E-ISSN: 2962-6110
P-ISSN: 2964-7843
JSRET
jsret.knpub.com
Journal of Scientific Research, Education, and Technology

Photo-to-Cartoon Image Translation Using CartoonGAN with a Joint Learning Approach

Muhammad Shiddiq, Ahmad Tri Hidayat
Univeristas Teknologi Yogyakarta, Indonesia

ABSTRACT

Photo-to-cartoon translation is a non-photorealistic rendering task that generates illustrative visuals while preserving fundamental object structures. This study proposes a CartoonGAN-based approach employing a joint learning scheme that integrates a lightweight denoising module into the generator. Trained end-to-end alongside the stylization process, this module suppresses noise and irrelevant textures without losing critical semantic information from input photographs. Using unpaired photo and cartoon images from the Hugging Face platform, the model is trained with a combination of adversarial and L1-based content losses to balance style generation and structural preservation. Experimental results indicate a stable and convergent training process, achieving an average content loss of 0.0286 and a generator adversarial loss of 0.3982 at epoch 50. Qualitatively, the generated images exhibit sharper contours, uniform color regions, and reduced fine textures compared to the original photographs. These findings demonstrate that integrating a denoising module via joint learning significantly improves visual consistency and training stability, providing an effective deep learning-based solution for photo-to-cartoon translation.

Keywords: *Photo-to-Cartoon, CartoonGAN, Joint Learning, CNN, Image Stylization*

Corresponding author
Name: Muhammad Shiddiq
Email: shiddiq066@gmail.com

INTRODUCTION

The development of deep learning-based digital image processing technologies in recent years has brought significant changes to how computers understand, represent, and manipulate visual information. Rapid advancements in computational hardware, the availability of large-scale datasets, and innovations in neural network architectures have driven performance improvements in computer vision systems across various complex tasks. Various deep neural network-based approaches have been successfully applied to image processing tasks, such as image classification, object detection, image segmentation, image restoration, and cross-domain image translation. Beyond focusing merely on accuracy and precision in visual representation, these developments also open massive opportunities for applications that emphasize aesthetics and visual creativity. For instance, CartoonGAN has been utilized to translate photographic images into cartoons by preserving the main object structures while generating illustrative-style visual representations,

demonstrating the potential of deep learning for creative and non-photorealistic applications (Chen, Lai, & Liu, 2018).

One rapidly growing application is the translation of photographic images into cartoon-style illustrations, commonly known as photo-to-cartoon translation. This translation falls under the domain of Non-Photorealistic Rendering (NPR), a branch of image processing aimed at generating non-realistic visual representations by emphasizing shape abstraction, color simplification, and contour enhancement. Unlike photorealistic rendering, which attempts to mimic the real world accurately, NPR prioritizes the artistic interpretation of the input image. This approach is widely utilized in various fields, such as the animation and entertainment industries, digital avatar development, social media, graphic design, and artificial intelligence-based creative applications. The popularity of photo-to-cartoon applications continues to rise along with user demand for more expressive and aesthetically pleasing visual content (Huang, 2024). Modern deep learning-based approaches, such as the diffusion models developed by Chen et al. (2025), have demonstrated effective capabilities in translating photographic images into cartoon styles while maintaining semantic structures and crucial visual details, thereby expanding the perspective of aesthetic applications beyond mere image recognition or analysis.

Several recent studies show that Generative Adversarial Network (GAN)-based approaches are capable of producing photo-to-cartoon translations with high visual quality, for both human portraits and natural landscapes. These models can effectively learn the visual distribution differences between the photo domain and the cartoon domain, even when utilizing unpaired data. Nevertheless, this cross-domain translation still faces significant challenges, particularly regarding differences in geometric structures, texture simplification, and visual style consistency across domains (Men et al., 2022). Other studies, such as ECGAN, indicate that an expanded generative architecture can improve the quality of translating real images into cartoon styles despite using unpaired data, while multi-style models like MS-CartoonGAN and unpaired photo-to-caricature approaches address cross-domain challenges and broad visual variations in the application of GANs for cartoonization (Tang, 2023; Zheng et al., 2019). These challenges render image cartoonization a complex problem that requires adaptive learning approaches.

Furthermore, the process of photo cartoonization is not a trivial task. Photographic images naturally contain complex texture details, non-uniform lighting variations, and noise resulting from sensor limitations and image acquisition conditions. These details often conflict with the visual characteristics of cartoon images, which generally feature simple textures, flatter colors, and bold contours. If such details and noise are not handled properly, the cartoonization output may contain visual artifacts, inconsistent contours, or even distortions in the main object's shape, thereby degrading the aesthetic quality and visual readability of the image. Several studies have shown that, within the context of GAN-based image-to-cartoon translation, the ability to extract and preserve texture details and contour structures is critical to the final quality. Various techniques, such as adaptive attention to textures, can help resolve the conflict between the complex textures of

photographs and the visual simplification of cartoons (Gao, Zhang, & Tian, 2022; Thakur, Rizvi, & Satish, 2021).

Traditional approaches to image cartoonization generally rely on rule-based image processing techniques, such as edge detection, bilateral filtering, color quantization, or region-based segmentation. These methods are relatively simple and computationally efficient, but they heavily depend on precise parameter selection. Additionally, rule-based approaches tend to have limited generalization capabilities, especially when faced with complex variations in texture, lighting, and backgrounds. Recent studies indicate that conventional cartoonization methods often fail to produce consistent cartoon styles across diverse visual scenarios (Wang, 2022). Several recent reviews demonstrate that traditional methods, such as edge detection and color simplification, frequently fail consistently when dealing with lighting variations, complex textures, and noise, because the smooth visual characteristics and sharp contours of cartoons differ significantly from real-world photographic images in terms of statistical features and color distributions (Xu, Xia, Hu, Zhou, & Weng, 2025).

Along with the development of deep learning, Convolutional Neural Network (CNN)-based approaches have increasingly been used to overcome the limitations of conventional methods. CNNs possess the ability to extract visual features hierarchically and in a data-driven manner, making them more adaptive to input image variations. One of the most influential deep learning paradigms in image translation tasks is the Generative Adversarial Network (GAN). GANs leverage an adversarial training mechanism between two neural networks—a generator and a discriminator—which compete to produce synthetic images that resemble the target data distribution. CNN and GAN-based approaches have proven effective in image-to-image translation; for example, conditional GANs like pix2pix can learn the mapping between image domains directly based on input conditions and generate high-quality outputs across various visual translation tasks (Isola, Zhu, Zhou, & Efros, 2017). In the context of image-to-image translation, GANs have been proven capable of generating cross-domain transformations with high visual quality.

Although CartoonGAN and its variants have shown promising cartoon stylization capabilities, several research gaps remain inadequately addressed. First, most existing GAN-based cartoonization models treat photographic images as clean inputs, without explicitly considering the impact of noise and excessive textures inherently present in real-world photographs. However, uncontrolled noise can cause the generator to learn irrelevant visual patterns, resulting in contour inconsistencies and stylization artifacts in the output images (Jo, Chun, & Choi, 2021; Ulyanov, Vedaldi, & Lempitsky, 2020). Second, existing approaches generally apply denoising as an independent pre-processing stage separated from the stylization process; thus, noise reduction cannot adaptively adjust to the specific requirements of the cartoonization task itself. Separating these two stages risks removing structural details that are actually crucial for cartoon contour formation, or conversely, leaving behind noise that disrupts stylization quality. Third, few studies have explicitly explored the end-to-end integration of a denoising module into the CartoonGAN framework, meaning the potential of joint learning between denoising and cartoon

stylization remains suboptimally utilized. The concept of deep image prior suggests that CNN architectures inherently possess a strong bias toward representing natural image structures while suppressing noise, yet this principle has not been systematically integrated into the context of GAN-based cartoonization.

Based on this background, this study proposes a CartoonGAN-based photo-to-cartoon translation system combined with a lightweight denoising module via a joint learning approach. The denoising module is directly integrated into the generator and trained end-to-end alongside the cartoon stylization process, ensuring that noise reduction occurs adaptively and aligns with the goals of cartoon style generation. This approach aims to suppress noise and irrelevant texture details while preserving the primary semantic structures of the photographic images. Therefore, the proposed system is expected to generate cartoon images with sharper contours, more consistent colors, and improved training stability compared to cartoonization approaches lacking end-to-end denoising integration.

METHOD

Research Design

This study employed a quantitative experimental research method, aimed at developing an optimized photo-to-cartoon image translation system based on CartoonGAN and testing its training stability and convergence performance. The quantitative experimental method is a research approach used to systematically modify generative neural network architectures while objectively measuring the mathematical variations in performance metrics under controlled training environments (Karras, Laine, & Aila, 2021).

The experimental framework used in this study consists of four stages: dataset partitioning, joint learning configuration, adversarial training, and quantitative convergence evaluation. The dataset partitioning stage was conducted to acquire unpaired photo and cartoon images from the Hugging Face platform and systematically divide them into a controlled 80:20 training and testing ratio. The joint learning configuration stage involved normalizing input pixel values and integrating a lightweight denoising module directly into the generator network to adaptively suppress irrelevant texture features during the end-to-end training process (Huang, 2024). The adversarial training stage then involved executing the CartoonGAN framework using a Convolutional Neural Network (CNN) encoder–decoder generator and a PatchGAN discriminator to guide competitive optimization loops. The final stage, quantitative convergence evaluation, was conducted to mathematically monitor the reduction and stabilization of both average content loss and generator adversarial loss over 50 epochs while verifying the visual quality of the output images.

Dataset and Sample Configuration

The subjects of this study comprise a collection of digital images divided into two distinct visual domains: real-world photographs and illustrative cartoon images. Since this research focuses on cross-domain image-to-image translation, the target population of data

was sourced from public repositories available on the Hugging Face platform. The total sample size utilized for training and evaluating the deep learning model consists of 10,000 images, which includes 5,000 photographic images representing the source domain and 5,000 cartoon images representing the target domain. Due to the unsupervised nature of the translation task, these image samples are configured in an unpaired manner, meaning there is no direct one-to-one correspondence between individual photographs and cartoon illustrations during the adversarial learning process.

To ensure a rigorous quantitative evaluation and prevent overfitting, the dataset sample was systematically partitioned using an 80:20 ratio. Specifically, 80% of the total image population (equivalent to 8,000 images) was allocated to the training set to optimize the network weights of both the generator and discriminator, while the remaining 20% (equivalent to 2,000 images) was strictly reserved as an independent testing set to evaluate the model's generalization capabilities on unseen data. All selected image samples underwent a standardization preprocessing pipeline prior to the training phase. This pipeline involved resizing all images to a uniform resolution of 256×256 pixels and normalizing their pixel intensities to a standard numerical range of [-1, 1], thereby minimizing visual variance and ensuring computational stability during the training loops.

Proposed System Architecture and Training Procedure

The proposed system architecture utilizes CartoonGAN as the primary model to execute photo-to-cartoon image translation. CartoonGAN is a Generative Adversarial Network (GAN) variant explicitly designed for image cartoonization tasks by emphasizing sharp contour learning and color simplification (Zhang, Zhao, Li, Wu, & Sun, 2023). The system consists of two primary components, namely a generator and a discriminator, which are trained adversarially.

The CartoonGAN generator is constructed using a Convolutional Neural Network (CNN)-based encoder–decoder architecture. The encoder function extracts spatial features from the photographic images, while the decoder is tasked with reconstructing these features into cartoon-style output images. This encoder–decoder architecture enables the model to learn visual representations hierarchically and has proven effective across various image-to-image translation tasks.

A joint learning approach is applied to directly integrate a lightweight denoising module into the CartoonGAN generator architecture. Through this approach, the noise reduction and cartoon stylization processes are trained simultaneously in an end-to-end manner within a unified network architecture. This enables the model to adaptively suppress irrelevant texture details and noise from the input photographic images without eliminating crucial semantic and structural information required for cartoon contour formation. This joint learning integration distinguishes the proposed method from conventional approaches that separate the preprocessing and stylization stages, thereby enhancing the visual consistency of the image translation results while maintaining the stability of the model training process.

The discriminator employs a PatchGAN architecture, which performs classification at the local patch level instead of the entire image. The PatchGAN approach is widely utilized in modern image-to-image translation research due to its ability to emphasize local texture and style consistency, which is highly relevant to image cartoonization tasks. The overall system architecture is illustrated in Figure 1.

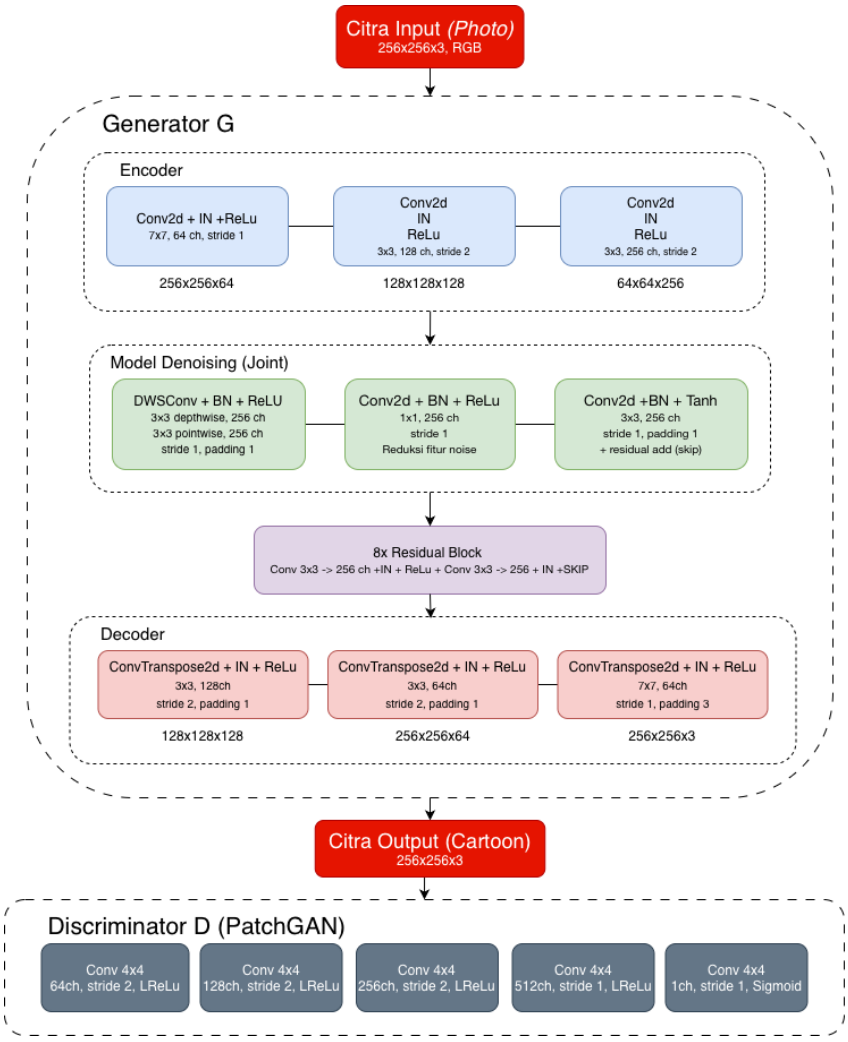

Figure 1: System Architecture of Photo-to-Cartoon

The execution of the entire training procedure is performed end-to-end over 50 continuous epochs, synchronizing the optimization of the denoising module and the stylization layers. The optimization process relies on competitive adversarial feedback loops driven by the Adam optimizer at a stable learning rate of 0.0002 to dynamically update network weights and minimize objective visual discrepancies across domains (Tang, 2023). By coupling the noise reduction and style translation parameters into a single mathematical runtime, the framework prevents the loss of vital edge boundaries that typically occurs in conventional, disconnected two-stage preprocessing workflows.

Data Analysis

A joint learning approach is implemented by integrating a lightweight denoising module into the CartoonGAN generator architecture. This module is positioned at the front end of the generator and functions to suppress noise and fine textures within the photographic images before the stylization process is executed.

Unlike conventional preprocessing approaches that are static, the denoising module in this study is trained end-to-end alongside the cartoonization process. This approach aligns with recent research findings indicating that integrating denoising directly into the primary learning framework can enhance the visual quality of results without requiring an isolated preprocessing stage (Jiménez-Gaona, Rodríguez-Alvarez, Escudero, Sandoval, & Lakshminarayanan, 2024; Wu, Chen, Xiang, Zhang, & Yang, 2023). Through end-to-end learning, the network can adaptively adjust the level of noise suppression according to the specific needs of the cartoon stylization.

Furthermore, the utilization of a lightweight residual structure within the denoising module supports the preservation of critical structural information and prevents over-smoothing, as demonstrated in modern CNN-based lightweight denoising network literature (Tiantian, Hu, & Guan, 2024). The joint learning process in this study is optimized using the same objective function as the cartoonization process, without adding a separate denoising target. This strategy ensures that the denoising module serves as a supporting component that reinforces the stylization process rather than acting as the primary training objective.

The model is trained end-to-end using a combination of Least Squares GAN (LSGAN)-based adversarial loss and L1-based content loss. The adversarial loss is utilized to drive the generator to produce images with a cartoon style that closely matches the target domain, while the content loss functions to preserve the primary semantic structures of the photographic images to prevent significant shape distortion. The lightweight denoising module operates as an integrated supporting component within the generator, trained via a joint learning scheme alongside the cartoonization process until the model achieves convergence and produces visually stable cartoon images.

The joint learning process in this study is optimized using a combined objective function consisting of adversarial loss and content loss. The adversarial loss is formulated using the Least Squares GAN (LSGAN) approach as follows:

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_{y \sim p_{cartoon}}[(D(y) - 1)^2] + \mathbb{E}_{x \sim p_{photo}}[(D(G(x)))^2]$$

The content loss is used to maintain semantic structural consistency between the input photographic images and the output cartoon images, which is formulated using the L1 loss as follows:

$$\mathcal{L}_{content}(G) = \mathbb{E}_{x \sim p_{photo}}[\|G(x) - x\|1]$$

The total objective function used to train the generator end-to-end is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{GAN} + \lambda \mathcal{L}_{content}$$

Here, λ acts as the weighting parameter that regulates the contribution of the content loss to the overall training process. This approach ensures that the denoising module acts as a supporting component that enhances the stylization process rather than serving as an independent training objective. The denoising module is not explicitly trained to minimize noise; instead, it is trained alongside the entire generator to minimize the total loss function, ensuring that the learned noise reduction level adaptively adjusts to the cartoon stylization requirements rather than independent denoising metrics. Furthermore, the λ parameter regulates the balance between two competing gradient signals: an excessively large λ value will cause the content loss gradient to dominate and inhibit stylization, whereas an excessively small value risks causing the generator to neglect structural consistency. Consequently, the residual connection at DN3 and within each residual block collectively ensure that signals from both loss functions reach the encoder parameters without significant degradation, thereby enabling the encoder to learn clean feature representations that are more easily stylized in subsequent stages.

FINDING AND DISCUSSION

RESEARCH RESULT

The training of the CartoonGAN model using the joint learning approach was conducted for 50 epochs utilizing an adversarial training scheme between the generator and the discriminator. This training process aimed to achieve an optimal equilibrium between the generator's ability to produce cartoon-style images and the discriminator's capability to differentiate between the generated outputs and real cartoon images. Throughout the training process, no indications of GAN training failure, such as loss divergence or the mode collapse phenomenon—which frequently present major challenges in generative model training—were observed.

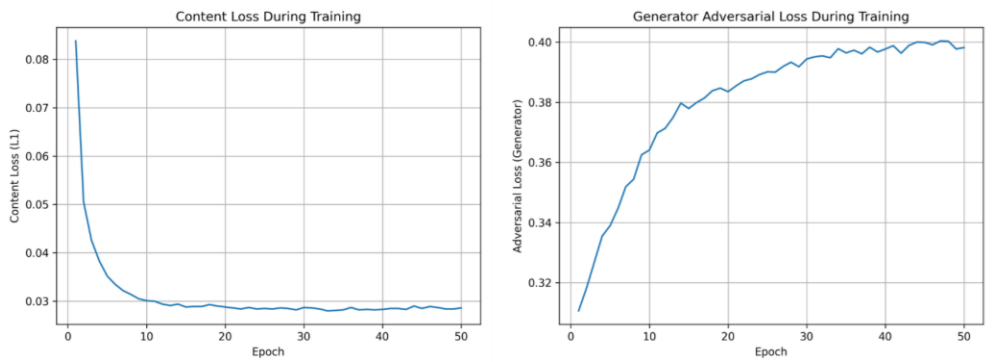

Figure 2: Grafik Content Loss dan Adversarial Loss dari Epoch 0-50

In the initial stages of training, the adversarial loss value of the generator was relatively high. This condition reflects that the generator was not yet capable of producing images with visual characteristics resembling the cartoon domain. At this stage, the discriminator could still easily distinguish between real cartoon images and the generated outputs. However, as the number of epochs increased, the generator began to learn the

primary visual patterns of the cartoon domain, such as color simplification and contour enhancement, leading to a gradual decrease in the adversarial loss value.

The stable decline in the adversarial loss value demonstrates that the generator became increasingly proficient at producing images that were difficult for the discriminator to distinguish. Conversely, the discriminator also continuously adapted by improving its ability to recognize subtle differences between real cartoon images and synthetic ones. This dynamic interaction resulted in a stable adversarial equilibrium, which serves as a critical indicator of successful GAN training. This condition aligns with the characteristics of an ideal GAN training process, where neither network excessively dominates the training procedure.

In addition to the adversarial loss, an L1-based content loss was employed to maintain semantic structural consistency between the input photographic images and the output cartoon images. Throughout the training process, the content loss value exhibited a consistent downward trend, reaching a relatively low value by the end of training. This decrease indicates that the generator did not solely focus on adapting the cartoon visual style, but was also capable of preserving the shapes and structures of the primary objects within the photographs. This preservation is crucial to ensure that the cartoonization results remain recognizable and do not undergo significant structural distortion.

At the end of the training phase, precisely at epoch 50, the discriminator loss value stabilized around 0.1957, while the generator adversarial loss value reached approximately 0.3862. These values indicate that the discriminator and the generator successfully achieved a well-balanced state equilibrium. Meanwhile, the content loss value, which reached approximately 0.0286, signifies that the structural discrepancy between the input photographic images and the output cartoon images was relatively minimal. The combination of these loss metrics demonstrates that the model successfully achieved convergence and was prepared for further qualitative visual evaluation.

Table 1: Loss Function Result

Epoch	Discriminator_loss	Content_loss	Adversarial_loss
50	0,1957	0,0286	0,3982

Overall, the model training results demonstrate that the proposed joint learning approach effectively maintains the training stability of CartoonGAN. The integration of the lightweight denoising module did not disrupt the adversarial learning process; instead, it supported the generator in learning cleaner visual representations focused on the characteristics of the cartoon style. This finding aligns with prior research indicating that the end-to-end integration of supporting components can enhance the stability and performance of generative models (Karras et al., 2021).

Visual Analysis of Cartoonization Results

The visual results of the photo-to-cartoon image translation serve as the primary indicator of the proposed approach's success. An example of the cartoonization results is

presented in Figure 2, which illustrates the comparison between the original source photograph and the generator's output cartoon image. Generally, the cartoonization results exhibit visual characteristics consistent with the cartoon style, characterized by sharper contours, color simplification, and a reduction in fine texture details.

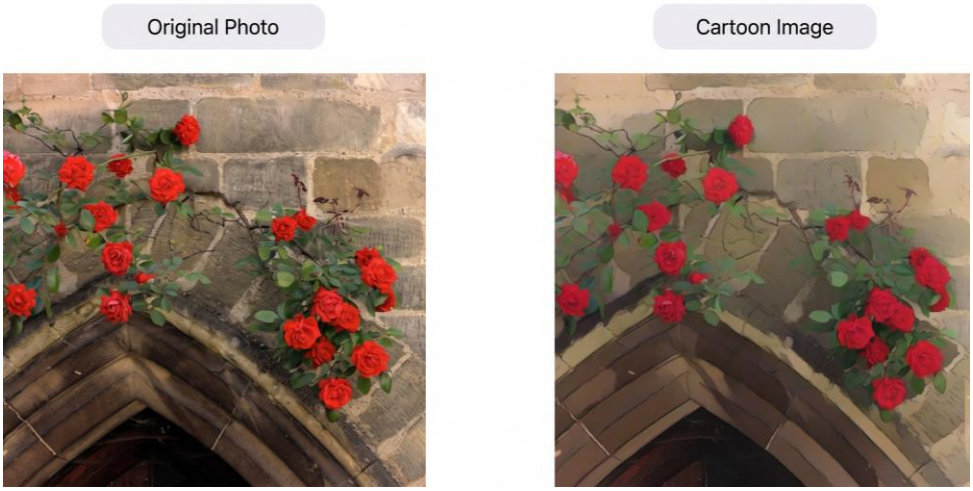

Figure 3: Visual results of cartoonization

One of the most prominent visual aspects of the cartoonization results is the enhancement of object contours. The edge boundaries of objects in the cartoon images appear more defined than in the original photographs, making object shapes more easily recognizable. This contour enhancement is a core characteristic of cartoon images and constitutes a primary objective of utilizing CartoonGAN. The obtained results demonstrate that the generator is capable of effectively learning object edge and shape representations through the adversarial training mechanism.

Beyond contours, color simplification serves as another critical characteristic of the cartoonization results. The cartoon images produced by the generator display flatter color regions with fewer gradient variations compared to the original photographs. This color simplification delivers a strong illustrative impression and minimizes unnecessary visual complexity. Concurrently, the color distinctions between primary objects are preserved, ensuring that the cartoon images maintain robust visual contrast.

The reduction of fine textures and noise represents another observed aspect of the cartoonization outputs. Photographic images typically contain noise and texture details originating from lighting conditions, object surfaces, and image acquisition processes. In the cartoonization results, these details are successfully suppressed, yielding a cleaner visual appearance. This noise reduction contributes to the consistency of the cartoon style and minimizes visual artifacts that could compromise image aesthetics.

Despite undergoing a significant stylization process, the cartoonization results successfully preserve the primary semantic structures of the original photographs. Object shapes, proportions, and spatial relationships between visual elements do not undergo drastic alterations. This indicates that the utilization of the L1-based content loss plays an

effective role in maintaining structural consistency between the input and output images. Consequently, the identities of objects within the cartoon images remain highly recognizable. When compared to a cartoonization approach lacking denoising integration, the results obtained via the joint learning approach demonstrate more consistent visual quality, where contours appear smoother and more stable, and artifacts caused by excessive texture are minimized.

DISCUSSION

This study demonstrate that integrating a lightweight denoising module via a joint learning framework successfully addresses the inherent trade-off between style abstraction and structural preservation in image cartoonization. The achievement of a well-balanced adversarial state—marked by a content loss of 0.0286, a generator adversarial loss of 0.3862, and a discriminator loss of 0.1957 at epoch 50—indicates that the generator can adaptively suppress high-frequency photographic noise without degrading the semantic edge boundaries necessary for clear cartoon representation. This stable convergence confirms that noise reduction and non-photorealistic rendering operate as complementary tasks when optimized under a shared objective function. Compared to conventional two-stage approaches where denoising is executed as an isolated preprocessing step, the proposed end-to-end system significantly reduces texture-driven visual artifacts and prevents over-smoothing. This outcome reinforces the assertions of recent generative learning literature which states that simultaneous parameter optimization yields higher aesthetic consistency and prevents training errors typical of disconnected pipelines (Jiménez-Gaona et al., 2024; Wu et al., 2023). Moreover, the successful retention of sharp contours and flat color distributions mirrors the structural high-fidelity achievements observed in modern CNN-based artistic style transfer frameworks (Tiantian et al., 2024; Zhang et al., 2023).

Despite the robust performance and training stability demonstrated by the modified CartoonGAN framework, certain limitations must be acknowledged. This study evaluated the proposed model using a specific configuration of 10,000 unpaired images from the Hugging Face repository at a standardized resolution of 256×256 pixels. Consequently, the network's behavior when processing ultra-high-definition imagery or handling highly intricate backgrounds under extreme variations in real-world illumination remains unexamined and may introduce additional computational overhead. Furthermore, because the denoising layers are explicitly tailored to cooperate with the cartoonization loss signals, the model may exhibit sub-optimal generalization if applied to alternative style transfer tasks, such as oil painting or sketch synthesis, without comprehensive architectural retraining.

These constraints offer valuable directions for future research and practical applications. Future studies should focus on scaling the joint learning architecture to accommodate larger image dimensions and integrating spatial attention mechanisms to better manage complex background dependencies. Practically, the proposed architecture holds substantial implications for the automated animation industry, digital content

creation tools, and mobile graphic applications, offering a highly efficient, reproducible pipeline that effectively bridges the gap between real-world computer vision tasks and automated digital art synthesis.

CONCLUSION

This study examines the application of the CartoonGAN model using a joint learning approach to execute photo-to-cartoon image translation. The proposed approach integrates a lightweight denoising module directly into the generator architecture, which is trained end-to-end alongside the stylization process. The primary objective of this approach is to suppress noise and irrelevant texture details in photographic images while preserving the core semantic structure of objects during cartoonization.

Experimental results demonstrate that the constructed system achieves a stable and convergent training process. This is evidenced by a low average content loss at the end of training, reaching 0.0286 at epoch 50, which indicates that the structural integrity of the photographic images is successfully maintained. Furthermore, an average adversarial loss of 0.3982 signifies that a healthy equilibrium is achieved between the generator and the discriminator, which is a critical characteristic in Generative Adversarial Network model training.

Qualitatively, the cartoonization results exhibit enhanced visual quality through sharper contours, flatter color regions, and a reduction in the fine textures typically found in photographs. The integration of the lightweight denoising module via a joint learning approach plays a pivotal role in generating cleaner feature representations, thereby supporting a more effective and consistent cartoon style learning process.

Despite the robust performance demonstrated by the results, this study still possesses certain limitations. The evaluation conducted has not yet included quantitative perceptual visual metrics, such as Fréchet Inception Distance (FID) or Learned Perceptual Image Patch Similarity (LPIPS), and remains restricted to a single cartoon style domain. Therefore, further development is required to test the model's generalizability across various cartoon styles and more diverse image conditions.

As directions for future research, this study can be extended by incorporating quantitative evaluation metrics, expanding the variety of cartoon styles, and optimizing the network architecture for higher computational efficiency. Additionally, exploring multi-style cartoonization approaches and integrating attention mechanisms hold the potential to enhance the flexibility and visual quality of the cartoonization outputs. Consequently, the proposed joint learning-based CartoonGAN approach can serve as a robust foundation for the future development of artificial intelligence-based image cartoonization systems.

REFERENCES

Chen, Y., Lai, Y.-K., & Liu, Y.-J. (2018). CartoonGAN: Generative adversarial networks for photo cartoonization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9465–9474.

https://openaccess.thecvf.com/content_cvpr_2018/html/Chen_CartoonGAN_Generative_Adversarial_CVPR_2018_paper

Chen, Y., Zhou, H., Chen, J., Yang, N., Zhao, J., & Chao, Y. (2025). Diffusion model-based cartoon style transfer for real-world 3D scenes. *ISPRS International Journal of Geo-Information*, *14*(8), 303. https://doi.org/10.3390/ijgi14080303

Gao, X., Zhang, Y., & Tian, Y. (2022). Learning to incorporate texture saliency adaptive attention to image cartoonization. *arXiv preprint arXiv:2208.01587*. https://doi.org/10.48550/arXiv.2208.01587

Huang, M. (2024). A survey on image style transfer based on deep learning. *Journal of Computing and Electronic Information Management*, *15*(3), 66–70. https://doi.org/10.54097/mxgtcj89

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5967–5976. https://doi.org/10.1109/CVPR.2017.632

Jiménez-Gaona, Y., Rodríguez-Alvarez, M. J., Escudero, L., Sandoval, C., & Lakshminarayanan, V. (2024). Ultrasound breast images denoising using generative adversarial networks (GANs). *Intelligent Data Analysis*, *28*(6), 1661–1678. https://doi.org/10.3233/IDA-230631

Jo, Y., Chun, S. Y., & Choi, J. (2021). Rethinking deep image prior for denoising. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 5067–5076. https://doi.org/10.1109/ICCV48922.2021.00504

Karras, T., Laine, S., & Aila, T. (2021). A style-based generator architecture for generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(12), 4217–4228. https://doi.org/10.1109/TPAMI.2020.2970919

Men, Y., Yao, Y., Cui, M., Lian, Z., Xie, X., & Hua, X.-S. (2022). Unpaired cartoon image synthesis via gated cycle mapping. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3491–3500. https://doi.org/10.1109/CVPR52688.2022.00349

Shu, Y., Yi, R., Xia, M., Ye, Z., Zhao, W., Chen, Y., . . . Liu, Y.-J. (2022). GAN-based multi-style photo cartoonization. *IEEE Transactions on Visualization and Computer Graphics*, *28*(10), 3376–3390. https://doi.org/10.1109/TVCG.2021.3067201

Tang, Y. (2023). ECGAN: Translate real world to cartoon style using enhanced cartoon generative adversarial network. *Computers, Materials & Continua*, *76*(1), 1195–1212. https://doi.org/10.32604/cmc.2023.039182

Thakur, A., Rizvi, H., & Satish, M. (2021). White-box cartoonization using an extended GAN framework. *International Journal of Engineering Applied Sciences and Technology*, *5*(12). https://doi.org/10.33564/IJEAST.2021.v05i12.049

Tiantian, W., Hu, Z., & Guan, Y. (2024). An efficient lightweight network for image denoising using progressive residual and convolutional attention feature fusion. *Scientific Reports*, *14*(1), 9554. https://doi.org/10.1038/s41598-024-60139-x

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2020). Deep image prior. *International Journal of Computer Vision*, *128*(7), 1867–1888. https://doi.org/10.1007/s11263-020-01303-4

Wang, L. (2022). Cartoon-style image rendering transfer based on neural networks. *Computational Intelligence and Neuroscience*, *2022*, 1–10. https://doi.org/10.1155/2022/2958338

Wu, W., Chen, M., Xiang, Y., Zhang, Y., & Yang, Y. (2023). Recent progress in image denoising: A training strategy perspective. *IET Image Processing*, *17*(6), 1627–1657. https://doi.org/10.1049/ipr2.12748

Xu, Y., Xia, M., Hu, K., Zhou, S., & Weng, L. (2025). Style transfer review: Traditional machine learning to deep learning. *Information*, *16*(2), 157. https://doi.org/10.3390/info16020157

Zhang, F., Zhao, H., Li, Y., Wu, Y., & Sun, X. (2023). CBA-GAN: Cartoonization style transformation based on the convolutional attention module. *Computers and Electrical Engineering*, *106*, 108575. https://doi.org/10.1016/j.compeleceng.2022.108575

Zheng, Z., Wang, C., Yu, Z., Wang, N., Zheng, H., & Zheng, B. (2019). Unpaired photo-to-caricature translation on faces in the wild. *Neurocomputing*, *355*, 71–81. https://doi.org/10.1016/j.neucom.2019.04.032