Jin, A. Sharma, and R. T. Tan, "Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network," IEEE/CVF International Conference on Computer Vision, 2021, pp. 5027–5036.

[30] Z. Liu, H. Yin, X. Wu, Z. Wu, Y. Mi, and S. Wang, "From shadow generation to shadow removal," IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4927–4936.

[31] Y. Zhu, J. Huang, X. Fu, F. Zhao, Q. Sun, and Z.-J. Zha, "Bijective mapping network for shadow removal," IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5627–563.