# Exploring Content Prompt Integration in Creative Adversarial Networks for Art Generation

Suweyba Abdiriziq
190281368
Supervisor : Simon Colton
MSc Artificial Intelligence

***Abstract*—**

***This research explores the integration of content prompts into Creative Adversarial Networks (CANs) to generate AI-driven art that balances creative diversity with conceptual consistency. By incorporating OpenAI's CLIP (Contrastive Language-Image Pretraining) model into the CAN framework, the study addresses the challenge of aligning generated artwork with specific prompts while preserving creative freedom. The experimental results indicate that the integration of CLIP successfully improved contextual alignment, as evidenced by the higher CLIP similarity scores, which show that the images generated more closely adhered to the provided prompts. After developing the basic CAN, additional loss components were incorporated into the training process to enhance image diversity and visual appeal. The Advanced CAN subsequently produced more varied and engaging outputs, demonstrating its improved capacity to blend creativity with prompt interpretation. Demonstrating a method of AI art generation that highlights prompt adherence, content consistency, and continues to push the boundaries of creative expression.***

***Keywords— Generative art, Creative Adversarial Networks (CAN), CLIP model, AI art, content consistency, prompt adherence***

## I. INTRODUCTION

The field of AI-generated art has seen remarkable growth and transformation over the past few decades. Its origins trace back to the 1960s and 1970s [17] when artists and computer scientists began exploring algorithmic art, using early computer technology to create innovative patterns and designs. For instance, Google's DeepDream algorithm[6] uses convolutional neural networks to enhance and reveal patterns in images, creating dream-like visuals. Additionally, Karl Sims' use of evolutionary algorithms[7] in the program demonstrated how these algorithms could evolve artistic forms through simulated evolutionary processes, generating visually compelling graphics and exploring complex design possibilities. As technology progressed, both in terms of computational power and algorithmic inovation, the potential for AI to generate more interesting and imaginative works expanded.

A pivotal advancement came in 2014 with the introduction of Generative Adversarial Networks (GANs) by Ian Goodfellow [3] and his colleagues. GANs represented a significant advance in AI art, enabling machines to generate images resemble the original images and explore diverse artistic possibilities. Causing a breakthrough that sparked widespread interest in GANs, making them essential tools for creating and experimenting with new artistic styles.

This led researchers to develop the Creative Adversarial Network (CAN) [1], an extension of the Generative Adversarial Network (GAN) designed to generate creative artworks by intentionally deviating from established artistic styles. While CAN successfully fostered creativity, it lacked control over the thematic content of the generated images, resulting in outputs without a specific theme or coherent idea. The research advances the field by tackling the challenge of achieving conceptual consistency in AI-generated art while preserving creative diversity. Aiming to direct the generative process to produce images that are both creatively unique and semantically aligned with specific prompts or themes.

The hypothesis of this research is that by integrating semantic prompts directly into the CAN via the OpenAI's CLIP model, making it possible to produce AI-generated art that is both creatively diverse and thematically consistent. This integration allows for guidance during the image generation process, ensuring that the output adheres to specific artistic instructions while still exploring creative variations.

## II. PROBLEM STATEMENT

This research tackles the challenge of traditional CANs, which often struggle with thematic coherence despite their creativity. By incorporating CLIP into the CAN framework, the goal is to generate images that are both creatively diverse and thematically aligned with specific textual prompts. Aiming to refine the generative process, ensuring that the resulting artwork is not only imaginative but also contextually relevant and consistent with the given themes.

## III. CONTRIBUTION

In the field of AI art generation, this research makes several novel contributions:

- Exploring Contextual Coherence and Creative Diversity: By merging the creative versatility of

CANs with the contextual accuracy of CLIP, this approach fosters the creation of art that is both imaginative and conceptually consistent, showcasing AI's potential to generate relevant, distinctive artistic outputs.

- Enhanced Image Diversity and Artistic Quality**:** The addition of extra components to the CAN model significantly improved the diversity and visual appeal of the generated images. This enhancement resulted in artwork that better resembles genuine art, showcasing the model's capability to produce varied and engaging images.
- Preserved content consistency across different prompts: The research validated the approach across various prompts, revealing that the enhanced models excelled with certain themes like mountain landscapes, where higher CLIP scores indicated better alignment. This suggests that integrating CLIP effectively enhances AI art quality in specific thematic areas, providing a targeted approach for creative projects.

## IV. Background

### *A. Generative Adversarial Networks (GANs)*

Generative Adversarial Networks (GANs), first introduced by Ian Goodfellow et al. in 2014 [3], have significantly advanced the field of generative artificial intelligence, particularly in the realm of image synthesis. The core concept of GANs is adversarial training, where two neural networks—a generator and a discriminator—are trained simultaneously, each with opposing objectives. These networks are trained simultaneously in a minimax game, where the generator tries to create images indistinguishable from real ones, and the discriminator aims to correctly classify images as real or fake. This adversarial interaction can be mathematically framed as a minimax game, where the generator seeks to minimize the discriminator's ability to correctly identify generated images, and the discriminator aims to maximize its classification accuracy. The objective function for this setup is typically expressed as:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

In this equation (1) from [3], the generator G(z) takes a noise vector z and outputs an image, while the discriminator D(x) predicts whether an image x is real or generated. Through iterative training, the generator improves its ability to produce images that are increasingly difficult for the discriminator to distinguish from real ones.

Since the introduction of GANs, several variants have emerged to broaden their applications. Wasserstein GAN (WGAN) [8] was developed to address the training instability and mode collapse often seen in traditional GANs. The Wasserstein distance used in WGAN ensures meaningful gradients throughout training, avoiding vanishing gradients and resulting in higher-quality, more diverse images. Another variant, StyleGAN [9], features a generator architecture that distinctly separates high-level attributes from stochastic variations, offering improved control over image synthesis and enhancing both image quality and attribute disentanglement. These advancements have significantly expanded the utility of GANs in generating diverse and high-quality images.

Traditional GANs, while effective in generating realistic images, often focus on mimicking the appearance of real-world data. This focus limits their ability to foster creativity or adhere to specific artistic concepts, as the generated outputs are constrained by the training data distributions.

### *B. Creative Adverserial Networks (CANs)*

Unlike traditional GANs that prioritize realism, CAN aims to produce art that is not easily categorized into any specific style, fostering greater creativity. The CAN model builds upon the foundational structure of the generative adversarial network by integrating stylistic ambiguity as a new objective.

The CAN model's process is as follows: the generators G start with random noise input z sampled from a distribution $p_z(z)$ and process it through the model's convolution layers to generate the image. This can be represented as G(z), where G is a function parameterized by the generator's weights.

The discriminator D evaluates the generated image G(z) to determine if it is real or fake, outputting a probability $D_r(G(z))$for realness. Additionally, the discriminator classifies the style of the image into one of K possible style categories, providing a probability distribution $D_r(C_k|G(z))$over these styles. The discriminator's loss function maximizes its ability to distinguish real from fake images and accurately classify styles by minimizing: $-E_{x\sim pdata}\,[\log D_r(x) + \log D_c(c = \hat{c}|x)]$ for the real images and $-E_{z\sim p_z}\,[\log(1 - D_r(G(z))]$ for the generated images as shown in [1]. The generator aims to minimize the following loss:

$$E_{z\sim p_z}[\log(1 - D_r(G(z)) - \sum_{k=1}^{k}\left(\frac{1}{k}\log(D_c(c_k|G(z)))\right) + \left(1 - \frac{1}{K}\right)\log(1 - D_c(c_k|G(z)))] \quad (2)$$

The first term in the equation (2) from [1], encourages the generator to produce images that the discriminator will classify as fake, pushing the generator to create more convincing images. The second term aims to increase the uncertainty about the style of the generated images by minimizing the cross-entropy between the predicted style distribution and a uniform distribution. The generator is penalized if the discriminator is confident about any specific style, promoting stylistic ambiguity.
Each time, the generator generates a new image, which the discriminator evaluates and provides feedback for further improvement.

The overall objective can be framed as a minimax game:

$$\min_{G}\max_{D} V(D,G) = E_{x,\hat{c}\sim pdata}[\log D_r(x) + \log D_c(c = \hat{c}|x)] + E_{z\sim p_z}\left[\log\left(1 - D_r(G(z))\right) - \sum_{k=1}^{k}\left(\frac{1}{k}\log(D_c(c_k|G(z)))\right) + \left(1 - \frac{1}{K}\right)\log(1 - D_c(c_k|G(z)))\right]$$

as outlined in Equation (3) from [1]. Through iterative adversarial training, the generator seeks to minimize its loss by producing images that are both realistic and ambiguous in style, making it challenging for the discriminator to classify them accurately. At the same time, the discriminator aims to maximize its accuracy by correctly distinguishing real images from generated ones and classifying styles correctly. This minimax game results in CANs producing increasingly realistic and creatively ambiguous artworks.

Creative Adversial Networks provide a significant advancement in artwork creation by utilising a dual-role discriminator to promote artistic creativity in generated art. But even though the CAN approach is successful at creating distinctive diverse art, it currently lacks thematic coherence. My goal is to close this gap by integrating the generating process with an artistic consistent theme, in terms of the content of the generated imagery.

### *C. Contrastive Language-Image Pretraining (CLIP)*

A key component of this research is text-to-image synthesis using OpenAI's CLIP (Contrastive Language-Image Pretraining) model [4], which aligns textual descriptions with images. CLIP accomplishes this by embedding both text and images into a shared latent space through contrastive learning, ensuring that related descriptions and images are positioned close together. Trained on 400 million (image, text) pairs, CLIP uses a ViT-like transformer for visual features and a causal language model for text, projecting both into a common space for direct comparison. The model calculates the dot product between image and text embeddings to measure their similarity, making it highly effective for tasks like image-text similarity. For instance, CLIPScore, a metric from Hessel et al [2]., uses CLIP to see how well captions match images by checking the similarity of their embeddings.

Integrating CLIP with the Creative Adversarial Network (CAN) will enhance its ability to produce images that align with specific textual prompts. CLIP will quantify how well the generated images match the text, addressing CAN's current lack of thematic coherence. This integration ensures that the images not only exhibit creativity but also adhere to the intended textual themes, thus improving both the relevance and artistic direction of the generated artworks.

### *D. Related work*

An inspiration for this work was the paper by Elgammal et al. [1], which advances GANs by generating art through understanding artistic styles and purposefully deviating from those norms to encourage creativity. In one of their experiments, human subjects rated the art generated by the CAN model higher than art created by humans. In their experiments, art generated by the CAN model was rated higher than human-created art, showcasing its innovative potential. This paper not only guided the approach of this study but also influenced Hall & Yaman [2**],** who employed interactive evolutionary approaches to navigate the latent space. The collaborative interactive evolution approach not only significantly preferred the evolved images over randomly generated ones but also produced creative, diverse artworks that frequently received higher satisfaction ratings from users, showcasing the effectiveness of CANs in generating novel art. The field of AI-generated art has advanced significantly due to these studies, but there is still a gap in integrating semantic prompts directly into the art generation process. Previous research has primarily focused on guiding the training process or using post-hoc evaluations. This research aims to fill that gap by embedding semantic prompts within the generator itself, enabling real-time guidance of the creative process.

Text-to-image generation has historically required specialized models and extensive training. Recent progress includes the VQGAN-CLIP approach developed by Crowson et al. [5], which integrates the Vector Quantized Generative Adversarial Network (VQGAN) [11] with the Contrastive Language-Image Pre-training (CLIP) model to create high-quality, semantically accurate images from text descriptions. This method surpasses previous models such as minDALL-E and GLIDE[12], with human evaluators rating it 4.6 out of 5, compared to 2.7 for minDALL-E and 3.3 for GLIDE. VQGAN-CLIP demonstrates superior performance by utilizing pretrained models enabling image generation with high fidelity. This study investigates the integration of semantic prompts into the art generation process to boost creativity and generate more dynamic outputs.

VQGAN-CLIP's success in using textual guidance to produce high-quality images and its efficient use of pretrained models offer valuable insights. Unlike VQGAN-CLIP, which excels in generating semantically accurate images, our work here focuses on integrating semantic prompts into the art generation process to enable more creative and dynamic outputs.

## V. METHOD

### *A. Datasets*

The primary dataset for this research is the WikiArt dataset, as introduced by Saleh et al. [13], which encompasses a wide range of historical artworks across various styles. The dataset comprises approximately 80,000 images classified into 27 distinct artistic styles and genres. Due to limitations in computational resources, a subset of 9,000 images, representing 9 different artistic styles, was selected for training. This smaller, yet diverse subset allows the Creative Adversarial Network (CAN) model to learn from a broad spectrum of artistic expressions, enhancing its capacity to generate unique and innovative outputs.

## *B. Model Architecture*

### 1. Generator network

The model architecture of the Creative Adversarial Network (CAN) used in this research consists of two primary components: the generator and the discriminator. The generator takes two inputs: a noise vector of size 100 and a semantic prompt embedding derived from CLIP, text as a 512-entry vector. These embeddings are processed to match the required input dimensions. The generator employs a series of transposed convolutional layers to up sample the combined input from the low-dimensional latent space to the desired image size, typically 3 x 64 x 64 for RGB images. Each transposed convolutional layer is followed by Batch Normalization and ReLU activation functions, ensuring stability and allowing the network to learn complex patterns. The final layer of the generator utilizes a Tanh activation function to scale the output values to the range of [-1, 1], suitable for image data. The CLIP-derived prompt embedding is projected using a linear layer to align with the input dimensions of the generator, allowing the generator to incorporate semantic information from the prompts into the generated images.

### 2. Discriminator Network

The discriminator evaluates the realism of the generated images and classifies their styles. It takes images as input and uses a series of convolutional layers to down sample the input, extracting hierarchical features. Each convolutional layer is followed by Leaky ReLU activations and Batch Normalization to stabilize the learning process. The discriminator has two output heads: one for binary classification (real vs. fake) using a Sigmoid activation function, and another for multi-class style classification, outputting a vector of class scores used with a softmax-based loss function. This dual-head approach ensures that the generated images are both realistic and stylistically diverse, advancing the capabilities of AI-generated art by leveraging the strengths of both transposed and standard convolutional layers, normalization, and activation functions.

### 3. Style Extractor

The VGG16 network, as originally developed by Simonyan et al.[14], is typically employed to extract basic style information from images, such as textures and patterns, using its early convolutional layers. In the CAN with CLIP model, however, VGG16 is repurposed to manage style deviations specifically.

During training, VGG16 is used to capture and average the stylistic features of a diverse set of reference styles drawn from nine distinct artistic styles in the WikiArt dataset. These reference styles represent various artistic genres and techniques, providing a comprehensive style profile. The generated images are then analysed by VGG16 to extract their stylistic features, which are compared to the precomputed reference style profile using a Mean Squared Error (MSE) loss function.

This loss function is modified to promote stylistic divergence by rewarding differences between the generated images and the reference styles. This adjustment encourages the generator to create images that exhibit creative variations and distinct stylistic properties, while still respecting the overarching artistic themes and goals.

### 4. Path Length Regularizer

In the CAN with CLIP model, the path length regularizer, like the approach used in StyleGAN2[18] , stabilizes the latent space mapping by analysing how random noise affects the generated images. Specifically, it looks at how introducing small, random changes in the latent space influences the gradients of the generated images. The regularizer measures the length of these gradients, ensuring that they remain consistent across different directions and magnitudes of noise. This helps in maintaining a stable relationship between the latent space and image space, preventing excessive or erratic changes in the generated images. As a result, the model can produce images that not only creatively deviate from the reference styles but also adhere to the intended artistic constraints, balancing creativity with reliability.

## *C. Training*

The training process optimized both the generator and discriminator networks for effective model performance. The dataset was accessed via Google Colab from Google Drive. Images were resized to 64x64 pixels with 3 RGB colour channels, and the batch size was 64. The noise vector size (nz) was 150, with feature map sizes of 64 (ngf) for the generator and 32 (ndf) for the discriminator. Training was performed over 250 epochs using the Adam optimizer.

To further enhance the networks during training, several advanced strategies were implemented. The Binary Cross-Entropy Loss (BCE) function was applied to both the generator and discriminator, effectively differentiating between real and generated images. Additionally, the discriminator's style classification was defined using Cross-Entropy Loss, ensuring the generated images accurately reflected the intended artistic style. A crucial component of the training was the CLIP-based prompt adherence loss, which measured cosine similarity between the generated image's CLIP embedding and that of the prompt, ensuring alignment with textual descriptions. Path Length Regularization was also employed to maintain smoothness in the latent space and encourage diversity in the generated outputs. Several learning rates were tested; increasing them caused divergence, leading to the final choice of learning rates 0.00022 and 0.00023 for the generator and discriminator, respectively. The AdamW optimizer was used for its effective weight decay handling, with β parameters set to (0.5, 0.999) for stable convergence.

## VI. Experiments

The objective of this study was to investigate the efficacy of integrating semantic prompts into Creative Adversarial Networks (CAN) to achieve thematic consistency in generated artwork. The experiments were designed to explore the effectiveness of various models, tracking the progression from a basic GAN with CLIP to a more sophisticated CAN with CLIP. For a consistent basis of comparison, the prompt "A portrait of a woman with long black hair" was employed across the initial three models.

The experimental models in this study are detailed as follows:

- **Basic GAN with CLIP:** This model serves as a baseline to evaluate the fundamental integration of GANs with semantic prompts.
- **Basic CAN with CLIP:** Employs a Creative Adversarial Network with CLIP integration, using the same prompt to assess the effectiveness of creative image generation.
- **Advanced CAN with CLIP:** Builds on the basic CAN model by incorporating VGG16 as a style extractor and a path length regularizer. This model uses the same prompt as the previous ones to demonstrate the improvements made through the addition components.
- **CAN with Additional Prompts:** Extends the advanced CAN model by testing with additional prompts such as "flowers" and "mountains," to evaluate its ability to maintain thematic coherence and artistic quality across diverse themes.

### A. Performance Metrics

The model's performance is quantitatively evaluated using three key metrics: generator and discriminator losses, similarity score, and visual representations. The losses of both the generator and discriminator are tracked over training iterations and plotted to monitor convergence. The similarity score, computed as the cosine similarity between generated images and textual prompts, provides insight into how well the images align with the given prompts, with higher scores indicating better adherence. Additionally, visual inspections of the generated art are conducted to assess stylistic accuracy and prompt adherence, offering a qualitative assessment of how closely the artwork matches the desired style and textual descriptions. For each model, a set of five representative images will be shown. The images were selected to illustrate the range of outputs produced by the models.

## VII. Results

### A. Model 1: Basic GAN with CLIP


*[ 1 ]0.211*


*[ 2 ] 0.2115*


*[ 3 ] 0.2300*


*[ 4 ]0.2003*


*[ 5 ] 0.2493*

The images with higher CLIP similarity scores closely align with the prompt, “A portrait of a woman with long black hair," particularly the third image, which best captures the described features.

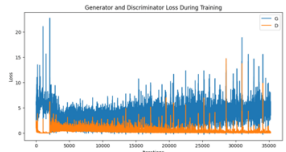

*Figure 1: Generator and discriminator loss during training in Model 1*

*Figure 2: Average similarity score trend for Model 1*

During training, the generator's loss (Loss_G) began high because the model initially struggled to produce realistic images. Over time, this loss decreased as the generator learned to create better images. However, this decrease was not uniform; it fluctuates which is typical behaviour for generator as it refines the output.

The discriminator’s loss (Loss_D) was more stable, fluctuating between 0 and about 0.5 with occasional spikes. These variations suggest that the discriminator's ability to distinguish between real and generated images varied throughout the training process. As the generator improved and started producing more convincing images, Loss_D occasionally increased, reflecting the increased difficulty for the discriminator. The interplay between these losses highlights a balanced training dynamic, where the generator and discriminator continuously adjusted to each other’s advancements.

This early increase in average similarity scores might not directly reflect improved alignment but rather a transition from random, noisy outputs to more structured and recognizable images as the generator begins to learn. As training progresses, the generator stabilizes and refines its output, which could explain the gradual decrease in similarity scores after reaching a peak.

### B. Model 2: ***Basic CAN with CLIP***


*[ 6 ]0.2030*


*[ 7 ]0.2583*


*[ 8 ]0.2505*


*[ 9 ] 0.2514*


*[ 10 ] 0.2069*

In analysing the generated images, several notable variations emerged. For instance, Picture 7 closely resembles a black-and-white version of Picture 8, while Pictures 6 and 9 appear to be variations of each other. Despite each image having unique characteristics, many of them seem to be slight modifications or variations rather than entirely distinct creations. This pattern suggests that while the generator produces a range of outputs, the images share common features and themes, reflecting a consistent underlying style or approach in the generated content.

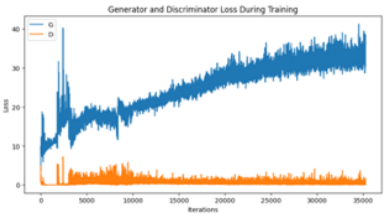


*Figure 3: Generator and discriminator loss during training in Model 2*

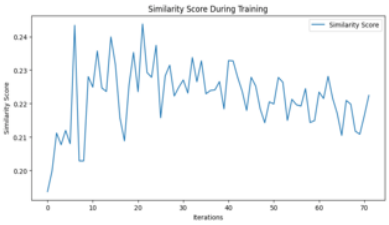


*Figure 4: Average similarity score trends for model 2*

Initially, both losses start at similar levels, but the generator's loss tends to oscillate more dramatically and generally increases over time. This behaviour highlights the generator's difficulty in aligning images with CLIP prompts while maintaining quality and style deviation. In contrast, the discriminator's loss remains relatively stable, oscillating between 0 and 0.5, reflecting its more consistent task of distinguishing between real and generated images, similar to the basic GAN-CLIP.

The average similarity graph for the basic CAN CLIP follows a trend similar to that of the basic GAN-CLIP, with a peak slightly higher at 0.24, indicating a modest improvement in average similarity.

## C. Model 3: ***Advance CAN with CLIP***


*[ 11 ] 0.2185*


*[ 12 ] 0.2388*


*[ 13 ] 0.2258*


*[ 14 ] 0.2306*


*[ 15 ] 0.2222*

The generated images are more abstract and diverse, with a noticeable variation between them. Although the CLIP scores are lower, the results are still impressive, demonstrating a wide range of creative outputs.

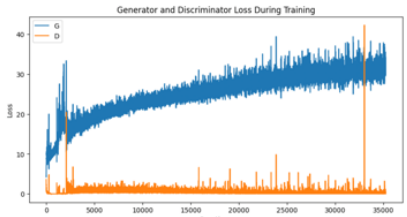


*Figure 5: Generator and discriminator loss during training in Model 3*


*Figure 6: Average similarity score trends for Model 3*

Both the loss graphs and similarity scores for the enhanced CAN CLIP model closely resemble those of the basic CAN CLIP model. This indicates that the improvements made primarily affected the quality of the generated images rather than altering the overall training dynamics. Notably, there was a significant spike in the loss for the discriminator near the end, which stands out as an outlier.

## D. *Advance CAN with Additional Prompts*

To assess how different prompts influence image generation, Model 3 was tested with alternative prompts. The first prompt chosen was "A bouquet of flowers."


*[ 16 ] 0.2039*


*[ 17 ] 0.2062*


*[ 18 ] 0.1899*


*[ 19 ] 0.2190*


*[ 20 ] 0.2140*


*Figure 7: Training Losses for Model 3 with "A bouquet of flowers" Prompt*


*Figure 8: Average Similarity Score Trends for Model 3 with "A bouquet of flowers" Prompt*

The second prompt, "A landscape with mountains," yielded the following results:

*[ 21 ] 0.2556*

*[ 22 ] 0.2983*

*[ 23 ] 0.2548*

*[ 24 ] 0.2375*

*[ 25 ] 0.2763*

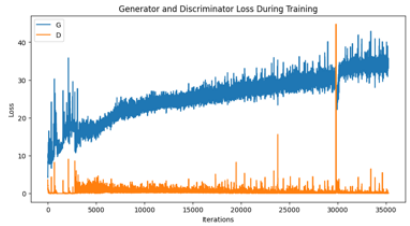


*Figure 9: Training Losses for Model 3 with "A landscape with mountains" Prompt*

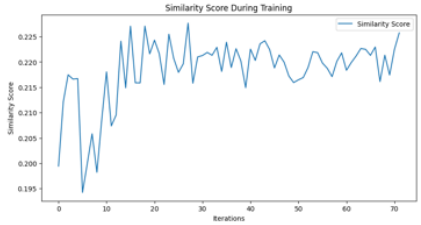


*Figure 10: Average Similarity Score Trends for Model 3 with "A landscape with mountains" Prompt*

Overall, the loss graphs for both the generator and discriminator closely resemble each other, including the spike where the discriminator's loss temporarily reaches the generator's loss level. However, the average similarity score graphs differ, indicating variations in how well the model aligns with the prompts. The results suggest that the model performs better at generating images for the "landscape with mountains" prompt, as evidenced by the higher CLIP scores, compared to the "a bouquet of flowers" prompt.

## VIII. Discussion

These results demonstrate the success of the experiment, revealing that the generated images are not only creative but also uniquely independent of existing styles. This observation is based on visual assessment rather than a formal measurement of style adherence, as no specific metrics were used to quantify stylistic adherence. Nonetheless, the CLIP scores clearly indicate that the images faithfully followed the prompts, producing art with a distinct thematic coherence and creative expression. The generated artwork does not merely replicate existing styles but forges its own unique form, which is a noteworthy achievement.

The integration of CLIP was central to this study, and the effectiveness of this approach is evident from the CLIP scores, with Image 22 demonstrating the highest alignment with the prompt. The observed trends in CLIP scores varied depending on the prompt, as illustrated in Figures 2, 4, and 6, where similar trends are seen, though exact values differ. Prompts with well-defined features, such as "mountains," resulted in images with more recognizable attributes and higher CLIP scores. In contrast, more abstract prompts produced images that were less clearly aligned with the prompt, leading to lower scores. This variation highlights how well the model adapts to different textual descriptions and interprets the provided information.

Both Model 2 (Figure 4) and Model 3 (Figure 6) exhibited initial CLIP score peaks at 0.24, followed by a decrease to 0.21. Model 2 showed greater fluctuations in scores, while Model 3, incorporating VGG16 and path length regularization, displayed more stability despite the overall decrease. The additional components in Model 3 likely contributed to this stability. However, the lower CLIP scores in Model 3 may also be attributed to the specific prompt rather than the added components. The Advanced CAN model, although generating more visually diverse and abstract images, exhibited a trade-off between creativity and strict adherence to prompts. This is reflected in the lower CLIP scores and higher generator loss.

The increase in generator loss in the Advanced CAN model, as shown in Figure 3 compared to Figure 1, can be attributed to the model's incorporation of style deviation mechanisms. By prioritizing creative variations outputs, the generator produced images that diverged more from the styles, leading to higher loss values. The difference between the loss graphs highlights the success of this approach, showing the distinct behaviours of GAN and CAN models. In a typical GAN, generator and discriminator losses are balanced, reflecting an equilibrium where the generator produces convincing images, and the discriminator effectively distinguishes real from fake. In contrast, the CAN's generator loss increases as it emphasizes creativity, indicating the model's focus on producing more innovative outputs.

The addition of VGG16 and path length regularization introduced added complexity to the training process, as seen in Figure 5, where a noticeable surge in discriminator loss appears toward the end of training. Initially, I considered this surge to be an anomaly; however, after reviewing the loss graphs in Figures 7 and 9, it became clear that this pattern consistently emerged after incorporating these additional components into the CAN model. This spike likely reflects the discriminator's challenge in adapting to the generator's evolving outputs under the new constraints. Although this occurrence does not significantly alter the overall loss trends, it underscores an area that warrants further investigation. The visual differences in the images generated by Model 3 indicate that these components do affect the model's creative output, though the precise impact on training dynamics remains to be fully understood.

The findings reveal a notable trade-off between creativity and prompt adherence. The Advanced CAN model, while generating more abstract and varied outputs, exhibited lower CLIP scores and higher generator loss, reflecting a compromise between artistic freedom and thematic accuracy. This trade-off illustrates the challenge of balancing innovation with strict alignment to prompts, emphasizing the need for ongoing refinement in future models to better manage these competing objectives.

## IX. Limitation and Further Work

This project successfully integrated the CLIP model with a Creative Adversarial Network (CAN) to achieve its primary objectives of generating artistic images. However, several areas could benefit from further improvement. One notable limitation was the partial use of the dataset due to computing resource constraints. Utilizing the entire dataset might enhance the creative quality of the generated art by introducing greater diversity and richness into the training process. Additionally, the architecture employed in this study may be considered simple. Future research could improve CLIP integration with Creative Adversarial Networks (CANs) by using residual blocks, which may enhance image quality and alignment with CLIP embeddings.

Additionally, the images presented in the Results section were handpicked based on their CLIP scores, how well they resembled the prompts, and my subjective preference. The remaining images generated during the experiment, which may offer additional insights, are included in Appendix Figures 12-15. While many of the generated images exhibited creativity, they often lacked strict adherence to the prompts. This is a known challenge with GANs, as they do not always generate every image with precise alignment to the intended goal. Future work could focus on improving prompt adherence to ensure that each generated image more consistently aligns with the desired outcomes.

While additional components like VGG16 and path length regularizer were added, their individual impacts on image generation are unclear. An ablation study to evaluate each component's effect and fine-tuning of parameters would help optimize performance.

The study also lacked a quantitative metric to objectively assess the deviation of generated images from the intended style. Although visual differences were evident, incorporating objective measures in future work could improve the evaluation of style adherence.

To address the trade-off between prompt adherence and promoting creativity through style deviation, consider separating the processes: one model could focus on generating content that strictly follows the prompts, while another model could apply stylistic deviations to this content. Alternatively, experimenting with different approaches to implementing CLIP embeddings might also help achieve a better balance. Essentially, various strategies can be explored to bridge this gap effectively.

## Acknowledgment

I would like to thank my lecturer, Simon Colton, for his helpful guidance and support in understanding key aspects of this work. His insights were valuable in refining the project.

## References

[1] Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M. CAN: Creative Adversarial Networks Generating "Art" by Learning About Styles and Deviating from Style Norms. arXiv:1706.07068.

[2] Hall, O., & Yaman, A. Collaborative Interactive Evolution of Art in the Latent Space of Deep Generative Models. Vrije Universiteit Amsterdam.

[3] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, 27, 2672–2680.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *arXiv:2103.00020*, 2021.

[5] Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L. and Raff, E., 2021. *VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance*. [online] Available at: https://arxiv.org/abs/2111.05267

[6] Mordvintsev, A., Olah, C., & Tyka, M. (2015, June 18). Inceptionism: Going deeper into neural networks. Google AI Blog. https://research.google/blog/inceptionism-going-deeper-into-neural-networks/

[7] Karl Sims. 1991. Artificial evolution for computer graphics. SIGGRAPH Comput. Graph. 25, 4 (July 1991), 319–328. https://doi.org/10.1145/127719.122752 casualy meantion as a ai generated art

[8] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17). JMLR.org, 214–223.

[9] Karras, T., Laine, S. and Aila, T., 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401-4410).

[10] Hessel, J., Holtzman, A., Forbes, M., Bras, R.L. and Choi, Y. (2022). CLIPScore: A Reference-free Evaluation Metric for Image

Captioning. *arXiv:2104.08718 [cs]*. [online] Available at: https://arxiv.org/abs/2104.08718.

[11] Esser, P., Rombach, R., and Ommer, B.: Taming Transformers for High-Resolution Image Synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vi- sion and Pattern Recognition, pp. 12873–12883 (2021)

[12] Nichol,A.,Dhariwal,P.,Ramesh,A.,Shyam,P.,Mishkin,P.,McGrew,B., Sutskever, I., and Chen, M.: GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, (2021). arXiv: 2112.10741v3 [cs.CV]

[13] B. Saleh and A. Elgammal, "Large-scale classification of fine-art paintings: Learning the right metric on the right feature," *arXiv preprint arXiv:1505.00855*, 2015.

[14] Simonyan, K. and Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. [online] arXiv.org. Available at: https://arxiv.org/abs/1409.1556.

[15] Wang, Y., Guo, X., Liu, P. and Wei, B. (2021). Up and Down Residual Blocks for Convolutional Generative Adversarial Networks. *IEEE Access*,9,pp.26051–26058. doi:https://doi.org/10.1109/access.2021.3056572.

[16] Tepencelik, O., Ocak, I., Lu, D. and Liu, S., *Creating Art with Various GAN Architectures*. GitHub. Available at: https://github.com/otepencelik/GAN-Artwork-Generation/tree/master

[17] Magazine, S. and Cengel, K. (2024). *The First A.I.-Generated Art Dates Back to the 1970s*. [online] Smithsonian Magazine. Available at: https://www.smithsonianmag.com/innovation/first-ai-generated-art-dates-back-to-1970s-180983700/.

[18] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J. and Aila, T. (2020). *Analyzing and Improving the Image Quality of StyleGAN*. [online] IEEE Xplore. doi:https://doi.org/10.1109/CVPR42600.2020.00813.

# Appendix A: Generated images

*Figure 11: Generated Images from the Final Epoch of Model 1 (Basic GAN with CLIP Integration). The prompt used was "A portrait of a woman with long black hair".*

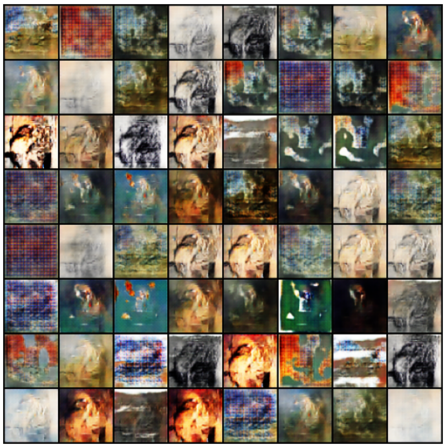

*Figure 12 Generated Images from the Final Epoch of Model 2 (Basic CAN with CLIP Integration). The prompt used was "A portrait of a woman with long black hair".*

*Figure 13: Generated Images from the Final Epoch of Model 3 (Advance CAN with CLIP Integration). The prompt used was "A portrait of a woman with long black hair".*

*Figure 14: Generated Images from Epoch 45 and Final Epoch of Model 3 (Advanced CAN with CLIP Integration). The prompt used was 'A bouquet of flowers'.*

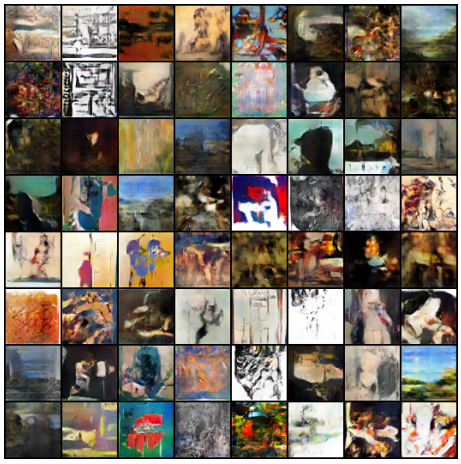

*Figure 15 :Generated Images Final Epoch of Model 3 (Advanced CAN with CLIP Integration). The prompt used was 'A landscape with mountains'.*