BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF INTELLIGENT SYSTEMS
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

DIFFUSION MODELS AND THEIR IMPACT ON CYBERSECURITY

DIFFUSION MODELY A ICH DOPAD NA POČÍTAČOVÚ BEZPEČNOSŤ

BACHELOR'S THESIS
BAKALÁŘSKÁ PRÁCE

AUTHOR **PATRIK DVORŠČÁK**
AUTOR PRÁCE

SUPERVISOR **Ing. TOMÁŠ LAPŠANSKÝ**
VEDOUCÍ PRÁCE

BRNO 2024

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Bachelor's Thesis Assignment

155153

Institut:	Department of Intelligent Systems (DITS)
Student:	**Dvorščák Patrik**
Programme:	Information Technology
Title:	**Diffusion Models and their Impact on Cybersecurity**
Category:	Security
Academic year:	2023/24

Assignment:

1. Learn about the technology of diffusion models and the tools that use them for image and video synthesis.
2. Identify the categories of images and videos (face synthesis, face swap, etc.) that these models produce, and for each category, find and become familiar with at least one tool that uses them for their creation.
3. For all the identified categories, find the tools that use generative adversarial networks (or other) technologies and select at least one of them to use.
4. Design experiments to test the difficulty of recognizing media generated by diffusion models compared to models based on previous technologies.
5. Perform the designed experiments for each identified category of images and videos.
6. Evaluate the potential security implications of diffusion models.

Literature:

- Koo, H., & Kim, T. E. (2023). A Comprehensive Survey on Generative Diffusion Models for Structured Data. *ArXiv*. /abs/2306.04139
- R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano and L. Verdoliva, "On The Detection of Synthetic Images Generated by Diffusion Models," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10095167.
- Kim, G., Kwon, T., & Ye, J. (2022). DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2426-2435).
- Nichol, A., & Dhariwal, P. (2021). Improved Denoising Diffusion Probabilistic Models. In Proceedings of the 38th International Conference on Machine Learning (pp. 8162–8171). PMLR.

Requirements for the semestral defence:
Items 1. to 4. of the assignment.

Detailed formal requirements can be found at https://www.fit.vut.cz/study/theses/

Supervisor:	**Lapšanský Tomáš, Ing.**
Head of Department:	Hanáček Petr, doc. Dr. Ing.
Beginning of work:	1.11.2023
Submission deadline:	9.5.2024
Approval date:	6.11.2023

Abstract

This thesis explores the performance of diffusion models (DMs) and generative adversarial networks (GANs) in creating AI-generated visual content across multiple applications, including face synthesis, text-to-image generation, artistic rendering, image-to-image translation, video synthesis, and super-resolution. Through comparative experiments, this research evaluates the models' ability to generate detailed, realistic, and artistically compelling visuals from textual and image prompts.

The results reveal that DMs excel in producing highly detailed images that closely follow text prompts, particularly effective in face synthesis and text-to-image tasks. In contrast, GANs are more adept at rendering realistic environmental scenes, suitable for applications requiring immersive visuals. Both model types are competent in artistic rendering, though they differ in style adaptation and creativity.

The thesis concludes with future research directions aimed at enhancing model efficacy and integrating these technologies more effectively into practical applications.

Abstrakt

Táto práca skúma výkonnosť difúznych modelov (DM) a Generative Adversarial Network (GAN) - Generatívna sieť súperiacích komponentov, pri vytváraní vizuálneho obsahu generovaného umelou inteligenciou vo viacerých aplikáciách vrátane syntézy tváre, generovania textu na obraz, umeleckého renderovania, prekladu obrazu na obraz, syntézy videa a superrozlíšenia. Prostredníctvom porovnávacích experimentov sa v tomto výskume hodnotí schopnosť modelov generovať podrobné, realistické a umelecky presvedčivé vizuály z textových a obrazových vstupov.

Výsledky ukazujú, že DM vynikajú pri vytváraní vysoko detailných obrazov, ktoré presne nasledujú textové vstupy, pričom sú obzvlášť účinné pri úlohách syntézy tváre a prevodu textu na obraz. Naproti tomu GAN sú zručnejšie pri vykresľovaní realistických scén prostredia, ktoré sú vhodné pre aplikácie vyžadujúce pohlcujúce vizuály. Oba typy modelov sú kompetentné v umeleckom vykresľovaní, hoci sa líšia v prispôsobovaní štýlu a kreativite.

V závere práce sú uvedené budúce smery výskumu zamerané na zvýšenie účinnosti modelov a efektívnejšiu integráciu týchto technológií do praktických aplikácií.

Keywords

GAN, DDPM, DDIM, Diffusion Models, Categories, Face Synthesis, Video Synthesis, Text-to-Image, Image-to-Image, Artistic Rendering, Super-Resolution, Upscaling, Generative Artificial Intelligence, Photorealism, Digital Content Generation, Cybersecurity, Content Authenticity, Structured Data Synthesis

Kľúčové slová

GAN, DDPM, DDIM, difúzne modely, kategórie, syntéza tváre, syntéza videa, text-na-obraz, obraz-na-obraz, umelecké vykreslenie, superrozlišenie, zväčšenie rozlíšenia, ESRGAN, generatívna umelecká inteligencia, fotorealizmus, generovanie digitálneho obsahu, kybernetická bezpečnosť, autenticita obsahu, syntéza štruktúrovaných údajov

Reference

DVORŠČÁK, Patrik. *Diffusion models and their impact on cybersecurity.* Brno, 2024. Bachelor's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Ing. Tomáš Lapšanský

Rozšírený abstrakt

Táto práca poskytuje komplexnú analýzu nedávneho pokroku v oblasti umelej inteligencie (AI) so zameraním na úlohu difúznych modelov (DM) ako výkonných nástrojov pri vytváraní digitálneho vizuálneho obsahu. Na rozdiel od tradičných generatívnych modelov, ako sú generatívne siete súperiacich komponentov (GAN), ktoré začínajú od náhodného rozdelenia šumu na vytvorenie komplexných údajov, difúzne modely začínajú od samotných údajov a postupne zavádzajú šum, aby tento proces obrátili. Tento reverzný proces, známy ako proces reverznej difúzie, je dôkladne preskúmaný s cieľom pochopiť jeho použitie pri syntéze rôznorodého a vysokokvalitného vizuálneho obsahu v rôznych oblastiach.

V úvodných kapitolách práce sú predstavené základné princípy difúznych modelov s osobitným dôrazom na pravdepodobnostné difúzne modely (DDPM) a deterministické difúzne implicitné modely (DDIM). Tieto modely sa vyznačujú jedinečným prístupom k spracovaniu a transformácii údajov, ktorý umožňuje generovať obrázky, videá a iné formy štruktúrovaných údajov s pozoruhodnou presnosťou a vizuálnou kvalitou. Teoretické skúmanie zahŕňa zhrnutie o procesoch pridávania šumu a iteračného spresňovania, ktoré charakterizujú difúzne modely, čím sa pripravuje pôda pre ich praktické aplikácie.

Experimentálne analýzy tvoria hlavnú zložku tohto výskumu, v rámci ktorého sa difúzne modely testujú v porovnaní s GAN, pokiaľ ide o ich schopnosť generovať syntetické médiá. Experimenty sú starostlivo navrhnuté tak, aby pokrývali celý rad aplikácií od syntézy tvárí a umeleckého vykresľovania až po syntézu videa. Pri syntéze tvárí sa difúzne modely ukázali ako lepšie pri preklade textových opisov na fotorealistické ľudské tváre, pričom preukázali vyššiu vernosť nuansám opísaným vo vstupných textoch. Táto presnosť v detailoch a realizme sa zdôrazňuje najmä prostredníctvom používateľských štúdií, v ktorých boli vygenerované obrazy konzistentne hodnotené vyššie ako obrazy vytvorené tradičnými GAN.

Umelecké vykresľovanie je ďalšou oblasťou, v ktorej difúzne modely vykazujú významné sľuby. V práci sú podrobne opísané experimenty, v ktorých tieto modely úspešne zachytávajú a replikujú rôzne umelecké štýly, od klasických až po súčasné, pričom sa presne držia štylistických prvkov uvedených v textových vstupoch. Táto schopnosť nielenže zdôrazňuje všestrannosť difúznych modelov v kreatívnych aplikáciách, ale poukazuje aj na ich potenciál revolučne zmeniť odvetvia, ako je digitálny marketing a tvorba obsahu vo virtuálnej realite.

Syntéza videa sa kriticky skúma ako vyvíjajúca sa aplikácia difúznych modelov. Hoci tieto modely vykazujú potenciál pri generovaní krátkych videoklipov zo statických obrázkov a textových opisov, výskum identifikuje obmedzenia v plynulosti a časovej nadväznosti generovaného obsahu. Tento poznatok poukazuje na potrebu ďalšieho zdokonaľovania algoritmov modelov s cieľom zlepšiť prirodzený tok a realizmus syntetizovaných videí.

Práca obsahuje aj zamyslenie z pohľadu etiky, ktoré poukazuje na možné zneužitie difúznych modelov pri vytváraní klamlivých alebo škodlivých médií, ako sú napríklad deepfakes. Schopnosť týchto modelov generovať fotorealistické ľudské tváre a realistické prostredia možno využiť na vytvorenie vizuálne presvedčivých, ale úplne fiktívnych scenárov, čo vyvoláva značné obavy zo šírenia dezinformácií a porušovania súkromia. Na riešenie týchto problémov sa v práci navrhuje viacstranný prístup na zmiernenie rizík spojených s nasadením generatívnych technológií umelej inteligencie. Zahŕňa to technologické riešenia, ako je implementácia sofistikovaných techník digitálneho vodoznaku na zabezpečenie sledovateľnosti a autenticity obsahu generovaného umelou inteligenciou. Práca sa zaoberá aj o regulačných opatreniach, pričom sa obhajuje vytvorenie spoľahlivých právnych rámcov na reguláciu používania a uplatňovania technológií umelej inteligencie pri tvorbe obsahu, čím sa zabezpečí dodržiavanie etických noriem. Okrem toho sa odporúčajú vzdelávacie iniciatívy na zvýšenie informovanosti verejnosti o možnostiach a potenciálnom zneužití technológií

umelej inteligencie. Cieľom týchto iniciatív je kultivovať informovanejšiu verejnosť, ktorá dokáže kriticky posúdiť a pochopiť obsah vytvorený umelou inteligenciou, čím sa podporí etickejšie uvedomelá komunita používateľov a vývojárov.

V závere práce sa rozpracúvajú nielen technické a praktické silné stránky difúznych modelov, ale rozoberajú sa aj spoločenské dôsledky ich používania. Predstavuje vyvážený pohľad a navrhuje budúce smery výskumu, ktoré sa zameriavajú na zvýšenie výkonnosti týchto modelov, najmä pri syntéze videa, a na vývoj integrovaných riešení, ktoré spájajú silné stránky difúznych modelov a GAN. Takéto hybridné prístupy by potenciálne mohli vyriešiť súčasné obmedzenia a otvoriť nové možnosti uplatnenia umelej inteligencie pri tvorbe digitálneho obsahu.

Diffusion models and their impact on cybersecurity

Declaration

I hereby declare that this Bachelor's thesis was prepared as an original work by the author under the supervision of Mr. Tomáš Lapšanský I have listed all the literary sources, publications and other sources, which were used during the preparation of this thesis.

.....
Patrik Dvorščák
May 8, 2024

Acknowledgements

I would like to extend my gratitude to Ing. Tomáš Lapšanský for his guidance throughout the development of this thesis.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

The field of artificial intelligence (AI) has seen remarkable advancements in the synthesis of images and videos driven by innovative generative models. Among these, diffusion models have emerged as a promising approach for generating high-quality and diverse visual content. Diffusion models, characterized by their ability to model complex distributions, have found applications in various domains, from face synthesis to artistic rendering.

At the core of diffusion models lies the concept of the reverse diffusion process, a distinctive mechanism that distinguishes them from other generative approaches. Unlike traditional generative models that generate samples from a simple distribution and transform them into complex data, diffusion models operate in the opposite direction. The reverse diffusion process starts with a complex data point and progressively transforms it into simpler distributions through a series of steps.

1.2 Research Objectives

This thesis aims to explore and analyze diffusion models' capabilities in synthesizing images and videos, with a particular focus on categories such as face synthesis, style transfer, Image-to-Image and Text-to-Image translation applications. The overarching goals can be summarized as follows:

1. Get acquainted with diffusion model technology and the tools that utilize them to synthesize images and videos.
2. Identify categories of images and videos (such as face synthesis, faceswap, etc.) produced by these models, and for each category, find and familiarize yourself with at least one tool that uses them for creation.
3. For all identified categories, find tools that use generative adversarial network technologies (or others) and choose at least one to operationalize.
4. Design experiments to verify the difficulty of recognizing media generated by diffusion models compared to models based on previous technologies.
5. Carry out the proposed experiments for each identified category of images and videos.
6. Evaluate the potential security impacts of diffusion models.

Through this research, we aim to contribute to a deeper understanding of the capabilities, challenges, and security considerations of diffusion models, shedding light on their potential applications and impact on the field of generative artificial intelligence.

Chapter 2

Diffusion models

Diffusion models have emerged as a groundbreaking approach in the field of deep generative models, demonstrating remarkable capabilities in the synthesis of structured data, including tabular and time series data. This paradigm shift is notably discussed in the comprehensive survey by Koo and Kim (2023), which highlights the advancements and potential of diffusion models in handling various forms of structured data, a domain that had received less attention compared to visual and textual data in deep learning research [16].

At their core, diffusion models work by gradually adding noise to an image or data point and then learning to reverse this process. This reverse diffusion process essentially involves starting with a noise-corrupted version of the data and iteratively denoising it to recover the original data. This process is characterized by its ability to model complex distributions, making it highly effective in generating high-quality and diverse visual content.

2.1 Denoising Diffusion Probabilistic Models (DDPMs)

Denoising Diffusion Probabilistic Models (DDPMs) are a key development in this area. Introduced by Nichol and Dhariwal (2021), DDPMs demonstrate how diffusion models can achieve competitive log-likelihoods while maintaining high sample quality with a few modifications. Their work reveals that learning variances of the reverse diffusion process allows for sampling with significantly fewer forward passes, enhancing practicality and efficiency [18].

2.1.1 Forward Diffusion Process

In a diffusion model's forward process, a sample from the data distribution, $\mathbf{x}_0$, is incrementally transformed by adding Gaussian noise over a sequence of steps, resulting in a progression $\mathbf{x}_1, \ldots, \mathbf{x}_T$. The equation formalizes this process:

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

where β_t denotes the variance schedule. As t increases, the original sample $\mathbf{x}_0$ becomes less distinct, approaching a pure noise distribution when t approaches infinity.

2.1.2 Reverse Diffusion Process

The reverse diffusion process in diffusion models is a pivotal mechanism where the model learns to reverse the diffusion process, transforming a simple noise distribution back into the

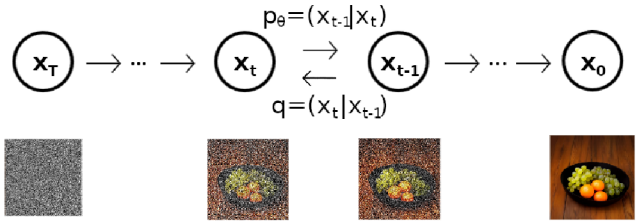

Figure 2.1: A sequential representation (Markov chain) of the forward (reverse) diffusion process, where a data sample undergoes a gradual transformation by incrementally introducing (and subsequently subtracting) noise to synthesize a final output.

complex original data distribution. This process is integral in generative diffusion models, especially when dealing with structured data [16].

Mathematically, the reverse diffusion process can be described by a sequence of conditional distributions, with the aim of reconstructing the original data point x from the latent variable z. The process is often represented as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

Here, x_{t-1} and x_t are the data points at steps $t-1$ and t in the reverse process, respectively. The function $\mu_\theta(x_t, t)$ defines the mean, and $\Sigma_\theta(x_t, t)$ defines the covariance, both parameterized by θ, the parameters of the model. The goal is to reach x_0, the original data point, starting from a distribution of noise at step T.

In the context of structured data, as highlighted by Koo and Kim (2023), this reverse process enables the generation of data that retains the complex patterns and relationships inherent in the original dataset, making it a powerful tool for data synthesis and analysis [16].

2.2 Deterministic Diffusion Implicit Models (DDIMs)

Deterministic Diffusion Implicit Models (DDIMs) present a variant of the conventional diffusion models, characterized by their deterministic nature. Unlike stochastic diffusion models, DDIMs follow a deterministic path in the data generation process, offering a more predictable and controlled approach to image synthesis [20].

Operational Principle

DDIMs operate by defining a deterministic sequence of latent variables, significantly reducing the randomness in the generation process. The core idea is to utilize an iterative procedure, where each step is a deterministic function of the previous step's output and the model's parameters. This approach simplifies the generative process and allows for more direct control over the generated output.

Advantages

One of the primary advantages of DDIMs is their ability to produce consistent and stable outputs, which is particularly important in applications requiring high precision. Further-

more, the deterministic nature of DDIMs simplifies the training process and can lead to faster convergence compared to stochastic models. Their predictable behavior also makes them suitable for tasks where reproducibility is key.

Applications

DDIMs have shown promise in various applications, especially in fields that benefit from controlled and predictable image synthesis. These include medical imaging, where consistent results are crucial, and artistic endeavors where specific outcomes are desired. The deterministic aspect of DDIMs opens up new possibilities in these areas, allowing for tailored and precise image generation.

GPT Certainly! Here's a concise section comparing DDIM and DDPM, using the already cited sources:

2.3 Comparison between DDIM and DDPM

Deterministic Diffusion Implicit Models (DDIMs) and Denoising Diffusion Probabilistic Models (DDPMs) are powerful tools in generative modeling. While DDPMs, as discussed by Nichol and Dhariwal (2021), leverage stochastic processes for sample generation, ensuring high sample quality and diversity, DDIMs offer a deterministic approach, providing consistency and predictability in outputs, as outlined by Song and Ermon (2021). This fundamental difference in stochasticity versus determinism defines their respective applications and advantages in various generative tasks [[18]; [20]].

2.4 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) represent a novel approach in deep learning, especially significant in the realm of computer vision. Introduced by [12], GANs comprise two neural networks, the generator and the discriminator, engaged in a unique adversarial training methodology. This structure enables GANs to generate highly authentic data for various tasks like image generation, style transfer, etc.

2.4.1 Adversarial Training Methodology

At the core of GANs lies the adversarial training process, wherein the generator network creates samples and the discriminator evaluates them. The generator's objective is to fabricate indistinguishable data from genuine data, while the discriminator aims to differentiate between real and synthetic data. This process is akin to a game between the two networks, with the generator constantly refining its data generation tactics to deceive the discriminator.

2.4.2 Challenges and Progress

Despite their efficacy, GANs encounter challenges like model collapse and unstable training. Recent advancements address these issues by innovating new training methodologies and architectures. For instance, the BasisGAN introduces a technique to stochastically generate basis elements for convolutional filters, effectively reducing the complexity of modeling the parameter space without compromising on image diversity or fidelity [23].

2.4.3 Computer Vision Applications

GANs have profoundly impacted computer vision, offering advancements in data enhancement, domain transfer, high-quality sample generation, and image restoration. They have redefined methods for feature learning and image generation, outperforming conventional machine learning algorithms in several aspects [14].

2.4.4 Comparison with Diffusion Models

While GANs and diffusion models share the goal of generating realistic data, they differ significantly in their approach. GANs generate samples by optimizing a generator network to produce indistinguishable data from real data according to a discriminator. In contrast, diffusion models leverage a reverse process to unfold complex data progressively.

Diffusion models have been shown to excel in capturing long-range dependencies and maintaining diversity, offering more stability during training. This distinction is crucial in understanding their applications in cybersecurity, where the detection of synthetic media, as explored by Corvi et al. (2023), becomes a significant challenge [11]. Koo and Kim (2023) also highlight the broader applications of diffusion models in structured data, underscoring their versatility compared to GANs [16].

Theoretical comparisons between GANs and diffusion models have highlighted the advantages of diffusion models in capturing long-range dependencies, preserving diversity, and providing better stability during training. However, both approaches have unique strengths and weaknesses, making them suitable for different applications.

In the subsequent chapters, we delve into practical implementations and experiments to explore and validate the theoretical foundations outlined in this chapter further.

Chapter 3

Categorization of AI-Generated Visual Content

3.1 Introduction

The evolution of generative models has ushered in a new era of artificial intelligence, with the synthesis of visually compelling content at its forefront. Amidst this burgeoning landscape, the categorization of AI-generated visual content emerges as a crucial undertaking, particularly with the spotlight on diffusion models. This chapter delves into the nuanced task of categorizing content generated by diffusion models, elucidating their applications, challenges, and the underlying mechanisms that define their unique contribution to visual creativity.

3.1.1 Background and Significance

The landscape of AI-generated visual content has been reshaped by the advent of diffusion models, offering a distinctive approach to image and video synthesis. From intricate face synthesis to complex style transfer, diffusion models have demonstrated a capacity to capture intricate patterns and dependencies within the data. Categorizing the outcomes of these models becomes pivotal, unlocking a deeper understanding of their applications and implications.

The significance of categorization extends beyond organizational principles; it serves as a conduit for unraveling the intricacies of diffusion model-generated content. Such categorization becomes a cornerstone for the exploration of novel applications, the identification of limitations, and the harnessing of the creative potential inherent in these models.

3.2 Face Synthesis

3.2.1 Defining Face Synthesis in AI

Face synthesis in artificial intelligence (AI) refers to the advanced process of generating hyper-realistic human faces using various AI techniques. This technology is pivotal in fields like digital media, gaming, and security. Key attributes essential for high-quality face synthesis include symmetry, clarity, and photorealism. Symmetry is crucial as human perception tends to favor facial balance, enhancing the realism of generated faces. Clarity is imperative to ensure that the synthesized faces are distinct and detailed, avoiding any

blurriness or ambiguity. Lastly, photorealism against ground truth images is fundamental; the generated faces should closely resemble real human faces, making distinguishing them from actual photographs challenging. Additionally, the ability to replicate a range of facial expressions adds depth and authenticity, allowing for dynamic and versatile face synthesis.

3.2.2 Applications and Ethical Considerations

The applications of face synthesis using diffusion models span various domains, including entertainment, where they are used to create realistic characters in movies and video games. They also find applications in virtual reality and digital marketing, where lifelike avatars or personalized content creation is required.

However, the advent of these technologies raises significant ethical considerations. The most prominent concern is the creation of deepfakes, which are synthetic media where a person's likeness is replaced with someone else's, often without consent. This has profound implications for misinformation, privacy violations, and even political manipulation.

From a cybersecurity perspective, the misuse of face synthesis can lead to identity theft, fraudulent activities, and the spread of false information. Hence, while these models open new avenues in digital media, they also necessitate stringent ethical guidelines and robust security measures to prevent their misuse.

The rise of diffusion models in face synthesis represents a significant leap in AI capabilities, offering unparalleled realism and potential. However, it's crucial to balance these technological advancements with responsible use and ethical considerations to mitigate potential risks in cybersecurity and societal impacts.

3.3 Artistic Rendering

3.3.1 Artistic Style Transfer with Diffusion Models

Artistic style transfer has become a significant application in the field of AI and computer vision, particularly with the advent of diffusion models. These models capture and replicate artistic styles with remarkable precision and diversity. One of the critical requirements in artistic style transfer is the faithful representation of the chosen style. Whether the style is that of a renowned artist or a specific art movement, the generated image must convincingly embody the characteristics and nuances of that style.

Furthermore, contemporary tools offer the flexibility to generate images either from textual descriptions or from existing images. This versatility allows for a wide range of creative outputs, from reimagining photographs in the style of famous paintings to bringing written artistic visions to life.

3.4 Image-to-Image Translation

3.4.1 Fundamentals of Image Translation

Image-to-image translation marks a pivotal advancement in the field of computer vision and artificial intelligence, primarily utilizing diffusion models. These models have revolutionized the way we approach image transformation, enabling the conversion of images from one domain to another with unprecedented accuracy and detail. Diffusion models work by gradually transforming an image through a series of steps, each adding a layer of complexity

or alteration. This process allows for a controlled and highly customizable transformation, making it ideal for various applications ranging from artistic expression to practical image enhancement.

3.5 Super-Resolution

3.5.1 High-Resolution Image Synthesis

In image processing, synthesizing high-resolution images using diffusion models like the Stable Diffusion x4 Upscaler [6], found on Hugging Face, represents a significant technological advancement. These models leverage iterative refinement processes to transform low-resolution images into detailed, high-resolution outputs. The practical applications of high-resolution images are vast and varied, encompassing fields such as medical imaging, where enhanced image clarity can lead to more accurate diagnoses, satellite imagery for improved environmental analysis, and digital art restoration, enabling the recovery of fine details in historical artworks. The Stable Diffusion x4 Upscaler stands out for its ability to maintain image integrity while enhancing resolution, which is crucial in applications where precision and detail are paramount.

3.6 Video Synthesis

The use of diffusion models in video synthesis is a significant milestone in digital content creation, thanks to artificial intelligence. Unlike traditional video production techniques, AI-generated video relies on advanced algorithms to synthesize video content, often from minimal inputs like images or text. Among the leading tools in this space are Stability AI's Stable Video Diffusion [9] and CogVideo [13] by the AI Lab at the Computer Vision Center (CVC).

These tools represent a paradigm shift in video creation, offering capabilities to generate short video clips, typically up to 4 seconds, from static images or text prompts. Stability AI's tool is notable for its ability to create videos from either an image or a text prompt, while CogVideo specializes in text-prompt-based video generation. The application of diffusion models in these tools showcases a significant advancement in AI's ability to understand and interpret visual data.

One of the primary issues with current AI video synthesis, especially those using diffusion models, is the lack of smoothness in the generated content. This is particularly evident in videos involving human expressions, where distortions are often noticeable. The complexity of human facial expressions and movements poses a significant challenge for these models, as accurately capturing and reproducing such dynamic details requires sophisticated understanding and processing of temporal and spatial data.

Another limitation is the duration of the videos produced. Tools like Stability AI's Stable Video Diffusion and CogVideo are restricted to generating short clips, typically no longer than 4 seconds. This constraint highlights the challenges in processing and synthesizing longer sequences of video, which would require more advanced handling of temporal continuity and narrative coherence.

3.7 Text-to-Image Synthesis

3.7.1 Bridging Text and Visuals with AI

Generating images from textual descriptions using diffusion models is a groundbreaking development in AI. Through a carefully orchestrated process of iterative refinement, these models translate textual inputs into detailed visual outputs. The underlying mechanism involves understanding and interpreting the nuances of the text, followed by a gradual development of images through a series of diffusion and reverse-diffusion steps. This method stands out for its ability to capture subtle details and the context of the text, allowing for a more accurate and representative visual representation of the described scene or object.

Chapter 4

Tools and Technologies

4.1 Tools for Face Synthesis

Various tools and technologies have been developed for face synthesis, each leveraging different AI models to achieve realistic results.

4.1.1 Generative Adversarial Networks (GANs)

BigGAN: One of the underlying technologies Artbreeder uses is BigGAN, which stands for Big Generative Adversarial Networks. Introduced in the paper „Large Scale GAN Training for High Fidelity Natural Image Synthesis“ [10], BigGAN is known for its ability to generate high-resolution, high-quality images. It utilizes a large-scale approach to training, employing massive datasets and increased batch sizes, which significantly improves the quality and diversity of the generated images. This model is particularly effective in synthesizing detailed and realistic facial features when used for face synthesis.

StyleGAN: Additionally, Artbreeder employs StyleGAN, another advanced version of GAN introduced by Tero Karras, Samuli Laine, and Timo Aila in their paper „A style-based generator architecture for generative adversarial networks“ [15]. StyleGAN is renowned for its novel approach of controlling the synthesis process through styles, which adjust features at different levels of detail. This allows for unprecedented control over facial attributes, making it highly effective for creating nuanced and diverse facial images.

Artbreeder: Artbreeder itself is a platform that allows users to explore and manipulate images through a simple interface, combining aspects of BigGAN and StyleGAN. It offers users the ability to blend and evolve images, effectively 'breeding' new visuals from existing ones [1]. This platform is particularly noted for its ability to merge facial features seamlessly, resulting in highly customizable and unique outputs. Artbreeder's utilization of text-to-image generation capabilities further enhances its versatility in creative applications.

Properties of Artbreeder: Artbreeder excels in user interactivity and ease of use, providing tools that allow even novices to create professional-level images. It also supports a collaborative environment where users can share and evolve images collectively, fostering a community of creators enhancing each other's work.

4.1.2 Diffusion Models

A type of generative models called diffusion models has greatly advanced image synthesis, particularly in generating realistic human faces from textual descriptions. These models

work by gradually transforming a random noise distribution into a coherent image through a process that effectively reverses diffusion, a concept derived from statistical physics.

Stable Diffusion: A prominent example of diffusion models is Stable Diffusion, which has been widely adopted for its efficiency and high-quality outputs. Developed by Stability AI, Stable Diffusion allows for generating photorealistic images based on textual descriptions. The model's latest iteration, Stable Diffusion V3, has further refined the synthesis process, enhancing the clarity and realism of generated faces [21].

Implementation in Web Interfaces: Tools such as AUTOMATIC1111's Stable Diffusion WebUI and NightCafe Creator Studio allow users to generate images from text prompts with added styles. It also provides the option to add a negative prompt, choose the size of the output batch, and adjust the size of the generated images. AUTOMATIC1111's WebUI provides an open-source interface that simplifies interaction with the Stable Diffusion model, enabling users to generate images directly from their browsers [8]. NightCafe Creator Studio leverages Stable Diffusion V3 to offer enhanced capabilities in face synthesis, allowing users to create highly detailed and expressive facial images from simple text prompts [19].

4.2 Tools for Artistic rendering

4.2.1 Generative Adversarial Networks (GANs)

Artbreeder, an innovative platform for artistic rendering, merges the powerful capabilities of BigGAN and StyleGAN to offer users an interactive canvas for creative expression. This tool excels in the art of visual synthesis, enabling users to blend and morph images to 'breed' entirely new visual forms. Artbreeder stands out for its seamless fusion of facial features, facilitating the creation of uniquely customized artistic outputs. This is achieved through its intuitive interface, which simplifies the complex algorithms into user-friendly controls for artistic manipulation. The integration of text-to-image generation techniques further broadens its application, allowing for the transformation of verbal ideas into stunning visual representations [1].

4.2.2 Diffusion Models

In the artistic domain of image generation, diffusion models like Stable Diffusion [8] and Midjourney [3] excel in transforming creative concepts into visually striking artworks. Stable Diffusion, renowned for its flexibility and open-source accessibility, facilitates intricate style transfers that cater to diverse artistic preferences. It supports a variety of artistic styles, including Cinematic visuals, Analog Film aesthetics, Fantasy Art creations, and Impressionist paintings, each offering a unique pathway to visual storytelling. Midjourney further extends these capabilities, adeptly converting text and images into artistically styled outputs that resonate with both contemporary and traditional artistic tastes. These tools showcase the profound impact of diffusion models in producing images that are not only stylistically diverse but also rich in creative expression and visual appeal.

4.3 Tools for Image-to-Image Translation

4.3.1 Exploration of Tools

Stable Diffusion Model

The Stable Diffusion img2img model, a standout in the domain of latent diffusion processes, offers advanced capabilities for image-to-image translation. This model excels in interpreting and transforming input images based on directive text prompts, allowing for nuanced adjustments that maintain the original context while infusing new artistic or realistic elements. It is particularly effective for applications in digital art, photo restoration, and creative content generation, where contextual coherence and aesthetic enhancement are crucial [4].

Artbreeder GAN Model

The Artbreeder Mixer employs a Generative Adversarial Network (GAN) to facilitate dynamic image manipulation and blending. Users can adjust the influence of input images and associated text prompts to produce a range of outputs from subtle modifications to drastic transformations. This model is widely used in fields such as character design, landscape modification, and other creative endeavors that benefit from high degrees of customization and iterative exploration [2].

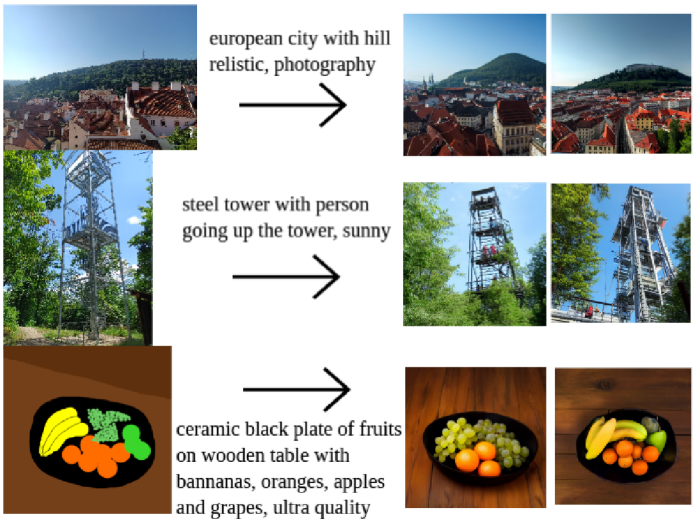

Figure 4.1: Images generated from images guided by prompt by AUTOMATIC1111's Stable Diffusion web UI

4.4 Tools for Super-Resolution

4.4.1 Real-ESRGAN

Real-ESRGAN [22] is an advanced Generative Adversarial Network designed to tackle the challenge of enhancing the perceptual quality of images that have undergone real-

world degradations. It utilizes robust training methodologies involving synthetic data to simulate a variety of image quality issues commonly found in real-world scenarios. While Real-ESRGAN significantly improves the sharpness and clarity of images, it may introduce specific artifacts like ringing or overshooting, which is particularly evident in scenarios involving complex degradations. This makes it both a powerful and a delicately balanced tool for super-resolution tasks [7].

4.4.2 Stable Diffusion x4 Upscaler

Stable Diffusion x4 Upscaler [17], specifically designed to work with images produced by Stable Diffusion models, this upscaler employs diffusion-based techniques to enhance image resolution by up to four times. It maintains the artistic integrity and detail of the original outputs while scaling them up to higher resolutions. The upscaler's effectiveness is tailored to images with particular characteristics typical of diffusion model outputs, making it somewhat specialized. Additionally, this upscaling process is resource-intensive, requiring significant computational power, which might limit its accessibility or practicality in resource-constrained environments [5].

4.5 Tools for Video Synthesis

4.5.1 Stable Video Diffusion Tool

Stable Video Diffusion, developed by Stability AI and detailed in the paper *Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets* [9], innovates in video generation from textual descriptions by implementing a three-stage training process. The initial stage involves image pretraining using a text-to-image diffusion model, preparing the system for more complex video tasks. The second stage, video pretraining, leverages a large dataset of videos to adapt the model to dynamic content. The final stage, video finetuning, polishes the model by training on a curated subset of high-quality videos to enhance resolution and motion realism. Despite its sophisticated architecture, it may still exhibit occasional abrupt transitions between video frames.

4.5.2 CogVideo Tool

CogVideo is model for short video generation from text, described in the paper *CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers* [13]. This model harnesses transformer-based architectures for converting text into video sequences, ensuring coherence over extended durations and varied content. It is distinguished by its efficient management of complex, dynamic video content generation, though it currently faces limitations in output resolution and duration, which may impede usage in more demanding scenarios.

4.6 Tools for Text-to-Image Synthesis

4.6.1 Stable Diffusion

Stability AI's Stable Diffusion is a groundbreaking open-source tool that leverages diffusion models for text-to-image synthesis [8]. This tool enables image creation using guided and unguided prompts, offering users extensive control over the image generation process. By

incorporating specific terms such as „photography“, „realistic“, and „ultra quality“ into prompts, users can guide the model towards producing exceptionally photo-realistic images. This mimics the output one might expect from high-end photography equipment, thereby elevating the realism and fidelity of the generated images.

4.6.2 Artbreeder GAN Model

Artbreeder utilizes Generative Adversarial Networks (GANs) to facilitate a highly interactive form of image creation. By manipulating genes — analogous to image characteristics — users can subtly influence the appearance and style of the generated images. This model excels in producing variations of existing images or combining multiple images into a new, unique output, offering a high level of detail and customization in response to user input [2].

4.6.3 NightCafe Diffusion Model

NightCafe Studio employs diffusion models to transform detailed textual prompts into vivid images. It supports a wide array of artistic styles and realism levels, enabling users to specify exactly what kind of artistic output they desire, from abstract art to ultra-realistic portraits. This model's effectiveness in generating high-quality images from complex prompts makes it an invaluable tool for artists and creators exploring the intersection of AI and art [19].

Chapter 5

Integration and Experimentation

5.1 Integration of Selected Tools

5.2 Experimental Design

These experiments aim to evaluate the performance of selected tools across various categories of AI-generated visual content. These categories include face synthesis, video synthesis, text-to-image, image-to-image, artistic rendering and super-resolution (upscaling). The following subsections outline the experimental setup and define metrics for assessing the generated content's quality, diversity, and realism.

5.2.1 General Methodology

Each experiment will involve generating content using the specified tools and assessing the outputs based on predefined metrics. The experiments will be conducted in controlled environments to ensure consistency in the evaluation process.

5.2.2 Face Synthesis

Task

Generate photorealistic human faces with different facial expressions, nationalities, genders, and ages. Then, compare the outputs of two tools that use GAN and Diffusion models, respectively, by using text prompts.

Metrics

- Quality: Assess the clarity and photorealism against ground truth images.
- Diversity: Evaluate the variety in facial features and expressions using a diversity index.
- Realism: Conduct a user study for subjective assessment of realism.

Text prompts

Through tools like Artbreeder (GAN) and Stable Diffusion, utilized by Creator Nightcafe, images are generated from text prompts using default style. Examples of used prompts:

„elderly European man portrayed with a wise and content expression, his thinning white hair and beard adding to his dignified appearance, wrinkles, photorealistic.“

„teenage Hispanic girl with an expression of surprise, her long, wavy brown hair cascading, realism in her raised eyebrows and slightly open mouth highlights her youthful spontaneity, photorealism“

„young Middle Eastern man with a composed expression, thick wavy hair and detailed stubble, photorealistic“

Data Collection

In this analysis, the users were given 12 images that were generated from six different prompts. Both AI platforms created images for each prompt, ensuring a balanced comparison. Users rated each image on a scale of 1 to 5, based on its realism, adherence to the prompt, and artistic quality. In summary, 54 responses were considered that contained some variety in answers and had non-algorithmic responses.

Figure 5.1: Results of human face generation with different age groups, nationalities and expressions, comparison of generation by chosen GAN model and Diffusion model

Statistical Analysis

The user ratings were analyzed using descriptive statistics, including mean, standard deviation, and confidence intervals, to determine which model performed better in various assessed categories.

Quantitative Results

The average scores for each model across different categories were as follows:

Model	Accuracy and Detail	Realism
Diffusion (DF)	4.07	3.80
Generative Adversarial Network (GAN)	3.31	3.61

Table 5.1: Average user ratings for the Diffusion and GAN models

Statistical results indicated differences in performance between the two models, as detailed in Table 5.1 and Graph 5.2.

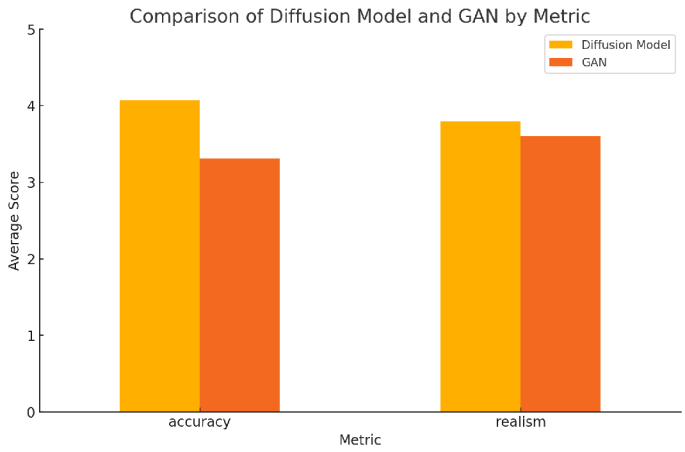

Figure 5.2: Comparison of scores between GAN and Diffusion models for generating face images using chosen metrics from user survey in the range of 1 to 5.

Qualitative Feedback

User comments suggested preferences for specific models based on factors like realism and artistic interpretation, which provided deeper insights into the strengths and weaknesses of each model.

Conclusion

The analysis revealed that the diffusion model generally outperformed the generative adversarial network model in both accuracy and realism. This suggests that diffusion models may be more adept at generating more detailed and realistic facial images according to user evaluations.

This study provides valuable insights into the capabilities of advanced AI art models in creating diverse and realistic facial images. Future research could explore further with a larger dataset or additional models to enhance the robustness of the findings.

5.2.3 Video Synthesis

Task

Generate short video clips from images or text prompts.

Metrics

- Smoothness: Measure frame-to-frame continuity using optical flow algorithms.
- Realism: Subjective assessment through user surveys.
- Fidelity: Compare synthesized videos with source material using Structural Similarity Index (SSIM).

Stable Video Diffusion

Stable Video Diffusion, developed by Stability AI [9], is designed to generate realistic and scalable video sequences from textual descriptions. The tool generates short videos from prompt text, with occasional abrupt transitions that can undermine realism and deform shapes.

The result videos have parameters of 7 frames per second, dimensions 768x768 px, and length 4 seconds. They were generated approximately in 1-2 minutes.

Figure 5.3: Selected frames from the video generated through stable diffusion.

CogVideo

CogVideo [13], employs transformer-based architectures to convert text into corresponding video content. Its constrained output resolution and duration are the primary challenges, which may affect its practical applications in more demanding scenarios.

Result videos have parameters of 5 frames per second and dimensions 480x480 px and length 3 seconds, videos were generated approximetely in 9 minutes.

Figure 5.4: Selected frames from the video generated through CogVideo.

In the preview of the image sequence referenced as 5.3 and 5.4, there are examples of human faces where the movement is focused on the head and flowing hair. There is also an example of a person walking in the mountains where the movement is focused on the movement of hands and legs. Another example is of nature, where the main movement is zooming in on a deer.

5.2.4 Text-to-Image

Task

Generate images from textual descriptions, text prompts should be as detailed and specific as possible.

Metrics

- Accuracy: Evaluate the correspondence between text and generated images.
- Creativity: Assess the novelty of the generated images using a creativity score.
- Detail: Analyze the intricacy and clarity of details in the images.

Text prompts

Through tools like Artbreeder (GAN) and Stable Diffusion, utilized by Creator Nightcafe, images are generated from text prompts using default style. Examples of used prompts:

„nighttime scene of the bustling streets of Tokyo, showcasing neon lights and towering digital billboards. The image includes a detailed depiction of diverse pedestrians ranging from businessmen to young fashionistas, vendors selling street food, and a glimpse of a passing high-speed train with the city's skyline in the background“

„detailed image of a quaint Armenian small city during sunset, the scene includes traditional stone houses with carved wooden balconies, cobblestone streets bustling with vendors selling fresh produce, and an ancient stone church with a distinctive khachkar (cross-stone) in front. The background features the rugged Caucasus Mountains tinted in the orange glow of the setting sun“

„hotorealistic image of an early morning in a dense, misty forest. The scene captures the dew on vibrant green ferns and wildflowers, a small deer peering through the foliage, and

rays of sunlight breaking through the high treetops, illuminating a narrow stream winding through the forest floor“

Data Collection

In this analysis, users evaluated 10 images generated from several prompts using both AI platforms. Each image was rated on a scale from 1 to 5, focusing on the accuracy, detail, and realism of the generated images. In summary, 54 responses were considered that contained some variety in answers and had non-algorithmic responses.

Figure 5.5: Results of different environments image generated by chosen GAN model and Diffusion model

Statistical Analysis

User ratings were analyzed using mean scores and standard deviations to determine which model performed better in terms of detailed accuracy and realism.

Quantitative Results

The average scores for each model across the categories were as follows:

Model	Accuracy and Detail	Realism
Diffusion (DF)	3.54	2.42
Generative Adversarial Network (GAN)	2.82	3.37

Table 5.2: Average user ratings for the Diffusion and GAN models in environmental image generation

Statistical results suggest that the diffusion model outperforms the GAN in accuracy and detail, whereas the GAN scores higher in realism.

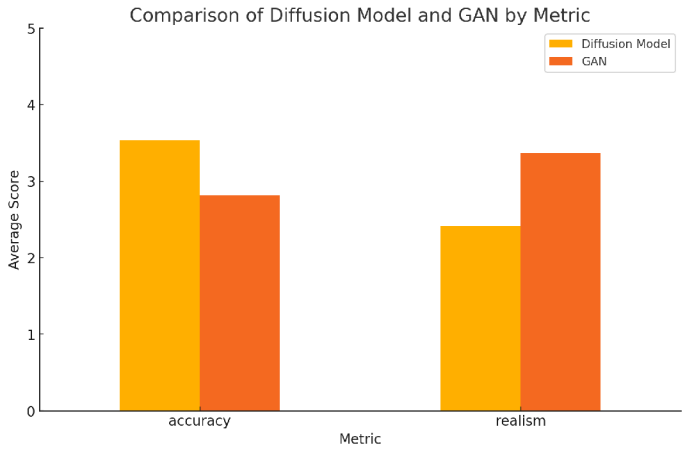

Figure 5.6: Comparison of scores between GAN and Diffusion models for generating images of detailed environment using chosen metrics from user survey in the range of 1 to 5.

Qualitative Feedback

User comments provided further insights, indicating preferences for specific models based on their perceived accuracy in reflecting the prompts and the realism of the environments depicted.

Conclusion

The study offers valuable insights into how different AI models perform in generating environmental images from text prompts, with each model showing particular strengths. Further research with more varied prompts and larger datasets could provide more definitive conclusions.

5.2.5 Image-to-Image Translation

Task

Transform images from one domain to another using two different models: the Stable Diffusion model and a GAN model from Artbreeder.

Metrics

- Fidelity: Compare with original images for content preservation.
- Quality: Assess resolution and clarity.
- Aesthetic Value: Rate the artistic appeal of the translated images.

Evaluation of Results

Stable Diffusion Model The Stable Diffusion model excelled in composition and detail retention between input and output images. However, it generally produces less realistic images, making it apparent that the images are AI-generated. In images with more than two people, facial features often lose precision.

Artbreeder GAN Model The GAN model used in Artbreeder closely replicated the composition of the input image, though not as precisely as the diffusion model. When the text prompt is given more weight, this model can produce more realistic lines, but at the expense of diminishing the influence of the original image composition.

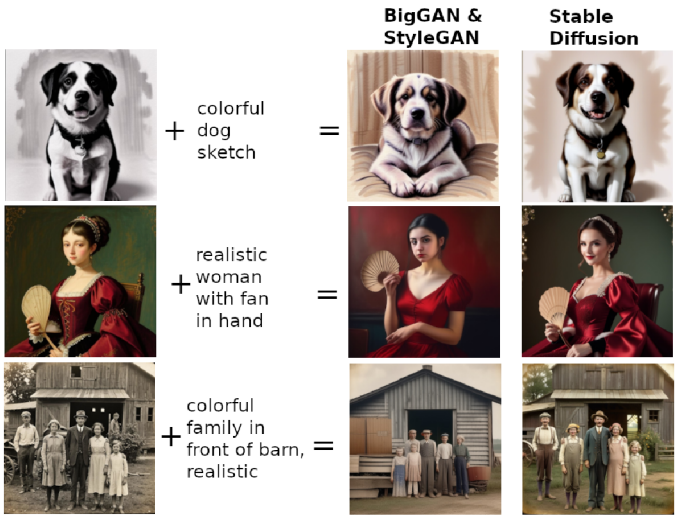

Figure 5.7: Results of image + text to image generation by chosen GAN model and Diffusion model

Refer to Figure 5.7 for a visual comparison of the results generated by the Stable Diffusion and Artbreeder GAN models.

5.2.6 Artistic Rendering

Task

Perform style transfer to render artistic effects on images.

Metrics

- Style Alignment: Evaluate how closely the generated image aligns with the intended artistic style using a style-match score.
- Content Preservation: Measure the degree to which the original content of the image is preserved.
- Aesthetic Appeal: Rate the visual appeal and artistic quality through expert reviews.

This section evaluates the performance of two AI models, the diffusion model used on Stable Diffusion Web and the generative adversarial network model (GAN) used on Artbreeder, in generating artistic style images. The evaluation focuses on how accurately the images contain all details of the text prompt (Accuracy and Details) and how well the images align with the chosen artistic style (Style Alignment).

Data Collection

Users rated 10 images generated from different artistic prompts, using both AI platforms. Each image was rated on a scale from 1 to 5, where 1 means least accurate or best style alignment and 5 means most accurate or best style alignment. In summary, 54 responses were considered that contained some variety in answers and had non-algorithmic responses.

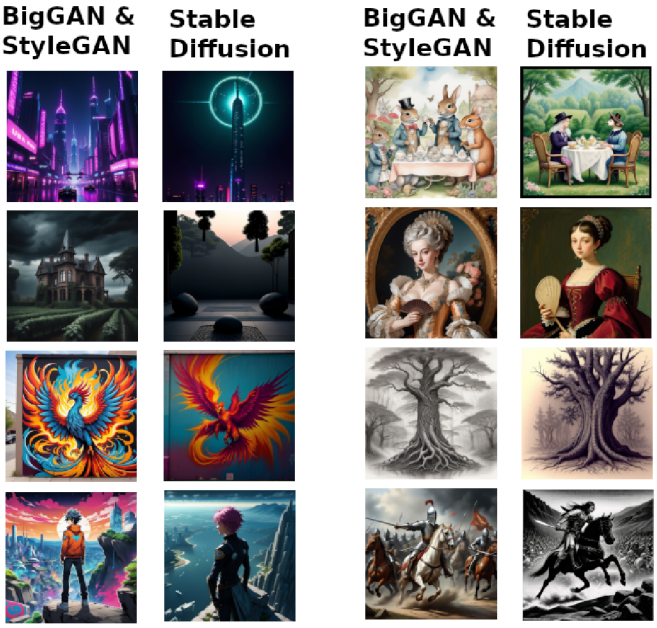

Figure 5.8: Results of images with different artistic styles generated by chosen GAN model and Diffusion model

Statistical Analysis

User ratings were analyzed to determine which model performed better in terms of detailed accuracy and style alignment.

Quantitative Results

The average scores for each model in the categories of Accuracy and Details and Style Alignment were as follows:

Model	Accuracy and Details	Style Alignment
Diffusion (DF)	3.69	3.71
Generative Adversarial Network (GAN)	3.72	3.49

Table 5.3: Average user ratings for the Diffusion and GAN models in artistic style image generation

The results show a very close competition between the two models in terms of detailed accuracy, with the GAN model slightly leading. However, the diffusion model performs slightly better in aligning with the artistic style as perceived by the users.

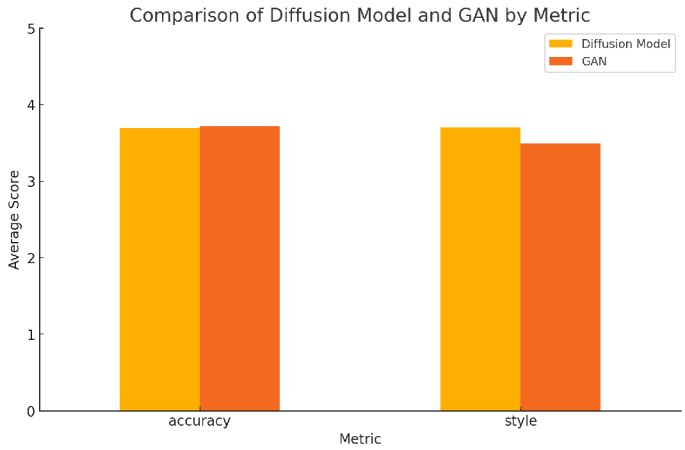

Figure 5.9: Comparison of scores between GAN and Diffusion models for generating images in specific artistic style using chosen metrics from user survey in the range of 1 to 5.

Qualitative Feedback

User comments highlighted specific preferences for model outputs, noting which model better captured the essence of the artistic prompts and maintained style consistency.

Conclusion

The analysis provides insights into the nuanced capabilities of the diffusion and GAN models in artistic image generation. The slight differences in scores suggest that both models have their strengths, with the GAN model excelling slightly in detail accuracy and the diffusion model in style alignment.

This comparative study helps to understand the effectiveness of different AI models in artistic style image generation. It underscores the importance of choosing the right model based on the specific needs of detail accuracy and style alignment. Further exploration with a broader range of artistic styles and prompts might provide more definitive conclusions.

5.2.7 Super-resolution (upscaling)

Task

Enhance the resolution of images. Analyze upscaled images and compare them to original images, downscaled images and upscaled images from other model.

Metrics

- Resolution Enhancement: Quantify the increase in resolution using Peak Signal-to-Noise Ratio (PSNR).
- Image Quality: Evaluate clarity and detail preservation post-upscaling using SSIM.
- Artifact Measurement: Assess the presence of any upscaling artifacts such as blurring or pixelation.

Models and Performance

- **Real-ESRGAN** [22]: This GAN-based model is optimized for real-world degradations and enhances the perceptual quality of images.
- **Stable Diffusion x4 Upscaler** [17]: Designed specifically for images generated by Stable Diffusion models, this diffusion-based upscaler enhances image resolution..

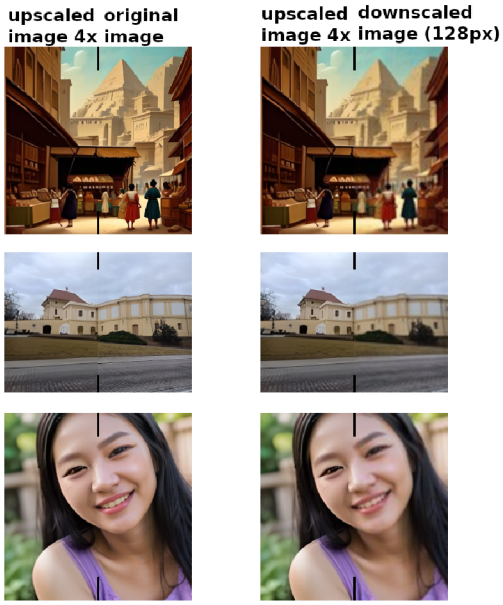

Figure 5.10: Upscaled image using Real-ESRGAN.

- **nightmareai/real-esrgan**: Produces better results with smoother lines and quicker generation time, approximately 0.64 seconds on Nvidia T4 GPU hardware. Input image parameters: 128x128 pixels, upscaling scale = 4.
- **lucataco/stable-diffusion-x4-upscaler**: Produces a grainier image and takes approximately 4.21 seconds longer on Nvidia A100 GPU hardware. Input image parameters: 128x128 pixels, upscaling scale = 4.

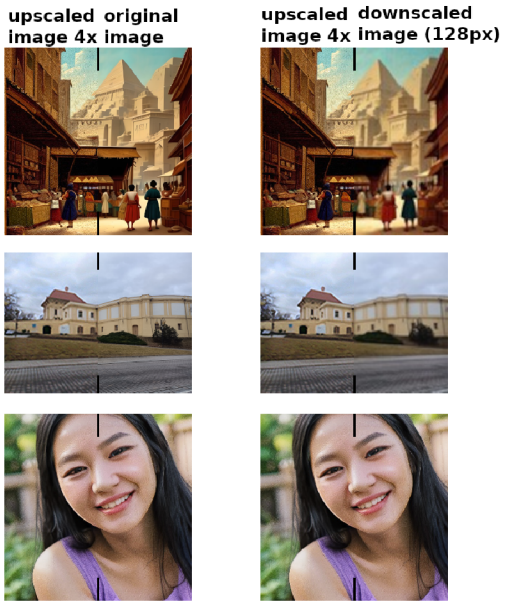

Figure 5.11: Upscaled image using Stable Diffusion x4 Upscaler.

Edge Detection of Upscaled Images

In this section, we examine the effect of image upscaling techniques on the preservation and enhancement of details. Specifically, we use edge detection algorithms to evaluate how different upscaling methods like Stable Diffusion x4 Upscaler and Real-ESRGAN affect the perceived detail in images. Edge detection is a crucial step in identifying the amount of detail an upscaler can extract from a low-resolution image. The detection of edges in an image highlights areas of rapid intensity change, which are typically indicative of important features and details.

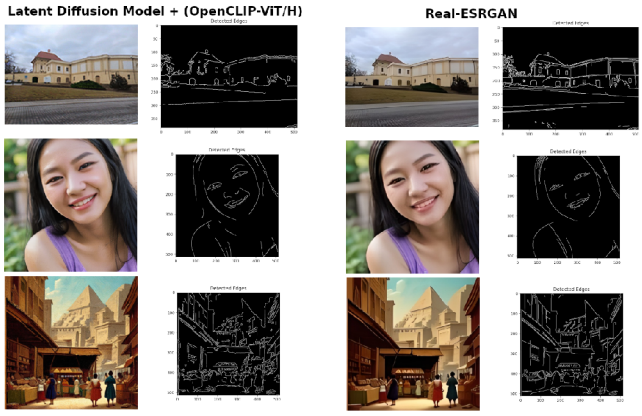

Figure 5.12: Detected edges of upscaled images by Stable Diffusion x4 Upscaler and Real-ESRGAN.

The images processed with Real-ESRGAN generally exhibit a higher number of detected edges compared to those processed with Stable Diffusion x4 Upscaler. This observation

suggests that Real-ESRGAN may be more effective at enhancing or preserving fine details in low-resolution images during the upscaling process.

Python Code for Edge Detection The following Python code is used to detect edges in the upscaled images. This method utilizes the Canny edge detector, which is effective at highlighting significant edges by detecting sharp changes in intensity across the images.

```
import cv2
import numpy as np
import matplotlib.pyplot as plt

def detect_edges(image_path, result_path):
    # Load image in grayscale
    img = cv2.imread(image_path, cv2.IMREAD_GRAYSCALE)

    # Apply Gaussian blur to reduce noise
    img_blurred = cv2.GaussianBlur(img, (3, 3), 0)

    # Detect edges using Canny edge detector
    edges = cv2.Canny(img_blurred, threshold1=100, threshold2=200)

    plt.imshow(edges, cmap='gray')
    plt.title("Detected Edges")
    plt.savefig(result_path)

    return edges
```

This script is designed to be executed for each upscaled image to assess the effectiveness of the upscaling method. It outputs images with detected edges, which are then analyzed to determine the quality of detail enhancement.

Chapter 6

Conclusion and Summary

6.1 Evaluation of Experiments

This section consolidates and interprets the experiments results to evaluate the performance of diffusion models (DM) and generative adversarial networks (GAN) across various visual content generation tasks, including face synthesis, text-to-image, and artistic rendering. The experiments were designed to assess both the accuracy and detail of the images generated from text prompts and their realism, as well as style alignment in artistic rendering.

6.1.1 Face Synthesis

In the face synthesis category, diffusion models demonstrated a higher capability for generating detailed and accurate facial features from text prompts compared to GANs. The DM average user rating for accuracy and detail was significantly higher, indicating a more precise interpretation of the textual descriptions into visual content. However, for realism, while diffusion models were slightly better, the difference was less pronounced, suggesting that both model types have substantial capabilities, but diffusion models may edge out slightly in capturing the nuanced expressions and features specified in the prompts 5.1.

6.1.2 Text-to-Image (Environment generation)

Similar trends were observed in the text-to-image synthesis category. Diffusion models consistently outperformed GANs in terms of detailed accuracy, closely matching the textual descriptions with high fidelity in the generated images. However, when evaluating the realism of environments depicted in the images, such as cities, towns, and natural landscapes, GANs received higher user ratings. This indicates that while DMs excel in detail precision, GANs may offer a more realistic portrayal of broader environmental scenes 5.2.

6.1.3 Artistic Rendering

Artistic rendering experiments showed a more balanced performance between diffusion models and GANs. Both types of models scored similarly in terms of style alignment, indicating their effective capacity to adapt to different artistic styles as dictated by the prompts. However, in accuracy and detail of capturing all elements described in the text, diffusion models slightly led, reinforcing their strength in closely following complex textual instructions to produce visually detailed outputs 5.3.

6.1.4 Super-resolution (upscaling)

Quantitative Results

Real-ESRGAN showed better performance in enhancing the resolution of images while maintaining smoother lines and less graininess compared to the Stable Diffusion x4 Upscaler. The PSNR and SSIM metrics indicated better performance of Real-ESRGAN regarding both resolution enhancement and image quality. The edge detection results further supported these findings, with Real-ESRGAN images displaying more detected edges, suggesting better detail preservation and less artifact introduction during upscaling.

6.1.5 Image-to-image

The comparative study of the Stable Diffusion model and the Artbreeder GAN model in the task of image-to-image translation with textual prompts highlighted distinct strengths and limitations of each technology.

The experiments conducted demonstrated that the Stable Diffusion model excels in retaining the composition and detail of the source images more effectively than the Artbreeder GAN model. This capability is particularly evident when the input images contain complex arrangements or detailed elements that need to be preserved in the translated output. The Stable Diffusion model's ability to maintain these aspects offers significant advantages for applications where fidelity to the original image's composition is critical.

Conversely, while the Artbreeder GAN model also performed well in replicating the composition, it did not match the precision of the diffusion model. However, it was noted that the GAN model could produce images with more realistic appearances when the text prompt emphasized creative or interpretative outputs over strict adherence to the original composition. This suggests that the GAN model may be preferable when a balance between creative flexibility and composition fidelity is desired.

In conclusion, the Stable Diffusion model is more suited for image-to-image translation tasks that require high fidelity to the source image's composition, especially when supplemented by text descriptions. It provides a robust framework for ensuring that the input's critical elements and overall structure are accurately reflected in the output. This finding is crucial for applications in fields such as digital art restoration, medical imaging, and any other area where maintaining the integrity of the original image is paramount.

6.1.6 Video generation from text

The experiments conducted with the Stable Video Diffusion and CogVideo models aimed to explore the capabilities of contemporary AI technologies in generating short video clips from textual descriptions. Despite the technological advancements represented by these models, the video synthesis task highlighted significant challenges, particularly with the diffusion model, in terms of achieving realistic and continuous video outputs.

While superior in several aspects such as frame rate and resolution compared to CogVideo, the Stable Video Diffusion model still faced difficulties in maintaining realism and smooth transitions between frames. Users often noted that the video sequences, although impressive in their ability to generate coherent scenes from text, suffered from abrupt transitions and occasionally distorted forms, which significantly detracted from the realism of the videos. These issues were particularly evident in sequences requiring complex movements or detailed animations, such as flowing hair or walking motions.

Furthermore, the short length of the videos, typically around 4 seconds, and the low frame rate of 7 frames per second imposed by the current capabilities of the Stable Video Diffusion model limit the practical applications of this technology in scenarios where longer and more fluid video content is required. This is a considerable limitation for use cases such as filmmaking, animated content creation, or any digital media production that seeks to convey a narrative or detailed action over time.

In conclusion, while the Stable Video Diffusion model demonstrates promising advancements in the field of AI-driven video synthesis, the results remain somewhat experimental and showcase the infancy of this technology in dealing with the complexities of realistic, continuous video generation. Future developments must focus on enhancing the smoothness of transitions, extending the length of the video output, and improving the frame rate to meet the demands of more realistic and practical applications. The field of AI video synthesis, although rapidly evolving, still has significant hurdles to overcome before it can produce consistently realistic and commercially viable video content from textual or image inputs.

6.2 User Studies

User studies were conducted to gather subjective assessments of the generated images across all experiments. Participants rated images on a scale from 1 (least accurate or realistic) to 5 (most accurate or realistic), providing insights into the perceived quality of the content produced by diffusion models and GANs.

6.2.1 Overall User Feedback

Across all categories, users appreciated the high level of detail and accuracy 5.2 5.1 in images generated by diffusion models. These models were particularly noted for their ability to translate detailed text prompts into clear, concise visual representations. In contrast, GANs were often preferred for their ability to render environments with a greater sense of realism, suggesting that they might be better suited for tasks where atmospheric or holistic scene composition is critical.

6.2.2 Implications for Model Selection

The user feedback highlights an essential consideration for selecting between diffusion models and GANs based on a task's specific needs. Diffusion models are preferable for applications requiring high fidelity to detailed prompts. Conversely, for projects where overall realism and environmental feel are more important, GANs might be the better choice.

This comparative analysis underscores the importance of aligning the choice of model with the specific objectives and requirements of the visual content generation task at hand. Future research could explore hybrid models that combine the strengths of both DMs and GANs to enhance both detail accuracy and environmental realism.

6.3 Security Implications of Diffusion Models

With the increasing capabilities of AI in synthesizing realistic images and videos, diffusion models pose unique security challenges. This section discusses potential risks and measures to mitigate them.

6.3.1 Potential Risks

- **Deepfakes:** The ability of diffusion models to generate photorealistic human images and videos can be exploited to create misleading or harmful content, such as fake news or impersonation.
- **Data Privacy:** There are concerns regarding the use of personal data in training these models without consent, leading to privacy violations.
- **Misinformation:** Enhanced capabilities in generating realistic environments and scenarios can contribute to the spread of misinformation.

6.3.2 Mitigation Strategies

- **Watermarking:** Implementing digital watermarks to trace and authenticate AI-generated content.
- **Regulation:** Establishing clear legal frameworks to govern the use and application of generative models.
- **Public Awareness:** Educating the public about the nature of AI-generated content and potential abuses.

6.4 Summary

This thesis has presented a comprehensive analysis of diffusion models (DMs) and generative adversarial networks (GANs) across a range of applications in AI-generated visual content. The experiments were strategically designed to compare these technologies on their ability to generate realistic and accurate images, videos, and artistic renderings from textual and image prompts. Here, we summarize the core findings and their broader implications.

6.4.1 Key Findings

- **Detail and Accuracy:** Diffusion models consistently demonstrated superior performance in generating detailed and accurate representations, especially when transforming detailed text prompts into high-fidelity images. This was notably evident in face synthesis and text-to-image translations, where DMs excelled in capturing intricate details that closely matched the input descriptions.
- **Realism:** In tasks requiring a realistic depiction of broader environmental scenes, GANs often outperformed DMs. This suggests GANs' strength lies in rendering complex textures and life-like scenarios, making them particularly effective for applications requiring high levels of environmental realism.
- **Artistic Rendering:** Both model types showed comparable capabilities in artistic style transfers, indicating their effective application in creative industries. However, diffusion models slightly edged out in maintaining textual detail within artistic interpretations.
- **Video Synthesis:** Despite the potential shown by diffusion models in generating short video clips from textual prompts, significant challenges remain in achieving

realistic and smooth transitions. These shortcomings highlight the developmental infancy of AI in producing dynamic and continuous video content.

- **Super-resolution:** Among upscaling technologies, the Real-ESRGAN emerged as a superior choice, offering enhanced image quality with fewer artifacts, which is critical for applications like digital restoration or medical imaging where detail preservation is paramount.

6.4.2 Implications

The findings from this thesis contribute valuable insights into the evolving landscape of AI technologies in visual content generation. By delineating the strengths and weaknesses of diffusion models and GANs across various contexts, this research not only aids in model selection but also informs future developments in AI. Practitioners and researchers can leverage these insights to select the appropriate technology based on the specific needs of their projects, whether they prioritize detail accuracy, realism, or artistic flexibility.

Furthermore, the experiments underscore the importance of ongoing improvements in AI technology to address the limitations observed, particularly in video synthesis and realistic environmental rendering. As AI continues to advance, it is imperative to refine these models to enhance their practicality and effectiveness in real-world applications.

6.4.3 Technological and Ethical Considerations

The exploration of security implications underscores the need for ethical considerations and regulatory measures as AI technologies become increasingly capable of producing photorealistic and persuasive media. Strategies such as watermarking, public awareness campaigns, and stringent regulations will be crucial in mitigating the risks associated with AI-generated content, particularly in preventing misinformation and protecting data privacy.

6.4.4 Conclusion

Overall, this thesis illustrates the significant potential and current limitations of diffusion models and generative adversarial networks in generating AI-driven visual content. The nuanced understanding of each model's capabilities and their appropriate applications sets the stage for their informed use and continuous improvement, paving the way for more realistic, ethical, and practical AI applications in the future.

6.5 Future Research Directions

The experiments conducted in this thesis have shed light on the capabilities and limitations of diffusion models and generative adversarial networks in generating AI-driven visual content. These findings not only highlight the advances in the field but also underscore the challenges that need to be addressed. The following areas represent promising directions for future research that could further enhance the performance and applicability of these technologies.

6.5.1 Enhancing Realism and Continuity in Video Synthesis

Future research should focus on overcoming the challenges related to the realism and continuity of video content generated by AI models, particularly diffusion models. Developing

algorithms that can handle complex dynamics and maintain temporal consistency without abrupt transitions would significantly advance video synthesis technologies. This could involve integrating more sophisticated temporal coherence techniques or exploring hybrid models that combine the strengths of both diffusion and GAN technologies.

6.5.2 Advanced Applications of Artistic Rendering

Research into the artistic rendering capabilities of AI models can be expanded to include more nuanced interpretations of artistic styles, potentially incorporating elements of historical art movements or individual artist signatures. This can broaden the applicability of AI in creative industries such as digital art, film, and gaming.

6.5.3 Hybrid Models for Enhanced Performance

The strengths and weaknesses observed in diffusion models and GANs suggest the potential benefits of hybrid models that leverage the advantages of both. Developing models that can seamlessly integrate the detailed accuracy of diffusion models with the creative flexibility of GANs could yield superior results in a wider range of applications.

6.5.4 Ethical Use and Security Measures

As AI technologies become more capable of producing photorealistic and persuasive content, it is crucial to develop robust frameworks for their ethical use simultaneously. Future research should not only focus on enhancing the technological aspects but also on implementing effective watermarking techniques, developing better detection methods for AI-generated content, and establishing clear ethical guidelines and regulations to govern their use.

6.5.5 Interactive and Real-Time AI Systems

Investigating the feasibility of real-time AI systems for content generation could open up new avenues in interactive media, virtual reality, and live broadcasts. This involves not only improving the speed and efficiency of model training and inference but also ensuring that these systems can adapt to real-time input variations and user interactions.

6.5.6 Cross-Disciplinary Studies

Finally, fostering cross-disciplinary studies that integrate insights from cognitive science, psychology, and art could enhance the design of AI models that are better attuned to human aesthetics and perceptual criteria. This approach can lead to more intuitive and user-friendly AI tools that cater to diverse artistic and cultural preferences.

Bibliography

[1] *Artbreeder*. Available at: https://www.artbreeder.com. Accessed: 2024-05-05.

[2] *Artbreeder Mixer*. Available at: https://www.artbreeder.com/create/mixer. Accessed: 2024-04-08.

[3] *Midjourney*. Available at: https://www.midjourney.com/. Accessed: 2024-01-07.

[4] *Stable Diffusion img2img* https://huggingface.co/docs/diffusers/api/pipelines/stable_diffusion/img2img. Accessed: 2024-04-08.

[5] AI, S. *High-Resolution Image Synthesis With Latent Diffusion Models*. 2022. Available at: https://huggingface.co/docs/stabilityai/stable-diffusion-x4-upscaler.

[6] AI, S. *Stable Diffusion x4 Upscaler*. 2023. Available at: https://huggingface.co/stabilityai/stable-diffusion-x4-upscaler. Accessed: 2024-01-05.

[7] Authors, V. *Improving Real-World Blind Super-Resolution*. 2022. Available at: https://ar5iv.labs.arxiv.org/html/2307.16169.

[8] AUTOMATIC1111. *Stable Diffusion Web UI*. 2023. Available at: https://github.com/AUTOMATIC1111/stable-diffusion-webui. Accessed: 2024-01-05.

[9] Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M. et al. *Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets* https://creator.nightcafe.studio/. Accessed through NightCafe Creator Studio.

[10] Brock, A.; Donahue, J. and Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In:. 2019.

[11] Corvi, R. et al. On The Detection of Synthetic Images Generated by Diffusion Models. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2023, p. 1–5.

[12] Goodfellow, I. J. et al. Generative adversarial networks. *Communications of the ACM*, 2020, vol. 63, p. 139 – 144. https://consensus.app/papers/networks-goodfellow/04a0119e913b549b90b3c0b6c6f42ac7/?utm_source=chatgpt.

[13] Hong, W.; Ding, M.; Zheng, W.; Liu, X. and Tang, J. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. In: *The Eleventh*

International Conference on Learning Representations. 2023. Available at: https://openreview.net/forum?id=rB6TpjAuSRy.

[14] Jin, L.; Tan, F. and Jiang, S. Generative Adversarial Network Technologies and Applications in Computer Vision. *Computational Intelligence and Neuroscience*. Hindawi, august 2020, vol. 2020, p. 1459107. Available at: https://doi.org/10.1155/2020/1459107.

[15] Karras, T.; Laine, S. and Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, p. 4401–4410. Available at: https://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html.

[16] Koo, H. and Kim, T. E. A Comprehensive Survey on Generative Diff usion Models for Structured Data. *ArXiv preprint arXiv:2306.04139*, 2023.

[17] Lucataco et al. *Stable Diffusion x4 Upscaler*. 2022. Available at: https://huggingface.co/lucataco/stable-diffusion-x4-upscaler.

[18] Nichol, A. and Dhariwal, P. *Improved Denoising Diffusion Probabilistic Models*. 2021.

[19] NightCafe Studio. *Create with AI in NightCafe Creator Studio* https://creator.nightcafe.studio/studio. Accessed: 2024-03-21.

[20] Song, J. and Ermon, S. Denoising Diffusion Implicit Models. *ArXiv preprint arXiv:2010.02502*, 2021.

[21] Stability AI. *Stable Diffusion V3* https://stability.ai/stable-image. Accessed: 2023-12-07.

[22] Wang, X. et al. *Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data*. 2022. Available at: https://ar5iv.labs.arxiv.org/html/2107.10833.

[23] Wang, Z. et al. Stochastic Conditional Generative Networks with Basis Decomposition. *ArXiv*, 2020, abs/1909.11286. https://consensus.app/papers/conditional-generative-networks-basis-decomposition-wang/f77f6a2bc2c65a71b9fd469479b4012f/?utm_source=chatgpt.