IJHSR
ijhighschoolresearch.org
REVIEW ARTICLE

Shadow Removal Based on Deep Learning

Heyuan Fang

Holy Trinity School, 11300 Bayview Avenue, Richmond Hill, ON, L4S 1L4, Canada; henryfang5908@gmail.com

ABSTRACT: This study evaluates ST-CGAN and RASM for shadow removal using the ISTD and ISTD+ datasets, comparing their accuracy, efficiency, and generalization. Results show ST-CGAN improves with training, reducing RMSE from 16.39 at epoch 1 to 9.64 at epoch 500, but gains plateau after 100 epochs. Training on ISTD+ lowers RMSE further, yet RASM significantly outperforms ST-CGAN, achieving an RMSE of 2.53 on ISTD+ compared to ST-CGAN's projected 5.02 at epoch 10,000. In addition to these two models, recent transformer-based methods such as ShadowFormer and HomoFormer have demonstrated state-of-the-art results on ISTD+ and SRD benchmarks. These findings highlight that RASM, which leverages a regional attention mechanism within a transformer framework, achieves superior accuracy and computational efficiency compared to earlier CNN-based approaches and other state-of-the-art transformer models, establishing it as a practical solution in the field of shadow removal.

KEYWORDS: Robotics and Intelligent Machines, Machine Learning, Computer Vision, Shadow Removal, ST-CGAN, ShadowFormer, HomoFormer, RASM.

Introduction

"Where there is light, there are shadows.":

Shadows occur when light is obstructed by objects–a natural and unavoidable phenomenon in human life that only rarely obstructs vision with any memorable significance. In robotics applications, however, shadows can cause frequent recognition errors in tasks such as object detection and tracking. These recurring failures, observed for example in competitive robotics systems, highlight the practical importance of effective shadow removal as a preprocessing step. Because robotic platforms typically operate under constrained computational resources, this motivates the exploration of lightweight shadow removal strategies that balance accuracy with efficiency.[1]

In the realm of computer vision, shadows present significant challenges for object detection, tracking, and segmentation. In finer image modification work, shadows impact light resolution—shadowed regions typically have lower luminance values, leading to altered color intensities and reduced accuracy in shadowed areas. Consequently, effective shadow removal is an essential prerequisite for nearly all computer vision applications and has been a subject of scholarly attention for decades.[2]

Artificial Intelligence models are poised to perform the majority of shadow removal in Image processing. While it is well known that AI models improve over time by incorporating new data and enhancing their predictive accuracy, the effectiveness of AI at shadow removal is constrained by several technical bottlenecks, most notably computational limitations within real-world constraints.[3] e.g., Autonomous vehicles will require exceptionally efficient image processing models to make sense of fast-moving objects in heterogeneous scenarios quickly. As of now, Deep Learning has achieved outstanding success in audio and speech processing, natural language processing (NLP), and numerical data analysis.[4] In recent years, deep learning-based shadow removal methods have demonstrated superior performance, primarily due to the availability of extensive training data.[5] With the rise in GPU capabilities, deep neural networks have become a central focus of modern shadow removal research. These models offer higher accuracy and efficiency compared to physical-model-based approaches. However, they also introduce new challenges, primarily the reliance on large and diverse datasets. Recent surveys[2,6] provide a comprehensive overview of shadow removal research from 2017 to 2023, illustrating a clear progression from early CNN-based methods toward transformer-based architectures[4,7] and diffusion-based approaches.[7,8] While these models achieve state-of-the-art results on widely used benchmarks such as ISTD+ and SRD, they generally require more computational resources than CNN or lightweight designs, making them less suitable for deployment in resource-constrained environments. This underscores the importance of evaluating new lightweight frameworks such as RASM not only against earlier CNN approaches but also within the broader trajectory of shadow removal research.

Although transformer- and diffusion-based architectures have recently advanced the field, embedded systems, mobile devices, and autonomous platforms still require methods that balance accuracy with efficiency. Within this context, two representative approaches illustrate distinct strategies. The first is the Stacked Conditional Generative Adversarial Network (ST-CGAN), an early CNN-based model that stacks two CGANs—one for shadow detection and one for shadow removal—thereby providing an end-to-end pipeline and demonstrating the benefits of multitask design under constrained computational budgets. By contrast, the Regional Attention Shadow Removal Model (RASM) reflects a newer, lightweight, region-aware paradigm: by enabling adaptive interaction between shadowed and non-shadowed areas, RASM leverages contextual correlation to restore shadowed content with improved accuracy while maintaining efficiency.

DOI: 10.36838/v8i3.63

While both models aim to improve shadow removal under limited resources, they embody different design philosophies—ST-CGAN as a canonical CNN-based baseline and RASM as a lightweight region-aware framework. This study compares these two approaches to analyze their respective strengths, limitations, and trade-offs in low-compute scenarios. The analysis begins with a brief review of shadow removal challenges, proceeds with a methodological comparison and experimental evaluation, and concludes with observations on practical implications.

Shadow Removal: A Scholarly Review:

Prior work has extensively examined shadow removal as a means to improve downstream computer vision tasks.[3,9] Traditional methods often relied on physical illumination and reflection models,[3] treating each pixel individually according to its lighting conditions. While these approaches can effectively restore images, they tend to be time-consuming and often require manual user interactions, limiting their practicality for large-scale or real-time applications. For example, breakthroughs in color detection for light restoration require a level of image comprehension that would be too costly to program into models running on home devices.[10] Patch-based shadow removal strategies[11] align more closely with deep learning models but would still be too slow for time-critical applications such as autonomous vehicles. In addition, filtering-based strategies like bilateral decomposition[12] were employed to separate base and detail layers for shadow removal. Arbel and Hel-Or[13] argued that shadow removal systems are inherently skewed towards inefficiency because "Shadows in images are typically affected by several phenomena in the scene, including physical phenomena such as lighting conditions, type and behavior of shadowed surfaces, occluding objects, etc.," suggesting profound detail orientation is necessary for complete artifact removal.

With the rise of deep learning, researchers began to address these inefficiencies. AI-assisted shadow removal has progressed beyond texture recognition and 3D modeling.[14,15] For instance, ST-CGAN[9] provided an end-to-end solution that simultaneously handled shadow detection and removal by stacking two conditional GANs, enhancing performance through mutual reinforcement of both tasks. Other studies emphasized context modeling: recent work highlighted concerns about artifact distortion at the shadow border, and proposed programs to model the correlation between shadowed and non-shadowed regions.[4,5] To further improve efficiency, several works[16] focused on increasing the number of training iterations and reducing computational demand. Wang, Li, and Yang[9] advanced this line of research by introducing a bijective mapping network, coupling the procedures of shadow removal and shadow generation in a unified parameter-shared framework. This approach effectively recovered the underlying background contents during the forward shadow removal process. However, their method still required additional programming layers to manage color-rich images, suggesting that dataset limitations remained unresolved.

Two primary strategies have been developed to overcome dataset limitations: Dataset Enhancement and Shadow Simulation Models. Dataset limitations. Dataset Enhancement involves creating shadow masks (binary images indicating shadowed and shadow-free areas) and shadow-free patches (manually edited versions of original images with shadows removed) to expand the dataset. While this method improves model performance, it requires significant human effort, particularly during the shadow removal and masking processes. Shadow Simulation Models artificially generate shadows on existing images to augment datasets. Though effective in increasing data volume, the quality of simulated shadows heavily depends on the diversity of the original dataset. A lack of variation in shadow patterns limits the effectiveness of this approach.

More recently, surveys[2, 17] have shown that shadow removal research from 2017 to 2023 has progressed from early CNN-based methods toward transformer-based architectures[4,6] and diffusion-based approaches[16]. These newer models achieve state-of-the-art performance on benchmarks such as ISTD+ and SRD, and their global context modeling and generative priors provide clear advantages in handling complex shadow patterns and boundary artifacts. While such models demonstrate clear advantages, this study restricts its scope—due to experimental constraints and the focus on robotics environments under limited computational resources—to two representative efficiency-oriented methods: ST-CGAN and RASM.

■ Methodology

This study evaluates the effectiveness of two shadow removal models: the Stacked-Conditional Generative Adversarial Network (ST-CGAN) and the Regional Attention Shadow Removal Model (RASM). The shadow removal process is divided into two main stages: (1) identifying shadowed regions and (2) reconstructing and refining these regions using different computational approaches. Each method is assessed independently to compare its effectiveness in deshadowing.

ST-CGAN is implemented using two separate Conditional Generative Adversarial Networks (CGANs) for the two stages of shadow removal. Python was chosen as the primary programming language due to its extensive machine learning libraries and the prevalence of prior implementations in Python.

As one of the transformers, the shadow removal model, using the Retinex-based model, ShadowFormer, introduced multiple channel-spatial attention mechanisms. Using the Shadow-Interaction Module along with the Shadow-Interaction Attention, ShadowFormer could build correlations between the shadowed and non-shadowed regions and use information from the shadow-free portion for the restoration of the image. ShadowFormer used the idea of a window, in which the transformer would only apply to a specific area (a channel), to compensate for the large calculation complexity and cost.[4]

As another development of implementing transformer use in the task of shadow removal, HomoFormer addressed the issue of the non-uniformity of the shadow in the given image.

Non-uniformity imposes a constraint on weight-sharing models, where they struggle to seek a compromise among regions of various degenerated degrees. HomoFormer implements a random shuffle mechanism in its encoding process to homogenize the degree of shadow degradation, while a de-shuffle in the decoder layer restores the image.[17]

RASM is a lightweight shadow removal model that leverages non-shadow areas to assist in restoring shadowed regions. By implementing a regional attention module, RASM has a regional attention module to continuously learning the correlation between adjacent shadow and non-shadow regions. Differentiating from ShadowFormer and HomoFormer, RASM took a step back and aggregated information from its adjacent non-shadowed regions in its regional attention module. This approach strikes a balance between model complexity and performance, optimizing both accuracy and computational efficiency.

To evaluate the performance of these models, we used the ISTD+ dataset throughout the training and testing phases. Model effectiveness was measured using Root-Mean-Square Deviation (RMSE) and PSNR, MAE, and SSIM. Additionally, we varied the number of training epochs to analyze the relationship between task complexity and image reconstruction quality.

ISTD Dataset:

The ISTD dataset is a widely used benchmark for shadow removal tasks, comprising 1,870 image triplets across 135 diverse scenes that feature various shadow shapes and lighting conditions. The dataset is divided into 1,330 triplets for training and 540 triplets for testing, facilitating model evaluation and generalization.

Each triplet in the dataset comprises the following components:

1. A shadowed image – the original input image containing natural shadows.
2. A shadow mask – a binary mask delineating the shadowed regions.
3. A shadow-free image – an image where shadows have been manually removed following the shadow mask.

The shadow removal in ISTD is performed manually, ensuring precise adherence to the shadow mask. This high-quality annotation enables Convolutional Neural Networks (CNNs) and Conditional Generative Adversarial Networks (CGANs) to effectively learn shadow localization from the shadow mask, while also identifying differences in luminance, texture, and resolution between shadowed and shadow-free images. By leveraging these structured triplets, the models can be trained to accurately detect and reconstruct shadowed areas, improving overall deshadowing performance.

Figure 1: Example of a triplet: Shadow-Free Image, Shadow Mask, and Original Image.[9]

Assessing the Accuracy of Models:

We use the root mean square error (RMSE) as the metric to quantify the discrepancy between the ground-truth shadow-free image and the recovered image, with lower values indicating higher accuracy or lower distortion.

The RMSE is determined by the equation:[18]

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

Figure 2: Root Mean Square Error (RMSE) Equation.[18]

We consider two input sets of images, I1 and I2, where each image consists of n pixels. For each corresponding pixel in the two images, we compute the difference between their RGB values, take the average of these differences, and then compute the square root of this average to determine the overall error. This results in the Root Mean Squared Error (RMSE), which quantifies the difference between the two images. The detailed RMSE calculation is provided in the Appendix.

Methodologies for Shadow Removal: ST-CGAN, ShadowFormer, HomoFormer, and RASM:

Shadow Removal Using ST-CGAN:

Shadow removal from a single image involves two fundamental tasks: shadow detection and shadow removal. The Stacked Conditional Generative Adversarial Network (ST-CGAN) is an architecture designed to perform both tasks jointly, enabling end-to-end learning for improved accuracy and efficiency.

The ST-CGAN architecture consists of two stacked Conditional Generative Adversarial Networks (CGANs):

1. Shadow Detection Network – Identifies shadowed regions and generates a shadow mask.
2. Shadow Removal Network – Uses the shadow mask along with the original image to reconstruct a shadow-free version.

Each CGAN comprises a generator and a discriminator, working in tandem to enhance the realism and accuracy of the shadow removal process. The generator aims to produce an image where shadows are effectively removed, while the discriminator evaluates the authenticity of the generated output, pushing the model toward higher-quality reconstructions.

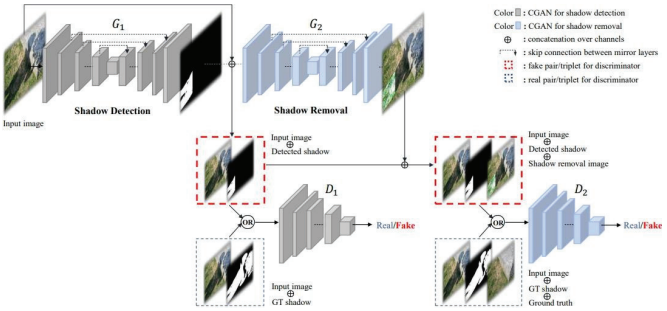

Figure 3: Architecture of the Proposed ST-CGAN: Stacked CGANs for Shadow Detection and Removal.[9]

Shadow Removal Using Shadowformer:

Shadowformer introduces a multiple-scale channel-spatial attention mechanism within a Transformer framework. This

DOI: 10.36838/v8i3.63

design addressed the challenge of other transformer models of integrating global semantics from deep feature layers with local details from shallow feature layers.

ShadowFormer started with the Retinex-based shadow model that handles shadow degradation to allow models to draw information from non-shadow regions to restore the shadow region. The restoration process is done through an encoder-decoder under a single-stage transformer framework via channel attention to efficiently multiple-stack the hierarchical information. Then, the correlation between the shadow vs non-shadow region is exploited by the Shadow-Interaction Module with Shadow-Interaction Attention. This allowed ShadowFormer to address the challenge of colour inconsistency and boundary trace in the restored shadow-free images.

Here, we present a quick overview of the architecture of ShadowFormer. The process begins with 2 inputs, the shadow image along with the shadow mask, which is linearly projected, which maps the inputs into a latent feature space. In the encoder stage, features are processed using a channel attention model and a standard transformer structure. Then, the features enter the decoder stage, which mirrors the encoder by using channel attention modules to reconstruct the spatial details. A linear projection layer transformation is connected to feature it back into image space.

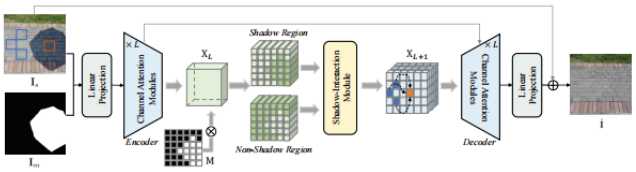

Figure 4: Architecture of the Proposed ShadowFormer.[4]

Shadow Removal Using HomoFormer:

The HomoFormer is an advanced shadow removal model that leverages a homography-inspired Transformer architecture to model the relationship between shadow and non-shadow regions. A foundation based on other transformer architectures that would focus on the non-uniform issue of shadow and as well as the classic self-attention method. The image shuffle strategy involves upsampling and downsampling to exchange information between channels and space, while preventing the spatial arrangements of pixels. At the end, with an implementation of self-attention to concentrate on regional attention, HomoFormer decreased computing complexity compared to other models.

This model compromises a conflict of previous transformer models: previous models suffer from either a quadratic increase in complexity as the resolution of the image increases, or the weight sharing when dealing with non-uniform shadow degradation. HomoFormer implemented the strategy to homogenize the non-uniform distribution.

Here, we present a brief overview of the HomoFormer structure. The process begins with the input project stage, where the images with shadow and the shadow masks are mapped into a feature space, which is the core parts: the HomoBlocks integrates Layer Normalization, local self-attention with random shuffle, and SMLP (standard transformer used). The regional attention is also implemented in this section to minimize complexity in calculation. During the down and then up-sampling layers, progressive reduction in the spatial resolution while increasing feature richness is applied with downsampling, while the upsampling strategy is used to restore the spatial resolution. This process is able to preserve important details lost during the traditional encoder-decoder architecture. The features would lastly pass through an output projection layer to generate the image.

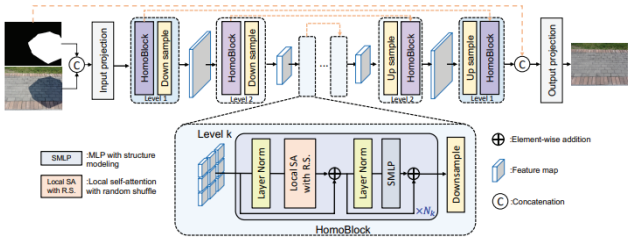

Figure 5: Architecture of the Proposed HomoFormer.[17]

Shadow Removal Using RASM:

The Regional Attention Shadow Removal Model (RASM) is a lightweight yet effective shadow removal approach that utilizes non-shadow regions to enhance the reconstruction of shadowed areas. By incorporating regional attention mechanisms, RASM enables context-aware interactions between shadowed and non-shadowed areas, facilitating a more accurate and natural restoration process.

This model is designed to strike an optimal balance between computational efficiency and accuracy, ensuring effective shadow removal while maintaining manageable model complexity. Specifically, the regional attention module could focus on a specific area of the matrix and therefore avoid being excessive yet non-informative. In doing so, RASM achieved only 1/4 of GFLOPs. Through its regionally contextual approach, RASM enhances the overall quality of shadow-free images while reducing computational overhead.

Here, we present a brief overview of the RASM structure. On the left side (a) was the process of model and shadow contraction area, which employs the Channel Attention Module (b). After allowing the global information to interact, Channel Attention enters and eventually concludes in a spatial information interaction.

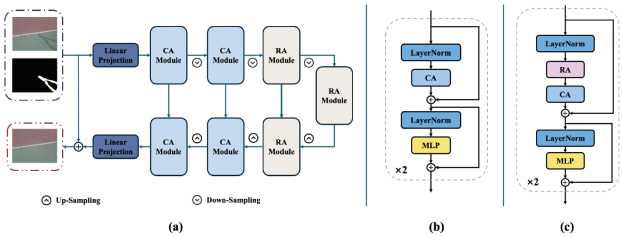

Figure 6: Demonstration of RASM.[19]

Experiment

For computational resources, we employed Google Colab as the primary cloud-based virtual machine. The training duration per epoch ranged from 3,087 to 3,850 seconds on a CPU and approximately 90.79 to 95.69 seconds on a GPU. However, GPU availability on Colab was limited, restricting the

number of training epochs to well below 500. To supplement this, we conducted additional tests on a consumer-grade laptop equipped with an NVIDIA GeForce GTX 1650 GPU with 16GB of memory. The GPU performance on this system was comparable to that of Colab, making it a reasonable benchmark for evaluating model feasibility on consumer hardware.

To systematically evaluate the performance of the shadow removal models, we conducted a series of experiments under varying computational environments and training durations. The experiments were performed using both Google Colab and a personal computing device, with training durations of 1, 100, and 500 epochs. The effectiveness of the models was assessed using both qualitative visual analysis and quantitative evaluation via Root-Mean-Square Deviation (RMSE).

To investigate the potential influence of computational environments on model performance, we trained the model for 100 epochs using both Google Colab and a personal laptop. Figure 4 and Figure 5 illustrate six representative examples from these experiments, where each row corresponds to a different test sample and each column represents the original shadowed image, the ground truth shadow-free image, the shadow-free image generated by the model, the ground truth shadow mask, and the shadow mask produced by the model. A qualitative analysis of the results revealed no significant differences between the outputs generated on the two platforms. Samples that were successfully processed on one platform were also successfully processed on the other, while failure cases remained consistent across both environments. This observation was further confirmed by the RMSE values, which indicated that the personal laptop achieved performance comparable to the Colab-based model. These findings suggest that the choice of computational environment does not substantially impact the model's performance, validating the robustness of the implementation across different hardware configurations.

Results and Discussion

Experiment Results:

To assess the impact of training duration on shadow detection and removal, we compared the results obtained after 1,100 and 500 epochs, all conducted on a personal laptop (Figures 6 and 7). After a single epoch of training, the model exhibited poor generalization, failing to accurately detect shadow boundaries or reconstruct shadow-free images. The generated shadow masks were imprecise, and the overall visual quality of the outputs was suboptimal. This was reflected in a high RMSE of 16.39, indicating a significant deviation from the ground truth. These results highlight the necessity of extended training for the model to learn meaningful representations of shadow regions and their corresponding shadow-free reconstructions.

After 100 epochs, the model demonstrated substantial improvements in both shadow detection and removal. The generated shadow-free images closely resembled the ground truth, with only minor imperfections in certain cases. This was quantitatively supported by a significant reduction in RMSE to 10.09, indicating enhanced accuracy and improved reconstruction quality. However, increasing the training duration further to 500 epochs yielded only marginal improvements. While the RMSE decreased slightly from 10.09 to 9.64, the visual differences were negligible. This suggests that beyond a certain point, additional training yields diminishing returns, as the model reaches a plateau in performance where further refinement offers minimal perceptible enhancement in shadow removal quality.

The experimental results lead to three key observations:

- The computational environment does not significantly affect model performance, as both Colab and a personal laptop produced comparable outputs and RMSE values.
- Training duration has a substantial impact on model effectiveness, particularly in the early stages, as demonstrated by the significant improvements between 1 and 100 epochs.
- Beyond 100 epochs, further training yields only incremental gains, with minimal reductions in RMSE and imperceptible improvements in visual quality.

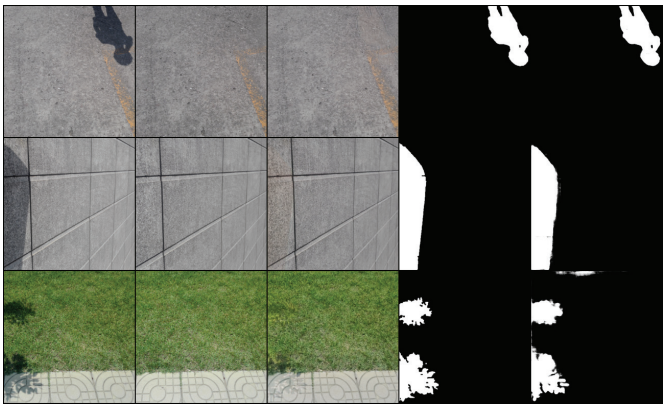

Figure 7: Completed with Colab, 100 epochs. Here, we demonstrate 3 examples (corresponding to 3 rows) of the output of our program. Each column, from left to right, corresponds to: the shadowed picture from the dataset, the shadow-free image from the dataset, the shadow-free image the code generated, the shadow mask the dataset contains, and the shadow mask the code generated.

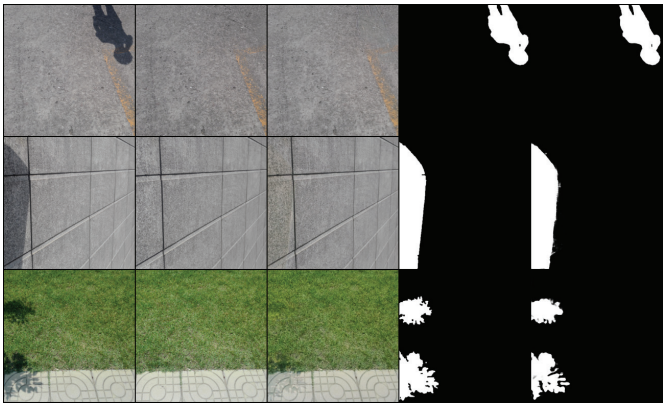

Figure 8: Completed on personal laptop, 100 epochs. As a result, between a personal laptop (Figure 8) vs Colab (Figure 7), we could eyeball and conclude that there is almost no difference. Therefore, we conclude that the computational environment does not significantly affect model performance.

DOI: 10.36838/v8i3.63

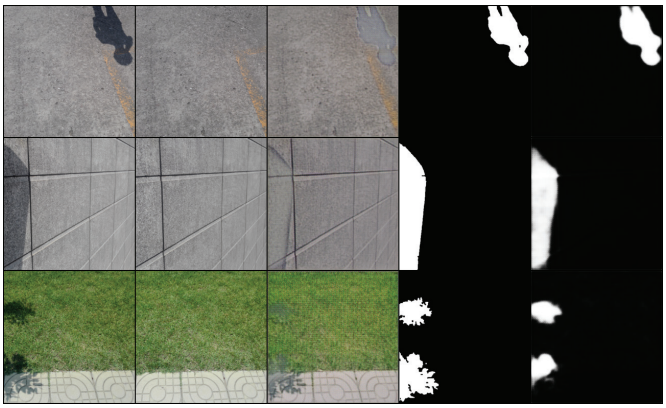

Figure 9: Completed on personal laptop, 1 epoch. Compared with the personal laptop after 100 epochs (Figure 8), the output after 100 epochs on a personal laptop (Figure 8) shows a significantly higher quality.

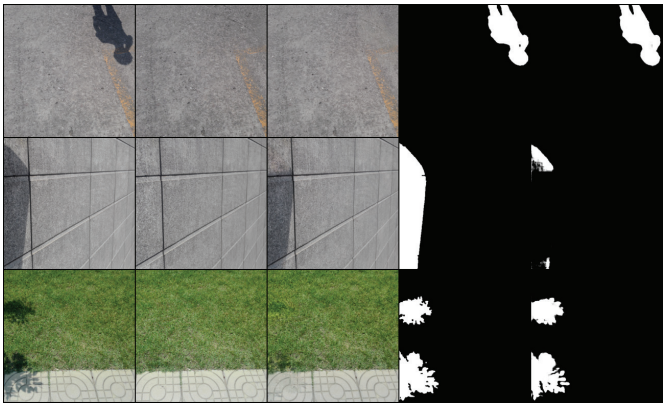

Figure 10: Completed on personal laptop, 500 epochs. At this point, it is difficult to eyeball any difference between 100 epochs (Figure 8) and 500 epochs (Figure 10), which means the improvements are gradually decreasing during this process; however, there could still be a non-obvious improvement.

Effect of Training Epochs on RMSE with ST-CGAN:

Table 1: RMSE was tested with different training epochs using ST-CGAN.

Dataset	epoch=1	epoch=10	epoch=100	epoch=300	epoch=500	epoch=10000
ISTD	16.39	12.83	10.09	9.88	9.64	7.47

Table 1 illustrates the Root Mean Squared Error (RMSE) values obtained at different training epochs for the ISTD dataset. RMSE serves as a quantitative measure of deshadowing effectiveness, where lower values indicate better performance. The results demonstrate a clear downward trend in RMSE as the number of training epochs increases, suggesting that extended training enhances the model's ability to remove shadows.

At epoch 1, the RMSE is 16.39, indicating a relatively high error in shadow removal. As training progresses to epoch 10, the RMSE decreases to 12.83, reflecting an improvement in the model's capacity to reconstruct shadow-free images. A more significant reduction is observed at epoch 100, where the RMSE drops to 10.09, indicating substantial progress in learning. However, after epoch 100, the reduction in RMSE becomes less pronounced. At epoch 300, the RMSE is 9.88, and at epoch 500, it further declines slightly to 9.64. This diminishing improvement suggests that the model approaches its performance plateau, where additional training yields only marginal enhancements.

For comparison, the ST-CGAN paper[9] reports an RMSE of 7.47 at epoch 10,000, which is significantly lower than the RMSE values achieved in our experiments. This discrepancy suggests that extended training beyond 500 epochs may further enhance the model's performance.

Effect of Training Epochs on RMSE between Models:

Table 2: RMSE compared with different training epochs between ST-CGAN and RMSE.

Dataset	Model	epoch=1	epoch=10	epoch=100	epoch=300	epoch=500	epoch=10000
ISTD	ST-CGAN	16.39	12.83	10.09	9.88	9.64	7.47
ISTD+	ST-CGAN	14.06	9.14	6.97	6.67	6.48	5.02(estimated)
ISTD+ RMSE / ISTD RMSE		0.858	0.712	0.691	0.675	0.672	0.672

To ensure a fair comparison between the ST-CGAN and RASM models, we used a consistent dataset for evaluation. The ST-CGAN model was originally trained and tested on the ISTD dataset. In contrast, the RASM model was evaluated using the ISTD+ dataset, a refined version of ISTD that addresses illumination inconsistencies between shadowed and shadow-free images. To facilitate direct comparison, we recalculated the RMSE values for the ST-CGAN model using the ISTD+ dataset at training epochs 1, 10, 100, 300, and 500. The recalculated RMSE values for ST-CGAN are derived from our experimental results, while the RMSE values for RASM are taken from "Regional Attention for Shadow Removal."[19]

According to Table 2, the comparison between the ST-CGAN model trained on ISTD vs. ISTD+ reveals notable differences in shadow removal effectiveness. At the initial training stage (epoch 1), the RMSE for ISTD+ is 14.06, slightly outperforming ISTD's 16.39, yielding an RMSE ratio of 0.858. A more pronounced improvement is observed at epoch 10, where the RMSE for ISTD+ decreases to 9.14, compared to 12.83 for ISTD, resulting in an RMSE ratio of 0.712. However, beyond epoch 10, the reduction in RMSE for ISTD+ becomes less significant in comparison to ISTD. By epoch 100, the ratio stabilizes at 0.691, then further decreases to 0.675 at epoch 300, and 0.672 at epoch 500. This trend suggests that the RMSE ratio is approaching a steady state. Assuming this ratio remains consistent up to epoch 10,000, we estimate the RMSE for ISTD+ at epoch 10,000 to be approximately 5.02.

According to the RASM paper,[19] the RMSE of the RASM model trained on ISTD+ is reported as 2.53, which is significantly lower than the projected RMSE of ST-CGAN from our experiments. This substantial difference indicates that RASM achieves superior performance in shadow removal compared to ST-CGAN, particularly in terms of quantitative accuracy. The results suggest that the regional attention mechanism employed in RASM is more effective in leveraging contextual information for shadow reconstruction, leading to a more precise removal process.

Comparison with Advanced Transformer Models:

Table 3: The performance of the 3 models mentioned, with measurements of PSNR, SSIM, and RMSE, using the ISTD+ dataset.[6]

Method	Shadow Region			Non-Shadow Region			All-Region		
	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE
ShadowFormer	39.48	0.992	5.23	38.82	0.983	2.30	35.46	0.971	2.78
HomoFormer	39.49	0.993	4.73	38.75	0.984	2.23	35.35	0.975	2.64
RASM	40.73	0.993	4.41	39.23	0.985	2.17	36.16	0.976	2.53

Table 4: The performance of the 3 models mentioned, with measurements of PSNR, SSIM, and RMSE, using the SRD dataset.[6]

Method	Shadow Region			Non-Shadow Region			All-Region		
	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE
ShadowFormer	35.55	0.982	6.14	36.82	0.983	3.54	32.46	0.957	4.28
HomoFormer	38.81	0.987	4.25	39.45	0.988	2.85	35.37	0.972	3.33
RASM	37.91	0.988	5.02	38.70	0.992	2.72	34.46	0.976	3.37

Table 5: The comparison of 4 models mentioned about the number of parameters and FLOPs.[4,20]

Method	#Params. (M)	FLOPs (G)
ST-CGAN	29.2	17.9
ShadowFormer	11.4	63.1
HomoFormer	17.8	38.4
RASM	5.2	25.2

Beyond CNN-based methods and lightweight designs such as RASM, recent studies have introduced transformer and diffusion architectures that achieve state-of-the-art results on benchmark datasets. For example, ShadowFormer reports an RMSE of 2.78, PSNR of 35.46 dB, and SSIM of 0.971 on the ISTD+ dataset.[4] Similarly, HomoFormer achieves an RMSE of 2.64, PSNR of 35.35 dB, and SSIM of 0.975,[7] reflecting its ability to model non-uniform degradation through homogenized attention blocks.

These figures illustrate that transformer-based models currently surpass CNN approaches in absolute accuracy, particularly in handling complex shadow boundaries and color inconsistencies. However, they are also characterized by significantly higher computational demands, often requiring multiple GPUs or extended training times.

RASM has a small number of parameters and low FLOPs, utilizing a negligible amount of computational resources while achieving superior performance, demonstrating that RASM effectively balances model complexity and model performance.

In this study, the experimental comparison is limited to ST-CGAN and RASM due to hardware constraints typical of robotics environments, where platforms must balance accuracy against strict efficiency requirements. Nevertheless, as robotic systems increasingly adopt more powerful GPUs, the incorporation of advanced architectures may become both feasible and advantageous, enabling higher-fidelity perception in dynamic real-world tasks.

■ Conclusion

This study compared ST-CGAN and RASM for shadow removal, evaluating their effectiveness, computational efficiency, and accuracy using the ISTD and ISTD+ datasets. Our experiments demonstrated that ST-CGAN benefits from extended training, with RMSE decreasing from 16.39 at epoch 1 to 9.64 at epoch 500 on ISTD. However, beyond 100 epochs, improvements became marginal, indicating a performance plateau. Additionally, results confirmed that the choice of computational environment (Google Colab vs. personal laptop) had no significant impact on model performance.

Training ST-CGAN on ISTD+ consistently resulted in lower RMSE values, highlighting the role of dataset refinement in improving shadow removal accuracy. However, RASM significantly outperformed ST-CGAN, achieving an RMSE of 2.53 on ISTD+, compared to ST-CGAN's projected 5.02 at epoch 10,000. This suggests that RASM's regional attention mechanism more effectively restores shadow-free images while maintaining computational efficiency.

Beyond these two models, transformer-based methods such as ShadowFormer, HomoFormer have recently set new benchmarks, with RMSE values under 2.8 and SSIM above 0.97. These results indicate that advanced architectures offer superior absolute accuracy, particularly for complex shadow boundaries. However, they remain computationally demanding, making them less feasible for current low-compute robotic platforms.

Taken together, our findings suggest that lightweight region-aware frameworks like RASM currently provide the best trade-off between efficiency and accuracy in resource-constrained settings. Future research should investigate strategies related to RASM approaches, emphasizing the use of regional channels and a continuously modifying window to minimize calculation complexity and thereby lower hardware requirements for its further use in applications under the current hardware level of robots. As robotics and embedded systems increasingly gain access to high-performance GPUs, it will become both feasible and advantageous to deploy these advanced models, enabling higher-fidelity visual perception in dynamic real-world environments.

■ Acknowledgments

I would like to sincerely thank all those who contributed to this paper, including all guidance teachers/professors for their invaluable support and for providing this exceptional environment. I would like to specifically thank Professor Yashtini from Georgetown University for her guidance, patience, and encouragement during the summer of 2024.

■ References

1. Alzayat Saleh, Alex Olsen, Jake Wood, Bronson Philippa, & Mostafa Rahimi Azghadi. (2025). FieldNet: Efficient real-time shadow removal for enhanced vision in field robotics. Expert Systems with Applications, 127442. https://doi.org/10.1016/j.eswa.2025.127442
2. Zhu, X., Chow, C.-O., & Chuah, J. H. (2024). From darkness to clarity: A comprehensive review of contemporary image shadow removal research (2017–2023). Image and Vision Computing, 148, 105100. https://doi.org/10.1016/j.imavis.2024.105100
3. Rubinger, L., Gazendam, A., Ekhtiari, S., & Bhandari, M. (2023). Machine learning and artificial intelligence in research and healthcare. Injury, 54(Suppl 3), S69–S73. https://doi.org/10.1016/j.injury.2022.01.046

DOI: 10.36838/v8i3.63

4. Guo, L., Huang, S., Liu, D., Cheng, H., & Wen, B. (2023, June). Shadowformer: Global context helps shadow removal. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 710–718). https://doi.org/10.1609/aaai.v37i1.25148
5. Liang, J., Wang, J., Xu, W., Hou, L., & Zhou, B. (2025). Towards hard and soft shadow removal via dual-branch separation network and vision transformer. arXiv. https://doi.org/10.48550/arXiv.2501.01864
6. Guo, L., Wang, C., Wang, Y., Yu, Y., Huang, S., Yang, W., Kot, A. C., & Wen, B. (2024). Single-Image Shadow Removal Using Deep Learning: A Comprehensive Survey (arXiv:2407.08865). arXiv. https://doi.org/10.48550/arXiv.2407.08865
7. Mei, K., Figueroa, L., Lin, Z., Ding, Z., Cohen, S., & Patel, V. M. (2023). Latent feature-guided diffusion models for shadow removal. arXiv. https://doi.org/10.48550/arXiv.2312.02156
8. Luo, J., Li, R., Jiang, C., Zhang, X., Han, M., Jiang, T., Fan, H., & Liu, S. (2024). Diff-Shadow: Global-guided diffusion model for shadow removal. arXiv. https://doi.org/10.48550/arXiv.2407.16214
9. Wang, J., Li, X., & Yang, J. (2018). Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1788–1797. https://doi.org/10.1109/CVPR.2018.00192
10. Finlayson, G. D., Hordley, S. D., Lu, C., & Drew, M. S. (2006). On the removal of shadows from images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(1), 59–68. https://doi.org/10.1109/TPAMI.2006.18
11. Zhang, L., Zhang, Q., & Xiao, C. (2015). Shadow remover: Image shadow removal based on illumination recovering optimization. IEEE Transactions on Image Processing, 24(11), 4623–4636. https://doi.org/10.1109/TIP.2015.2465159
12. Yang, Q., Tan, K.-H., & Ahuja, N. (2012). Shadow removal using bilateral filtering. IEEE Transactions on Image Processing, 21(10), 4361–4368. https://doi.org/10.1109/TIP.2012.2208976
13. Arbel, E., & Hel-Or, H. (2011). Shadow removal using intensity surfaces and texture anchor points. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(6), 1202–1216.https://doi.org/10.1109/TPAMI.2010.157
14. Finlayson, G. D., Drew, M. S., & Lu, C. (2009). Entropy minimization for shadow removal. International Journal of Computer Vision, 85(1), 35–57. https://doi.org/10.1007/s11263-009-0243-z
15. Liu, F., & Gleicher, M. (2008). Texture-consistent shadow removal. In D. Forsyth, P. Torr, & A. Zisserman (Eds.), Computer Vision – ECCV 2008 (Vol. 5305). Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-88693-8_32
16. Le, H., & Samaras, D. (2020). From shadow segmentation to shadow removal. In A. Vedaldi, H. Bischof, T. Brox, & J. M. Frahm (Eds.), Computer vision – ECCV 2020 (Vol. 12356). Lecture Notes in Computer Science. Springer, Cham. https://doi.org/10.1007/978-3-030-58621-8_16
17. Xiao, J., Fu, X., Zhu, Y., Li, D., Huang, J., Zhu, K., & Zha, Z.-J. (2024). HomoFormer: Homogenized Transformer for Image Shadow Removal. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 25617–25626. https://doi.org/10.1109/CVPR52733.2024.02420
18. Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. Geoscientific Model Development, 15(14), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022
19. Liu, H., Li, M., & Guo, X. (2024, October). Regional attention for shadow removal. In Proceedings of the 32nd ACM International Conference on Multimedia (pp. 5949–5957). https://doi.org/10.1145/3664647.3681126
20. Xiao, J., Wang, C., Mei, K., Ding, Z., Lin, Z., Pfister, H., & Wen, B. (2024). HomoFormer: Homogenized transformer for image shadow removal. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1-11). IEEE. https://doi.org/10.1109/CVPR52733.2024.02420

Authors

Henry Fang is a grade 11 student at Holy Trinity School in Canada. He is passionate about STEM subjects and hopes would obtain a degree in the field of Engineering. He is also typically interested in the function of AI in the fields of medicine and Engineering and hopes he will participate in related jobs.