الجامعة المصرية اليابانية للعلوم و التكنولوجيا
エジプト日本科学技術大学
EGYPT-JAPAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

THESIS REPORT
“COMPUTER ENGINEERING DEPARTMENT”

2023 / 2024

On 11/06/2024

Cross-Domain Image-to-Image Translation with VQGAN and BBDM

Presented by:

Mostafa Ahmed Kotb - 120200043
Abderahman Said Ahmed - 120200075
Ahmed Hamdy Rashed - 120200225
Mariam Ayman Mostafa - 120200094
Ashrakat Saeed Elfawal - 120200091

Advisor: Dr. Ahmed Fares

Abstract

Image-to-image translation remains one of the core challenges in computer vision and image processing. Diffusion models represent one of the most promising directions for high-quality image synthesis and have quickly displayed excellent performance on a variety of relevant tasks. Yet, existing diffusion models mainly consider image-to-image translation as conditional generation processes and face the challenge of domain shift.

In this thesis, we propose a novel image-to-image translation approach based on a concept called the Brownian Bridge Diffusion Model (BBDM). Our approach conceptualizes image translation as a stochastic Brownian bridge process and overcomes the limitations of conditional generation by learning the translation between different domains directly via bidirectional diffusion.

We present diverse experiments on different benchmarks to show that our proposed BBDM model is effective. Our method consistently outperforms other approaches under both qualitative visual inspection and quantitative evaluation metrics, showing potential to be used as a robust solution for image-to-image translation tasks.

The BBDM framework is a groundbreaking method in image-to-image translation. Using ideas from the definition of stochastic processes, our framework improves not only on generated image quality but also on generalization across different domains. Second, the natural bidirectionality of the diffusion process allows our model to capture intricate relationships between the source and target domains, leading to subtle and context-relevant translations.

Finally, in addition to its performance improvement, the diffusion of BBDM provides natural interpretability in the image translation process. Due to the decompensation of the translation task into a sequence of diffusion steps, our model reveals the dynamics of image translation over domains, enabling practitioners to gain an improved understanding of the underlying data manifold.

In conclusion, the BBDM framework opens up wide opportunities for the further development of research on image-to-image translation. Future directions can focus on improving the diffusion process to support complex data distributions and generalizing the framework to cover multimodal translation. In addition, reinforcement learning and self-supervised learning methods can be integrated to improve the generalization properties of the BBDM model, leading to its application in real-world problems in various domains.

Keywords: Brownian Bridge, Image to Image Translation, Diffusion Model

Contents

Chapter 1

Introduction

Image-to-image translation is the problem of learning a mapping function between two distinct image domains. It has a very wide range of applications, which include problems such as style transfer, semantic image analysis or data augmentation, and image restoration in super-resolution, colorization, inpainting, and sketch-to-image translation.[1] In recent months, image-to-image translation techniques have gained tremendous attention. One such common approach is Pix2Pix, where conditional adversarial neural networks learn the mapping between input and output images. Desired results are obtained by training with an adversarial loss function. Although these models have shown great performance and widespread use in many applications, model training of Pix2Pix is hard and often drops modes in the output distribution. Modest results are also obtained from this mapping as it does the job as a one-to-one mapping. In addition, the adversarial loss is under-constrained, which implies the existence of multiple possible mappings between the two domains, and that largely influences the instability of the training and makes the translation unsuccessful. Such a problem also occurs in CycleGANs, where the cycle consistency issue enforces the relationship between two domains to be bijective[1].

Recently, Diffusion Models (DMs) have shown competitive results and performance in image synthesis and translation.[2] DMs are taught through an iterative method of reverse diffusion image denoising. They have shown how, in producing high-quality mappings from randomly sampled Gaussian noise to target distributions that are very complicated, principled probabilistic diffusion modeling is so effective. Unlike GANs, they don't have training instabilities or mode-collapse.[3]However, conventional DMs are designed for picture synthesis, and diffusion operations are applied only on the whole images or feature maps, after which the reversal operations are executed. With this dense estimation framework, they can have considerably strong generation capabilities. Still, it also ensures countless numbers of iteration steps, typically ranging from 50 to 1000 steps, more so on the large denoising models, which consume a lot of computational power.[3]In addition, the structure of the typical DM, using some I2I tasks directly, is not suitable because in most of the cases, it leads to the production of unpleasant artifacts as well as having low efficiency and slow convergence.

On the other hand, the designed diffusion models are conditional, implying that the feature encoded by the reference image is integrated into the reverse process. This ensures very poor generalization and doesn't guarantee that the final diffusion result yields the desired conditional distribution. We can, therefore, explain the reasons why most of the conditional diffusion models show poor model generalization and are only suitable for a few applications, such as inpainting and super-resolution, where the conditional input and output show much similarity[4].

Since a multi-modal condition is projected and entangled via a complex attention mechanism, LDM improved model generalization by conducting a diffusion process in the latent space of some pre-trained models. But this makes LDM much harder to obtain such a theoretical guarantee. In the meanwhile, as the different latent feature levels bring about significant differences in the performance of LDM, it represents LDM is instable[3].

For our work, we propose a novel image-to-image translation framework based on the Brownian Bridge diffusion process. Our method models the mapping between the input and output domains through a Brownian Bridge stochastic process, directly without conditional generation, in contrast to existing diffusion methods. We perform the diffusion process in the same latent space and sample from latent space of VQGANs. However, it inherently distinguishes itself from LDM in the way we model the mapping between the two image domains.

The detailed frame of the Brownian Bridge Diffusion Model is shown in the second row of Figure 1. In the reverse diffusion process, we only condition on a reference image, which is randomly selected from the other domain B. But different from other related methods, we only use this conditioning as an initial point of the generation processing in our model, and we do not use this conditioning as input in the prediction network.

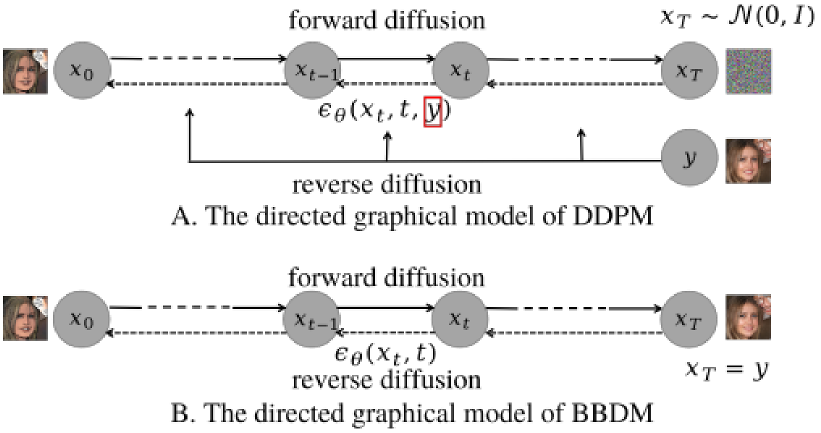
forward diffusion
$x_T \sim \mathcal{N}(0, I)$
x_0
x_{t-1}
x_t
x_T
$\epsilon_\theta(x_t, t, y)$
y
reverse diffusion
A. The directed graphical model of DDPM
forward diffusion
x_0
x_{t-1}
x_t
x_T
$\epsilon_\theta(x_t, t)$
reverse diffusion
$x_T = y$
B. The directed graphical model of BBDM

Figure 1.1: BBDM and DDPM Visualization

The major contributions of the paper are as follows:

- **Novel Methodology:** A novel image-to-image translation approach based on the Brownian Bridge diffusion process is proposed.

- **Direct Domain Translation:** For image-to-image translation tasks, we model it as statistical Brownian Bridge procedures to study domain translation directly. This strategy omits the conditional information leverage that related works exploit with conditional diffusion models.

- **Performance Evaluation:** The proposed BBDM technique is further validated on different image-to-image translation tasks through quantitative and qualitative experiments, and the competitive performance is demonstrated.

In particular, we apply the BBDM framework to the CelebAMask-HQ dataset to investigate its applicability to the problem of high-resolution image synthesis from masked inputs. In the context defined here, the BBDM framework translates masked images in the input domain to new, realistic facial images in the output domain. Now, with the utilization of the latent space, BBDM enables the diffusion process to occur at a faster rate, producing results that are much faster—free from noise, as well[5].

The experiments on the CelebAMask-HQ dataset show that BBDM is capable of obtaining diverse and visually appealing solutions while keeping important data introduced by the masks intact. This has strong practical implications in, for instance, data augmentation, semantic analysis of images, and high-end image restoration. All in all, the use of Brownian Bridge diffusion for the problem of image-to-image translation is a huge leap in this domain, making it a solid and efficient solution.

Chapter 2

Related Work

Image-to-image translation tasks have been around since the early 2000s, introducing simpler tasks like inpainting, super-resolution, colorization, and uncropping. It grew rapidly since the introduction of GANs (Generative Adversarial Networks). Since then, the tasks have become more complex and context-aware, such as style transfer and domain-to-domain transfer.

U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-instance Normalization for Image-to-image Translation

In April 2020, Junho Kim et al. proposed a novel architecture for unsupervised image-to-image translation. This new architecture incorporates an innovative attention module along with a new learnable normalization function [6]. The attention module identifies important regions in both the source and target images through attention maps provided by an auxiliary classifier. This enables the model to focus on these significant regions and assign less weight to less important areas. Unlike other attention-based architectures that struggle with geometric transformations between domains, this novel architecture allows the model to translate images requiring substantial shape changes. Additionally, the new normalization function, Adaptive Layer-Instance Normalization (AdaLIN), enables the attention module to flexibly control the extent of changes in shape and texture.

The architecture (Figure 2.1) consists of two generators and two discriminators, with the attention module integrated into both generators and discriminators. The attention module in the discriminator guides the generator to concentrate on regions crucial for generating realistic images. Each generator-discriminator pair handles a specific translation direction: one for translating from domain X to domain Y, and the other for the reverse direction. The attention module in the generator focuses on regions that distinguish one domain from the other. The generator's translation model comprises an encoder, a decoder, and an auxiliary classifier. The residual blocks are equipped with AdaLIN, with parameters γ and β dynamically computed by a fully connected layer based on the attention map.

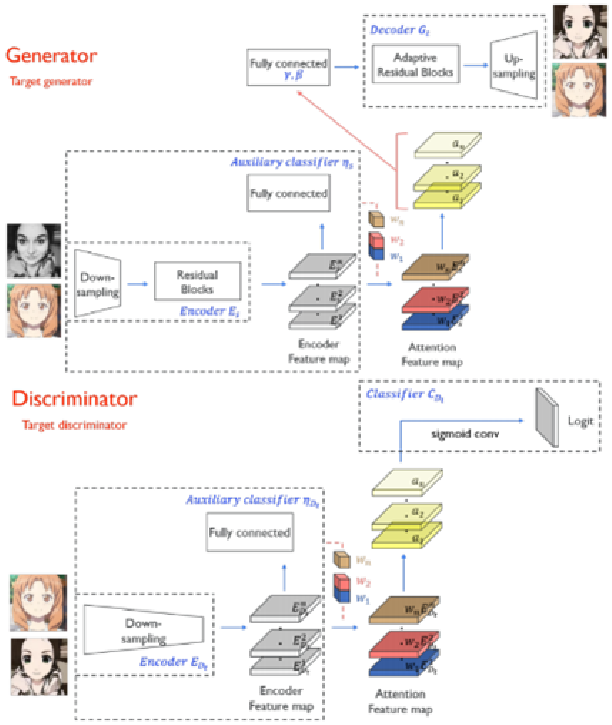

Figure 2.1: The model architecture of U-GAT-IT.

The loss function employs a combination of Adversarial Loss, Cycle Loss, Identity Loss, and CAM Loss to achieve optimal visual results.

Upon testing, the model demonstrates superior performance, outperforming many well-known architectures such as CycleGAN and AGGAN both qualitatively and quantitatively. This is evident across various datasets, including selfie2anime and horse2zebra. These advancements are attributed to the use of the attention module, which guides the model on the importance of different regions, and AdaLIN, which facilitates the translation process for datasets with varying degrees of geometric and style changes.

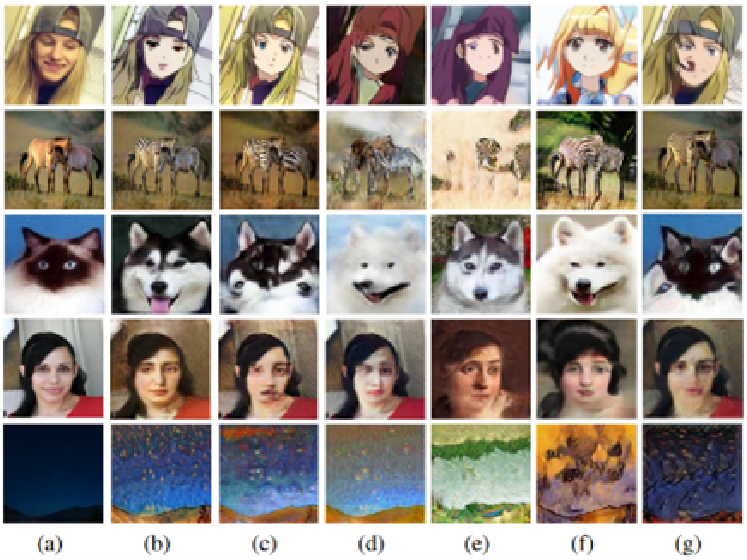

Figure 2.2: Visual comparisons on the 5 datasets: (a) Source images, (b) U-GAT-IT, (c) CycleGAN, (d) UNIT, (e) MUNIT, (f) DRIT, (g) AGGAN.

Dual Contrastive Learning for Unsupervised Image-to-Image Translation

In April 2021, Junlin Han et al. introduced a novel technique for unpaired image-to-image translation that achieved state-of-the-art results in unsupervised image-to-image translation [7]. This technique builds upon CUT (Contrastive learning for Unpaired image-to-image Translation) by incorporating a dual learning setting, which enhances the efficiency of mapping between unpaired data and addresses issues such as mode collapse. DCLGAN (Dual Contrastive Learning GAN) aims to maximize mutual information by learning the correspondence between input and output image patches using separate embeddings.

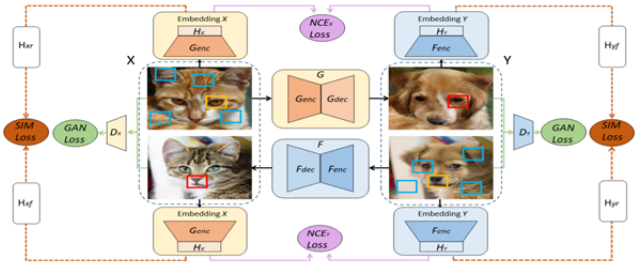

Figure 2.3: Overall architecture of DCLGAN.

The architecture of DCLGAN (Figure 2.3) consists of two generators and two discriminators, facilitating dual-direction translation. Each generator comprises an encoder followed by a decoder. Features extracted from four layers of the encoder are fed into a two-layer MLP to generate embeddings.

DCLGAN employs a combination of three different loss functions: Adversarial Loss, PatchNCE Loss, and Identity Loss. This combination aims to achieve the best qualitative results.

Testing results demonstrate the stability of the generated images, with DCLGAN outperforming several baseline models, including the original CUT model.

An extended version of the model, designed to address mode collapse, is known as SimDCL. Neither CUT nor DCLGAN effectively prevent mode collapse in tasks such as Photo → Label. SimDCL, however, shows superior results on the Facade → Label task, where outputs from both CUT and DCLGAN are nearly identical and less plausible. SimDCL proves to be more robust against mode collapse compared to other methods based on mutual information maximization.

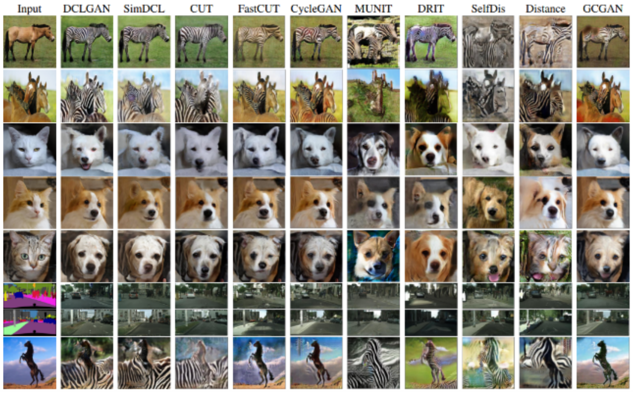

Figure 2.4: Comparison to all baselines on different datasets.

By utilizing a dual setting, DCLGAN achieves better results in contrastive learning for unsupervised unpaired image-to-image translation. SimDCL specifically addresses the mode collapse issue present in previous architectures.

StyleDiffusion: Controllable Disentangled Style Transfer via Diffusion Models

In August 2023, Zhizhong Wang et al. introduced StyleDiffusion, a novel framework for style transfer that focuses on disentangling content and style (C-S) using diffusion models [4]. This method addresses challenges in existing approaches that rely on explicit definitions (e.g., Gram matrix) or implicit learning (e.g., GANs), which are often not interpretable or controllable. StyleDiffusion provides a more interpretable and controllable C-S disentanglement without previous assumptions.

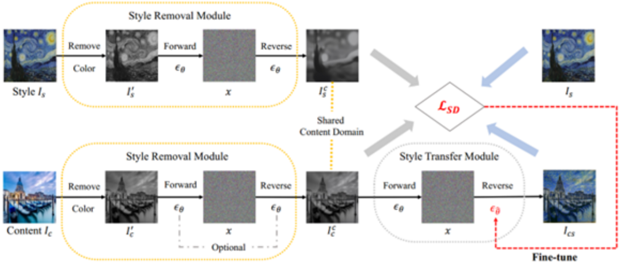

Figure 2.5: Overview of StyleDiffusion.

The architecture of StyleDiffusion consists of three main components:

1. **Diffusion-based Style Removal Module**: This module explicitly extracts content by removing style information from the input images. It uses a pre-trained diffusion model to eliminate domain-specific characteristics, aligning the content to a common domain while maintaining the structural integrity of the input images.

2. **Diffusion-based Style Transfer Module**: This component leverages the generative power of diffusion models to transfer the extracted style onto a new content image. The process allows for flexible control over the degree of style application.

3. **CLIP-based Style Disentanglement Loss**: Coordinated with a style reconstruction prior, this loss function ensures the effective disentanglement of content and style in the CLIP image space. It facilitates better control and quality in the style transfer process.

The method utilizes a combination of deterministic DDIM sampling for the reverse process and ODE approximation for the forward process to ensure efficient and accurate content extraction. This results in high-quality stylizations that are both interpretable and controllable, addressing the limitations of previous methods that often led to entangled representations and less satisfying results.

An ablation study demonstrates the control over C-S disentanglement by adjusting the return step of the style removal module, allowing for varying levels of style characteristics to be transferred. The framework also shows superiority in quality and flexibility compared to other models, achieving better stylizations and being more robust to issues like mode collapse.

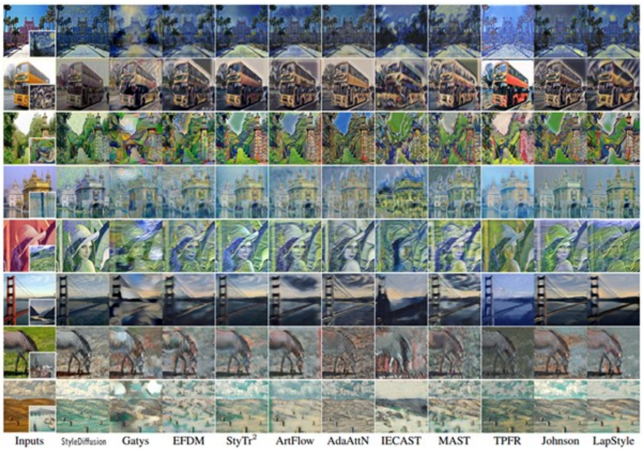

Figure 2.6: Qualitative comparisons with state of the art

Experimental results verify the effectiveness and superiority of StyleDiffusion against state-of-the-art methods, particularly in handling challenging styles. However, the framework still faces limitations, such as the need for fine-tuning for each style and inefficiency due to the use of diffusion models. Future work aims to address these issues and explore applications in other image translation and manipulation tasks.

DiffI2I: Efficient Diffusion Model for Image-to-Image Translation

In August 2023, Bin Xia et al. introduced DiffI2I, a novel approach to improving the efficiency and performance of diffusion models in image-to-image translation tasks [3]. While diffusion models have shown state-of-the-art results in image synthesis, their application to I2I tasks such as super-resolution and inpainting has been less effective due to inefficiencies and artifacts caused by extensive iterations and large denoising models.

DiffI2I addresses these challenges by introducing a compact and efficient diffusion model framework designed specifically for image-to-image tasks. The architecture consists of three key components: a compact image-to-image prior extraction network (CPEN), a dynamic image-to-image transformer (DI2Iformer), and a denoising network. The training process is divided into two stages: pretraining and diffusion model training.

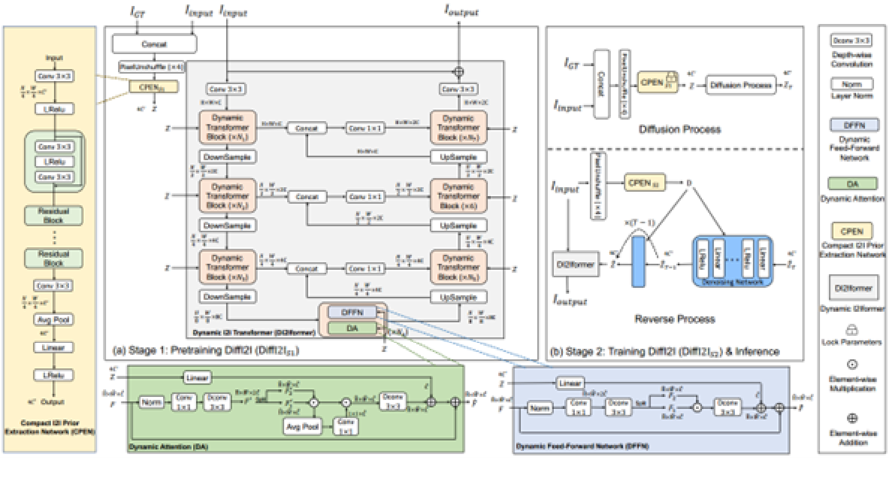

Figure 2.7: Overview of the proposed DiffI2I

During pretraining, both ground-truth and input images are fed into the CPEN to capture a compact image-to-image prior representation (IPR), which guides the DI2Iformer. In the diffusion model training stage, the model is trained to use only the input images to estimate the IPR, enabling the use of a lighter denoising network and fewer iterations compared to traditional diffusion models. This results in more accurate and efficient outcomes.

The DI2Iformer is equipped with a Dynamic Feed-Forward Network (DFFN) and Dynamic Attention (DA), which allow it to effectively utilize the IPR. Unlike traditional diffusion model-based methods that separate the training of the denoising network and the decoder, DiffI2I performs joint optimization, enhancing the robustness of error estimation.

Extensive experiments demonstrate that DiffI2I achieves state-of-the-art performance across various image-to-image tasks, including inpainting, super-resolution, and semantic segmentation, while significantly reducing computational costs. Notably, DiffI2I shows a remarkable efficiency improvement of 3500 times over existing methods like RePaint.

The authors highlight the generality and extendibility of DiffI2I, showing its applicability to real-world super-resolution, semantic segmentation, and depth estimation. The comprehensive evaluation and benchmark experiments underscore its superiority in both performance and runtime efficiency.

Image-to-Image Translation with Deep Reinforcement Learning

In February 2024, Xin Wang et al. introduced a novel approach to Image-to-Image Translation (I2IT) using deep reinforcement learning (DRL) [8]. Traditional I2IT methods typically generate images in a single run of deep learning models, which often leads to overfitting and high computational costs. Inspired by the analogy between diffusion models and reinforcement learning, the authors reformulate I2IT as an iterative decision-making problem, resulting in the RL-I2IT framework.

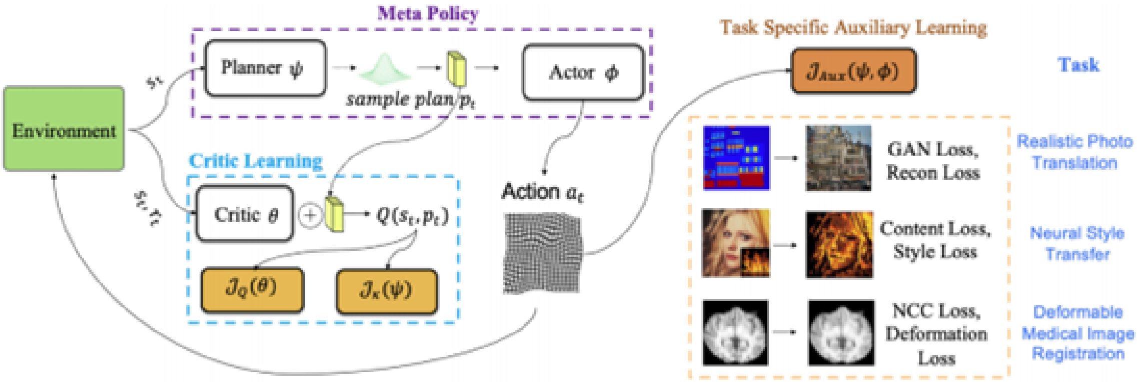

Figure 2.8: RL-I2IT framework with a Planner-Actor-Critic structure.

The RL-I2IT framework decomposes the monolithic learning process into smaller steps using a lightweight model to progressively transform the source image into the target image. This approach addresses the challenge of handling high-dimensional continuous state and action spaces, which is difficult for conventional RL frameworks. The core components of the RL-I2IT framework include a planner, an actor, and a critic, forming a meta-policy structure that divides decision-making into two steps: state to plan and plan to action.

The architecture includes three main neural networks:

1. **Planner**: Generates a high-level plan in low-dimensional latent space to guide the actor.
2. **Actor**: Uses the plan to generate high-dimensional actions that interact with the environment.
3. **Critic**: Evaluates the plan rather than the action, making it easier to learn the value function in the low-dimensional latent space.

The meta-policy and auxiliary learning strategy, which can incorporate advanced losses or objectives, stabilize the training process and improve performance across various I2IT tasks.

The RL-I2IT framework was tested on several I2IT tasks, including face inpainting, neural style transfer, and deformable image registration. Experimental results demonstrated the framework's effectiveness and robustness in handling high-dimensional continuous action spaces. The proposed method outperformed several baseline models in both stability and quality of generated images.

Key results include:

- **Face Inpainting**: The framework achieved superior PSNR and SSIM scores compared to existing state-of-the-art methods, producing more realistic and semantically consistent inpainted images.
- **Realistic Photo Translation**: The model showed significant improvements over traditional methods like pix2pix and pix2pixHD, with fewer parameters and lower computational complexity.
- **Neural Style Transfer (NST)**: RL-I2IT provided flexible control over the degree of stylization, producing high-quality results with fewer parameters and faster inference times.

While the RL-I2IT framework shows promising results, future work will focus on addressing limitations such as the need for fine-tuning for specific tasks. The authors plan to explore additional applications in other image translation and manipulation tasks, further refining and extending the capabilities of the RL-I2IT framework.

Chapter 3

Methodology

Study Design

Our study presents a novel approach to cross-domain image-to-image translation that takes advantage of the capabilities of Generative Adversarial Networks (GANs). We use the Vector Quantized Generative Adversarial Network (VQGAN)[9] architecture in conjunction with the Brownian Bridge Diffusion Model (BBDM) to achieve high-quality image translation. We use the VQGAN model architecture and the CelebAMask-HQ dataset to generate a discrete latent space representation. This latent space is then fed into the BBDM for further refinement and improvement of image quality. The integration of VQGAN and BBDM enables us to effectively explore and utilize the latent space for domain enhancement.

Participants

The CelebAMask-HQ dataset, which contains high-resolution images annotated with facial masks, is the primary data source. This dataset was chosen for its extensive and diverse collection of facial images, which are essential for training and evaluating generative models.

3.1 VQGAN Architecture

Overview

- Vector Quantized Variational Autoencoder (VQ-VAE): VQ-VAE is an important component of the VQGAN architecture because it encodes images into a discrete latent space. Unlike traditional autoencoders, which use continuous latent spaces, VQ-VAE uses a quantization process to limit each latent variable to a specific set of values. This quantization improves the model's ability to learn meaningful and distinct representations, which is necessary for creating high-quality images.

  1. Encoder: The encoder converts the input image to a latent space. This compression is carried out using convolutional neural networks (CNNs), which capture the image's key features and structures. The encoder produces a set of latent vectors that represent the image in lower-dimensional space. This representation retains the essential information needed for the subsequent decoding process, ensuring that critical details and patterns are preserved while reducing the dimensionality.

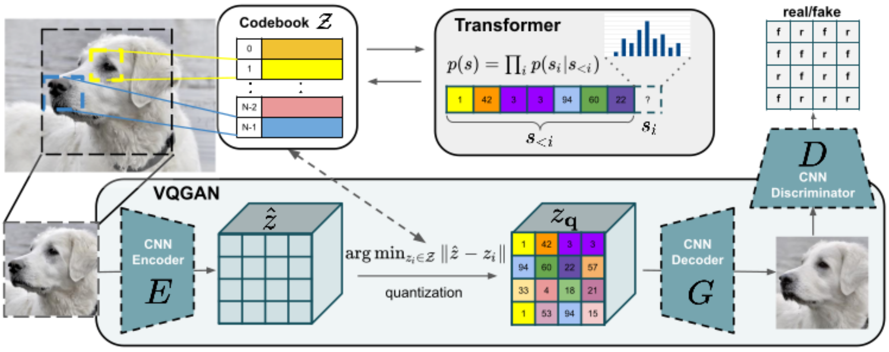

Figure 3.1: VQGAN Architecture

2. Quantizer: It takes the encoder's latent vectors and maps them to the nearest vectors in a predefined codebook. This discrete representation simplifies the latent space, making it easier for the model to learn and generate realistic images. The quantization procedure entails reducing the distance between the latent vectors and their corresponding codebook entries.

3. Decoder: It reconstructs the image using the quantized latent space. It employs transposed convolutional layers to convert the discrete latent vectors back into images. The decoder's goal is to generate images that are very similar to the original inputs, ensuring that the encoded and quantized representations capture the important features of the input images.

- Generative Adversarial Network (GAN): The GAN component of VQGAN enhances the realism and diversity of the generated images. It consists of two main parts: the generator and the discriminator.

  1. Generator: The generator takes the quantized latent codes and produces images from them. It aims to generate images that are indistinguishable from real images. It is then trained to minimize the adversarial loss, which measures the difference between the generated images and real images according to the discriminator.

  2. Discriminator: The discriminator's role is to differentiate between real images and those generated by the generator. It provides feedback to the generator, indicating how realistic the generated images are. The discriminator is trained to maximize the adversarial loss, effectively acting as an adversary to the generator. Through this adversarial training, the generator learns to produce increasingly realistic images.

- Adversarial Training Process: The adversarial training process is a key feature of the VQGAN architecture. It involves a minimax game between the generator and the discriminator where each of them has an object to achieve:

  1. Generator Objective: The generator aims to produce images that the discriminator cannot distinguish from real images. It is trained to minimize the adversarial loss, which includes a reconstruction loss and a GAN loss. The reconstruction loss ensures that the generated images are similar to the input images, while the GAN loss encourages the generator to produce realistic images.

  2. Discriminator Objective: The discriminator aims to accurately classify real images and generated images. It is trained to maximize the adversarial loss, which measures the accuracy of its classifications. By providing feedback to the generator, the discriminator helps improve the quality and realism of the generated images.

Advantages of VQGAN

VQGAN offers several advantages that make it unique and highly effective for image synthesis and translation tasks.

1. High-Quality Images: The combination of VQ-VAE's discrete latent space and GAN's generative capabilities results in images with high fidelity. The quantization process helps preserve fine details and structures, reducing blurriness and artifacts commonly seen in traditional GAN outputs.

2. Diverse Image Generation: The adversarial training process ensures that the generated images are not only high-quality but also diverse. The generator learns to capture a wide range of variations present in the real data, making VQGAN suitable for applications that require diverse image synthesis.

3. Efficient Representation: The discrete latent space provided by VQ-VAE allows for efficient and compact image representations. This efficiency is crucial for tasks that involve large datasets or require real-time image generation.

4. Robustness: VQGAN is particularly effective in scenarios where fine details and high-quality reconstructions are essential. For example, in the CelebAMask-HQ dataset, which contains high-resolution facial images, VQGAN excels at preserving intricate facial details and structures, making it ideal for applications in facial recognition, avatar creation, and more.

3.2 Brownian Bridge Diffusion Model (BBDM)

Overview

The Brownian Bridge Diffusion Model (BBDM) is a novel approach to image-to-image translation that takes advantage of the Brownian bridge process. Apart from traditional diffusion models[10][11][12], which frequently treat image-to-image translation as a conditional generation task, BBDM deals with translation as a stochastic Brownian bridge process. Instead of relying on a conditional input, this method learns the translation between two image domains directly via a bidirectional diffusion process. By performing the diffusion process in the latent space of a pre-trained VQGAN, BBDM achieves high-quality image synthesis while addressing the shortcomings of existing methods. The BBDM framework includes several key components and processes:

1. Latent Space Mapping: One of the key characteristics of BBDM is that it operates within the latent space of a pre-trained VQGAN. This latent space mapping dramatically improves learning efficiency and model generalization. By focusing on the critical features contained within the latent space, BBDM ensures that the diffusion process is both efficient and effective. The latent space representation captures the images' core attributes, allowing the model to learn meaningful mappings across image domains. The latent space, as a lower-dimensional representation of the images, reduces computational complexity and allows for more effective learning of the underlying patterns and structures. This

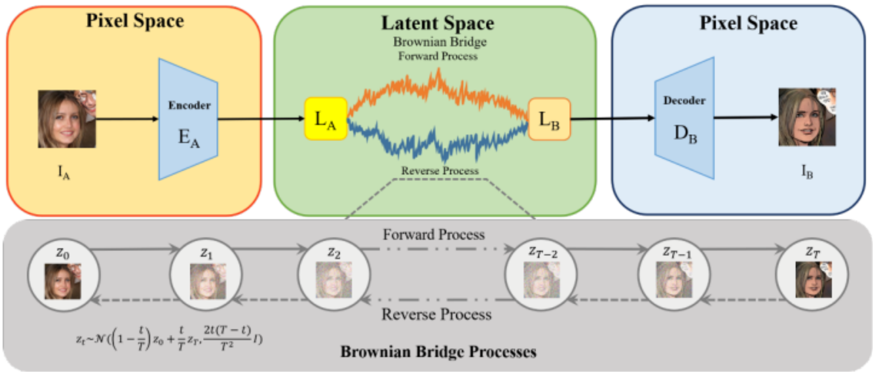

Figure 3.2: BBDM Architecture

approach differs from operating directly in the pixel space, where high dimensionality can introduce noise and make learning more difficult. Working within this compressed and abstract representation allows the model to focus on the most important features, resulting in improved performance and faster convergence

2. Forward Diffusion Process: The forward diffusion process in BBDM begins with a clean image from domain A and progresses to produce a corresponding image in domain B. This process is described as a Brownian bridge, a stochastic process with fixed starting and ending points. The forward diffusion process can be thought of as building a bridge between the source and target domains, with intermediate states created by gradually approaching the starting image to the target image. As the process progresses, the model gradually adjusts the image, adding controlled amounts of noise and transforming it so that it increasingly resembles the target domain. This step-by-step transformation is critical because it allows the model to learn the intricate details of the target domain while retaining the structure and content of the source image. At the start of the process, the state is identical to the initial image. As the process proceedes, the state transitions towards the target image, driven by noise reduction and process parameter adjustments. This ensures a smooth and continuous transition from the source domain to the target domain, resulting in an accurate representation of the target image.

3. Reverse Diffusion Process: The reverse diffusion process is the inverse of the forward process, attempting to reconstruct the source image from the target image by iteratively denoising the intermediate states. This procedure entails predicting and removing noise at each step, gradually refining the target image to a high-quality image that matches the original domain. This reverse process begins with the target image and employs a learned model to predict the previous state. This prediction entails estimating the noise component and subtracting it from the current state, thereby denoising the image. The model iteratively performs this denoising step, gradually refining the image and removing noise with each iteration. The iterative nature of the reverse diffusion process is important because it allows for fine-tuned adjustments at each stage. By removing noise and making precise corrections, the model ensures that the final image is a clear and high-quality representation of the source domain. This process continues until the image is completely transformed back into its original domain, yielding a high-fidelity reconstruction. The reverse diffusion process is guided by a neural network that has been trained to maximize the likelihood of correctly predicting the previous state given the current state and target image. This network learns to recognize patterns and correlations in the data, allowing it to make accurate predictions and effectively process the images. The iterative refinement ensures that noise is gradually removed, yielding a clear and high-quality image that is

consistent with the source domain.

4. The primary training goal for the Brownian Bridge Diffusion Model (BBDM) is to optimize the Evidence Lower Bound (ELBO). This objective is critical because it directs the model to accurately predict noise and reconstruct images at each stage of the diffusion process, resulting in high-quality output. The ELBO optimization ensures that the model learns a strong and consistent mapping between the two image domains.

Advantages of BBDM

The Brownian Bridge Diffusion Model (BBDM) outperforms traditional image-to-image translation methods, making it an effective tool for a variety of image synthesis tasks. These advantages are due to the Brownian bridge process's unique properties and the efficient use of a pre-trained VQGAN's latent space.

1. High-Quality Image Generation: One of the most notable benefits of BBDM is its ability to produce high-quality images. The Brownian bridge process, which underpins the BBDM framework, allows for seamless and continuous transitions between image domains. This process ensures that intermediate states are correctly transformed while maintaining the images' structural integrity and visual quality. As a result, the final output images are more realistic, and closely match the target domain's characteristics while preserving the key details from the source domain. This high-quality generation is especially useful in applications requiring visual accuracy and details, such as medical imaging, digital art, and photorealistic rendering.

2. Stable Training Process: Training Generative Adversarial Networks (GANs) can be difficult due to the adversarial nature of the loss functions used. GANs frequently suffer from mode collapse, in which the generator produces a limited number of outputs, and training instability, in which the generator and discriminator networks fail to converge properly. In contrast, BBDM provides a consistent training process by modelling image translation as a diffusion process. This method eliminates the need for adversarial loss functions, avoiding the associated training difficulties. BBDM's training process is stable, resulting in a higher performance and reliable convergence, making it easier to train and deploy in practical applications.

3. Efficient Use of Latent Space: BBDM conducts the diffusion process within a pre-trained VQGAN's latent space, taking advantage of the compact space's efficiency and representational power. The latent space of VQGAN captures the most important features of the images in a lower-dimensional representation, significantly reducing computational complexity. This efficient representation allows BBDM to concentrate on the most important aspects of the images, resulting in faster learning and better generalization across domains. By operating in the latent space, BBDM can effectively learn the mappings between source and target domains without being constrained by the raw pixel space's high dimension and noise.

4. Diversity of Outputs: The stochastic nature of the Brownian bridge process enables BBDM to produce a wide variety of outputs. Unlike deterministic models, which produce only one output for a given input, BBDM can capture the variations in the training data, resulting in multiple outputs for the same input image. This diversity is especially useful in creative applications like artwork generation, where a variety of styles and variations

are desired. It also improves the model's performance in tasks like data augmentation, where generating diverse examples can help machine learning models perform better.

3.3 BBDM with VQGAN: Advancing Image-to-Image Translation

Combining the Brownian Bridge Diffusion Model (BBDM) with the Vector Quantized Generative Adversarial Network (VQGAN) architecture is a significant step forward in image synthesis and translation. This integration builds on the strengths of both approaches, providing improved capabilities and performance in a variety of applications. We discuss some of these capabilaties and advatages in the following part:

1. Synergistic Model Fusion: BBDM and VQGAN work together to improve image synthesis by providing distinct advantages. VQGAN provides a compact and efficient latent space representation, capturing essential image features and allowing for effective domain mapping learning. Meanwhile, BBDM uses the Brownian bridge process to seamlessly translate images between domains, resulting in high-fidelity output with smooth transitions and fine-grained details. By combining these methodologies, BBDM with VQGAN achieves higher image synthesis quality and diversity than traditional approaches. The collaboration of these models enables high-quality image synthesis with sharp details and smooth transitions, ensuring that the generated images retain critical features and details from the original domain, whereas BBDM's stochastic process ensures smooth and natural transformations between image domains.

2. Cross-Domain Adaptability: One of the most important features of BBDM with VQGAN is its ability to adapt to a variety of image domains. Whether translating between different artistic styles, improving medical imaging modalities, or creating photorealistic scenes, the combined model excels at capturing the characteristics of each domain. BBDM with VQGAN's cross-domain adaptability makes it a powerful tool for a variety of creative, scientific, and practical applications, allowing users to explore and manipulate visual data in a wide range of contexts. For example, in artistic style transfer, the model can effectively translate images between various artistic styles, preserving the original content while adopting the target style's characteristics. In medical imaging, the model can improve images from various modalities, such as translating MRI scans to CT images, improving diagnostic accuracy, and assisting with medical research. The model can generate extremely detailed and lifelike images for tasks that require high realism, such as virtual reality environments or CGI in films.

3. Enhanced Learning Dynamics: By operating within the latent space of a pre-trained VQGAN, BBDM with VQGAN achieves improved learning dynamics and convergence properties. The compact latent representation enables efficient gradient propagation and parameter optimization, resulting in faster training convergence and better model generalization. Furthermore, the structured nature of the latent space facilitates better disentanglement of latent factors, allowing the model to learn meaningful representations and semantic relationships between images. These improved learning dynamics allow BBDM with VQGAN to outperform standalone models on challenging image synthesis tasks. The model's improved learning dynamics allow it to perform better on difficult

image synthesis tasks than standalone models. VQGAN's structured latent space reduces overfitting and ensures that the model generalizes well to new, previously unseen data.

4. Robustness and Stability: BBDM with VQGAN inherits the robustness and stability properties of both BBDM and VQGAN, increasing its practical reliability. The diffusion-based approach of BBDM ensures stable training dynamics, mitigating issues like mode collapse and training instability, which are common in traditional GAN frameworks. Similarly, the structured latent space provided by VQGAN promotes robust learning and efficient representation of image semantics, lowering the risk of overfitting while improving model generalization. This combination of robustness and stability makes BBDM with VQGAN ideal for real-world applications requiring consistent performance and reliability. The combination of robustness and stability makes BBDM with VQGAN ideal for real-world applications requiring consistent performance and reliability.

Chapter 4

Results

In our study, we tackled complex image-to-image translation tasks, specifically focusing on semantic synthesis using the CelebAMask-HQ dataset[5]. We conducted a thorough performance comparison of our proposed method, the BBDM, against state-of-the-art baselines across various image-to-image translation tasks.

Semantic synthesis involves the generation of photorealistic images based on semantic layout. We presented the experimental results in the table below, highlighting the performance of BBDM and other baseline methods. Our findings revealed that while Pix2Pix yields reasonable results with paired training data, CycleGAN's[2] performance diminishes on small-scale datasets. DRIT++ demonstrates superior performance among GAN-based methods, but the translated images tend to be overly smooth and deviate from the true distribution of the target domain.

Our comparison also showed that diffusion-based methods exhibit competitive performance compared to GAN-based approaches. Both Conditional Diffusion Models (CDE) and Latent Diffusion Models (LDM) face difficulties in effectively integrating conditional information during the diffusion process, particularly when dealing with irregular occlusions.

In contrast, BBDM directly learns the diffusion process between the two domains and circumvents the challenges including leveraging conditional information. By leveraging the stochastic nature of the Brownian Bridge, our method can generate high-fidelity images and low LPIPS.

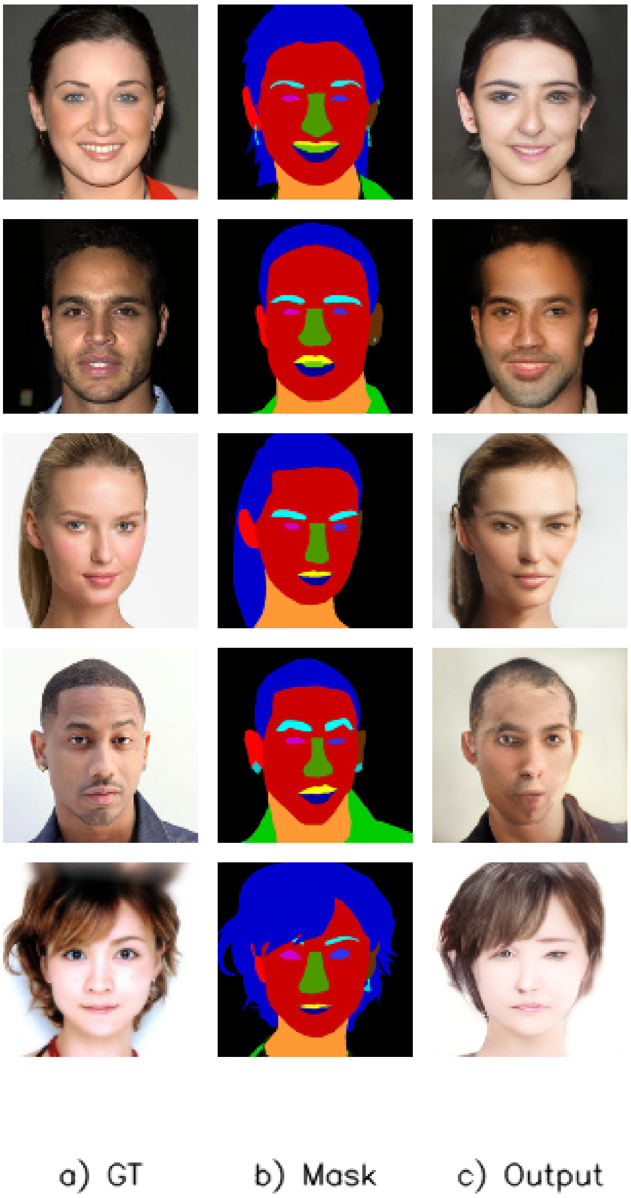

Figure 4.1: BBDM Sample Result

Table 4.1: Model Performance on CelebAMask-HQ Dataset

Model	**FID ↓**	**LPIPS ↓**
Pix2Pix	56.997	0.431
CycleGAN	78.234	0.490
DRIT++	77.794	0.431
SPADE	44.171	0.376
OASIS	27.751	0.384
CDE	24.404	0.414
LDM	22.816	0.371
BBDM	18.273	0.338

Chapter 5

Discussion

This report demonstrates substantial progress in image-to-image translation in semantic image synthesis, by examining the semantic synthesis task using the CelebAMask-HQ dataset. Previous research has shown the shortcomings of conventional techniques and outlined the difficulties in producing realistic, high-quality photographs using semantic patterns. Based on Generative Adversarial Networks (GANs), methods like Pix2Pix, CycleGAN, and DRIT++ have shown varying results. However, GANs are prone to mode collapse, lack of diversity, and over-smoothing, especially when working with complex, high-resolution images.

The use of the Brownian Bridge Diffusion Model (BBDM) represents a major change from these traditional techniques. Recent research by Dhariwal and Nichol (2021)[13] has demonstrated that diffusion models may produce high-fidelity images with fewer artifacts than GANs, and this has led to their increasing popularity. While using the ideas of conventional diffusion models, the BBDM takes a novel approach by using the Brownian Bridge process to translate images directly. In addition to addressing the limitations of GANs, this approach makes use of diffusion models' advantages, including increased image quality and variety.

Our findings from tests show that the BBDM regularly generates varied, high-quality images, especially in the CelebAMask-HQ dataset's semantic synthesis task. This dataset gave our model a reliable testing environment because of its intricacy and high-resolution face imagery. The findings demonstrated that in terms of quantitative indicators like FID and LPIPS, BBDM greatly beat numerous baseline models. While the LPIPS values verify that the perceptual quality of these pictures is on par with or better than state-of-the-art models like the Latent Diffusion Model (LDM), the lower FID scores show that BBDM produces images that are more akin to actual images.

It is important to recognize a few of this study's limitations, despite the encouraging findings. Firstly, BBDM training and inference demand a significant amount of processing power. A trade-off between computational cost and performance is highlighted by the requirement for 200 sample steps to balance quality and efficiency during the inference stage. Applications with low processing power, including real-time image synthesis or deployment on edge devices, may find this need to be prohibitive. This significant computational load highlights the necessity for resource-saving optimization strategies that don't sacrifice image quality.

The CelebAMask-HQ dataset, although extensive, only includes a particular subset of possible semantic layouts and image types, which is an additional limitation. Further testing is needed to see whether the BBDM is generalizable to additional datasets and domains with distinct characteristics. Although CelebAMask-HQ serves as a useful benchmark for face image synthesis, expanding the analysis to additional datasets—such as those with distinct objects, sceneries, or more diverse semantic layouts—would yield a more thorough knowledge of the model's strengths and limitations.

Moreover, even if the evaluation criteria employed in this work are generally acknowledged in the literature, it's possible that they don't fully capture all facets of image quality and variety. Although FID and LPIPS are strong measures of perceptual similarity and visual realism, they fall short of capturing user satisfaction and perception. Additional information regarding the perceived quality of the synthesized images might be obtained through subjective human evaluation, which was not included in this study. It would be easier to determine how well the generated images satisfy user expectations and aesthetic criteria if human judgments were included in subsequent research.

The implications of this research are diverse. Firstly, diffusion-based models may be a competitive and even better option than GAN-based models for high-fidelity image production, as evidenced by the proven effectiveness of BBDM in the semantic synthesis challenge on CelebAMask-HQ. This discovery may motivate additional investigation into the enhancement and implementation of diffusion models in diverse image synthesis uses, encompassing virtual reality content production and medical imaging, where superior synthetic data might be quite beneficial.

The capacity of BBDM to generate a variety of outputs from a single semantic input emphasizes its potential for creative fields where realism and unpredictability are essential, such as game design and animation. In these domains, the ability to produce an extensive variety of high-quality images from semantic layouts can greatly improve the creative process, allowing designers and artists to experiment with new ideas and optimize their processes.

Furthermore, BBDM's enhanced performance measures imply that it can improve on current facial recognition and enhancement applications, leading to more reliable and accurate systems. Realistic facial picture generation, for example, can help train more successful models in facial recognition systems, especially in situations when data augmentation is required to increase model robustness.

Conclusion

In conclusion, the Brownian Bridge Diffusion Model (BBDM) was presented as a unique image-to-image translation technique. The CelebAMask-HQ dataset was specifically used to address the semantic synthesis challenge. By using the Brownian Bridge diffusion process, BBDM uses a direct translation technique as opposed to conventional models, which rely on conditional generation processes. This model outperformed baseline models in important quantitative criteria like LPIPS and FID, producing high-fidelity and diversified images.

In this thesis, the main study issue was whether, in comparison to other models, the BBDM substantially improves the quality and diversity of images generated in semantic synthesis tasks. Based on the results, BBDM regularly outperformed contemporary diffusion models and conventional GAN-based models, obtaining equivalent LPIPS values and lower FID scores, indicating superior perceptual quality and realism in the generated images. Within the framework of the semantic synthesis challenge on the CelebAMask-HQ dataset, BBDM exhibited its resilience and effectiveness by generating various photorealistic images.

The significance of this research is that it has the potential to completely change a number of industries that depend on producing high-quality images. This research creates new possibilities for applications in creative industries, such as game design and animation, where the capacity to generate realistic and diverse visuals is essential. It does this by offering a model that yields superior outcomes in terms of both quality and diversity. Furthermore, BBDM can greatly increase the resilience and accuracy of these systems by offering high-quality synthetic data for training and augmentation, which has important implications for facial recognition and enhancement technologies. Furthermore, BBDM's proven performance in the semantic synthesis challenge points to potential uses in the healthcare sector, especially when it comes to creating training sets of synthetic medical images. This capacity can help with concerns about privacy and data scarcity, which will aid in the creation of more effective medical professional training programs and diagnostic tools.

Bibliography

[1] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2020.

[2] Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong, and C. Xu, "Inversion-based style transfer with diffusion models," 2023.

[3] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, R. Timotfe, and L. V. Gool, "Diffi2i: Efficient diffusion model for image-to-image translation," 2023.

[4] Z. Wang, L. Zhao, and W. Xing, "Stylediffusion: Controllable disentangled style transfer via diffusion models," 2023.

[5] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[6] J. Kim, M. Kim, H. Kang, and K. Lee, "U-GAT-IT: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," *CoRR*, vol. abs/1907.10830, 2019. [Online]. Available: http://arxiv.org/abs/1907.10830

[7] J. Han, M. Shoeiby, L. Petersson, and M. A. Armin, "Dual contrastive learning for unsupervised image-to-image translation," *CoRR*, vol. abs/2104.07689, 2021. [Online]. Available: https://arxiv.org/abs/2104.07689

[8] X. Wang, Z. Luo, J. Hu, C. Feng, S. Hu, B. Zhu, X. Wu, X. Li, and S. Lyu, "Image-to-image translation with deep reinforcement learning," 2024.

[9] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 12 873–12 883.

[10] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "ILVR: conditioning method for denoising diffusion probabilistic models," *CoRR*, vol. abs/2108.02938, 2021. [Online]. Available: https://arxiv.org/abs/2108.02938

[11] G. Batzolis, J. Stanczuk, C. Schönlieb, and C. Etmann, "Conditional image generation with score-based diffusion models," *CoRR*, vol. abs/2111.13606, 2021. [Online]. Available: https://arxiv.org/abs/2111.13606

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *CoRR*, vol. abs/2112.10752, 2021. [Online]. Available: https://arxiv.org/abs/2112.10752

[13] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf