BBDM: Image-to-Image Translation with Brownian Bridge Diffusion Models

Bo Li, Kaitao Xue, Bin Liu
School of Mathematics and Information Science, Nanchang Hangkong University, Nanchang, China

Yu-Kun Lai
School of Computer Sciences and Informatics, Cardiff University, Cardiff, UK

Abstract

Image-to-image translation is an important and challenging problem in computer vision and image processing. Diffusion models (DM) have shown great potentials for high-quality image synthesis, and have gained competitive performance on the task of image-to-image translation. However, most of the existing diffusion models treat image-to-image translation as conditional generation processes, and suffer heavily from the gap between distinct domains. In this paper, a novel image-to-image translation method based on the Brownian Bridge Diffusion Model (BBDM) is proposed, which models image-to-image translation as a stochastic Brownian Bridge process, and learns the translation between two domains directly through the bidirectional diffusion process rather than a conditional generation process. To the best of our knowledge, it is the first work that proposes Brownian Bridge diffusion process for image-to-image translation. Experimental results on various benchmarks demonstrate that the proposed BBDM model achieves competitive performance through both visual inspection and measurable metrics.

1. Introduction

Image-to-image translation [14] refers to building a mapping between two distinct image domains. Numerous problems in computer vision and graphics can be formulated as image-to-image translation problems, such as style transfer [3,9,13,22], semantic image synthesis [21,24,34,36,37,40] and sketch-to-photo synthesis [2, 14, 43].

A natural approach to image-to-image translation is to learn the conditional distribution of the target images given the samples from the input domain. Pix2Pix [14] is one of the most popular image-to-image translation methods. It is a typical conditional Generative Adversarial Network (GAN) [26], and the domain translation is accomplished by learning a mapping from the input image to the output image. In addition, a specific adversarial loss function is also trained to constrain the domain mapping. Despite the high fidelity translation performance, they are notoriously hard to train [1, 10] and often drop modes in the output distribution [23, 27]. In addition, most GAN-based image-to-image translation methods also suffer from the lack of diverse translation results since they typically model the task as a one-to-one mapping. Although other generative models such as Autoregressive Models [25, 39], VAEs (Variational Autoencoders) [16, 38], and Normalizing Flows [7, 15] succeeded in some specific applications, they have not gained the same level of sample quality and general applicability as GANs.

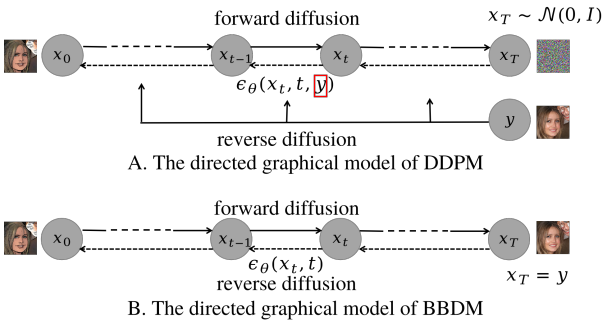

Figure 1. Comparison of directed graphical models of BBDM (Brownian Bridge Diffusion Model) and DDPM (Denoising Diffusion Probabilistic Model).

Recently, diffusion models [12, 31] have shown competitive performance on producing high-quality images compared with GAN-based models [6]. Several conditional diffusion models [2, 4, 28–30] have been proposed for image-to-image translation tasks. These methods treat image-to-image translation as conditional image generation by integrating the encoded feature of the reference image into the U-Net in the reverse process (the first row of Figure 1) to guide the diffusion towards the target domain. De-

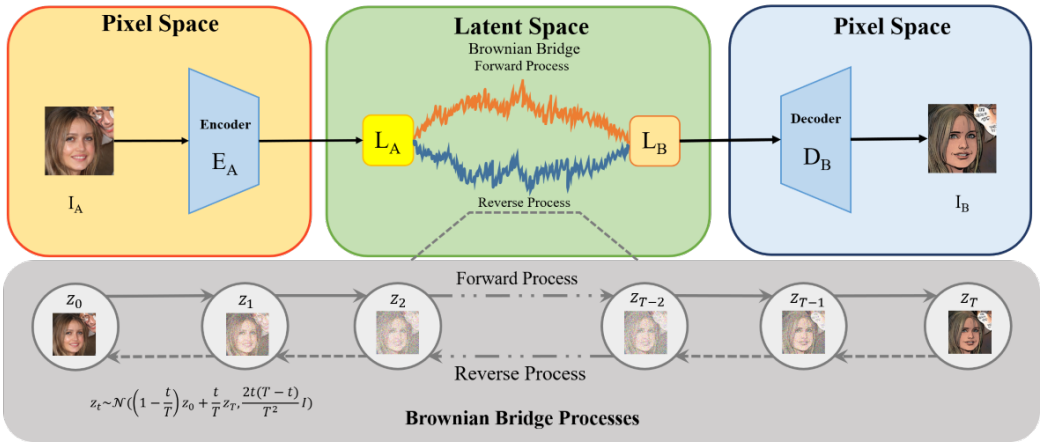

Figure 2. Architecture of our proposed Brownian Bridge Diffusion Model (BBDM).

spite some practical success, the above condition mechanism does not have a clear theoretical guarantee that the final diffusion result yields the desired conditional distribution. Therefore, most of the conditional diffusion models suffer from poor model generalization, and can only be adapted to some specific applications where the conditional input has high similarity with the output, such as inpainting and super-resolution [2, 4, 30]. Although LDM (Latent Diffusion Model) [28] improved the model generalization by conducting diffusion process in the latent space of certain pre-trained models, it is still a conditional generation process and the multi-modal condition is projected and entangled via a complex attention mechanism which makes LDM much more difficult to get such a theoretical guarantee. Meanwhile, the performance of LDM differs greatly across different levels of latent features showing instability.

In this paper, we propose a novel image-to-image translation framework based on Brownian Bridge diffusion process. Compared with the existing diffusion methods, the proposed method directly builds the mapping between the input and the output domains through a Brownian Bridge stochastic process, rather than a conditional generation process. In order to speed up the training and inference process, we conduct the diffusion process in the same latent space as used in LDM [28]. However, the proposed method differs from LDM inherently in the way the mapping between two image domains is modeled. The framework of BBDM is shown in the second row of Figure 1. It is easy to find that the reference image $\boldsymbol{y}$ sampled from domain B is only set as the initial point $\boldsymbol{x}_T = \boldsymbol{y}$ of the reverse diffusion, and it will not be utilized as a conditional input in the prediction network $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ at each step as done in related works [2, 4, 28, 30]. The main contributions of this paper include:

1. A novel image-to-image translation method based on Brownian Bridge diffusion process is proposed in this paper. As far as we know, it is the first work of Brownian Bridge diffusion process proposed for image-to-image translation.

2. The proposed method models image-to-image translation as a stochastic Brownian Bridge process, and learns the translation between two domains directly through the bidirectional diffusion process. The proposed method avoids the conditional information leverage existing in related work with conditional diffusion models.

3. Quantitative and qualitative experiments demonstrate the proposed BBDM method achieves competitive performance on various image-to-image translation tasks.

2. Related Work

In this section, we briefly review the related topics, including image-to-image translation, diffusion models and Brownian Bridge.

2.1. Image-to-image Translation

Isola *et al*. [14] firstly proposed a unified framework Pix2Pix for image-to-image translation based on conditional GANs. Wang *et al*. [40] extended the Pix2Pix framework to generate high-resolution images. Unpaired translation methods like CycleGAN [43] and DualGAN [41] used two GANs separately on two domains and trained them together with dual learning [11], which allows them to learn from unpaired data. However, these one-to-one mapping translation methods fail to generate diverse outputs. With the aim of generating diverse samples, Lee *et al*. [19] proposed DRIT++, but it requires that the condition image and result image must have high structural similarity. Several other GAN-based techniques have also been proposed

for image-to-image translation such as unsupervised cross-domain method [35], multi-domain method [5], few-shot method [20]. Nevertheless, GAN-based techniques suffer from the training instabilities and mode collapse problems. In addition to GAN-based models, diffusion models [31] have also achieved impressive results on image generation [6, 12], inpainting [29], super-resolution [29, 30], and text-to-image generation [28].

2.2. Diffusion Models

A T-step Denoising Diffusion Probabilistic Model (DDPM) [12] consists of two processes: the forward process (also referred to as diffusion process), and the reverse inference process.

The forward process from data $\boldsymbol{x}_0 \sim q_{data}(\boldsymbol{x}_0)$ to the latent variable $\boldsymbol{x}_T$ can be formulated as a fixed Markov chain:

$$q(\boldsymbol{x}_1, ..., \boldsymbol{x}_T|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \tag{1}$$

where $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I})$ is a normal distribution, β_t is a small positive constant. The forward process gradually perturbs $\boldsymbol{x}_0$ to a latent variable with an isotropic Gaussian distribution $p_{latent}(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

The reverse process strives to predict the original data $\boldsymbol{x}_0$ from the latent variable $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ through another Markov chain:

$$p_\theta(\boldsymbol{x}_0, ..., \boldsymbol{x}_{T-1}|\boldsymbol{x}_T) = \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \tag{2}$$

The training objective of DDPM is to optimize the Evidence Lower Bound (ELBO). Finally, the objective can be simplified as to optimize:

$$\mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{\epsilon}} ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)||_2^2$$

where $\boldsymbol{\epsilon}$ is the Gaussian noise in $\boldsymbol{x}_t$ which is equivalent to $\nabla_{\boldsymbol{x}_t} \ln q(\boldsymbol{x}_t|\boldsymbol{x}_0)$, $\boldsymbol{\epsilon}_\theta$ is the model trained to estimate $\boldsymbol{\epsilon}$.

Most conditional diffusion models [2, 4, 28–30] maintain the forward process and directly inject the condition into the training objective:

$$\mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{\epsilon}} ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{y}, t)||_2^2$$

Since $p(\boldsymbol{x}_t|y)$ does not obviously appear in the training objective, it is difficult to guarantee the diffusion can finally reach the desired conditional distribution.

Except for the conditioning mechanism, Latent Diffusion Model (LDM) [28] takes the diffusion and inference processes in the latent space of VQGAN [8], which is proven to be more efficient and generalizable than operating on the original image pixels.

2.3. Brownian Bridge

A Brownian bridge is a continuous-time stochastic model in which the probability distribution during the diffusion process is conditioned on the starting and ending states. Specifically, the state distribution at each time step of a Brownian bridge process starting from point $\boldsymbol{x}_0 \sim q_{data}(\boldsymbol{x}_0)$ at $t = 0$ and ending at point $\boldsymbol{x}_T$ at $t = T$ can be formulated as:

$$p(\boldsymbol{x}_t|\boldsymbol{x}_0, \boldsymbol{x}_T) = \mathcal{N}\big((1 - \frac{t}{T})\boldsymbol{x}_0 + \frac{t}{T}\boldsymbol{x}_T, \frac{t(T-t)}{T}\boldsymbol{I}\big) \tag{3}$$

It can be easily found that the process is tied down at both two ends with $\boldsymbol{x}_0$ and $\boldsymbol{x}_T$, and the process in between forms a bridge.

3. Method

Given two datasets $\mathcal{X}_A$ and $\mathcal{X}_B$ sampled from domains A and B, image-to-image translation aims to learn a mapping from domain A to domain B. In this paper, a novel image-to-image translation method based on stochastic Brownian Bridge diffusion process is proposed. In order to improve the learning efficiency and model generalization, we propose to accomplish the diffusion process in the latent space of popular VQGAN [8]. The pipeline of the proposed method is shown in Figure 2. Given an image $\mathbf{I}_A$ sampled from domain A, we can first extract the latent feature $\mathbf{L}_A$, and then the proposed Brownian Bridge process will map $\mathbf{L}_A$ to the corresponding latent representation $\mathbf{L}_{A\to B}$ in domain B. Finally, the translated image $\mathbf{I}_{A\to B}$ can be generated by the decoder of the pre-trained VQGAN.

3.1. Brownian Bridge Diffusion Model (BBDM)

The forward diffusion process of DDPM [12] starts from clean data $\boldsymbol{x}_0 \sim q_{data}(\boldsymbol{x}_0)$ and ends at a standard normal distribution. The setup of DDPM is suitable for image generation, as the reverse inference process naturally maps a sampled noise back to an image, but it is not proper for the task of image translation between two different domains. Most of the existing diffusion-based image translation methods [2, 4, 28, 30] improved the original DDPM model by integrating the reference image as a conditional input in the reverse diffusion process.

Different from the existing DDPM methods, a novel image-to-image translation method based on Brownian Bridge diffusion process is proposed in this section. Instead of ending at the pure Gaussian noise, Brownian Bridge process takes the clean conditional input $\boldsymbol{y}$ as its destination. We take similar notations as DDPM [12], and let $(\boldsymbol{x}, \boldsymbol{y})$ denote the paired training data from domains A and B. To speed up the training and inference process, we conduct diffusion process in the latent space of popular VQGAN [8]. For simplicity and following notations as in DDPMs, we

still use $\boldsymbol{x}, \boldsymbol{y}$ to denote the corresponding latent features ($\boldsymbol{x} := \mathbf{L}_A(\boldsymbol{x}), \boldsymbol{y} := \mathbf{L}_B(\boldsymbol{y})$). The forward diffusion process of Brownian Bridge can be defined as:

$$q_{BB}(\boldsymbol{x}_t|\boldsymbol{x}_0, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}_t; (1-m_t)\boldsymbol{x}_0 + m_t\boldsymbol{y}, \delta_t\boldsymbol{I}) \quad (4)$$
$$\boldsymbol{x}_0 = \boldsymbol{x}, \quad m_t = \frac{t}{T}$$

where T is the total steps of the diffusion process, δ_t is the variance. It is noticed that if we take the variance of original Brownian Bridge as shown in Eq.(3), $\delta_t = \frac{t(T-t)}{T}$, the maximum variance at the middle step, $\delta_{\frac{T}{2}} = \frac{T}{4}$, will be extremely large with the increase of T, and this phenomenon will make the BBDM framework untrainable. Meanwhile, it has been mentioned in DDPM [12] and VPSDE [33] that the variance of middle steps should be preserved to be identity, if the distribution of $\boldsymbol{x}_0$ is supposed to be a standard normal distribution. Therefore, assuming that $\boldsymbol{x}_0, \boldsymbol{y} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ are relatively independent, with the aim of preserving variances, a novel schedule of variance for Brownian Bridge diffusion process can be designed as

$$\begin{aligned}\delta_t &= 1 - \left((1-m_t)^2 + m_t^2\right) \\ &= 2(m_t - m_t^2)\end{aligned}$$

It is easy to find that at the start of the diffusion process, i.e., $t = 0$, we can have $m_0 = 0$, and the mean value is equal to $\boldsymbol{x}_0$ with probability 1 and variance $\delta_0 = 0$. When the diffusion process reaches the destination, $t = T$, we get $m_T = 1$, and the mean is equal to $\boldsymbol{y}$ while the variance $\delta_T = 0$. During the diffusion process, the variance δ_t will first grow to the biggest value at the middle time $\delta_{max} = \delta_{\frac{T}{2}} = \frac{1}{2}$, and then it will drop until $\delta_T = 0$ at the destination of the diffusion. According to the characteristic of Brownian Bridge diffusion process, the sampling diversity can be tuned by the maximum variance at the middle step $t = \frac{T}{2}$, therefore, we can scale δ_t by a factor s to control the sampling diversity in practice:

$$\delta_t = 2s(m_t - m_t^2) \quad (5)$$

We set $s = 1$ by default, and we will further discuss the influence of different s values for sampling diversity in Section 4.5.

3.1.1 Forward Process

According to the transition probability shown in Eq.(4), the forward diffusion of Brownian Bridge process only provides the marginal distribution at each step t. For training and inference purpose, we need to deduce the forward transition probability $q_{BB}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{y})$.

Given initial state $\boldsymbol{x}_0$ and destination state $\boldsymbol{y}$, the intermediate state $\boldsymbol{x}_t$ can be computed in discrete form as follows:

$$\boldsymbol{x}_t = (1-m_t)\boldsymbol{x}_0 + m_t\boldsymbol{y} + \sqrt{\delta_t}\boldsymbol{\epsilon}_t \quad (6)$$
$$\boldsymbol{x}_{t-1} = (1-m_{t-1})\boldsymbol{x}_0 + m_{t-1}\boldsymbol{y} + \sqrt{\delta_{t-1}}\boldsymbol{\epsilon}_{\mathbf{t-1}} \quad (7)$$

where $\boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_{t-1} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The transition probability $q_{BB}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{y})$ can be derived by substituting the expression of $\boldsymbol{X}_0$ in Eq.(6) by the corresponding formula in Eq.(7)

$$\begin{aligned}q_{BB}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}_t; &\frac{1-m_t}{1-m_{t-1}}\boldsymbol{x}_{t-1} \\ &+(m_t - \frac{1-m_t}{1-m_{t-1}}m_{t-1})\boldsymbol{y}, \delta_{t|t-1}\boldsymbol{I})\end{aligned} \quad (8)$$

where $\delta_{t|t-1}$ is calculated by δ_t as:

$$\delta_{t|t-1} = \delta_t - \delta_{t-1}\frac{(1-m_t)^2}{(1-m_{t-1})^2}$$

According to Eq.(8), when the diffusion process reaches the destination, i.e., $t = T$, we can get that $m_T = 1$ and $\boldsymbol{x}_T = \boldsymbol{y}$. The forward diffusion process defines a fixed mapping from domain A to domain B.

3.1.2 Reverse Process

In the reverse process of traditional diffusion models, the diffusion process starts from a pure noise sampled from a Gaussian distribution, and eliminates the noise step by step to get the clean data distribution. In order to model the conditional distribution, the existing methods [2, 4, 28, 30] take the condition as an additional input of the neural network in the reverse diffusion process.

Different from the existing diffusion-based image-to-image translation methods, the proposed Brownian Bridge process directly starts from the conditional input by setting $\boldsymbol{x}_T = \boldsymbol{y}$. Based on the main idea of denoising diffusion methods, the reverse process of the proposed method aims to predict $\boldsymbol{x}_{t-1}$ based on $\boldsymbol{x}_t$:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \tilde{\delta}_t\boldsymbol{I}) \quad (9)$$

where $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ is the predicted mean value of the noise, and $\tilde{\delta}_t$ is the variance of noise at each step. Similar to DDPM [12], the mean value $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ is required to be learned by a neural network with parameters θ based on maximum likelihood criterion. Although the variance $\tilde{\delta}_t$ does not need to be learned, it plays an important role in high-quality image translation. The analytical form of $\tilde{\delta}_t$ will be introduced in Section 3.1.3.

It is important to notice that the reference image $\boldsymbol{y}$ sampled from domain B is only set as the start point $\boldsymbol{x}_T = \boldsymbol{y}$ of the reverse diffusion, and it will not be utilized as a conditional input in the prediction network $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ at each step as done in related works [2, 4, 28, 30] (Figure 1).

3.1.3 Training Objective

The training process is performed by optimizing the Evidence Lower Bound (ELBO) for the Brownian Bridge diffusion process which can be formulated as:

$$\begin{aligned} ELBO = -\mathbb{E}_q\big(&D_{KL}(q_{BB}(\boldsymbol{x}_T|\boldsymbol{x}_0,\boldsymbol{y})||p(\boldsymbol{x}_T|\boldsymbol{y})) \\ &+\sum_{t=2}^{T} D_{KL}(q_{BB}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0,\boldsymbol{y})||p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{y})) \\ &-\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1,\boldsymbol{y})\big) \end{aligned} \tag{10}$$

Since $\boldsymbol{x}_T$ is equal to $\boldsymbol{y}$ in Brownian Bridge, the first term in Eq.(10) can be seen as a constant and ignored. By combining Eq.(4) and Eq.(8), the formula $q_{BB}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0,\boldsymbol{y})$ in the second term can be derived through Bayes' theorem and the Markov chain property:

$$\begin{aligned} q_{BB}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0,\boldsymbol{y}) &= \frac{q_{BB}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{y})q_{BB}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0,\boldsymbol{y})}{q_{BB}(\boldsymbol{x}_t|\boldsymbol{x}_0,\boldsymbol{y})} \\ &= \mathcal{N}(\boldsymbol{x}_{t-1};\tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t,\boldsymbol{x}_0,\boldsymbol{y}),\tilde{\delta}_t\boldsymbol{I}) \end{aligned} \tag{11}$$

where the mean value term is:

$$\begin{aligned} \tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t,\boldsymbol{x}_0,\boldsymbol{y}) = &\frac{\delta_{t-1}}{\delta_t}\frac{1-m_t}{1-m_{t-1}}\boldsymbol{x}_t \\ &+(1-m_{t-1})\frac{\delta_{t|t-1}}{\delta_t}\boldsymbol{x}_0 \\ &+(m_{t-1}-m_t\frac{1-m_t}{1-m_{t-1}}\frac{\delta_{t-1}}{\delta_t})\boldsymbol{y} \end{aligned} \tag{12}$$

and the variance term is:

$$\tilde{\delta}_t = \frac{\delta_{t|t-1}\cdot\delta_{t-1}}{\delta_t} \tag{13}$$

As $\boldsymbol{x}_0$ is unknown in the inference stage, we propose to utilize a reparametrization method used in DDPM [12] by combining Eq.(4) and Eq.(12). Then $\tilde{\boldsymbol{\mu}}_t$ can be reformulated as:

$$\tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t,\boldsymbol{y}) = c_{xt}\boldsymbol{x}_t + c_{yt}\boldsymbol{y} + c_{\epsilon t}\big(m_t(\boldsymbol{y}-\boldsymbol{x}_0)+\sqrt{\delta_t}\boldsymbol{\epsilon}\big)$$

where

$$\begin{aligned} c_{xt} &= \frac{\delta_{t-1}}{\delta_t}\frac{1-m_t}{1-m_{t-1}} + \frac{\delta_{t|t-1}}{\delta_t}(1-m_{t-1}) \\ c_{yt} &= m_{t-1} - m_t\frac{1-m_t}{1-m_{t-1}}\frac{\delta_{t-1}}{\delta_t} \\ c_{\epsilon t} &= (1-m_{t-1})\frac{\delta_{t|t-1}}{\delta_t} \end{aligned}$$

Instead of predicting the whole $\tilde{\boldsymbol{\mu}}_t$, we just train a neural network ϵ_θ to predict the noise. For clarification, we can reformulate $\boldsymbol{\mu_\theta}$ in Eq.(9) as a linear combination of $\boldsymbol{x}_t$, $\boldsymbol{y}$ and the estimated noise $\boldsymbol{\epsilon}_\theta$:

$$\boldsymbol{\mu_\theta}(\boldsymbol{x}_t,\boldsymbol{y},t) = c_{xt}\boldsymbol{x}_t + c_{yt}\boldsymbol{y} + c_{\epsilon t}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t) \tag{14}$$

Therefore, the training objective ELBO in Eq.(10) can be simplified as:

$$\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{y},\boldsymbol{\epsilon}}[c_{\epsilon t}||m_t(\boldsymbol{y}-\boldsymbol{x}_0)+\sqrt{\delta_t}\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t)||^2]$$

Algorithm 1 Training

1: **repeat**
2: paired data $\boldsymbol{x}_0\sim q(\boldsymbol{x}_0)$, $\boldsymbol{y}\sim q(\boldsymbol{y})$
3: timestep $t\sim Uniform(1,...,T)$
4: Gaussian noise $\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})$
5: Forward diffusion $\boldsymbol{x}_t=(1-m_t)\boldsymbol{x}_0+m_t\boldsymbol{y}+\sqrt{\delta_t}\epsilon$
6: Take gradient descent step on
$\nabla_\theta||m_t(\boldsymbol{y}-\boldsymbol{x}_0)+\sqrt{\delta_t}\epsilon-\epsilon_\theta\big(\boldsymbol{x}_t,t\big)||^2$
7: **until** converged

Algorithm 2 Sampling

1: sample conditional input $\boldsymbol{x}_T=\boldsymbol{y}\sim q(\boldsymbol{y})$
2: **for** $t=T,\dots,1$ **do**
3: $\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$ if $t>1$, else $\boldsymbol{z}=\mathbf{0}$
4: $\boldsymbol{x}_{t-1}=c_{xt}\boldsymbol{x}_t+c_{yt}\boldsymbol{y}-c_{\epsilon t}\epsilon_\theta(\boldsymbol{x}_t,t)+\sqrt{\tilde{\delta}_t}\boldsymbol{z}$
return $\boldsymbol{x}_0$

3.2. Accelerated Sampling Processes

Similar to the basic idea of DDIM [32], the inference processes of BBDM can be accelerated by utilizing a non-Markovian process while keeping the same marginal distributions as Markovian inference processes.

Now, given a sub-sequence of $[1{:}T]$ of length S $\{\tau_1,\tau_2,...,\tau_S\}$, the inference process can be defined by a subset of the latent variables $\boldsymbol{x}_{1:T}$, which is $\{\boldsymbol{x}_{\tau_1},\boldsymbol{x}_{\tau_2},...,\boldsymbol{x}_{\tau_S}\}$,

$$\begin{aligned} q_{BB}(\boldsymbol{x}_{\tau_{s-1}}|\boldsymbol{x}_{\tau s},\boldsymbol{x}_0,\boldsymbol{y}) = \mathcal{N}\Big((1-m_{\tau_{s-1}})\boldsymbol{x}_0 + m_{\tau_{s-1}}\boldsymbol{y} + \\ \sqrt{\delta_{\tau_{s-1}}-\sigma_{\tau_s}^2}\frac{1}{\sqrt{\delta_{\tau_s}}}\big(\boldsymbol{x}_{\tau_s}-(1-m_{\tau_s})\boldsymbol{x}_0-m_{\tau_s}\boldsymbol{y}\big),\sigma_{\tau_s}^2\boldsymbol{I}\Big) \end{aligned}$$

A numerical experiment is conducted in Section 4 to evaluate the performance with different numbers of sampling steps. To balance the sampling quality and efficiency, we choose $S=200$ by default. The whole training process and sampling process are summarized in Algorithm 1 and Algorithm 2.

4. Experiments

4.1. Experiment Setup

Models and hyperparameters: The BBDM framework is composed of two components: pretrained VQGAN

Figure 3. Qualitative comparison on CelebAMask-HQ dataset.

model and the proposed Brownian Bridge diffusion model. For fair comparison, we adopt the same pretrained VQGAN model as used in Latent Diffusion Model [28]. The number of time steps of Brownian Bridge is set to be 1000 during the training stage, and we use 200 sampling steps during the inference stage with the considerations of both sample quality and efficiency.

We train the network by using the Adam optimizer on a PC with an Intel Core i9-9900K CPU @ 3.2 GHz, 24GB RAM, and a GeForce GTX 3090 GPU.

Evaluation: For the visual quality and fidelity, we adopt the widely-used Fréchet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS) metrics [42]. To evaluate the generation diversity, we adopt the diversity metric proposed in [2]. Specifically, we generate five samples $(\hat{\boldsymbol{x}}_{t=1}^{5})$ for a given conditional input $\boldsymbol{y}$, and calculate the average standard deviation for each pixel among the samples. Then, we report the average diversity over the whole test dataset.

Datasets and baselines: To demonstrate the capability of handling image-to-image translation on various datasets, We evaluate the BBDM framework on three distinct and challenging image-to-image translation tasks, including semantic synthesis task on CelebAMask-HQ dataset [18], sketch-to-photo task on edges2shoes and edges2handbags [14], and style transfer task on faces2comics dataset. The baseline methods include Pix2Pix [14], CycleGAN [43], DRIT++ [19], CDE [30] and LDM [28]. Among the baselines, Pix2Pix, CycleGAN and DRIT++ are image-to-image translation methods based on conditional GANs, while CDE and LDM conduct image translation by conditional diffusion models. We additionally compare BBDM with OASIS [34] and SPADE [24] on CelebAMask-HQ dataset.

4.2. Qualitative Comparison

In this section, we evaluate the performance of the proposed BBDM against the state-of-the-art baselines on several popular image-to-image translation tasks. Semantic synthesis aims to generate photorealistic images based on semantic layout, while edges-to-images aims at synthesizing realistic image with the constraint of image edges. As both semantic layout and edge images are abstract, another task referred to as faces-to-comics conducted on two domains with more similar distributions is involved.

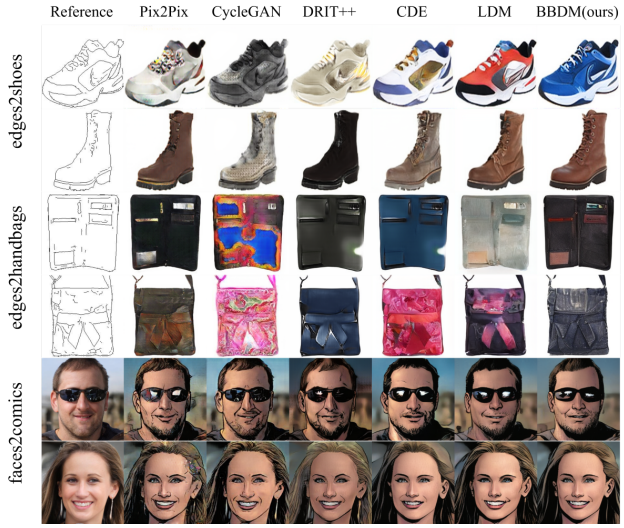

Figure 4. Qualitative comparison on different image-to-image translation tasks.

The experimental results of the proposed BBDM and other baselines are shown in Figures 3 and 4. Pix2Pix [14] can get reasonable results benefiting from the paired training data, while the performance of CycleGAN [43] drops on small scale datasets. DRIT++ achieves better performance among GAN-based method, however, the translated images are oversmoothed and far from the ground truth distribution of the target domain. Compared with methods with GANs, diffusion based methods gain competitive performance. However, as is discussed in the introduction section, both CDE [30] and LDM [28] are conditional diffusion models, and suffer from conditional information leverage during the diffusion process. For example, when there are irregular occlusions as shown in the first row of Figure 3, CDE and LDM cannot generate satisfactory results due to the mechanism of integrating conditional input into the dif-

Figure 5. Diverse samples of BBDM on different image-to-image translation tasks.

model	CelebAMask-HQ		
	FID ↓	LPIPS ↓	Diversity ↑
Pix2Pix	56.997	0.431	0
CycleGAN	78.234	0.490	0
DRIT++	77.794	0.431	35.759
SPADE	44.171	0.376	0
OASIS	27.751	0.384	39.662
CDE	24.404	0.414	**50.278**
LDM	22.816	0.371	20.304
BBDM(ours)	**21.350**	**0.370**	29.859

Table 1. Quantitative comparison on CelebAMask-HQ dataset.

fusion model. In contrast, the proposed BBDM conducts image-to-image translation by directly learning a diffusion process between these two domains, and avoids the conditional information leverage.

Benefiting from the stochastic property of Brownian Bridge, the proposed method can generate samples with high fidelity and diversity. Some examples are shown in Figure 5.

4.3. Quantitative Comparison

In this section, we compare the proposed BBDM against baselines with several popular quantitative metrics, including FID, LPIPS and diversity measurement [2]. The numerical results are shown in Tables 1 and 2. It is obvious that the proposed BBDM method achieves the best FID performance on all of the four tasks, and gains competitive LPIPS scores.

4.4. Other Translation Tasks

In order to further verify the generalization of BBDM, we conducted inpainting, colorization experiments on VisualGENOME [17] and face-to-label on CelebAMask-HQ [18]. The experimental results in Figure 6 show that BBDM can achieve comparable performance on various image translation tasks. More examples are shown in supplementary materials.

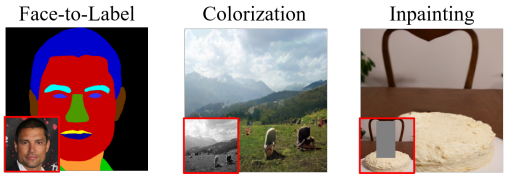

Figure 6. Face-to-label, colorization and inpainting results.

4.5. Ablation Study

We perform ablative experiments to verify the effectiveness of several important designs in our framework.

Influence of the pre-trained latent space: To speed up the training and inference process, the diffusion process of the proposed BBDM is conducted in a pre-trained latent space the same as the one used in LDM [28]. In order to demonstrate the influence of different latent spaces to the performance of the proposed method, we conduct an ablation study by choosing different downsampling factors for VQGAN model as done in LDM.

In this experiment, we compare our BBDM framework and LDM with downsampling factors $f \in \{4, 8, 16\}$ on CelebAMask-HQ. For fair comparison, We implemented BBDM based on the same network structure as LDM and used the same VQGAN-f4, VQGAN-f8, VQGAN-f16 checkpoints of LDM. The quantitative metrics are shown in Table 3. We can find that the proposed BBDM performs robustly w.r.t. different levels of latent features. The latent space learned with downsampling factor 16 leads to more abstract feature, and as a result, the performance of the LDM model drops dramatically especially with the FID metric. To further verify the image-to-image translation process during the diffusion of Brownian Bridge, we decode the latent code at each time step in the inference processes by the decoder of $VQGAN_B$. As shown in Figure 7, the input image is smoothly and gradually translated to the target domain within the Brownian Bridge.

Sampling Steps: To evaluate the influence of sampling steps in the reverse diffusion process to the performance of BBDM, we evaluate the performance with different numbers of sampling steps. In Table 4, we report the quantitative scores of semantic-to-image task with models trained on CelebAMask-HQ. We can find that when the number of

model	edges2shoes			edges2handbags			faces2comics		
	FID ↓	LPIPS ↓	Diversity ↑	FID ↓	LPIPS ↓	Diversity ↑	FID ↓	LPIPS ↓	Diversity ↑
Pixel2Pixel	36.339	0.183	0	32.994	0.273	0	49.964	0.282	0
CycleGAN	66.115	0.276	0	40.175	0.367	0	35.133	0.263	0
DRIT++	53.373	0.498	**23.552**	43.675	0.411	**30.169**	28.875	0.285	18.047
CDE	21.189	0.196	14.980	28.575	0.313	24.158	33.983	0.259	**19.532**
LDM	13.020	**0.173**	10.999	24.251	0.307	22.705	24.280	0.205	9.032
BBDM(ours)	**10.924**	0.183	12.226	**17.257**	**0.286**	15.656	**23.203**	**0.192**	10.046

Table 2. Quantitative comparison on different image-to-image translation tasks.

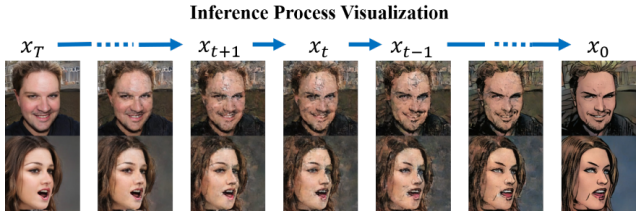

Figure 7. Latent space visualization.

model	FID ↓	LPIPS ↓	Diversity ↑
LDM-f4	22.816	0.371	20.304
LDM-f8	24.530	0.418	41.625
LDM-f16	56.404	0.416	22.112
BBDM-f4	21.350	0.370	29.859
BBDM-f8	21.966	0.392	38.978
BBDM-f16	22.061	0.391	40.120

Table 3. Quantitative scores of LDM and BBDM with different downsampling factors.

sampling steps is relatively small (fewer than 200 steps), the sample quality and diversity improve rapidly with the increase of sampling steps. When the number of sampling steps is relatively large (greater than 200 steps), the FID and diversity metrics get better slightly and the LPIPS metric almost remains the same as the sampling steps are raised.

Sampling Steps	FID ↓	LPIPS ↓	Diversity ↑
20 steps	33.409	**0.362**	17.587
50 steps	25.188	0.372	23.191
100 steps	23.503	0.378	26.157
200 steps	21.350	0.370	29.859
1000 steps	**21.348**	0.375	**29.924**

Table 4. Quantitative scores of different numbers of sampling steps on CelebAMask-HQ.

The Influence of maximum variance of Brownian Bridge. As shown in Eq.(5), we can control the diversity of Brownian Bridge through scaling the maximum variance of Brownian Bridge which can be achieved at $t = \frac{T}{2}$ by a factor s. In this section, we conduct several experiments taken on $s \in \{1, 2, 4\}$ to investigate the influence of s to the performance of our Brownian Bridge model. The quantitative metrics are shown in Table 5. With the increase of s, the diversity grows but the quality and fidelity decrease. This phenomenon is consistent with the observation in Section 3.1 that if we use the original variance design of Brownian Bridge, BBDM cannot generate reasonable samples due to the extremely large maximum variance.

s	FID ↓	LPIPS ↓	Diversity ↑
$s = 0.5$	22.627	0.387	27.791
$s = 1$	21.350	0.370	29.859
$s = 2$	23.278	0.380	37.063
$s = 4$	24.490	0.384	39.573

Table 5. Quantitative scores of different factor s on CelebAMask-HQ.

5. Conclusion and Future Work

We proposed a new method for image-to-image translation based on Brownian Bridge. Compared with other diffusion-based methods, the proposed BBDM framework learns the translation between two domains directly through the Brownian Bridge diffusion process rather than a conditional generation process. We showed that our BBDM framework can generate promising results on several different tasks. Nevertheless, there is still much room for improvement of BBDM, e.g., it would be interesting to apply our framework to various multi-modal tasks like text-to-image.

Acknowledgments

The work was funded by Natural Science Foundation of China (NSFC) under Grant 62172198, 61762064, Key Project of Jiangxi Natural Science Foundation 20224ACB202008, and the Opening Project of Nanchang Innovation Institute, Peking University.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017. 1

[2] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021. 1, 2, 3, 4, 6, 7

[3] Tian Qi Chen and Mark Schmidt. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337*, 2016. 1

[4] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 1, 2, 3, 4

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. 3

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1, 3

[7] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. *arXiv preprint arXiv:1605.08803*, 2016. 1

[8] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, June 2021. 3

[9] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016. 1

[10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in Neural Information Processing Systems*, 30, 2017. 1

[11] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. *Advances in Neural Information Processing Systems*, 29, 2016. 2

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 1, 3, 4, 5

[13] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 1

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2, 6

[15] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in Neural Information Processing Systems*, 31, 2018. 1

[16] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2017. 7

[18] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020. 6, 7

[19] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. DRIT++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, 128(10):2402–2417, 2020. 2, 6

[20] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10551–10560, 2019. 3

[21] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[22] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4990–4998, 2017. 1

[23] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016. 1

[24] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 1, 6

[25] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018. 1

[26] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1

[27] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *Advances in neural information processing systems*, 32, 2019. 1

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image

synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 3, 4, 6, 7

[29] Chitwan Saharia, William Chan, Huiwen Chang, Chris A Lee, Jonathan Ho, Tim Salimans, David J Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. *arXiv preprint arXiv:2111.05826*, 2021. 1, 3

[30] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2, 3, 4, 6

[31] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265, 2015. 1, 3

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[33] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 4

[34] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. *arXiv preprint arXiv:2012.04781*, 2020. 1, 6

[35] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016. 3

[36] Hao Tang, Song Bai, and Nicu Sebe. Dual attention GANs for semantic image synthesis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1994–2002, 2020. 1

[37] Hao Tang, Dan Xu, Yan Yan, Philip HS Torr, and Nicu Sebe. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7870–7879, 2020. 1

[38] Arash Vahdat and Jan Kautz. Deep hierarchical variational autoencoder, Dec. 23 2021. US Patent App. 17/089,492. 1

[39] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with PixelCNN decoders. *Advances in Neural Information Processing Systems*, 29, 2016. 1

[40] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 1, 2

[41] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. DualGAN: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2849–2857, 2017. 2

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6

[43] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 1, 2, 6