Comparative Study of Image Style Transfer Methods – VGG-19 vs CycleGAN vs Stable Diffusion Model

1. Introduction

1.1. Background and Motivation

In the quiet of a UK winter night, scrolling through TikTok for some midnight relaxation, I stumbled upon the enchanting paintings of Grimshaw. They stirred within me an ineffable sentiment, echoing the melancholy beauty of winter scenery in Durham, which was like a conversation beyond the time. What did Grimshaw perceive when he set up his easel under the Victorian moonlight? Capturing a winter night in England through my camera lens, I glimpsed an overlapping between the photograph and Grimshaw paintings – an ethereal sense of solitude and solace. Can technology be the artisan to weave them seamlessly together?

The potential realization of these scenarios might lie in the capabilities of various image style transfer models introduced on the course. Therefore, I embarked on a journey to use methods to train datasets of Grimshaw's oeuvre and night view photographs, which aimed to transform everyday scenes into Grimshaw-style paintings. Though my grasp of these techniques may be nascent, a comparative analysis of different models is provided, illustrating the alchemy of digital humanities. Here, art, emotion, and technology intertwine, opening new vistas for exploration and expression.

1.2. Project Objectives

- Gain a preliminary comprehension of VGG-19, CycleGAN, and Stable Diffusion Models.
- Collect and preprocess an appropriate dataset.
- Implement and evaluate three distinct image style transfer models.
- Transfer midnight photographs into Grimshaw-style paintings.

1.3. Project Structure

This project explores VGG-19, CycleGAN, and Stable Diffusion Models for image transfer and conducts a comparative analysis of these methods. Initially, it covers an

overview of the Art of John Atkinson Grimshaw and concepts for each model, followed by criteria for comparative analysis. The experiment section covers data scraping, preprocessing, and model fine-tuning. Then, it demonstrates the output for each method, qualitative and quantitative analysis. Finally, the conclusion discusses the pros and cons of each approach, and research limitations and offers suggestions for future exploration.

2. Art and Methodology

2.1. Art of John Atkinson Grimshaw

Figure 1. John Atkinson Grimshaw and Figure 2. A Cheshire Road 1883

John Atkinson Grimshaw (1836-1893) gained renown for his moonlit nocturne paintings, earning admiration even from artists like James McNeill Whistler, who remarked, "I considered myself the inventor of nocturnes until I saw Grimmy's moonlit pictures." (Dan Scott, 2019). Richard Dorment (2011) noted that without any formal painting training and family support, Grimshaw took a shortcut to commercial success by projecting photos or lantern slides onto a blank canvas and tracing over the outlines of instant composition, instead of painting from nature. Despite facing criticism, Grimshaw's paintings exuded a captivating allure, particularly in their portrayal of moonlit scenes, evoking mystery and eliciting poignant emotional responses from viewers.

Just as Grimshaw's unconventional painting methods sparked debate in his time, contemporary art infused with computer technology has also faced similar controversies, reflecting the meaning and tradeoff of the digital humanities field. The

paintings of Grimshaw present the moonlit, mist, and smoky fog of late Victorian industrial England with great poetry, achieved through his unique photo-projecting technique. It could be interesting and an echo to reproduce the sadness and emotional responses of Grimshaw's styles using nowadays techniques.

2.2. Neural Style Transfer - VGG-19

Leon A. Gatys, et al. (2015) first published the Neural Style Transfer, a technique that splits and recombines the content and style of arbitrary images using deep neural networks. Aman Kumar Mallik (2020) explains that this process extracts structural and stylistic features from the content and style images respectively, iteratively refining the image pixels through gradient descent. The model can prioritize either aspect in the generated image by adjusting the balance of content and style losses.

VGG-19, originally designed for image classification, finds a new purpose in style transfer by leveraging only its features component (Mallik, 2020). This study adopts VGG-19 as the primary method for Neural Style Transfer, using a pre-trained version to accelerate the image transformation.

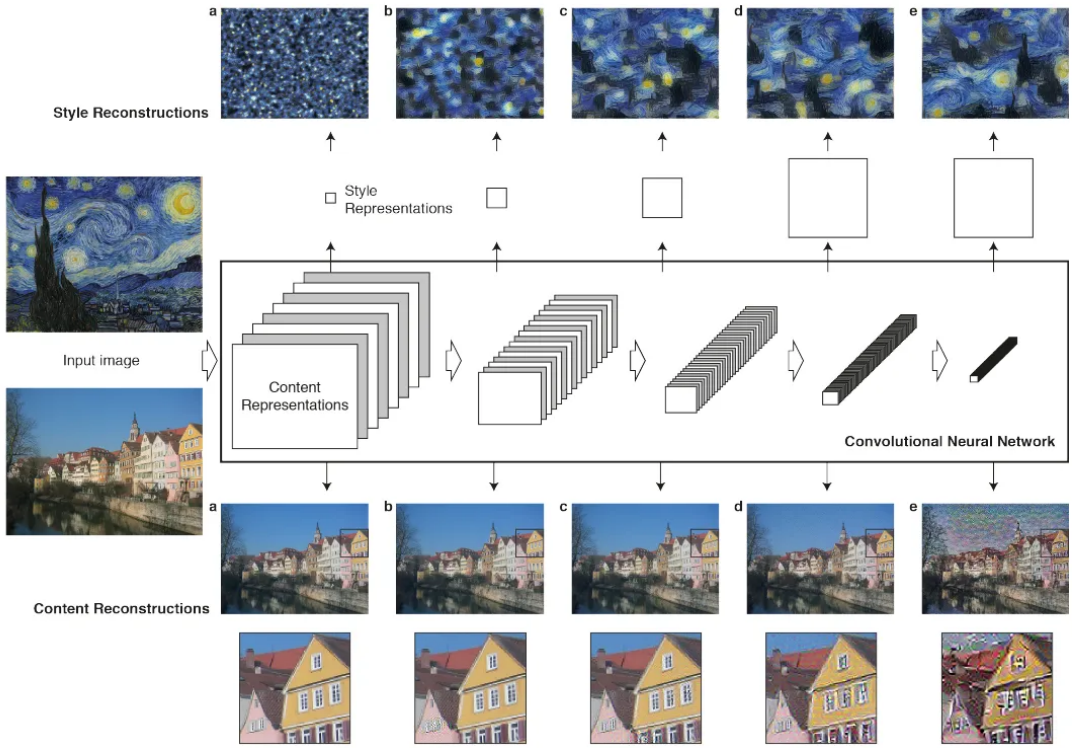

Figure 3. The Model Architecture of NST (Source: A Neural Algorithm of Artistic Style, 2015)

2.3. Generative Adversarial Network – CycleGAN

The Generative Adversarial Network (GAN) concept, developed by Ian Goodfellow in 2014, comprises two components: the generator and the discriminator. The generator produces image data, while the discriminator learns to distinguish between fake and real data. Through iterative backpropagation, the generator improves by fooling the discriminator, which is also constantly improving, resulting in realistic virtual data (Google, 2022). However, Nikolas Adaloglou (2020) highlighted the challenge of mode collapse in GANs, where either the generator or discriminator gains an overwhelming advantage, leading to a stagnant training process. To ensure model stability, it is crucial for both components to learn simultaneously, avoiding these pitfalls.

CycleGAN, the second method employed here, operates with unpaired datasets for image transfer. GANs leverage adversarial loss to force the generated images to be nearly indistinguishable from real images. To refine the mappings, two “cycle consistency losses” are introduced, when translated from one domain to the other and then reverted, a return to the starting point: forward and backward cycle-consistency loss: $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$, $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$ (Zhu et al., 2017). By minimizing the cycle-consistency losses (i.e., The result of the cycle conversion should be as close as possible to the original image), the model preserves original features while effectively transferring image style.

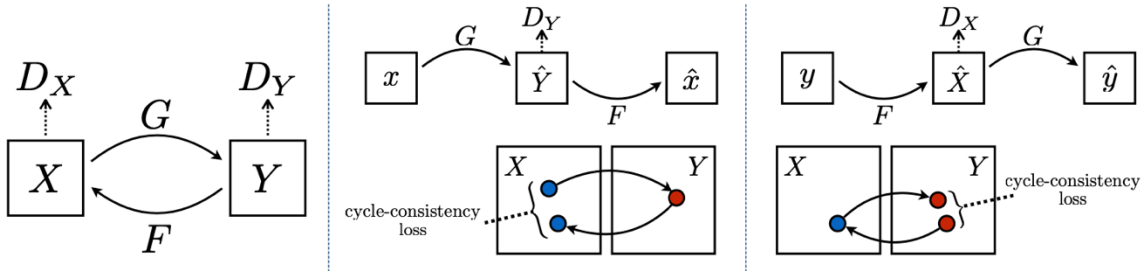

Figure 4. Cycle Consistency Loss (Source: Zhu et al., ICCV, 2017)

2.4. Stable Diffusion

The Stable Diffusion model, developed by researchers at Ludwig Maximilian University of Munich and Heidelberg University with funding from Stability AI, was published in 2022. However, Matt Growcoot (2023) argued that doubts have been raised regarding the educational background of Emad Mostaque, who is the founder and was CEO of Stability AI until 23 March 2024, according to Forbes. Additionally, the company has

not provided extensive details about the contributions of the original researchers. Despite the potential impacts of commercial controversies and dramas, this model is still utilized as the third method in the project due to its open-source nature.

The Stable Diffusion model comprises three key components: a variational autoencoder (VAE), U-Net, and an optional text encoder. The VAE encoder maps input data to latent space, where Gaussian noise is added during forward diffusion. The U-Net then denoises the output backwards, predicting noise samples to obtain latent information, with text information (token embeddings) incorporated via a cross-attention. Finally, the VAE decoder uses the information produced by U-Net to generate the final image (Jay Alammar 2022). In this project, the version of Stable Diffusion with Diffusers is employed, a text-to-image latent diffusion model trained on 512x512 images from a subset of the LAION-5B database (Suraj Patil et al., 2022).

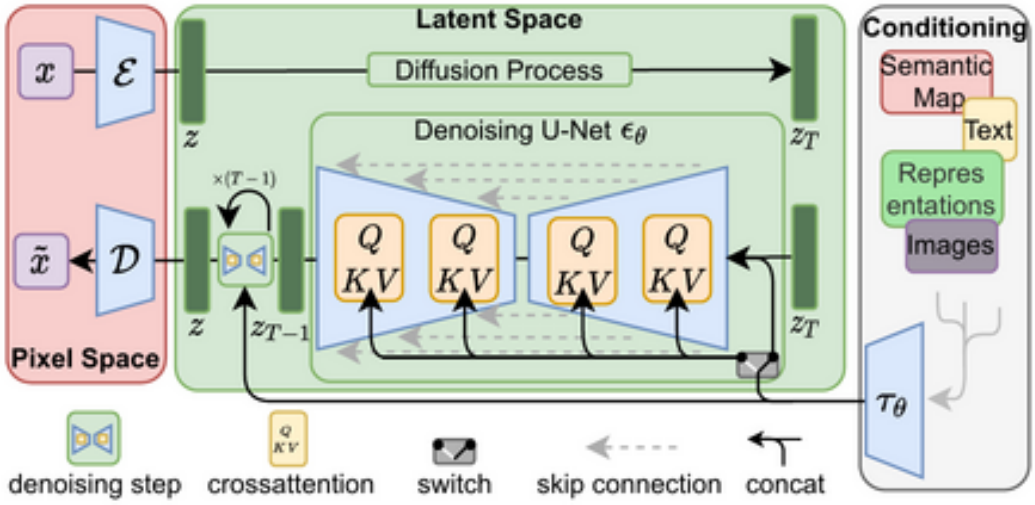

Figure 5. Diagram of the Latent Diffusion Architecture (Source: Jay Alammar, 2022)

2.5. Selection Criteria for Comparison

The comparative analysis focuses on two main aspects: the reproduction of style (qualitative) and the efficiency of model performance (quantitative). Style reproduction involves subjectively evaluating the ability of each model to transfer style, by comparing the features, color tone, brush strokes, atmosphere etc. Model efficiency compares the performance of the models based on their time consumption, providing insights into their computational efficiency.

3. Experimental Setup

3.1. Dataset Source and Pre-processing

Pre-trained models of VGG-19 and Stable Diffusion are utilized in this project. The dataset for training CycleGAN is collected separately to compare the performance between pre-trained and untrained models. Night view photographs and Grimshaw paintings are initially scraped from LAION, a German non-profit organization that received funding from Stability AI. However, Andy Baio (2023) updated the information in his paper that LAION took down its LAION-5B and LAION-400M datasets due to being accused that its dataset included child sexual abuse material. The ethical issues of data have always been a significant part of discussion and concern. The good news is that JSON files containing thousands of URLs downloaded in the previous term still function. Although some URLs are unavailable, 6180-night view photos and 4634 Grimshaw paintings are successfully downloaded. Due to limitations in the number of artworks available and duplications among the 4636 paintings, Grimshaw paintings are instead sourced from WIKIART. These 145 paintings depict various subjects and stages of Grimshaw's work, including figure painting and daytime scenery. Although they may not be relevant to night views, the model might learn some style, feature, or brushstrokes. Through data augmentation techniques such as rotation and flipping, the dataset is expanded to 1160 paintings. Combining these with 1160-night view photos extracted from 6180 photos, the datasets are divided into training and testing data. To meet the requirements of VGG-19 and Stable Diffusion models, images are standardized to 512x512 pixels and cropped into square shapes.

3.2. Training and Testing Procedures

Model	***VGG-19***	***CycleGAN***	***Stable Diffusion v1-4***
Neural network	*Neural Style Transfer*	*Generative Adversarial Network*	*Stable Diffusion*
Fine-tuning	*Adjust hyperparameters such as epoch, weight etc.*	*Custom training datasets*	*Adjust hyperparameters such as initial image, text prompt, guidance scale etc.*
Training Duration	*Pre-trained*	*8h26min*	*Pre-trained*
Training Dataset	*×*	*1150 content images + 1150 style images*	*×*
Testing Dataset	*10 content images + 10 style images*	*10 content images*	*10 content images*
Software	*Google Colab*	*Google Colab*	*Google Colab*
Hardware	*Colab GPU (T4)*	*Colab GPU (V100)*	*Colab GPU (T4)*
Time Consuming	*43 min*	*5 seconds*	*3 min*

Figure 6. Comparison Table of Procedures for Each Method

VGG-19

Notably, the pre-trained VGG-19 optimizer backpropagates and updates the generated image pixels, rather than adjusting the model parameters themselves. Hyperparameters are fine-tuned through multiple experiments to achieve optimal transfer output. Various weights of style and content images are compared to determine the best combination. When content and style images share similar features, such as buildings, the output appears satisfactory. A beta weight of 100 is found to be most suitable; a higher value blurs cars and buildings, while a lower value results in darker tones. In the formal transfer process, content and style images are paired based on similar features and transferred using 1000 epochs, 0.004 learning rate, 8 alpha, and 100 beta.

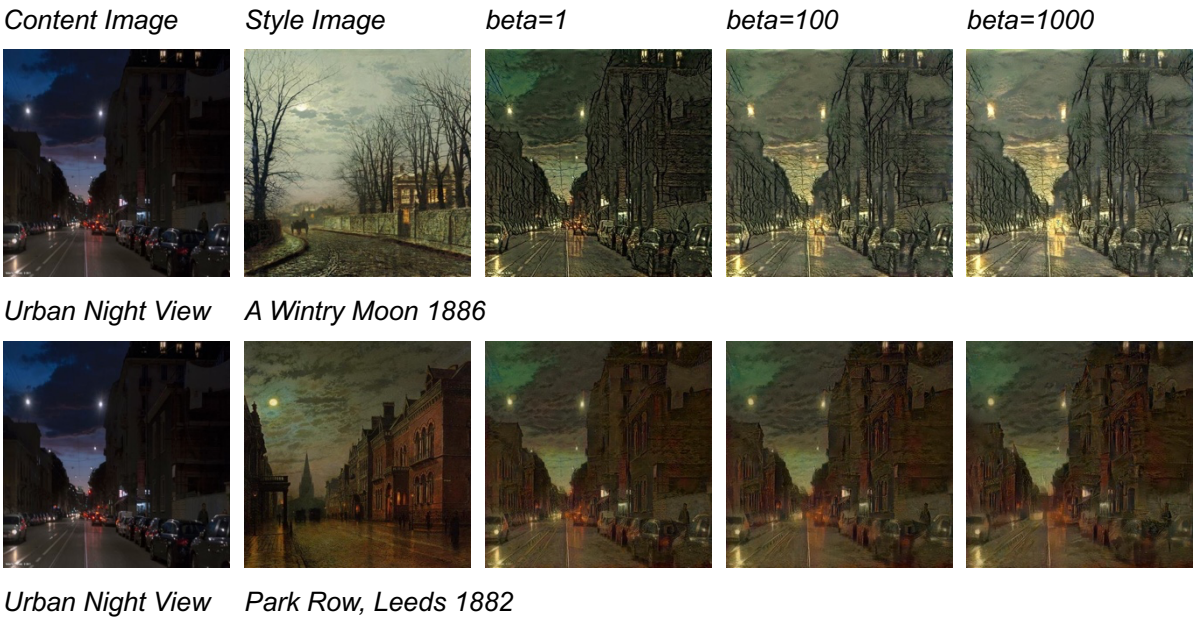

Figure 7. Experimental Results of the VGG-19 Model

CycleGAN

After numerous attempts and training across different environments including NCC SLURM and Google Colab, the CycleGAN successfully ran on Google Colab Pro with a V100 GPU, training for a total of 8 hours and 26 minutes. The model was trained on 1150 night view images and 1150 Grimshaw images (though the same number of images, were not paired). Due to the limited number of Grimshaw paintings, simple data augmentation techniques were used to increase the dataset size. The trained

model was then applied to testing data, consisting of 10 night view images, to transfer the image style.

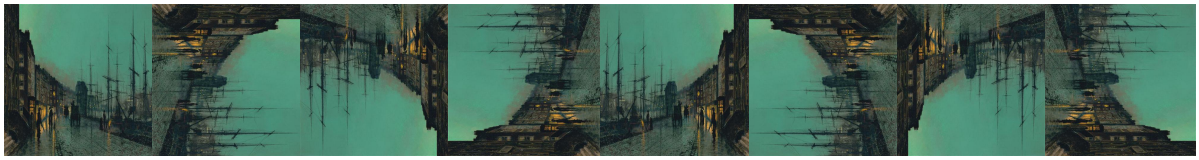

Figure 8. Data Augmentation (Rotation and Mirroring)

Stable Diffusion v1-4

Using Google Colab to train the Stable Diffusion model posed challenges, as the Stable Diffusion XL version easily exhausted the free virtual memory. Additionally, the output was less related to the input image and focused more on the provided text prompt, despite offering high resolution. Therefore, the older version v1-4 of Stable Diffusion was deemed more suitable for the project. A manual seed of 1012 was set to ensure consistent output. The prompt "a painting of night view in the style of John Atkinson Grimshaw" was used as text input. Three images were generated, indicating the efficiency of Stable Diffusion compared to other models. Various guidance scales are experimented with to find the best output close to the style of Grimshaw, which forces the generation to better match the prompt potentially at the cost of image quality or diversity (Suraj Patil et al., 2022). A scale of 5 yields the most favorable results. This scale will be used for the formal generation process.

Figure 9. Experimental Results of the Stable Diffusion V1-4 Model

4. Results

4.1. Visual Comparison of Style Transfer Outputs

Color Tone, Brightness and Contrast

The VGG-19 model accurately captures the color tone of paired style images. CycleGAN also successfully transfers images to a similar Grimshaw style. However, Stable Diffusion v1-4 maintains the color tone of input content images. In terms of brightness, VGG-19 effectively maps the brightness relationship into generated images. The brightness of images generated by CycleGAN appeared smooth and moderate, with both brighter and darker areas converging towards the medium tone. Both methods brighten the darker areas, resulting in soft contrast and hazy gray tones reminiscent of Grimshaw's style. Conversely, images created by Stable Diffusion exhibit strong contrast, inherited from the input images.

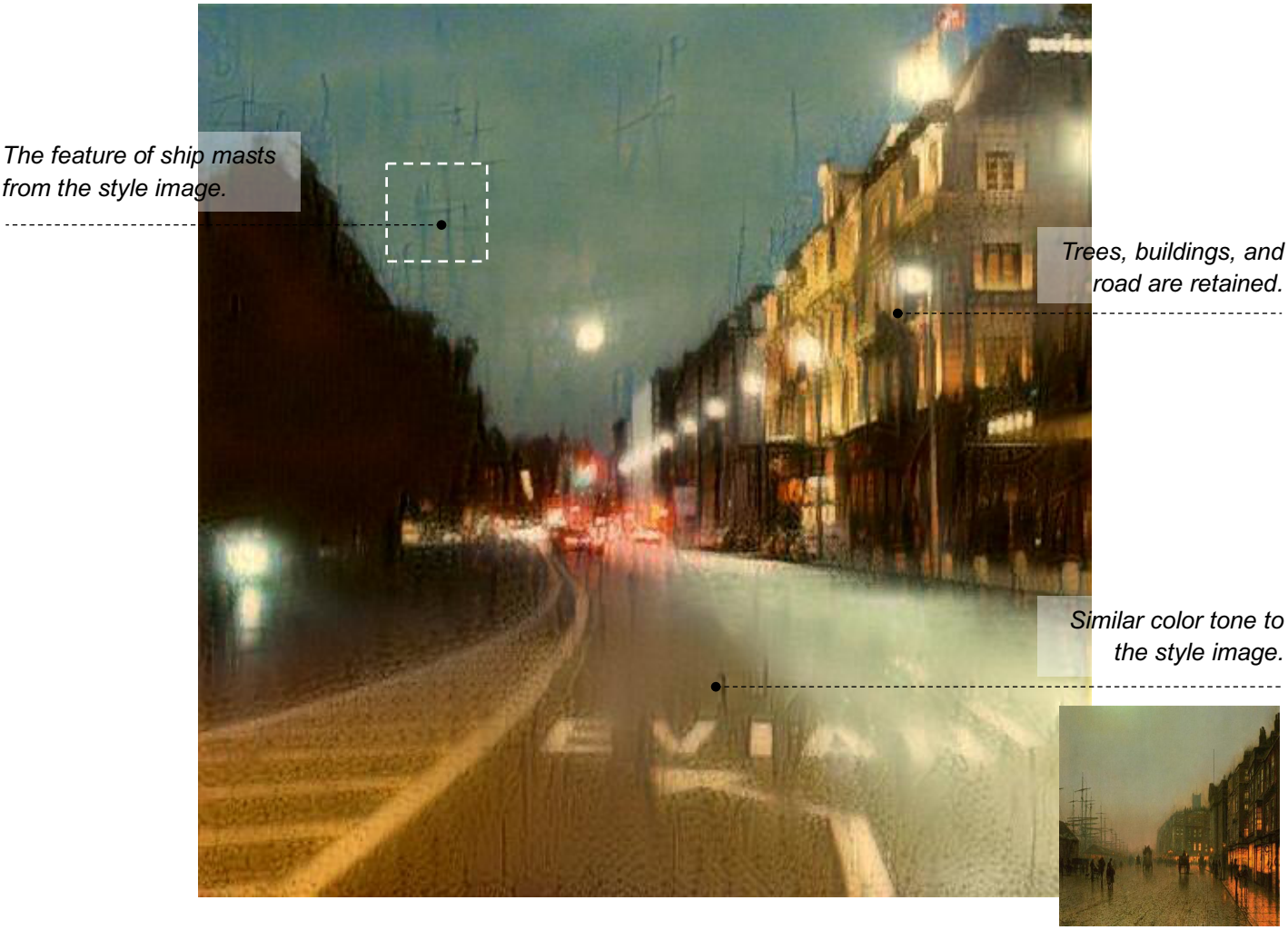

Figure 10. Analysis of the Output Generated by VGG-19

Features

VGG-19 effectively preserves the content features of input images due to the use of paired datasets. However, with unpaired datasets, there is a risk of buildings being transferred to trees, as observed in the experiments. Additionally, it tends to map the patterns or brush strokes of style images, such as tree branches, clouds, and ship masts, onto the generated images, resulting in some mottled and unusual textures. The CycleGAN model tends to neutralize and blur everything, as seen in the relative middle brightness parts of the images. And lights, moon, dark trees, and dark buildings become faded. Stable Diffusion v1-4 preserves the perspective relationship of the input images to some extent, such as one-point perspective, and retains the color tone and brightness. For instance, bright areas in the original image are also highlighted in the generated image. However, it struggles to recognize and maintain specific content details, often leading to the transfer of buildings to trees and vice versa. This could be influenced by the impact of text prompts and its underlying mechanism.

Figure 11. Analysis of the Output Generated by CycleGAN

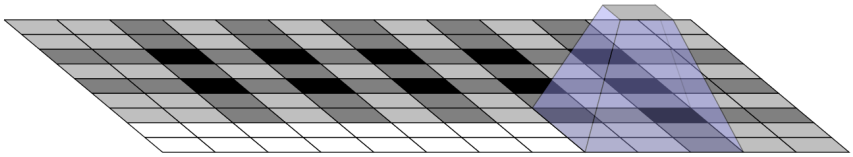

Figure 12. Checkerboard Patterns Caused by Uneven Overlap (Source: Odena et al., 2016)

Brushstrokes and Atmosphere

The images produced by VGG-19 closely resemble Grimshaw's style in terms of brushstrokes and atmosphere compared to other models. While CycleGAN generates images with a decent Grimshaw-style atmosphere, some peculiar checkerboard patterns are noticeable, which has been acknowledged by the author of CycleGAN. This issue is attributed to uneven overlap during deconvolution, particularly when the kernel size is not divisible by the stride (Odena, et al., 2016). The generated images from the Stable Diffusion model resemble crayon drawings, which is understandable considering that the model is pre-trained on mixed datasets, and models tend to learn more precise information about image style from image data rather than text data.

Figure 13. Analysis of the Output Generated by Stable Diffusion V1-4

Figure 14. Content and Style Images, and Comparison of Outputs for Each Method (Part 1)

Figure 15. Content and Style Images, and Comparison of Outputs for Each Method (Part 2)

4.2. Computational Efficiency Analysis

Since VGG-19 and Stable Diffusion v1-4 models are pre-trained, their training times are not recorded here. For generation time, VGG-19 took approximately 43 minutes to generate 10 paintings using paired content and style images as inputs, with 1000 epochs set. In contrast, Stable Diffusion v1-4 spent about 3 minutes to generate 3*10 higher-resolution images using content images and text prompts as inputs. Despite the potential influence of epoch numbers on VGG-19's generation speed, it's evident that Stable Diffusion is more efficient in terms of speed. CycleGAN, with a training time of 8 hours and 26 minutes, and testing time of 4 seconds, produces outputs with some semblance of Grimshaw's style, albeit with lower detail quality.

5. Discussion and Conclusion

5.1. Discussion

In summary, this project has extensively explored, compared, analyzed, and discussed image style transfer using three different models: neural style transfer, generative adversarial network, and stable diffusion. It provided background information on the target artist's style, outlined the concepts and structures of the three networks, and detailed the dataset and model operation process. Despite encountering challenges and undergoing multiple experiments, the results and performances of each model were compared and analyzed.

The qualitative and quantitative analyses revealed that the VGG-19 model exhibits a relatively slow generation speed, requiring several epochs to converge the loss gradient. However, it excels in capturing the features of content images and the style features and color tones of style images. The performance of CycleGAN is moderate, as it manages to learn some aspects of Grimshaw's style but lacks detail and quality. The analysis also indicates that Stable Diffusion v1-4 offers significant speed and high-resolution generation capabilities but struggles with the correlation between target content and style, although accurately capturing the color tone and brightness of original images.

5.2. Limitations

Methodologies

The parameter settings and operating environment of the model varied significantly and lacked rigor, primarily due to the inherent complexities of their different frameworks. As a result, achieving strict control becomes impractical. Additionally, the horizontal comparison of the three methods limited the depth of exploration for individual models due to constraints in space and time of this project.

Datasets

The pre-training of the VGG-19 and Stable Diffusion models utilized datasets that were not specifically tailored to Grimshaw's style, potentially impacting the accuracy of outputs. Furthermore, the Grimshaw dataset used in the project was limited in size, and the data augmentation process was simplistic, involving only rotation and flipping of images.

5.3. Future Plans

In future endeavors, it would be beneficial to train all three methods using the same datasets instead of relying on pre-trained models. Generating various images in different painting styles, such as those of van Gogh and Picasso, instead of solely focusing on Grimshaw's style, would allow for a more well-rounded and comprehensive comparison of the performance of each model. Implementing more complex data augmentation techniques could help increase the diversity of the datasets, such as cropping, color jittering, and noise injection (Connor Shorten et al., 2019). For more ambitious projects, the analysis could be more objective and scientific by diving into the structure of each model to gain valuable insights into the factors influencing their outputs. For instance, Odena, et al. (2016) suggest that adjusting the classes of deconvolution to upsampling can reduce checkerboard artifacts, which could be explored further to enhance the performance of the models.

6. References

References for the Artist

- Dan Scott (2019), “John Atkinson Grimshaw – The Inventor of Nocturnes”. Available at: https://drawpaintacademy.com/john-atkinson-grimshaw/ (Accessed on 4th April 2024).

- Richard Dorment (2011), “John Atkinson Grimshaw, Guildhall and Richard Green Galleries, review”, The Telegraph. Available at: https://www.telegraph.co.uk/culture/art/art-reviews/8776001/John-Atkinson-Grimshaw-Guildhall-and-Richard-Green-Galleries-review.html (Accessed on 4th April 2024).

References for Datasets

- Connor Shorten, et al. (2019). "A survey on Image Data Augmentation for Deep Learning". Mathematics and Computers in Simulation. 6. springer: 60. doi:10.1186/s40537-019-0197-0. Available at: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0 (Accessed on 13th April 2024).

- LAION. Available at: https://laion.ai (Accessed on 4th January 2024).

- WIKIART, “John Atkinson Grimshaw”. Available at: https://www.wikiart.org/en/john-atkinson-grimshaw/all-works/text-list (Accessed on 8th April 2024).

References for Neural Networks and Models

- Aman Kumar Mallik (2020), “Neural Style Transfer Using PyTorch”, Towards Data Science. Available at: https://towardsdatascience.com/implementing-neural-style-transfer-using-pytorch-fd8d43fb7bfa (Accessed on 24th March 2024).

- Andy Baio (2022) "Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator". Waxy.org. Archived from the original on January 20, 2023. Retrieved November 2, 2022. Available at:

https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/ (Accessed on 6th April 2024).

- Augustus Odena, et al. (2016), “Deconvolution and Checkerboard Artifacts”, Distill. Available at: https://distill.pub/2016/deconv-checkerboard/#citation (Accessed on 13th April 2024).

- Google (2022), “GAN”. Available at: https://developers.google.com/machine-learning/gan/gan_structure (Accessed on 23rd March 2024).

- Jay Alammar (2022), “The Illustrated Stable Diffusion”, jalammar.github.io. Archived from the original on November 1, 2022. Retrieved October 31, 2022. Available at: https://jalammar.github.io/illustrated-stable-diffusion/ (Accessed on 6th April 2024).

- Jun-Yan Zhu, et al. (2017), “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks”. Available at: https://openaccess.thecvf.com/content_ICCV_2017/papers/Zhu_Unpaired_Image-To-Image_Translation_ICCV_2017_paper.pdf (Accessed on 27th December 2023).

- Leon A. Gatys, et al. (2015), “A Neural Algorithm of Artistic Style”, arXiv:1508.06576 (cs). Available at: https://arxiv.org/pdf/1508.06576.pdf (Accessed on 5th April 2024).

- Matt Growcoot (2023), “‘So Many Things Don’t Add Up’: Stability AI Founder Accused of Exaggerations”, PetaPixel. Available at: https://petapixel.com/2023/06/05/so-many-things-dont-add-up-stability-ai-founder-accused-of-exaggerations/ (Accessed on 6th April 2024).

- Nikolas Adaloglou (2020), “GANs in computer vision - Introduction to generative learning”. Available at: https://theaisummer.com/gan-computer-vision/#vanilla-gan-generative-adversarial-networks-2014 (Accessed on 6th April 2024).

- Suraj Patil, et al. (2022), “Stable Diffusion with Diffusers”, Hugging Face. Available at: https://huggingface.co/blog/stable_diffusion (Accessed on 11th April 2024).

7. Appendices

- Python Code for Data Crawling and Pre-processing. Available at: https://github.com/DwayneHuang/DH_Assignment/blob/main/Data_Crawling_and_Pre_processing_(Night%20view).ipynb

- https://github.com/DwayneHuang/DH_Assignment/blob/main/Data_Crawling_and_Pre_processing_(Grimshaw).ipynb

- Python Code for VGG-19. Available at: https://github.com/DwayneHuang/DH_Assignment/blob/main/VGG-19.ipynb

- Python Code for CycleGAN. Available at: https://github.com/DwayneHuang/DH_Assignment/blob/main/CycleGAN.ipynb

- Python Code for Stable Diffusion. Available at: https://github.com/DwayneHuang/DH_Assignment/blob/main/Stable_Diffusion.ipynb