Modeling Global Illumination in Day-Night Image Translation Using Swin-CBAM CycleGAN

Nambiraj R.M , Nithish Kumar V, V. Subapriya and R. Sathya Bama Krishna
Department of Computer Science and Engineering,
Sathyabama Institute of Science and Technology,
Chennai, India

nambiraj25.meenakshi@gmail.com

Abstract. Global illumination variations make unpaired day–night image translation challenging. Under such conditions, convolution-based CycleGAN methods primarily focus on local texture translation, often neglecting global lighting consistency, which results in inconsistent sky–ground illumination and noisy outputs. To address this limitation, we propose an illumination aware CycleGAN that integrates global transformer-based attention with local convolutional attention mechanisms. The generator incorporates Swin Transformer blocks at the bottleneck to model long-range contextual dependencies, while Convolutional Block Attention Modules in the encoder–decoder enhance structural representation and suppress noise. A two stage training strategy is adopted, consisting of illumination focused pretraining on separate day and night domains, followed by fine-tuning on diverse real-world conditions. Experimental evaluations using FID, SSIM, PSNR, and LPIPS demonstrate consistent improvements over purely convolution based CycleGAN baselines, particularly in global lighting coherence and cycle consistency. These results indicate that incorporating global context modeling is crucial for realistic bidirectional day–night image translation.

Keywords: Unpaired image translation, Day–night image translation, CycleGAN, Swin Transformer, Global illumination

1 INTRODUCTION

Over the past few decades, image processing has undergone significant evolution, moving beyond traditional techniques such as filtering, edge detection, and handcrafted feature extraction. Early approaches relied on rule-based algorithms that were effective for specific tasks but lacked flexibility and adaptability. The introduction of machine learning marked a shift toward data-driven methods, and the integration of deep learning further accelerated progress by enabling models to learn complex visual patterns directly from data [1]. Modern image processing systems are now capable of capturing fine-grained textures and high-level semantic structures, supporting tasks that previously depended heavily on human expertise and manual interpretation.

© The Author(s) 2026
R. Vasanth Kumar Mehta et al. (eds.), *Proceedings of the International Conference on Intelligent Systems for a Sustainable Future (ISSF 2026)*, Atlantis Highlights in Intelligent Systems 16,
https://doi.org/10.2991/978-94-6239-693-7_84

One of the major breakthroughs in modern image processing is the development of techniques for image-to-image translation. This technique involves transforming an image from one domain to another, such as converting a black-and-white photograph to color, translating a daytime scene to a nighttime one, or applying artistic styles to photographs. The transformation is not limited to pixel-wise modifications; rather, it captures complex patterns and structures that allow for meaningful changes in the image representation. Unlike traditional image editing tools, which perform predefined transformations, deep learning models can learn these mappings directly from datasets that often contain thousands of training images across different domains, making them suitable for a wide range of real-world applications [2]. Recent studies have further advanced this area through attention-guided generative models and normalization-based frameworks that improve structural preservation and visual realism in challenging cross-domain translation tasks [3], [4], [5].

Despite significant progress in image-to-image translation, handling large illumination variations remains a fundamental challenge, particularly in day–night and night–day scenarios. Such transformations require coherent changes in global lighting while preserving local structural details, including roads, buildings, and scene geometry. In many outdoor scenes, illumination differences between day and night images can exceed 40–60% brightness variation, making consistent translation difficult. Conventional convolution-based architectures primarily emphasize local texture mapping and often struggle to enforce scene-wide illumination consistency, leading to artifacts such as uneven sky brightness or inconsistent lighting across regions. While cycle consistency loss helps preserve structural content in unpaired translation settings [6], it alone is insufficient for modeling long-range dependencies required for realistic illumination transitions. Recent transformer-based vision models have demonstrated strong capability in capturing long-range contextual relationships, motivating their adoption in generative image translation frameworks [7]. As a result, unpaired day–night translation remains a challenging problem, especially in the presence of extreme contrast, noise, and low-light conditions.

Day–night image translation is fundamentally an illumination-driven task rather than a purely texture-based transformation. Realistic translation requires coherent, scene-wide changes in lighting while preserving local structural details. Convolution-only architectures operate with limited receptive fields and often fail to capture long-range dependencies, leading to spatially inconsistent illumination and visual artifacts, particularly in the challenging Night→Day direction. Effective unpaired translation therefore requires explicit modeling of global context in addition to localized feature refinement.

To overcome the limitations of convolution-only models in unpaired day–night translation, a hybrid attention-based CycleGAN framework is introduced. Swin

Transformer blocks are incorporated at the generator bottleneck to model global scene dependencies and enforce coherent illumination changes [7], while Convolutional Block Attention Modules within the encoder–decoder enhance structural features and suppress nighttime noise [8]. A two-stage curriculum training strategy, consisting of illumination-focused pretraining followed by large-scale fine-tuning, further improves generalization. Quantitative and qualitative evaluations demonstrate improved illumination consistency and perceptual quality over conventional CycleGAN baselines [6], particularly for the Night→Day translation task.

2 RELATED WORK

Image-to-image translation advanced significantly with the introduction of Generative Adversarial Networks (GANs) [1]. Early approaches relied on paired datasets for supervised translation, such as Pix2Pix [2], which employed a conditional GAN with a U-Net generator to preserve structural details. However, the requirement for paired data limited applicability in real-world scenarios, including day–night translation.

Unpaired translation methods addressed this limitation by enforcing structural consistency across domains. CycleGAN [6] introduced cycle consistency loss to enable unpaired learning, while DualGAN [3] and DiscoGAN [4] provided alternative unsupervised frameworks applicable to scenarios where ground-truth correspondence is unavailable.

Subsequent research focused on improving perceptual quality and training stability through techniques such as adaptive learning rates [5] and multi-scale discriminators. Attention-based models further enhanced semantic consistency, with U-GAT-IT [9] demonstrating improved structural preservation at the cost of increased computational complexity.

Despite these advances, convolution-based architectures remain limited in modeling long-range dependencies required for globally consistent illumination changes. Transformer-based models address this limitation by capturing global contextual information. The Swin Transformer [7] enables efficient hierarchical self-attention, while Convolutional Block Attention Modules (CBAM) [8] refine local feature representations through channel and spatial attention. Combining global and local attention mechanisms therefore provides a promising direction for illumination-consistent unpaired day–night image translation.

3 METHODOLOGY

3.1 Data Preparation

Training was conducted using an unpaired day–night image dataset collected from publicly available sources [10], consisting of approximately 32,000 daytime images and 25,000 nighttime images. A two-stage curriculum training strategy was adopted to improve illumination modeling and training stability. In the first stage, a subset of visually distinct images (5,000 day and 5,000 night) was selected to emphasize strong illumination contrast and enable the model to learn coarse global lighting transformations. In the second stage, training was extended to the full dataset to improve generalization across diverse real-world lighting conditions. All images were resized to 256 × 256 pixels to ensure consistent preprocessing and manageable training complexity.

For quantitative evaluation, images from the validation split of the same dataset [10] were used, ensuring sufficient sample size for reliable computation of distribution-based metrics such as FID. The use of the validation set avoids data leakage while providing statistically stable evaluation across domains. Balanced sampling between day and night images was maintained to prevent bias during metric computation. An additional day–night dataset [11] was examined for qualitative inspection and cross-dataset reference; however, it was not employed for quantitative evaluation due to its limited size.

3.2 Hybrid Swin-CBAM Generator Architecture

The proposed method follows the standard CycleGAN framework for unpaired image translation, employing two generators for bidirectional domain mapping and two discriminators for adversarial supervision, as introduced in CycleGAN [6]. Architectural modifications are applied only to the generator, while standard PatchGAN discriminators are retained to maintain training stability, following the discriminator design used in earlier image-to-image translation frameworks [2].

Fig. 1. illustrates the baseline generator architecture commonly used in conventional CycleGAN models, consisting 2 of an encoder–decoder structure with residual blocks at the bottleneck

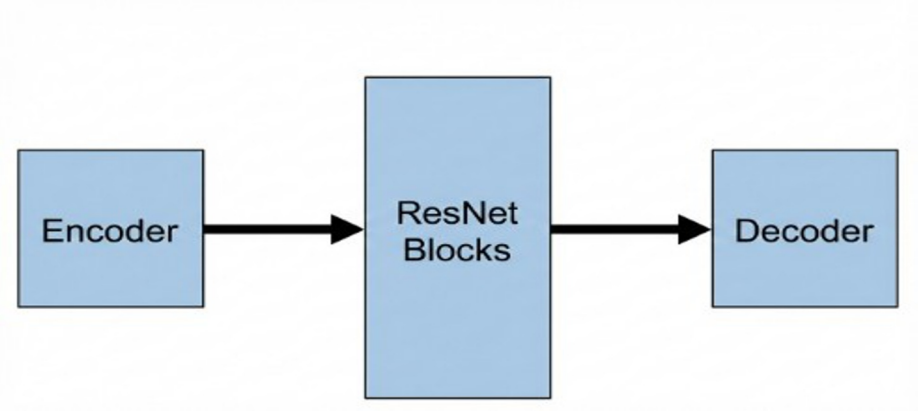

Fig. 1. ResNet-based CycleGAN generator

To improve global illumination consistency, a hybrid attention-based generator is introduced, as shown in Fig. 2. Compared to the baseline design, the architecture explicitly incorporates both global context modeling and localized feature refinement within the generator.

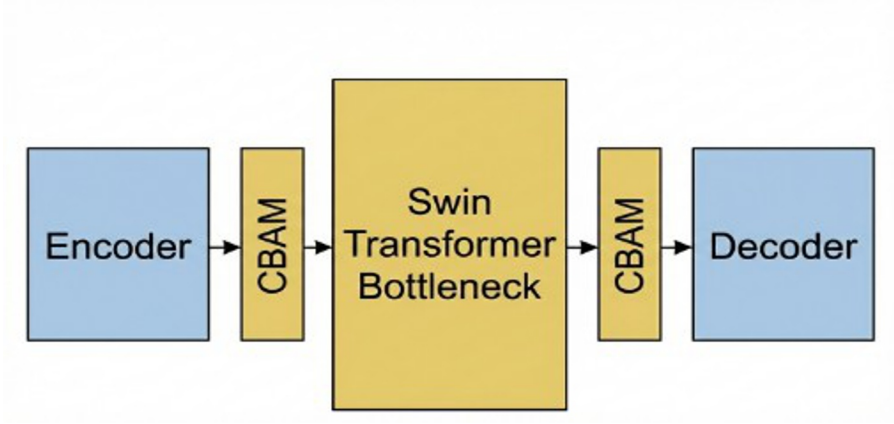

Fig. 2. Swin CBAM generator architecture

In the Swin-CBAM generator, Convolutional Block Attention Modules (CBAM) are integrated within the encoder and decoder to enhance informative channel responses and spatial feature localization, aiding noise suppression and structural preservation in nighttime scenes [8]. At the bottleneck, residual blocks are replaced with Swin Transformer blocks, which employ shifted-window self-attention to capture long-range dependencies and enforce coherent scene-wide illumination changes [7].

The combination of local convolutional attention and global transformer-based attention enables effective modeling of both fine structural details and scene-wide

lighting variations. This design balances localized feature refinement with long-range contextual awareness, which is essential for realistic day–night translation. Adversarial supervision is provided by PatchGAN discriminators operating on local image patches, encouraging high-frequency texture realism without introducing additional architectural complexity, consistent with prior image-to-image translation approaches [2].

3.3 Loss Functions

Various loss functions are used to train the proposed CycleGAN framework for realistic image generation and stable unpaired translation.

Adversarial Loss
The adversarial loss motivates the generators to render images that cannot be distinguished from real images when examined by PatchGAN discriminators.

Formulation [1]:

$$L(G,DY) = \mathbb{E}y[\log DY(y)] + \mathbb{E}x[\log (1 - DY(G(x)))] \quad (1)$$

Cycle Consistency loss
Cycle consistency loss is introduced to preserve the structure, which requires that an image in target space which is transported back again into input space can reconstruct the original one.

Formulation [6]:

$$Lcyc(G, F) = \mathbb{E}x[|| F(G(x)) - x ||1] + \mathbb{E}y[|| G(F(y)) - y ||1] \quad (2)$$

Identity Loss
Identity loss prevents unnecessary modifications of the input images that already belong to the target domain and hence can help preserve color distribution and structural information.

Formulation [6]:

$$Lid(G, F) = \mathbb{E}y[|| G(y) - y ||1] + \mathbb{E}x[|| F(x) - x ||1] \quad (3)$$

Total Loss
The generator objective is defined as the weighted sum of adversarial, cycle consistency, and identity losses [6].

$$\text{Total Loss} = \text{adversarial loss} + \text{cycle consistency loss} + \text{identity loss} \quad (4)$$

3.4 Evaluation Metrics

Model performance is evaluated using complementary quantitative metrics. Structural Similarity Index Measure (SSIM) assesses structural preservation [12], Peak Signal-to-Noise Ratio (PSNR) measures pixel-level fidelity, Learned Perceptual Image Patch Similarity (LPIPS) evaluates perceptual similarity using deep features [13], and Fréchet Inception Distance (FID) measures distributional similarity between generated and real images [14].

3.5 Training Process

Training is performed using the Adam optimizer for both generators and discriminators. A batch size of 8 is used during training. The proposed model is trained using a two-stage strategy. In the first stage, the network is pretrained for 50 epochs on a subset of approximately 5,000 visually distinct day and night images with a learning rate of 1×10^{-4}. In the second stage, the model is fine-tuned for 4 epochs on the full dataset containing approximately 50,000 images, using a reduced learning rate of 5×10^{-5} to improve stability and generalization. Mean squared error loss is applied to the discriminators, while the generators are optimized using adversarial, cycle consistency, and identity losses with weighting factors of 10 and 0.5, respectively, following standard CycleGAN practice [6]. The baseline CycleGAN model was implemented following the original configuration described in [6] and trained under similar conditions for comparison. All experiments for the proposed model were conducted on an NVIDIA Tesla T4 GPU.

4 RESULTS AND EVALUATION

4.1 Evaluation Protocol

The Swin-CBAM CycleGAN is evaluated using both quantitative and qualitative analyses to assess performance in unpaired day–night image translation. Evaluation is conducted separately for Day-Night and Night-Day directions. Structural Similarity Index Measure (SSIM) [12] and Peak Signal-to-Noise Ratio (PSNR) are used to evaluate structural and pixel-level fidelity, while Learned Perceptual Image Patch Similarity (LPIPS) [13] and Fréchet Inception Distance (FID) [14] assess perceptual similarity and distributional alignment with real images. FID is computed using balanced sampling across domains to account for dataset imbalance and ensure fair comparison.

Comparisons are conducted between a standard convolution-based CycleGAN and the Swin-CBAM CycleGAN, both trained under identical data preprocessing and evaluation protocols.

4.2 Quantitative Evaluation

Quantitative performance is reported for both translation directions to capture differences in task difficulty.

Day → Night Translation

Day→Night translation requires consistent global darkening while preserving structural content such as roads, buildings, and object boundaries. In addition to overall brightness reduction, the translation must maintain spatial coherence across large regions of the scene to avoid uneven illumination artifacts. Table I summarizes the quantitative comparison between CycleGAN and Swin-CBAM CycleGAN, highlighting differences in structural preservation, perceptual quality, and distributional alignment with real night-time images.

TABLE I. DAY→ NIGHT QUANTITATIVE RESULTS

Model	SSIM ↑	PSNR ↑ (dB)	LPIPS ↓	FID ↓
CycleGAN	0.8758	26.60	0.1501	55.11
Swin-CBAM-CG	0.8424	25.03	0.1863	18.72

Although the Swin-CBAM CycleGAN does not consistently outperform the convolution-based CycleGAN across all pixel-level metrics, it achieves a substantial reduction in FID, indicating significantly improved alignment with the real night-time image distribution. This suggests that the proposed architecture prioritizes global illumination realism over local pixel similarity, which is critical for perceptual quality in day-to-night translation tasks. Such behavior is commonly observed in generative image translation models, where improvements in perceptual realism and distributional consistency may occur even when pixel-level similarity metrics remain comparable.

Night → Day Translation

Night→Day translation is inherently more challenging due to the need for global brightness restoration and the suppression of low-light noise, which leads to unavoidable information loss in dark regions. Table II presents the quantitative comparison for this translation direction.

TABLE II. NIGHT→DAY QUANTITATIVE RESULTS

Model	SSIM ↑	PSNR ↑ (dB)	LPIPS ↓	FID ↓
CycleGAN	0.8679	28.62	0.1316	116.13
Swin-CBAM-CG	0.8800	28.17	0.1516	42.55

For Night-to-Day translation, the Swin-CBAM CycleGAN demonstrates clear improvements in FID, confirming superior global illumination recovery and perceptual alignment with real daytime images. While SSIM and PSNR improvements are more modest, the transformer-based global context modeling enables more coherent brightness restoration compared to the convolution-only baseline, which struggles with long-range illumination dependencies under low-light conditions. This trend further supports the importance of global contextual modeling for illumination-driven translation tasks.

4.3 Qualitative Results

In addition to quantitative metrics, qualitative results provide visual insight into the realism and consistency of the translated images.

Day → Night Visual Results

The Swin-CBAM CycleGAN effectively transforms daytime scenes into realistic night-time images by applying consistent global darkening while preserving structural elements such as buildings, roads, and object boundaries. The translated outputs exhibit coherent illumination across the scene, with reduced artifacts and stable brightness transitions between sky and foreground regions, resulting in visually plausible night-time representations

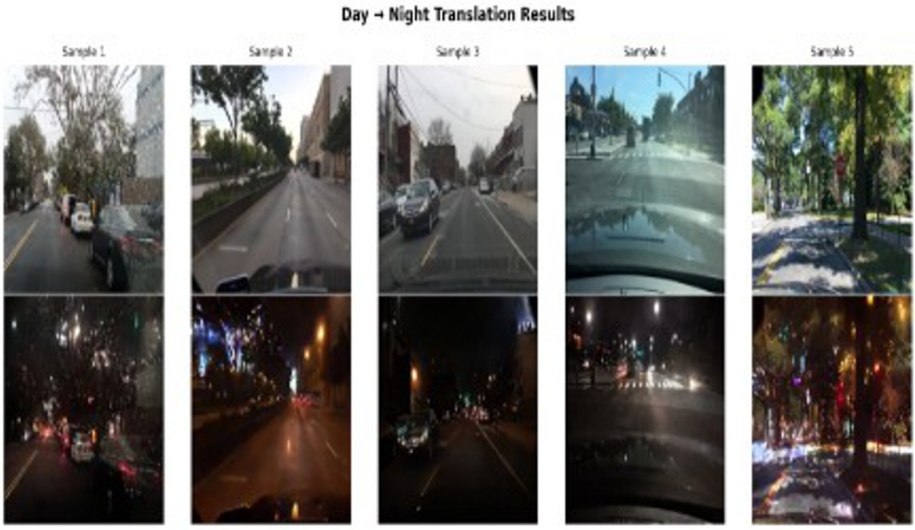

Fig. 3. Day-to-Night translation output.

Night → Day Visual Results

For Night→Day translation, the model restores daylight illumination by brightening the scene, improving visibility, and suppressing low-light artifacts while maintaining structural integrity

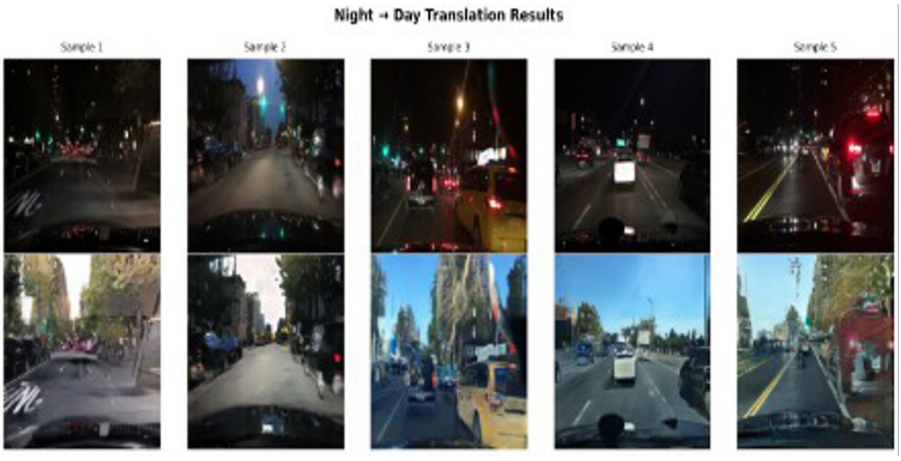

Fig. 4. Night-to-Day translation output.

As shown in Fig. 4. (Sample 3), the proposed model reconstructs a coherent daytime sky and improves scene visibility from near-dark inputs, a scenario where conventional CycleGAN models often exhibit uneven illumination. Although the generated daytime images exhibit improved brightness and visibility, certain regions may

still lack fine texture details due to the inherent difficulty of reconstructing illumination from low-light inputs.

4.4 Discussion

The combined quantitative and qualitative results demonstrate that Swin-CBAM CycleGAN significantly improves unpaired day–night image translation compared to CycleGAN. The integration of Swin Transformer blocks enables effective modeling of global illumination relationships, while CBAM enhances local structural refinement and noise suppression. Improvements are particularly pronounced in the Night-to-Day direction, confirming that architectural designs incorporating global context modeling are essential for achieving illumination consistent image translation. Furthermore, the results indicate that improvements in perceptual realism and illumination coherence may not always correlate with higher pixel-level similarity metrics, particularly for illumination-driven translation tasks.

5 CONCLUSION

This work presented a Swin-CBAM CycleGAN architecture for unpaired day–night and night–day image translation, with a particular focus on modeling global illumination changes while preserving local structural details. By integrating transformer-based global attention with convolutional local attention within the generator, the proposed approach effectively addresses the limitations of convolution-only CycleGAN models in handling large illumination variations.

Experimental results demonstrate that Swin-CBAM CycleGAN improves perceptual realism and illumination consistency compared to a standard CycleGAN, with especially notable gains in the challenging Night→Day translation task. Quantitative metrics such as SSIM, PSNR, LPIPS, and FID indicate improved distributional alignment and visually coherent illumination transitions, while qualitative analysis confirms enhanced global consistency and reduced visual artifacts across diverse lighting conditions. These results support the importance of combining global contextual modeling with local structural refinement for illumination-driven image translation.

Overall, the proposed framework demonstrates that incorporating transformer-based global context mechanisms can significantly enhance unpaired image translation performance without requiring paired supervision. Future work may explore temporal consistency for video-based day– night translation, further optimization of transformer components, and integration with downstream vision tasks such as autonomous perception and scene understanding.

REFERENCES

1. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley,S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2014.
2. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017.
3. H. Yi, H. Zhang, P. Tan, and M. Gong, “DualGAN: Unsupervised dual learning for image-to-image translation,” in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017.
4. T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, “Learning to discover cross-domain rela tions with generative adversarial networks,” in Proc. Int. Conf. Machine Learning (ICML), 2017.
5. J. Kim, M. Jung, and H. Lee, “Stabilizing training of generative adversarial networks t hrough adaptive learning rates,” 2021
6. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in Proc.IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017.
7. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang,S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE Int.Conf. Computer Vision (ICCV), 2021.
8. S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention mod ule,” in Proc. Eur. Conf. Computer Vision (ECCV), 2018.
9. J. Kim, M. Yoon, H. Lee, and S. Kim, “U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization,” in Proc. Int. Conf. Learning Repre sentations (ICLR), 2020.
10. Kaggle, “Night-to-day image translation dataset.”
11. H. Heon, “DayNight-CityView dataset,” Kaggle, 2021.
12. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., 2004.
13. R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effective ness of deep features as a perceptual metric,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2018.
14. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two-time-scale update rule converge to a local Nash equ

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.