Online ISSN: 3104-9338
Print ISSN: 3104-932X

Performance Comparison of AI Models for Image Shadow Removal: UNet, CGAN, and Swin-Transformer with a Note on Diffusion Models

Shangan Zhou*

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

Received: January 23, 2026

Revised: February 24, 2026

Accepted: February 5, 2026

Published online: February 11, 2026

To appear in: *International Journal of Advanced AI Applications*, Vol. 2, No. 3 (March 2026)

* Corresponding Author: Shangan Zhou

(3217432128@qq.com)

Abstract. This study conducts a comprehensive performance comparison of three prominent deep learning architectures—UNet, Conditional Generative Adversarial Network (CGAN), and Swin-Transformer—for the task of single-image shadow removal, with additional theoretical consideration given to Denoising Diffusion Probabilistic Models (DDPM). Evaluated on the ISTD benchmark dataset using quantitative metrics (PSNR, SSIM, RMSE, MAE) and qualitative visual assessment, the results establish a clear performance hierarchy. The Swin-Transformer model consistently achieves superior results, excelling in detail preservation, artifact reduction, and maintaining global illumination consistency, attributed to its hierarchical structure and shifted-window self-attention mechanism. The CGAN model demonstrates enhanced perceptual realism through adversarial training, while the UNet provides a computationally efficient baseline. The findings offer practical guidance for model selection based on specific application requirements and highlight the impact of architectural design. This analysis concludes by suggesting future research pathways, including the exploration of hybrid models and the empirical application of diffusion models for high-fidelity image restoration tasks.

Keywords: *Shadow Removal; UNet; Generative Adversarial Networks; Vision Transformer; Diffusion Models*

1. Introduction

1.1. Background and Motivation

The rapid advancement of artificial intelligence (AI), particularly in deep learning, has

catalyzed a paradigm shift across numerous scientific and industrial fields. Computer vision stands as one of the most profoundly affected disciplines. Algorithms powered by deep neural networks have demonstrated exceptional performance in critical applications such as medical diagnostics, autonomous driving, and industrial quality control. This success stems from their capability to autonomously learn hierarchical and complex feature representations directly from vast datasets, tackling tasks that often elude traditional methods reliant on hand-crafted features.

However, the robustness and accuracy of any computer vision system are fundamentally contingent upon the quality of its input data. In real-world, unconstrained environments, captured imagery is frequently degraded by various photometric challenges. Among these, shadows—arising from the occlusion of light sources—are particularly pervasive and problematic. They alter the perceived color, texture, and geometry of scenes, leading to significant information loss. This degradation directly impairs downstream vision tasks; for example, shadows can cause false positives in object detection or obscure boundaries in semantic segmentation. Therefore, developing effective shadow removal techniques is a crucial preprocessing step to enhance the reliability of automated visual analysis systems. This work positions itself as a direct investigation into the application of modern AI models to solve this persistent and practical challenge in computer vision.

1.2. Problem Formulation

The core technical challenge addressed in this study is single-image shadow removal. This task is an ill-posed inverse problem: given a single RGB image containing shadows, the goal is to generate a corresponding, realistic shadow-free image that faithfully represents the scene under uniform illumination.

The problem's ill-posed nature originates from its inherent ambiguity. For any observed shadowed region, multiple combinations of true surface albedo, texture, and lighting conditions could produce the same pixel values. An effective solution must not only accurately localize shadow regions but also synthesize photometrically and texturally plausible content for these areas, ensuring seamless consistency with the non-shadowed parts of the image.

Formally, the objective is to learn a mapping function F from the domain of shadowed images IsIs to the domain of shadow-free images Isf:

$$Isf = F(Is, \Theta)$$

where Θ represents the parameters of the AI model. The function F must implicitly infer and disentangle the intertwined factors of scene geometry, material properties, and lighting to

perform a coherent reconstruction. This paper conducts a systematic comparative analysis to evaluate how three distinct deep learning paradigms—UNet [1], Conditional Generative Adversarial Networks (CGANs) [2], and Swin-Transformer [5]—learn this complex mapping F, assessing their respective strengths and limitations. A discussion on the emerging Denoising Diffusion Probabilistic Models (DDPM) [3] is provided based on recent literature for context.

2. Introduction to Shadow Removal Models

This section provides a detailed description of the three deep learning architectures evaluated in this study. For each model, we discuss its fundamental structure, the specific loss functions used for training, and their application to the shadow removal task. A brief overview of the Diffusion Model paradigm is also included.

2.1. UNet Model

2.1.1. Model Architecture

The UNet, first introduced by Ranneberger et al. [1] for medical image segmentation, is a fully convolutional neural network characterized by its distinctive U-shaped architecture. Figure 1 illustrates its general structure. The architecture consists of two primary paths: an encoder and a decoder.

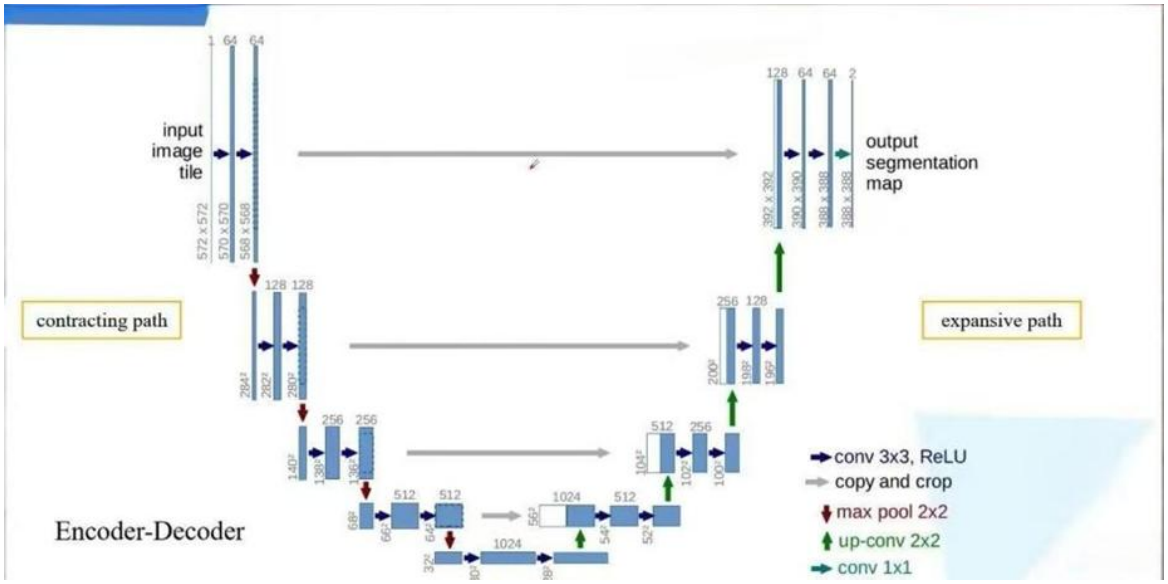


Figure 1. Basic structure diagram of UNet model [1]

- Encoder : The encoder follows the typical architecture of a convolutional network. It comprises a series of convolutional blocks, each containing two 3×3 convolutional layers followed by a ReLU activation function. After each block, a 2×2 max-pooling operation with stride 2 is applied for downsampling. With each downsampling step, the

spatial dimensions of the feature maps are halved, while the number of feature channels is doubled. This process allows the network to capture contextual information and learn hierarchical features from low-level edges and textures to high-level semantic content. The encoder effectively compresses the input image into a compact, high-dimensional feature representation.

- Decoder : The primary role of the decoder is to upsample the feature representation from the encoder back to the original image resolution, thereby generating per-pixel predictions. It consists of a series of upsampling blocks. Each block begins with an up-convolution , doubling the spatial dimensions and halving the number of feature channels. The upsampled feature map is then concatenated via skip connections with the corresponding feature map from the encoder path. This is the most critical feature of the UNet architecture. These skip connections provide the decoder with high-resolution feature information from the earlier encoder layers, which aids in recovering fine-grained spatial details that might have been lost during downsampling. Following concatenation, two 3×3 convolutional layers with ReLU activation are applied.
- Output Layer : The final layer is a 1×1 convolution that maps the feature channels of the last decoder block to the desired number of output channels (3 for a shadow-free RGB image). The complete architecture is symmetric, resembling the letter 'U'. This design is well-suited for image-to-image translation tasks like shadow removal, where the output must be spatially aligned with the input and contain precise local details.

2.1.2. Loss Function

To train the UNet model for shadow removal, a per-pixel reconstruction loss is typically employed to ensure the generated shadow-free image closely approximates the ground truth. The most common choice is the L_1 loss, also known as the Mean Absolute Error (MAE). The L_1 loss measures the average absolute difference between predicted pixel values and the ground truth pixel values. For image restoration tasks, L_1 loss is often preferred over L_2 (MSE) loss due to its lower sensitivity to outliers and its tendency to produce less blurry results.

The L_1 loss is defined as:

$$L_1 = \frac{1}{N}\sum_{i=1}^{N} |\, y_i - y_i^* |$$

where yi is the value of the i-th pixel in the ground truth shadow-free image, y_i^* is the value of the corresponding pixel in the model's output image, and N is the total number of pixels.

2.2. Conditional Generative Adversarial Network (CGAN) Model

2.2.1. Model Architecture

Generative Adversarial Networks (GANs), proposed by Goodfellow et al. [2], constitute a generative modeling framework based on a min-max game between two neural networks: a generator G and a discriminator D.

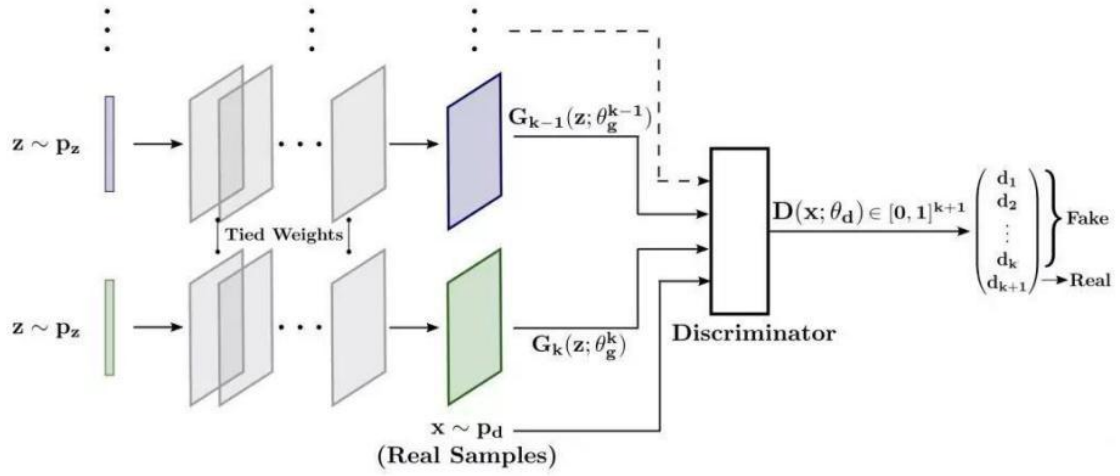


Figure 2. Basic structure diagram of GAN model [2]

- Generator : The generator's objective is to learn the mapping from the source image domain (shadowed images) to the target image domain (shadow-free images). In the context of shadow removal, the generator takes a shadowed image as input and outputs a corresponding shadow-free image. The generator's architecture is often based on an encoder-decoder structure, similar to UNet. This enables it to process the input image, comprehend its content, and generate an output of the same size. Skip connections are almost invariably used to ensure the preservation of low-level information, thereby avoiding the necessity for the model to relearn basic structures and textures already present in the non-shadowed regions of the input.
- Discriminator : The discriminator functions as a learned loss function. Its goal is to differentiate between "real" image pairs (input shadow image + ground truth shadow-free image) and "fake" image pairs (input shadow image + generator's output shadow-free image). It takes a conditional input (the shadow image) and a target image (either real or fake) and outputs a scalar value representing the probability of the target being real. This compels the generator to produce an output that is not only plausible on its own but also constitutes a plausible transformation for the specific input image. The

discriminator is typically a standard convolutional network that downsamples its input to produce a single classification output. The training process involves an adversarial game: the generator strives to create images realistic enough to "fool" the discriminator, while the discriminator continually improves its ability to distinguish them. This dynamic mechanism encourages the generator to create outputs perceptually indistinguishable from real images, often yielding sharper and more realistic textures compared to models trained solely with pixel-level losses like L_1.

2.2.2. Loss Function

The training objective for a CGAN combines two components:

- Adversarial Loss : This loss drives the adversarial game. The generator seeks to minimize it, while the discriminator aims to maximize it. For a conditional GAN, it is often formulated as:

$$L_{\text{adv}}(G, D) = E_{x,y}[log\ D(x, y)] + E_x[log\ (1 - D(x, G(x)))]$$

where x is the input image and y is the shadow-free ground truth image.

- Reconstruction Loss : To stabilize training and provide more direct gradients to the generator, an L_1 loss is typically added. This encourages the generator to produce outputs that are structurally aligned with the ground truth.

$$L_{L1}(G) = \mathrm{E}_{x,y}[\| y - G(x) \|_1]$$

The generator's final objective is a weighted sum of the two losses:

$$\mathcal{L}_G = \mathcal{L}_{adv}(G, D) + \lambda \mathcal{L}_{L1}(G)$$

where λ is a weighting hyperparameter.

2.3. Swin-Transformer Model

2.3.1. Model Architecture

The Swin-Transformer, proposed by Liu et al. [5], is a hierarchical vision Transformer designed to serve as a general-purpose backbone for various computer vision tasks. It introduces two key innovations that make Transformers computationally efficient for high-resolution images: window-based multi-head self-attention and a shifted window scheme. For shadow removal, the Swin-Transformer is typically integrated into an encoder-decoder framework, analogous to UNet.

- Encoder : The encoder progressively reduces the spatial resolution of the input while increasing the feature dimension. The process begins by partitioning the input image into

non-overlapping patches (e.g., 4×4 pixels), which are then projected into an embedding dimension. The core of the encoder consists of multiple "Swin Transformer Blocks". Unlike standard Vision Transformers that compute self-attention globally across all patches, the Swin-Transformer confines self-attention computation to local, non-overlapping windows (e.g., 7×7 patches). This dramatically reduces computational complexity from quadratic to linear with respect to the number of patches. To enable cross-window communication, consecutive Swin Transformer blocks use a shifted window partitioning scheme. The window grid is shifted by half a window size, allowing patches that were in different windows in one block to interact in the next. This mechanism enables the model to effectively learn both local details and long-range dependencies. Patch merging layers are used to downsample the feature maps between stages.

- Decoder : The decoder mirrors the encoder. It uses patch expansion layers to upsample the feature maps. Similar to UNet, skip connections are employed to concatenate the upsampled features with multi-scale features from the encoder. This fusion of deep semantic features from the decoder with shallow, high-resolution features from the encoder is crucial for reconstructing fine details. The decoder also utilizes Swin Transformer blocks to refine the upsampled features.
- Output : A final linear projection layer maps the features back to the RGB pixel space, generating the shadow-free image.

The Swin-Transformer's ability to model long-range dependencies is particularly beneficial for shadow removal. It can better comprehend the global illumination consistency of a scene, ensuring that the appearance of restored shadowed regions aligns with non-shadowed regions of similar material and orientation situated farther away.

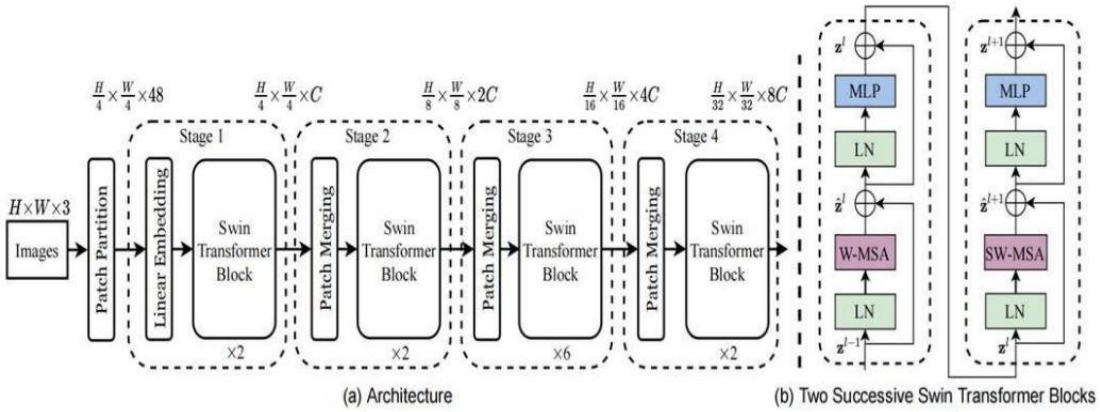


Figure 3. Basic structure diagram of Swin Transformer model [5]

2.3.2. Loss Function

Like UNet, Swin-Transformer models for shadow removal are primarily trained using a pixel-wise reconstruction loss to ensure high-fidelity output. The L_1 loss is the standard choice for this purpose.

$$L_1 = \frac{1}{n}\sum_{i=1}^{n} |\, y_i - y_i^* |$$

2.4. Note on Diffusion Model

2.4.1. Model Architecture

Denoising Diffusion Probabilistic Models (DDPMs) constitute a class of generative models that have recently demonstrated remarkable performance in image synthesis [3][10]. They operate by learning to reverse a gradual noise-addition process. While not trained in this study due to computational constraints, their state-of-the-art performance on the ISTD dataset, as reported in recent literature [9][13], is included for a broader perspective. Their core idea involves two processes:

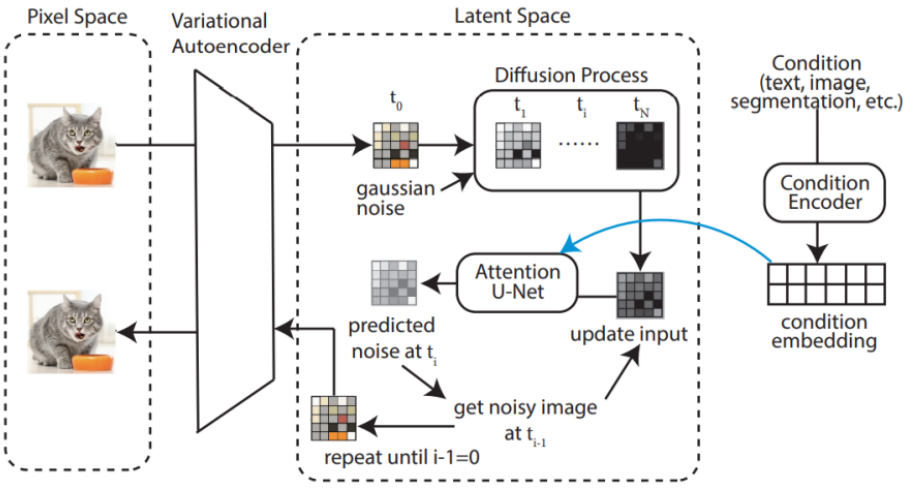


Figure 4. Basic structure diagram of Diffusion model [3]

- Forward Process (Fixed) : A fixed Markov chain across T timesteps gradually adds Gaussian noise to a clean image x_0:

$$q(\mathrm{x}_t|\mathrm{x}_{t-1}) = \mathcal{N}(\mathrm{x}_t; \sqrt{1-\beta_t}\mathrm{x}_{t-1}, \beta_t \mathrm{I})$$

  Where $\{\beta_t\}_{t=1}^{T}$is a variance schedule.

- Reverse Process (Learned) : A neural network (typically a U-Net) $\epsilon\theta(xt, t, c)$ is trained to predict the noise ϵ added at step t, conditioned on the noisy image xt, the timestep t, and the conditional input cc (the shadow image in our case). The simplified training

objective is [3]:

$$\mathcal{L}_{simple} = \mathbb{E}_{t,x_0,e}\left[\|\epsilon - \epsilon_t(\sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon, t, c)\|^2\right]$$

Where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$.

The inclusion of DDPMs in the title and discussion serves to contextualize the performance of the three empirically tested models within the broader landscape of contemporary image restoration architectures.

2.4.2. Loss Function

The simplified loss function is shown in Equation:

$$\mathcal{L}_{simple} = \mathbb{E}_{x_0,\epsilon,t}[\|\epsilon - \epsilon_\theta(x_t, t)\|^2]$$

3. Experimental Setup

This section details the environment, dataset, evaluation metrics, and implementation parameters used for the comparative analysis of the three trained models (UNet, CGAN, Swin-Transformer).

3.1. Experimental Environment

All experiments were conducted on a consistent hardware and software platform to ensure a fair comparison. The operating system was Windows 11, utilizing an Intel 12th Gen Intel(R) Core(TM) i5-12450H processor, an NVIDIA GeForce RTX 5060 Ti .

3.2. Disentangled Representations

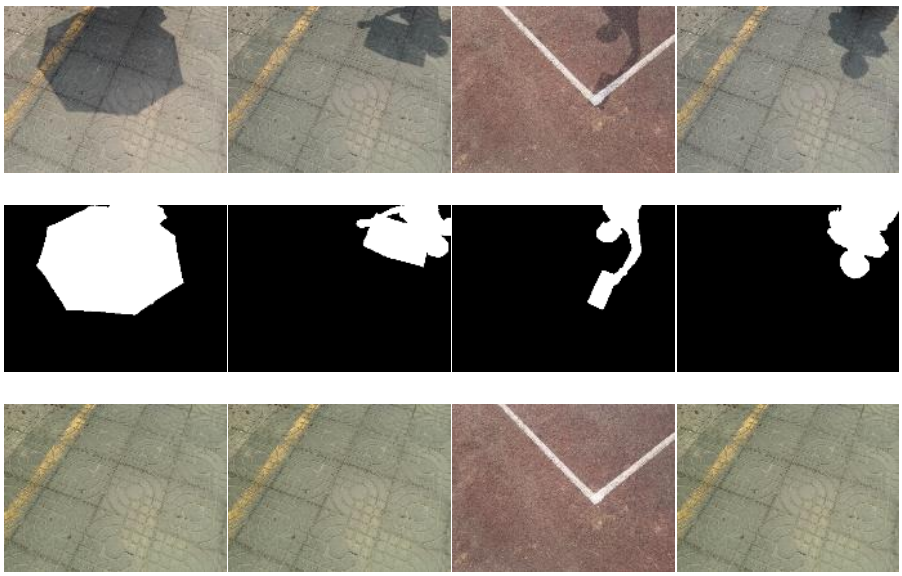


Figure 5. Example triplets from the ISTD dataset [8]: shadow image, shadow mask, and shadow-free ground truth.

The ISTD (Image Shadow Triplets Dataset) [8], a public benchmark dataset specifically designed for shadow-related tasks, was employed. This dataset is particularly suitable for our evaluation because each sample consists of a triplet: (1) a shadowed image, (2) a binary shadow mask, and (3) the corresponding shadow-free ground truth image. The dataset contains 2000 triplets captured in 145 different scenes, covering a wide variety of indoor and outdoor environments, lighting conditions, shadow shapes, and background textures. The diversity of the ISTD dataset provides a challenging and realistic testbed for evaluating the generalization capability of shadow removal models.

3.3. Evaluation Metrics

To comprehensively assess the performance and effectiveness of the models, this paper employs four quantitative metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

- SSIM : A perceptual metric that quantifies image quality degradation as perceived changes in structural information, considering luminance, contrast, and structure. SSIM values range from -1 to 1, with 1 indicating perfect similarity [11].

$$SSIM_{(x,y)} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

  where μ_x,μ_y are means, σ_x^2,σ_y^2 are variances, σ_{xy} is the covariance of images x and y, and C_1,C_2 are stability constants.

- PSNR : Measures the ratio between the maximum possible power of a signal and the power of noise that affects its fidelity. The unit is decibels (dB). A higher PSNR value indicates better reconstruction quality.

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX^2}{MSE}\right)$$

  where MAX is the maximum possible pixel value (255 for 8-bit images), and MSE is the Mean Squared Error.

- RMSE : Measures the square root of the average squared differences between predicted and actual pixel values, being more sensitive to larger errors. A lower RMSE is better.

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

- MAE : Measures the average magnitude of errors without considering their direction,

making it less sensitive to outliers than RMSE. A lower MAE is better.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - y_i^* |$$

In all equations, yi is the ground truth pixel value, y_i^* is the predicted pixel value, and n is the total number of pixels.

3.4. Implementation Details

The Adam optimizer [4] was used for all models due to its adaptive learning rate and robust performance. The batch size was set to 32, and all models were trained for 300 epochs. The performance was evaluated on the ISTD test set.

4. Experimental Results and Analysis

4.1. Quantitative Results

The quantitative performance of the UNet, CGAN, and Swin-Transformer models on the ISTD test set is summarized in Table 1. For completeness, the reported state-of-the-art performance of a Diffusion model (DDPM) on the same dataset, as cited from recent literature [9], is included in parentheses for reference. It is important to note that the DDPM result was not generated by our training but is presented to situate the performance of our tested models within the current research landscape.

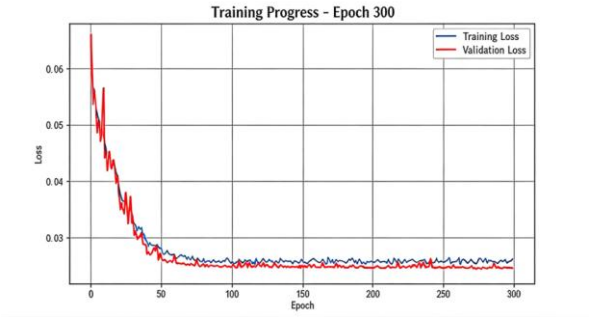
Table 1. Comparison of Evaluation Metrics for the Shadow Removal Task of Three Network Models

Model	MAE	RMSE	SSIM	PSNR
UNet	5.568707	8.345609	0.946751	30.336732
GAN	4.134678	6.012376	0.936753	33.698213
Swin-Transformer	3.812357	4.69824	0.953253	36.011241

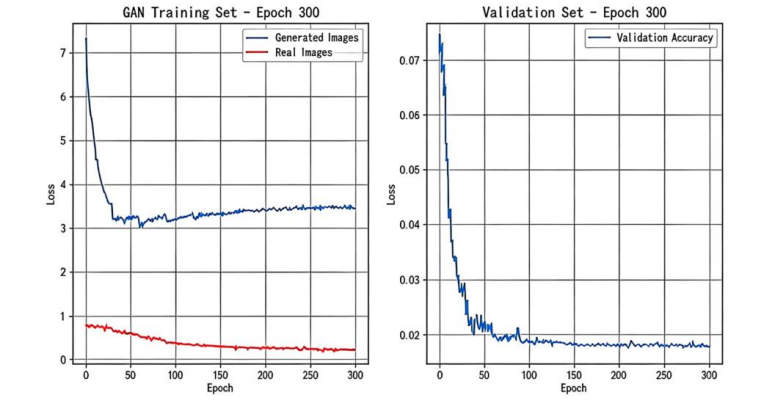
As evident from Table 1, the UNet model demonstrates robust but not outstanding performance. Among all tested models, it has the highest MAE and RMSE scores and the lowest PSNR score, indicating a larger average pixel-level error and lower objective quality. This suggests that while the basic encoder-decoder structure is effective, it may struggle to perfectly reconstruct complex textures and colors within shadow regions. The CGAN model showcases the power of adversarial training by producing outputs that are closer to the ground truth pixel-wise. Although it excels in error reduction, its SSIM score is slightly lower than the Swin-Transformer's, indicating that while the shadow removal result is closer to reality, its structural

and visual consistency is somewhat weaker. The Swin-Transformer achieves the best performance across all four metrics among the trained models. It has the lowest MAE and RMSE, as well as the highest SSIM and PSNR scores, demonstrating its effectiveness in avoiding obvious artifacts, preserving structural information, and maintaining lower overall distortion.

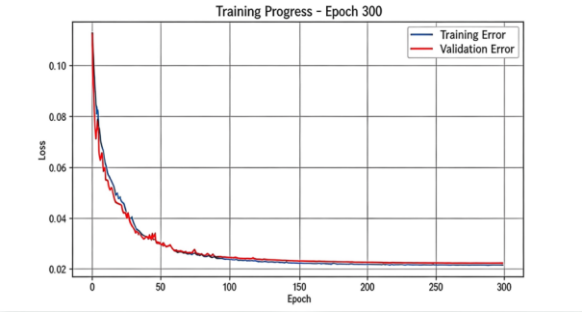
4.2. Training Dynamics



(a)



(b)



(c)

Figure 6 (a). UNet L_1 loss convergence. (b). CGAN generator loss (combination of adversarial and L_1 loss). (c). Swin-Transformer L_1 loss convergence.

Figure 6 shows the convergence behavior of the training loss for the three models. figure (a) shows the steady descent of the L_1 loss for UNet. figure (b) depicts the adversarial and L_1 losses for the CGAN, illustrating the dynamic interplay between the generator and discriminator. figure (c) shows the L_1 loss for the Swin-Transformer, which converges to a lower value than UNet, consistent with its superior final performance.

4.3. Qualitative Visual Analysis

Visual comparisons across five challenging scenes (Figure 7) reveal a clear performance progression from UNet to CGAN to Swin-Transformer.

- Scene (a) Tiled Path : UNet left noticeable dark and blurry patches. CGAN removed the shadow more effectively but over-smoothed the tile texture. Swin-Transformer excellently removed the shadow and reconstructed the tile pattern with subtle color variations.

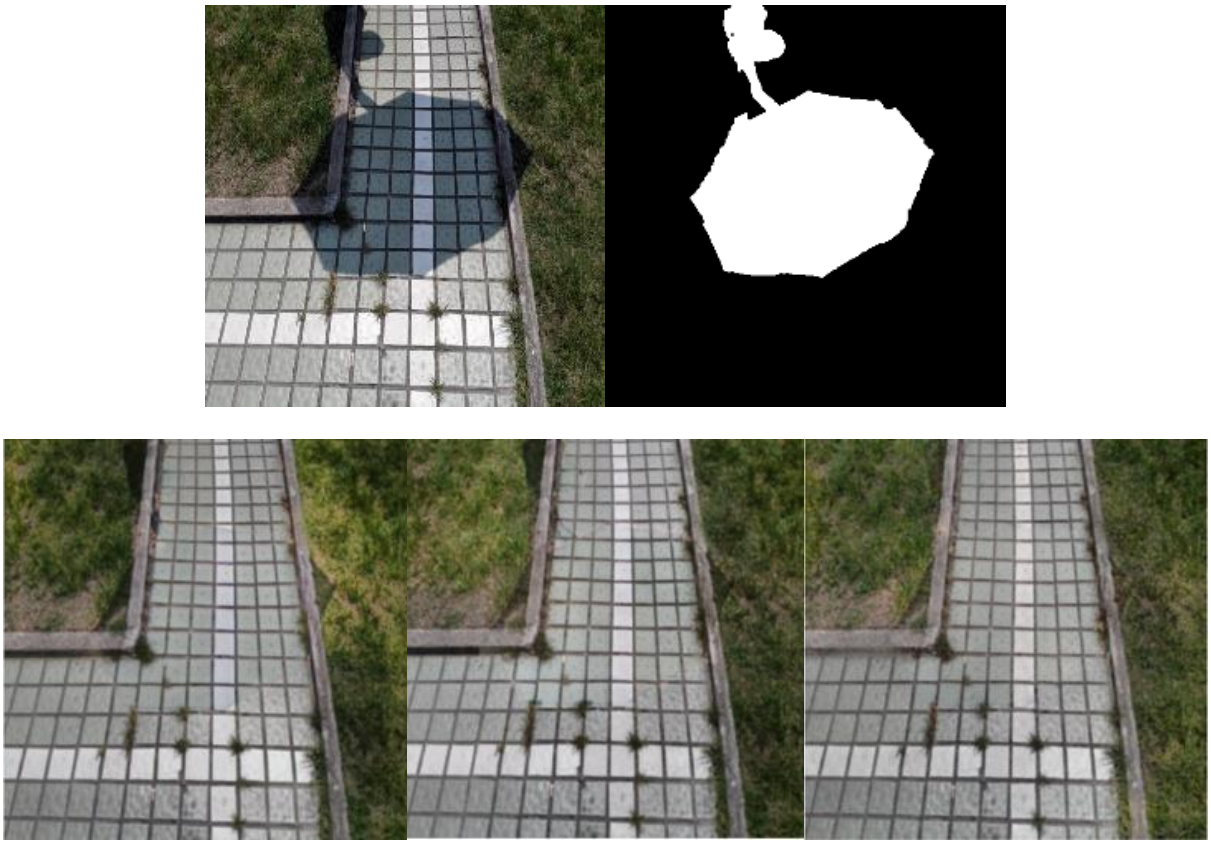


Figure (a)

- Scene (b) Sidewalk with Yellow Lines : UNet left blurred, faded traces. CGAN produced a cleaner result but with slightly softer pavement texture. Swin-Transformer perfectly removed the shadow while restoring the concrete texture with high fidelity.

Figure (b)

- Scene (c) Two-Toned Wall : UNet failed, leaving obvious shadow residue. CGAN removed the shadow but yielded a flat, textureless surface. Swin-Transformer eliminated the shadow while accurately preserving the wall's delicate texture.

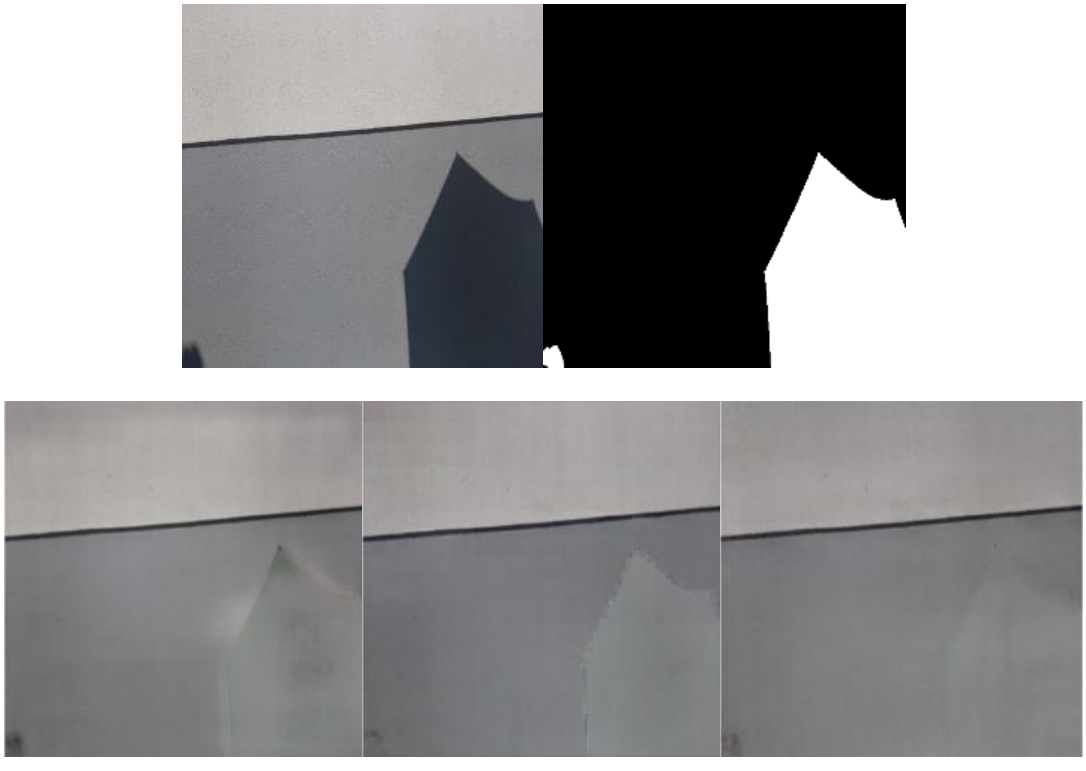


Figure (c)

- Scene (d) Checkered Pavement : UNet performed poorly, leaving a dark, chaotic area. CGAN removed the shadow but blurred the grid pattern. Swin-Transformer cleanly removed the shadow and meticulously reconstructed the individual tiles.

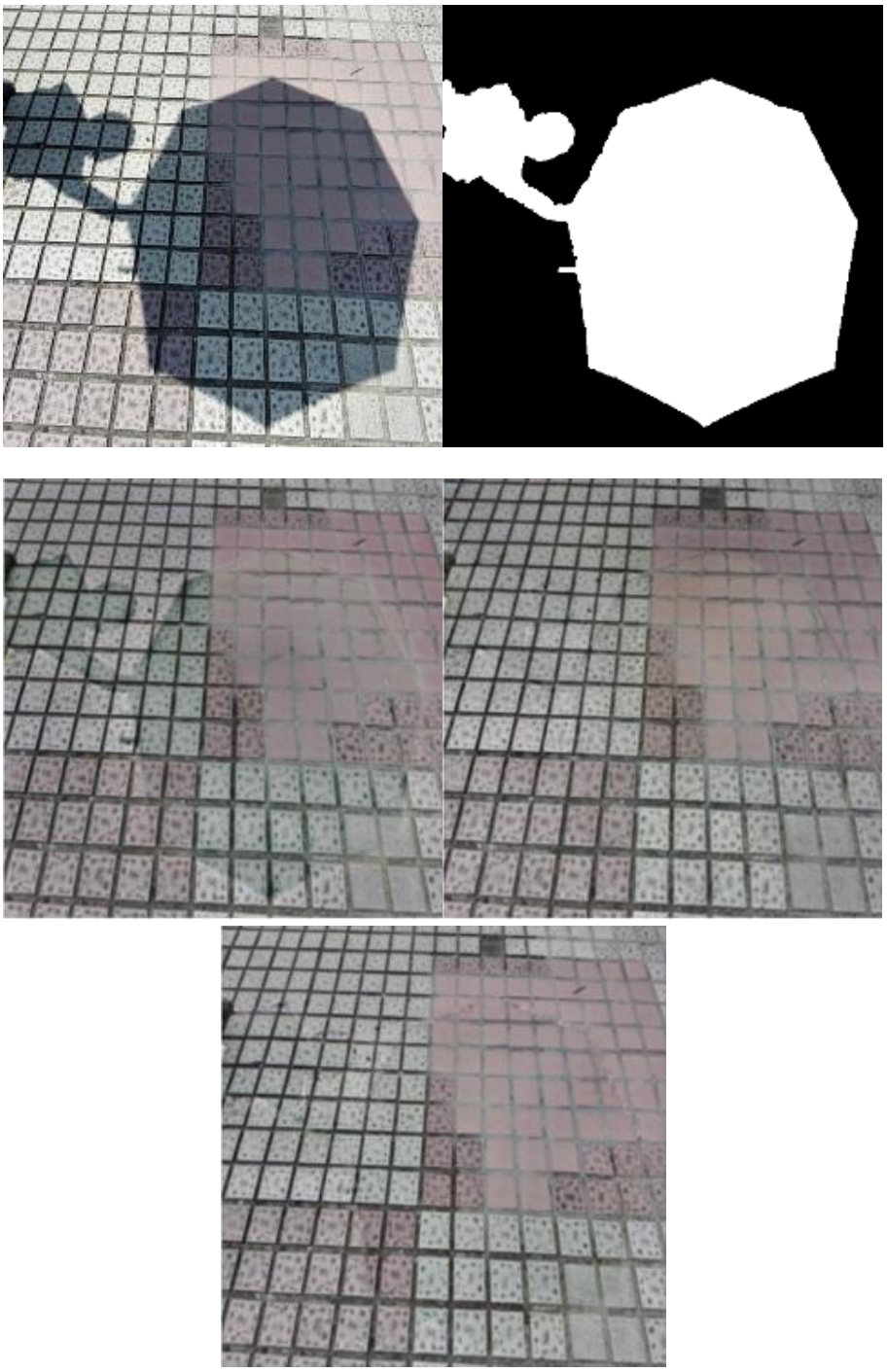


Figure (d)

- Scene (e) Information Sign : UNet left bright, blurred streaks over text. CGAN provided a clearer result but with a mottled background. Swin-Transformer completely eliminated the shadow, restored a uniform background, and kept the text sharp.

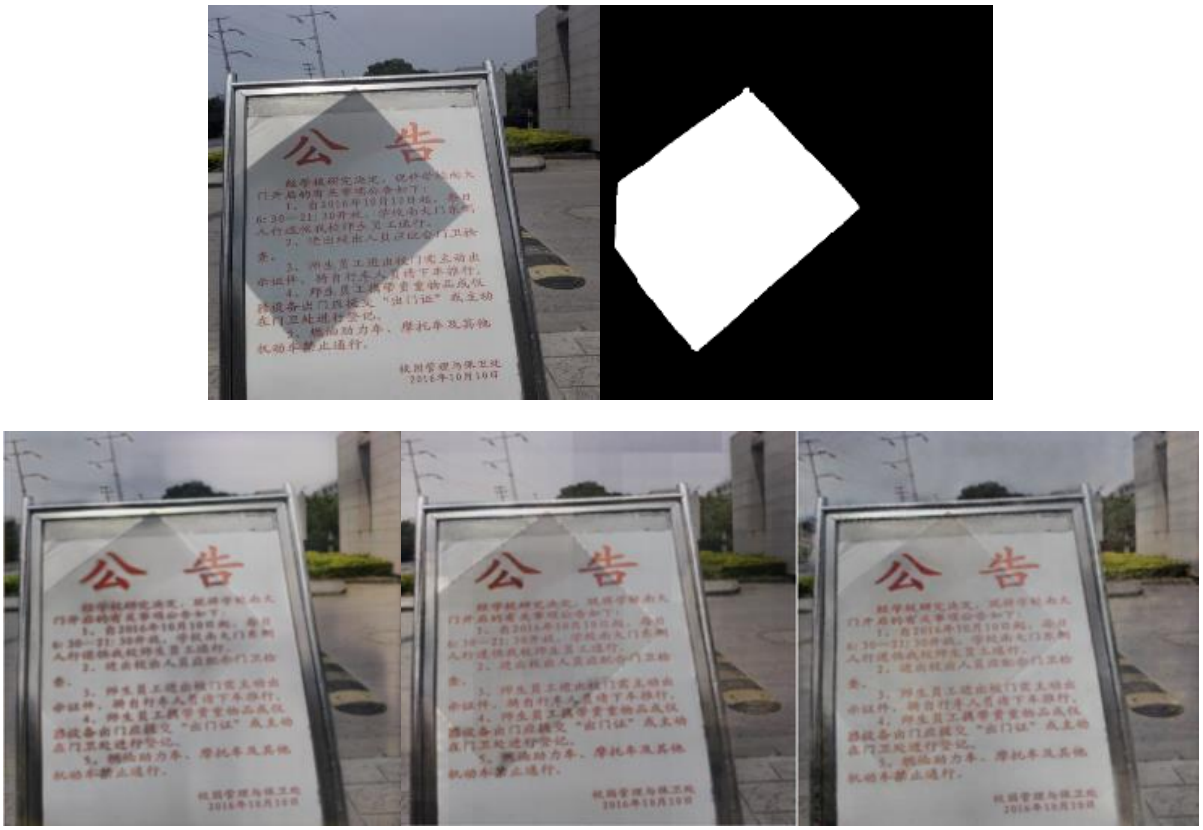


Figure (e)

Figure 7. Visual comparison for five scenes(figure(a)(b)(c)(d)(e)). Each row from top to bottom (or left to right as per original layout) shows: Original shadowed image → Shadow mask → UNet result → CGAN result → Swin-Transformer result.

These visual results highlight the Swin-Transformer's advantages, attributed to its hierarchical structure and shifted window self-attention mechanism [5]. This enables the model to capture long-range dependencies, understand the global lighting environment, and utilize information from distant unshaded areas to accurately reconstruct affected regions.

5. Conclusion

This study presented a comprehensive comparative analysis of the performance of UNet [1], CGAN [2], and Swin-Transformer [5] models in the ill-posed task of single-image shadow removal, supplemented by a contextual discussion on the emerging paradigm of Denoising Diffusion Probabilistic Models (DDPMs) [3][9]. Through rigorous quantitative evaluation on the ISTD benchmark dataset [8]—employing metrics such as PSNR, SSIM, RMSE, and MAE—and supported by detailed qualitative visual inspection across diverse scenes, a clear and consistent performance hierarchy was established. The experimental results unequivocally identify the Swin-Transformer as the superior-performing architecture among the three

empirically tested models, achieving the highest PSNR (36.01 dB) and SSIM (0.953), alongside the lowest MAE (3.81) and RMSE (4.70). This is followed by the CGAN, which leverages adversarial training to enhance perceptual realism, and finally the UNet, which provides a computationally efficient and robust baseline. Visual comparisons corroborate these quantitative findings: while UNet predictions often exhibit residual blurring and localized artifacts, and CGAN outputs can suffer from over-smoothed textures or minor structural inconsistencies, the Swin-Transformer excels in recovering fine-grained details, suppressing artifacts, and maintaining seamless photometric and textural consistency between shadowed and non-shadowed regions.

The performance disparities observed stem fundamentally from the distinct architectural inductive biases of each model. The UNet, with its localized convolutional operations and symmetric encoder-decoder design with skip connections, is highly effective at capturing hierarchical local features and preserving spatial details [1][6]. However, its inherent locality limits its capacity to model the long-range contextual dependencies and global illumination relationships that are crucial for coherent shadow removal across an entire scene. The CGAN framework addresses part of this limitation by introducing an adversarial loss, which trains the generator to produce outputs that reside within the manifold of natural, shadow-free images [2][7]. This leads to sharper textures and improved perceptual quality. Nevertheless, the min-max training dynamic is notoriously unstable and can lead to mode collapse or generate implausible details that degrade structural metrics like SSIM, as observed in our results. In contrast, the Swin-Transformer's success is attributed to its hierarchical vision transformer architecture incorporating shifted window-based self-attention [5]. This design enables it to efficiently model both local information within windows and long-range dependencies across windows through the shifting mechanism. For shadow removal, this translates to a superior ability to understand the global lighting context of a scene. The model can effectively propagate information from brightly lit, unshaded areas to guide the semantically aware and contextually consistent reconstruction of shadowed regions, a capability that is less pronounced in purely convolutional or locally constrained models [12].

Beyond establishing a performance ranking, this analysis yields practical insights for model selection tailored to specific application requirements. In resource-constrained or real-time deployment scenarios (e.g. mobile preview or video processing), the UNet offers a compelling balance of speed and acceptable quality. For applications prioritizing visual plausibility and sharpness in still image editing, the CGAN presents a strong candidate, despite its potential

training challenges. However, for high-stakes vision systems where accuracy, detail fidelity, and global consistency are paramount—such as in autonomous driving perception, historical document restoration, or photogrammetry—the Swin-Transformer emerges as the most reliable and high-performing architecture among those compared, justifying its increased model complexity.

This study is not without limitations. The evaluation was primarily conducted on the ISTD dataset; while diverse, its shadow scenarios may not encompass all real-world complexities like soft shadows, complex mutual illumination, or shadows in highly dynamic environments. Furthermore, due to significant computational resource requirements, we could not empirically train and compare a diffusion model, instead relying on reported literature values [9][13]. While DDPMs show remarkable potential, their iterative denoising process results in slow inference, posing a challenge for practical, time-sensitive applications.

Future work should explore several promising avenues to advance the field of image shadow removal. First, architectural hybridization is a key direction: designing models that integrate the efficient long-range modeling of Transformers with the generative sharpness of GANs could yield further improvements [10]. Second, a dedicated empirical investigation of diffusion models for this task is warranted, particularly focusing on developing more efficient conditional sampling techniques to make them viable for practical use [3][9]. Third, enhancing generalization and robustness requires validation on more challenging and varied datasets, including those with adverse weather conditions, synthetic-to-real domain gaps, and a wider variety of shadow types. Finally, moving beyond purely pixel-level metrics, incorporating perceptual user studies and task-driven evaluation (e.g. measuring the improvement in downstream tasks like object detection after shadow removal) would provide a more holistic assessment of model utility in real-world vision systems.

References

[1] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. *In International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Cham: Springer international publishing.

[2] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

[3] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, 6840-6851.

[4] Kinga, D., & Adam, J. B. (2015, May). A method for stochastic optimization. *In International conference on learning representations (ICLR)* (Vol. 5, No. 6).

[5] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *In Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012-10022).

[6] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2022, October). Swin-unet: Unet-like pure transformer for medical image segmentation. *In European conference on computer vision* (pp. 205-218). Cham: Springer Nature Switzerland.

[7] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *In Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).

[8] Wang, J., Li, X., Yang, J., & Lu, T. (2017). Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1788-1797).

[9] Wang, J., Li, X., & Yang, J. (2018). Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1788-1797).

[10] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., ... & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35, 36479-36494.

[11] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4), 600-612.

[12] Ali, A. M., Benjdira, B., Koubaa, A., El-Shafai, W., Khan, Z., & Boulila, W. (2023). Vision transformers in image restoration: A survey. *Sensors*, 23(5), 2385.

[13] Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., & Li, H. (2022). Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*.