Source-Agnostic Image Translation Based on Latent Aware Adaptive Masking

Tomislav Dobrički[1] and Byung-Woo Hong[1]

Chung-Ang University, Seoul, South Korea
{tomislav, hong}@ai.cau.ac.kr

Abstract. In this work, we propose a source-agnostic framework that dynamically refines a binary mask throughout the reverse diffusion process by computing the discrepancies of a pretrained diffusion model's prediction for each latent time step. Rather than relying on a fixed threshold, our method introduces a time-dependent statistical thresholding scheme derived from the empirical mean and standard deviation of prediction discrepancies across the latent noisy images from the target distribution. This allows the mask to adapt to the model's varying predictive confidence at different noise levels, effectively isolating domain-specific regions while preserving global structural coherence. Experimental results on the AFHQ and Celeba-HQ datasets demonstrate that our approach outperforms state-of-the-art unsupervised Image-to-Image methods in both realism (FID, KID) and faithfulness (SSIM, LPIPS). By requiring only a pretrained model of the target domain, our approach enables precise, automated localization and seamless translation across diverse source distributions without any specialized training. The project source code is available at: `https://github.com/dtoma95/PM-Edit`

Keywords: diffusion model · image-to-image translation · image editing

1 Introduction

Image-to-Image (I2I) translation is a fundamental task in computer vision that has many possible applications in fields such as medical imaging and industrial anomaly detection. However, traditional I2I translation methods rely on predefined pairs of source and target data samples for training [18, 49, 61], making them highly impractical for real world applications where image pairs are often unavailable. For this reason, unsupervised I2I translation methods [17, 29, 37] present great potential, as they do not require corresponding pairs from the source and target domains, alleviating the need for extensive data collection.

Previous approaches leverage the power of Generative Adversarial Networks (GANs) [12, 21]. Methods such as CycleGAN [60] use GANs to learn a bidirectional mapping between a source and target domain. Later, CUT [35] leverages contrastive learning to maximize the mutual information between the input and output images. Allowing the neural network to isolate domain-specific features of the target dataset and ignore the shared features. Later methods improve on these works [14, 28, 44, 47], but are limited by the capabilities of GANs.

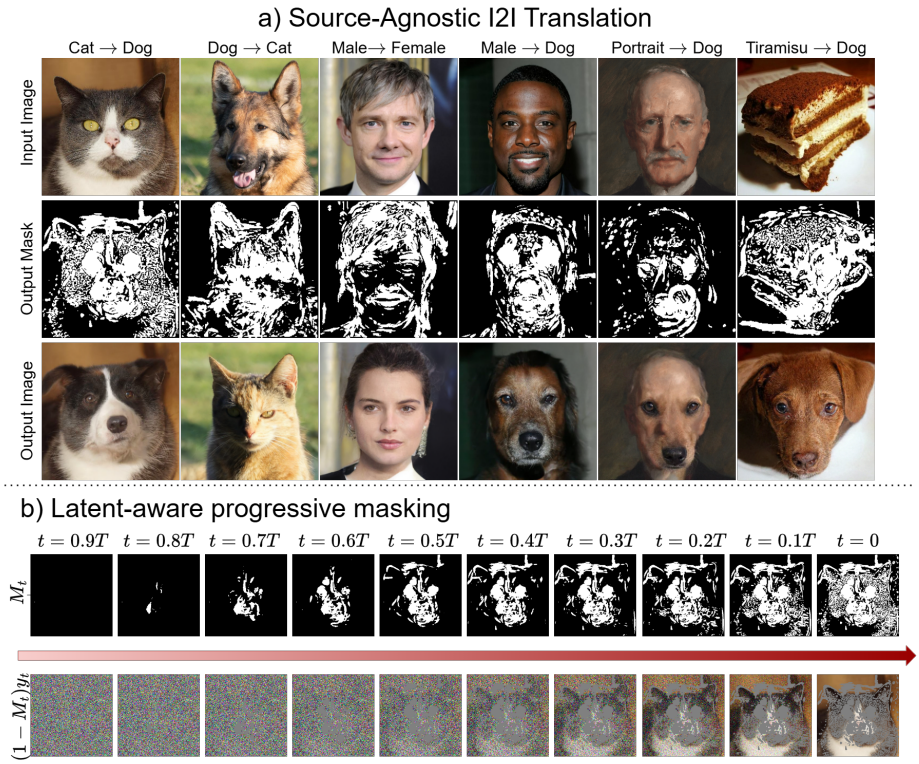

Fig. 1: a) For a given source image (top), masks are computed at each time step (middle; only final mask is shown) and the translated image (bottom) is generated by inpainting. The generative network is only trained on the target domain in each example. Note that the displayed masks do not directly correspond to the image modifications because the masks are slightly dilated and blurred during inpainting. b) The mask at each time step (top) adapts to the latent noise level. When applied to the noised image (bottom) the mask conceals features foreign to the target domain.

Denoising Diffusion Probabilistic models (DDPMs [8,16,34,38,42] are a class of generative methods that achieve state-of-the-art results in image generation. They achieve this via an iterative process where images are perturbed using Gaussian noise in a forward process, and then a neural network is trained to perform a reverse denoising process to generate new images that belong to the desired distribution. Because of their superior generative capabilities, there has been a lot of interest in utilizing DDPM for unsupervised I2I translation [39]. Recent methods such as EGSDE [59] and CycleDiffusion [52] aim to guide the denoising process of pretrained DDPM networks by injecting knowledge of the differences between the features of the source and target domains. Older diffusion guidance methods such as SDEdit [32] and ILVR [5] are source-agnostic, meaning that they can be applied to images of any domain given a generator from the

target. However, they often fail in either faithfulness or realism, as they are not adaptive to different input images.

Mask guided diffusion has been used by several methods in other domains, such as image editing [7] and anomaly detection [55]. These approaches automatically designate a masked region by comparing the discrepancy in the neural networks prediction in the latent space. This discrepancy is then binarized using a fixed threshold. An output image is then generated by diffusion inpainting, where the unmasked regions guide the denoising process to create a cohesive final image. However, the masks in these methods are computed at a single latent time step and remain constant afterward. Additionally, the neural network is typically pretrained using knowledge about the source distribution.

We propose a novel, source-agnostic masking approach for unsupervised I2I translation. In our method, a mask is computed using the latent variable of a pretrained DDPM neural network at each noise level; meaning that it progressively grows larger as more details of the image are revealed. Our approach assumes that for every time step, there exists a corresponding minimal mask required to occlude domain-specific regions from the noise estimator. Translation is then achieved via inpainting, where the appropriate mask is applied at its corresponding time step in the DDPM backward process. This means that features requiring large modifications are altered early, while finer details can be altered at the end of the generation. Our approach relies only on a denoising neural network that is trained to generate images of the target domain, and the training process is not different from the typical DDPM methods.

Similar to the previously mentioned mask guidance-based methods, we compute the mask by comparing the difference in the denoising network's prediction between the noised source image and our generated sample. This is achieved using a discrepancy metric applied to the model's estimation of the denoised image. Unlike previous works, we compute the binary mask using a threshold that corresponds to each noise level. This is achieved by pre-computing the discrepancy statistics of the pretrained network at each time step, using samples from the target domain. To evaluate the effectiveness of this approach, we evaluate our method on three standard image translation tasks using the CelebaHQ [20], and AFHQ [6] datasets. Our method outperforms the state-of-the-art in unsupervised I2I translation on several metrics of both faithfulness and realism. A demonstration of the source-agnostic I2I translation capabilities, together the latent noise aware masking process is shown in Figure 1. In summary, the primary contributions of this work are as follows:

- We introduce Progressive Mask Editing (PM-Edit): an unsupervised I2I translation approach based on an adaptive masking framework that guides the diffusion model by applying a progressively growing minimal mask at each latent time step. This approach is source agnostic, in that the denoising neural network is trained solely on the target distribution. The masks are computed using the prediction discrepancy of the network.
- A time-dependent statistical thresholding scheme. In order to compute the masks, we first calculate the statistical properties of the neural networks

prediction discrepancies on samples from the target distribution. This allows for robust masking at each latent noise level and can be tuned using a single hyperparameter.
– Experiments that demonstrate superior results over state-of-the-art methods both qualitatively and quantitatively, evaluated on standard I2I translation tasks on the AFHQ, CelebaHQ and PIE-Bench datasets. In addition, we provide ablation studies for different hyperparameter settings of our method.

2 Related Work

2.1 Diffusion Models

Diffusion probabilistic models are a type of generative model that can generate realistic images of a given domain by reversing an iterative noise degradation process [41]. Denoising Diffusion Probabilistic models (DDPM) [16] simplified the formulation by training a denoising neural network to reverse the diffusion process. Later works improved on this formulation and achieved impressive results on a variety of image synthesis tasks [8,34,38]. Denoising diffusion implicit models (DDIM) [42] improve the sampling efficiency of diffusion models by employing a non-Markovian degradation process. This also introduces a deterministic alternative for DDPM, as previous approaches relied on a stochastic backwards process where random noise is injected at each sampling step. This deterministic approach allows for the inverse mapping of a given image from the target domain to its corresponding noise latent. Other than image generation, diffusion models have been adapted to several other computer vision problems, such as image editing [7, 10, 22], anomaly detection [51, 55, 57] or image inpainting [11, 31].

2.2 Unsupervised Image to Image Translation

Early works have shown reasonable results in unsupervised (or unpaired) I2I translation by utilizing GANs. Works such as CycleGAN [60], DualGAN [56] and DiscoGAN [25]define the translation as a bidirectional mapping between the source and target domains via cycle-consistency, and subsequent methods improve on this approach [24, 25, 28]. To avoid the limitations of bijection transformation, subsequent works aim for a one-sided mapping approach. GCGAN [9] enforces geometric consistency while DistanceGAN [1] learn a mapping to maintain a distance between the input and output images. Finally, CUT [35] achieves stat-of-the-art the results for GAN-based methods by using contrastive learning to ensure patch-wise mutual information.

Given the superior generative capabilities of Diffusion-based frameworks, many unpaired I2I translation approaches aim to take advantage of them to create realistic translations. UNIT-DDPM [39] and CycleDiffusion [52] utilize two denoising neural networks using cycle-consistency. EGSDE [59] guides the diffusion model using an energy function pretrained on the source and target domains. At the same time, Brownian Bridge Diffusion Models (BBDM) [27]

and Unpaired Schrödinger Bridge (UNSB) [23] deviate from standard diffusion by mapping the forward and backward process between the source and target domain directly, instead of starting the inference process and Gaussian noise. In terms of source agnostic approaches ILVR [5] and SDEdit [32] use reference models to guide the DDPM generative process. However these methods require careful tuning of hyperparameter and modify image regions indiscriminately, without the consideration of domain-specific and unspecific features. Our approach is different as we utilize an adaptive masking scheme to achieve source-agnostic I2I translation. And unlike previous masking-based approaches like DiffEdit [7], we do not use a constant threshold or mask.

3 Preliminaries

3.1 Denoising Diffusion Models

Denoising diffusion probabilistic models (DDPM) [16, 34, 41] are a family of iterative generative models that consist of a forward and a backward process. In the forward process, data samples are gradually perturbed over T time steps until they are indistinguishable from Gaussian noise. For a data sample x_0, its noised version at any arbitrary time step t is given by:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t}\epsilon, \tag{1}$$

where the coefficient α_t corresponds to the noise level as defined by a noising schedule, and ϵ is randomly sampled as $\epsilon \sim \mathcal{N}(0, I)$. The noising schedule is a decreasing function with respect to t, such that $\alpha_0 = 1$ and $\alpha_T \approx 0$.

The goal of the backward process is to reverse the perturbation by iteratively denoising the sample, starting from the final time step. This is modeled using a deep neural network trained to predict the added noise according to the current time step. During training, the neural network parameters θ are optimized to minimize the objective:

$$L(\theta) = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta\left(x_t, t\right)\|_2^2\right]. \tag{2}$$

Inference is performed by first sampling $x_T \sim \mathcal{N}(0, I)$ and then iteratively denoising the image for T time steps until the generated sample x_0 is provided. This generative process is notoriously slow, as diffusion models require a large value for T to generate realistic images.

Denoising Diffusion Implicit Models (DDIM) [42] improve the sampling of DDPM by computing a prediction of x_0 at every t. For convenience, we denote this projection as the function f_θ:

$$f_\theta(x_t, t) := \frac{x_t - \sqrt{1 - \alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}}. \tag{3}$$

Noise is then re-injected to correspond to the next step of the backward process:

$$x_{t-1} = \sqrt{\alpha_{t-1}} f_\theta(x_t, t) + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \epsilon_\theta(x_t, t) + \sigma_t \epsilon_t, \tag{4}$$

where σ_t is the standard deviation of the stochastic noise added at each step, and if it is set to 0, the generative process becomes deterministic. DDIM introduces a more flexible inference process, allowing for faster sampling because eq. 4 does not have to be tied to the pre-defined number of time steps.

3.2 Mask Guided Diffusion

We consider a pre-trained generative denoising neural network ϵ_θ designed to generate images of a given target domain $\mathcal{X} \subset \mathbb{R}^{C \times H \times W}$. For a reference image y from a potentially unknown source distribution $\mathcal{Y} \subset \mathbb{R}^{C \times H \times W}$, we assume that there is a mask $M \in \{0, 1\}^{H \times W}$ that identifies regions of y that are inconsistent with $\mathcal{X}$. Regions where $M = 1$ contain evidence that the image does not belong to the target domain; $M = 0$ signifies the domain-consistent regions.

In SDEdit [32], the generated image x_0 is conditioned to resemble y by initializing the inference process at an intermediate time step t_{start}. Subsequently, DiffEdit [7] utilizes a mask to guide the diffusion process and minimize unwanted modifications to the reference image. This is equivalent to diffusion inpainting, where the update step is expressed as:

$$\tilde{x}_t = M x_t + (1 - M) y_t, \tag{5}$$

where y_t is obtained by applying the forward diffusion from eq. 1 on the reference image, and x_t is sampled from the denoising network inference process. In DiffEdit, M is computed at the predetermined starting time step t_{start} by comparing the discrepancy between the noise predictions of ϵ_θ for different class condition inputs and binarized according to a fixed threshold.

4 Methodology

In this section we first introduce our latent aware adaptive masking framework, then we specify the precomputation process for the dynamic threshold. And at the end we detail the inpainting methodology used to generate realistic images. An overview of our method is given in Figure 2.

4.1 Latent Aware Adaptive Masking

We propose an adaptive masking framework to address the limitations inherent in using a static, single-step computation for M. Our approach is motivated by the observation that the optimal mask is not constant across the diffusion process; rather, for every noise level t, there exists a corresponding minimal mask M_t required to effectively occlude domain-specific regions from the noise estimator. By iteratively refining this mask, we ensure that the guidance remains sensitive to the evolving structural details of the image as it emerges from noise.

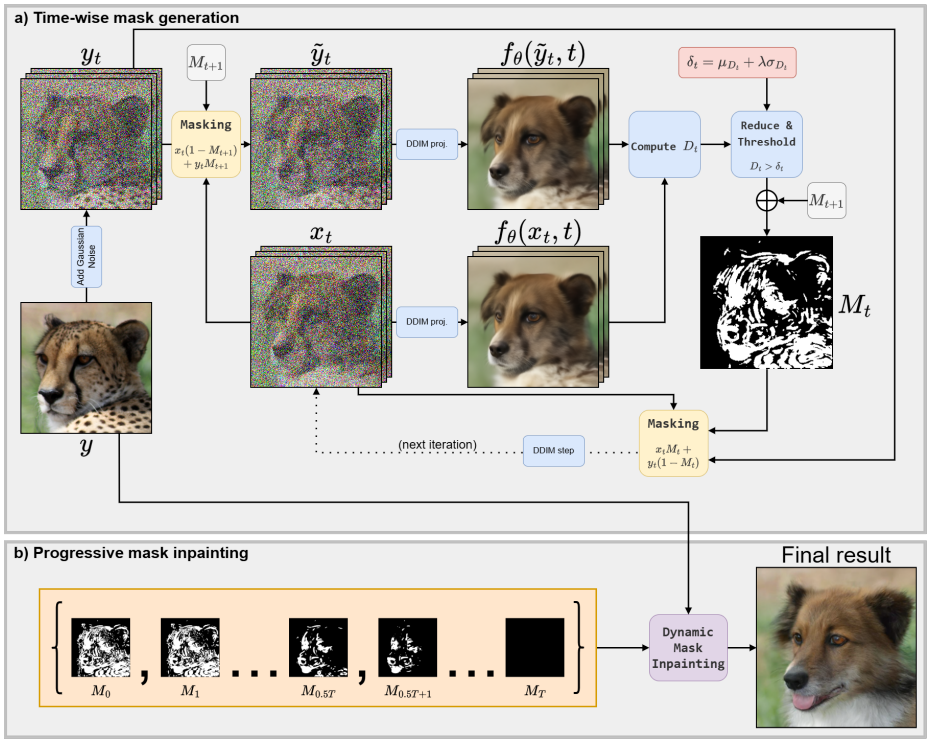

Fig. 2: Overview of our image editing method. a) For a given reference image y an appropriate mask is computed at each time-step of the DDIM forward process. The current mask M_t is computed by the union between the previous mask and the discrepancy between the model's prediction for the noisy masked reference image $\tilde{y}_t$ and the generated sample x_t. b) The output list of masks is used to perform inpainting to generate a more realistic final result. Any diffusion-based inpainting algorithm can be used. At each time step t the appropriate mask M_t is applied to guide generation.

To determine which regions to mask at a given time step t, we rely on the difference in the predictions of the pretrained denoising neural network. Projections of the reference and reconstructed images are compared using a pixel wise difference metric. For simplicity, we define this difference D_t to be the element-wise L1 distance:

$$D_t(x_t, y_t) := |f_\theta(x_t, t) - f_\theta(y_t, t)|\,. \tag{6}$$

The latent variable x_t is generated using the deterministic DDIM sampling applied to its masked version in the previous time step:

$$\begin{aligned} \tilde{x}_{t+1} &= M_{t+1}x_{t+1} + (1 - M_{t+1})y_{t+1}, \\ x_t &= \sqrt{\alpha_t} f_\theta(\tilde{x}_{t+1}, t+1) + \sqrt{1-\alpha_t}\epsilon_\theta(\tilde{x}_{t+1}, t+1). \end{aligned} \tag{7}$$

Finally, we define the mask M_t as the union of the previously accumulated mask M_{t+1} (with M_T initialized as 0 for all pixels) and the mask that demarcates the pixels that need to be obscured in the current step ΔM_t:

$$\begin{aligned} \tilde{y}_t &= M_{t+1} y_t + (1 - M_{t+1}) x_t, \\ \Delta M_t &= \left(D_t\left(x_t, \tilde{y}_t\right) > \delta_t\right), \end{aligned} \tag{8}$$

where δ_t is a scalar time-dependent threshold value. Finally, we define the mask M_t as the union of the previous mask M_{t+1} and computed discrepancy:

$$M_t = (M_{t+1} + \Delta M_t \geq 1). \tag{9}$$

Because x_t is derived from the masked state $\tilde{x}_{t+1}$, the two images differ only in regions bounded by M_{t+1}. By comparing their respective projections via f_θ, we isolate the model's prediction error caused specifically at the transition from $t+1$ to t. This allows the framework to ignore discrepancies in regions previously identified as anomalous and focus on newly emerging deviations. To ensure that the discrepancies are limited to the denoising network's prediction error, we sample the initial fake image x_T and the noised reference images at every time step y_t using the same noise $\epsilon_{const} \sim \mathcal{N}(0, I)$. However, even with a deterministic sampling process, the diffusion model predictions can still have large outlier values, especially at larger time steps where noise dominates and the network is less certain in its predictions. To counteract this, we determine D_t as an average over N batches of images with different initial noise.

4.2 Time-Dependent Threshold

Determining an adequate value for the threshold δ_t is not trivial, as it depends on many factors, such as the noise schedule and training procedure of the neural network ϵ_θ. Previous works often use an arbitrary, predetermined value for the threshold. This is not ideal, as it is expected that the statistical properties of D_t largely vary between time-steps. This makes the choice of a single constant threshold highly undesirable, as it limits the flexibility of the masking approach.

For this reason, we establish the optimal values of δ_t empirically by computing the expected prediction deviation at every time step. We sample images from the target dataset $z \sim \mathcal{X}$ in order to compute the expected discrepancy of the networks prediction on latents it was trained on (or generated itself). This is different from the formulation in section 4.1 as we do not run the full sampling procedure until time t, the sampled image $\hat{z}_t$ is instead computed by a single DDIM sampling step from the previous noised image z_t:

$$\hat{z}_t = \sqrt{\alpha_t} f_\theta(z_{t+1}, t+1) + \sqrt{1 - \alpha_t} \epsilon_\theta(z_{t+1}, t+1), \tag{10}$$

where both z_t and z_{t+1} are the noisy versions of the original image via Eq. 1. For simplicity, we define the discrepancy as the pixel-wise L1 distance and compute the discrepancy statistics as:

$$\begin{aligned} \mu_{D_t} &= \mathbb{E}_{z \sim \mathcal{X}}\left[D_t(z_t, \hat{z}_t)\right], \\ \sigma_{D_t} &= \mathbb{E}_{z \sim \mathcal{X}}\left[\sqrt{(D_t(z_t, \hat{z}_t) - \mu_{D_t})^2}\right]. \end{aligned} \tag{11}$$

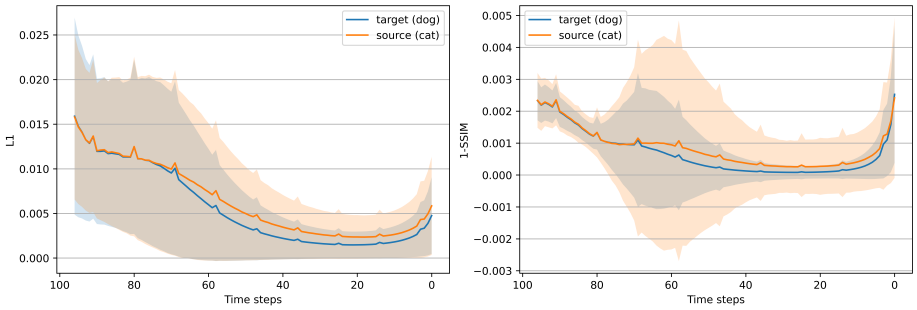

Fig. 3: Mean and standard deviations for L1 (left) and 1-SSIM (right) discrepancy metrics. We compare statistics for reference from the target dataset (AFHQ-Dog) and a source dataset that is unknown to the diffusion model (AFHQ-Cat). Note that source dataset statistics are not used in our method, and are displayed only for presentation purposes.

Note that $\mu_{D_t}$and σ_{D_t} are both scalar values and represent the average expectation over all dimensions. The time-dependent threshold is subsequently defined as:

$$\delta_t = \mu_{D_t} + \lambda \sigma_{D_t}, \tag{12}$$

where λ is a tunable hyperparameter that is constant across all time-steps.

This allows for a dynamic thresholding scheme that is easily adjustable by a single parameter. In practice, we compute μ_{D_t} and σ_{D_t} once for a given generative model and target distribution. Once computed, the distance statistics can be applied to mask generation for reference images from any source distribution. To ground our approach, in Figure 3, we display plots of μ_{D_t} and σ_{D_t} computed over the target distribution and a source distribution that the model was not trained on. It can be observed that predictions deviate the most around the middle of the sampling process. This is expected, as it is known that the middle time-steps represent the critical phase where global semantic structure and coarse geometry are resolved.

4.3 Progressive Inpainting

Although the masking framework described in Section 4.1 employs the full DDIM generative process, it is not adequate to generate realistic samples from the target domain $\mathcal{X}$. After computing the sequence of masks $[M_T, M_{T-1}, ...M_1]$, we apply a modified version of the RePaint [31] inpainting algorithm. In Repaint, time is periodically reversed so that denoising steps are repeated several times to allow the generative process to create a globally consistent image. However, instead of a constant mask, at each time-step t of the backward diffusion process, we apply the corresponding mask M_t. As in previous works, we apply a slight dilation and Gaussian blurring filter to the mask, as it has been shown to improve image coherence at the mask boundaries [7]. Additionally, for better generation results,

Table 1: Quantitative comparison of our results with baseline methods on AFHQ and Celeba-HQ I2I translation tasks. ILVR, SDEdit, EGSDE, CycleDiffusion all use the same backbone generator model, the best result among these methods is denoted in bold. Method denoted with † require specialized training on the source domain. Our method outperformed baseline methods while remainging source agnostic.

Method	FID ↓	KID×10^3 ↓	SSIM ↑	LPIPS ↓
		Wild → Dog		
CUT† [35]	92.94	48.0	0.592	-
Santa† [54]	74.93	31.6	0.453	-
UNSB† [23]	86.72	40.2	0.524	-
SDDM† [43]	57.38	-	0.328	-
ILVR [5]	81.34 ± 1.44	36.6 ± 1.56	0.289 ± 0.001	0.530 ± 0.001
SDEdit [32]	69.38 ± 1.10	27.6 ± 1.1	0.343 ± 0.001	0.515 ± 0.001
EGSDE† [59]	57.89 ± 0.62	19.2 ± 0.98	0.363 ± 0.001	0.504 ± 0.001
CycleDiffusion† [52]	56.10 ± 0.59	19.5 ± 1.02	**0.479 ± 0.001**	0.465 ± 0.001
PM-Edit (Ours)	**55.64 ± 0.61**	**17.2 ± 0.92**	**0.479 ± 0.001**	**0.445 ± 0.001**
		Cat → Dog		
NOT† [26]	161.54	-	0.566	-
DIOTM† [4]	74.70	-	0.363	-
ASBM† [13]	91.40	-	0.463	-
DSBM† [40]	100.08	-	0.532	-
EGEOT† [33]	53.29	-	0.349	-
ILVR [5]	75.38 ± 1.49	31.5 ± 1.95	0.359 ± 0.001	0.507 ± 0.001
SDEdit [32]	73.72 ± 1.01	28.1 ± 1.13	0.418 ± 0.001	0.495 ± 0.001
EGSDE† [59]	63.37 ± 0.71	21.4 ± 1.21	0.437 ± 0.001	0.471 ± 0.001
CycleDiffusion† [52]	**58.87 ± 0.69**	20.4 ± 1.10	**0.557 ± 0.001**	0.426 ± 0.001
PM-Edit (Ours)	61.31 ± 0.73	**18.3 ± 1.80**	0.551 ± 0.001	**0.415 ± 0.001**
		Male → Female		
ILVR [5]	60.17 ± 0.33	47.30 ± 1.22	0.510 ± 0.001	0.390 ± 0.001
SDEdit [32]	54.91 ± 0.47	47.70 ± 1.34	0.577 ± 0.001	0.361 ± 0.001
EGSDE† [59]	48.50 ± 0.24	45.96 ± 0.96	0.579 ± 0.001	0.357 ± 0.001
CycleDiffusion† [52]	**45.04 ± 0.23**	45.11 ± 0.98	0.564 ± 0.001	0.369 ± 0.001
PM-Edit (Ours)	50.99 ± 0.25	**42.10 ± 0.91**	**0.586 ± 0.001**	**0.356 ± 0.001**

the inpainting can be run for a different number of total sampling steps τ. In this case, we adapt the time schedule by utilizing the closest corresponding mask to the current time step t as $M_{\lfloor t\frac{T}{\tau}\rfloor}$.

5 Experiments

In this section, we first introduce the Implementation details and setup for our experiments. Then, we present qualitative and quantitative comparisons with previous methods in unsupervised image translation. Lastly, we conduct ablation studies on the individual components and parameter values of our method.

Table 2: Results on the PIE-Bench (Prompt-driven Image Editing Benchmark). Our method outperform previous methods in both fidelity in prompt adherence.

	Structure	Fidelity		CLIP Similarity	
Method	**Distance$_{\times 10^3}$ ↓**	**LPIPS$_{\times 10^3}$ ↓**	**SSIM↑**	**Whole↑**	**Edited↑**
MasaCtrl. [3]	24.70	87.94	0.813	24.38	21.35
Wang et al. [48]	22.40	84.45	0.816	25.15	22.12
LatentEdit [30]	22.13	-	0.808	25.45	22.51
PnP [46]	28.22	113.46	0.791	25.41	22.55
PnPInversion [19]	24.29	106.06	0.797	25.41	22.62
PM-Edit (Ours)	**19.68**	**83.51**	**0.819**	**26.43**	**22.81**

5.1 Implementation Details

Datasets. We conduct I2I translation experiments using two common benchmark datasets: AFHQ [6] and Celeba-HQ [20]. AFHQ contains high quality images divided into three categories: cat, dog, and wild; we evaluate on the Cat→Dog and Wild→Dog translation tasks. We only use the test split of the dataset; it contains 500 test images for each category. Celeba-HQ consists of high-resolution images split into two categories: male and female. We evaluate on the Male→Female task, using the validation split of 1000 images. For both datasets, images are resized to 256×256. We also evaluate our method on the PIE-Bench [19] dataset for text prompt guided multi-aspect image editing. It consists of 700 images, each assigned a source and editing prompt, encompassing a diverse range of manipulations, from object edits to style transfer tasks.

Evaluation Metrics. To properly evaluate image translation, we measure two aspects: realism and faithfulness. For realism, we compute Frechet Inception Distance (FID) [15] and Kernel Inception Distance (KID) [2] scores. Following CUT [35], on AFHQ they are computed between the generated images and the target test split, while for Celeba-HQ we use the target train split like in EGSDE [59]. To evaluate faithfulness, we report Structural Similarity Index Measure (SSIM) [50] and Learned Perceptual Image Patch Similarity (LPIPS) [58] between the generated samples and their respective reference images. In this setting, LPIPS and KID are considered more reliable metrics, as FID is known to be inaccurate for small datasets, and SSIM does not capture human level perception as well as LPIPS. For PIE-Bench we evaluate prompt-image consistency using CLIPSIM [53] and fidelity via structure distance [45].

Implementation Details. On the unconditional I2I translation tasks, we utilize the same pretrained diffusion backbones as in previous works. We use the AFHQ dog generator network provided by ILVR [5] and the Celeba-HQ network provided by the authors of EGSDE. Both networks were trained only on the target domain, and we do not fine tune either model. When computing the masks, we set the threshold coefficient to be $\lambda = 1.2$ and use $N = 10$ image

Fig. 4: Qualitative comparison of our method and other diffusion-based approaches.

batches. During inpainting, we first dilate and then blur the masks with a kernel size of 11 in both, and set the gaussian blur sigma to be 5.0. We generate the mask over 100 steps of a DDIM backward process but perform inpainting over the full 1000 steps. We integrate Repaint [31] by reverting back for 10 time steps at every 20th step, and repeat this action twice. We use $1 - SSIM$ for the discrepancy metric D_t, and have found μ_{D_t} and σ_{D_t} to be robust across models and datasets. We use μ_{D_t} and σ_{D_t} computed on a model trained on AFHQ-dog in all presented results. Other configurations are explored in the Supp. Material.

On PIE-Bench, we used Sable Diffusion v1.5 [36] as a backbone and applied our method in the latent space of the VAE with L1 as the discrepancy metric. Performing inpainting in an LDM's latent space does not require extra RePaint steps thanks to the robustness of the VAE decoder; this significantly lowers

Fig. 5: Examples of text prompt guided Image translation on the PIE-Bench dataset. The left image and text for each sample represents the source image-prompt pair. With the result given by our method, according to the target prompt, shown on the right.

execution time, which we explore in Sec. D.2 of the Supplementary Material. Furthermore, in our PIE-Bench experiment, we first perform DDIM inversion using the source prompt of a given sample; this improves faithfulness and is standard in prompt guided image translation methods.

5.2 Image to Image Translation

We provide a quantitative comparison of our method with previous baseline works in Table 1. Our progressive masking and editing via inpainting pipeline is denoted as PM-Edit. It outperforms the other source-free approaches in ILVR [5] and SDEdit [32]; and achieves competitive results with methods that rely on training models using the source domain. Notably, we observe lower scores in both KID and LPIPS, which are considered to be the more accurate metrics. KID is a better metric than FID when used on small datasets; and LPIPS is considered to have better correlation with human perception than SSIM. Qualitative comparisons are available in Figure 4, showing that our masking approach has an advantage in preserving class-irrelevant details without hindering realism.

Quantitative comparisons for PIE-Bench are shown in Table 2. Our method represents a solid balance between prompt adherence and faithfulness to the source image, reflected by superior results across all metric compared to other recent methods. Qualitative results in Figure 5 show examples of text guided image editing. Thanks to the masking-based approach our method performs minimal changes to the source image by preserving image areas that are not relevant to the editing prompt.

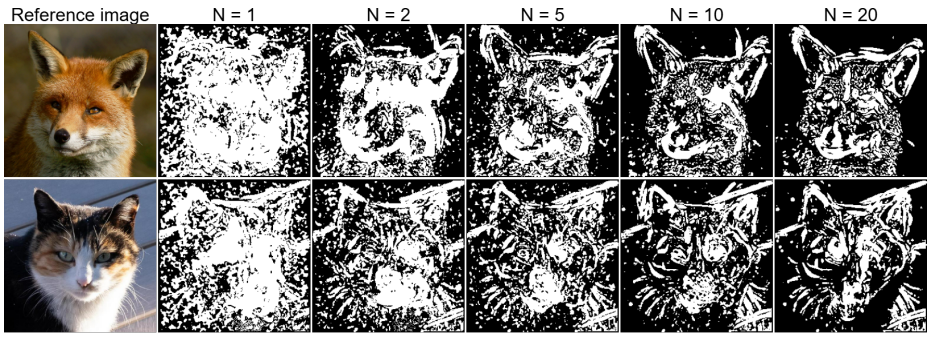

Fig. 6: Demonstration of the final mask for different numbers of mask-computation batches N. Computing discrepancies over of average of several images yields cleaner masks with less noise interference.

5.3 Ablation Study

In Table 4, we present results for different values of the threshold coefficient λ. A higher threshold value leads to stricter masking, which yields higher SSIM and LPIPS scores. Conversely, lower λ values lead to better FID and KID scores because the generation process is less restricted. This means that by adjusting λ we gain a tradeoff between realism and faithfulness in the translation. Figure 6 shows the effect of varying values of the mask computation batch size N. Computing the prediction discrepancy over only one random noise seed is sensitive to outliers, and yield to a noisy mask. Using a larger number of images in a batch increases our method's robustness to these outliers, but comes at the cost of higher computation time. We include a detailed analysts of run-time efficiency in section D.2 of the Supplementary Material.

In Table 3 we present ablations for the different components of our method. First we we compute the mask normally, but do not use dynamic masking during inpainting. In the *"constant $M_{0.5T}$"* the mask computed at the middle, and in *"constant M_0"* only the final mask is used. These settings are similar to the constant masking in DiffEdit [7], but neither show comparable results to our dynamic mask inpainting approach. This is because a larger mask at smaller t unnecessarily covers region of an image at early steps, this means that some structural features might not be preserved, or insignificant details are lost. In contrast, computing the mask too early limits the generation too much, which leads to unrealistic final image that does not resemble the target domain.

Table 3 also contains ablations that justify our adaptive masking and inpainting approach by experimenting with a constant threshold set to the equivalent of $\lambda = 1.2$ at $t = 0.5T$ across the entire sampling process, labeled as *"constant δ_t"*. This fails to generate translations faithful to the original images, as discrepancies at the early steps make the mask too large. Additionally, we ablate for when a single batch is used during mask computation ($N = 1$). Which shows that as statistical errors dominate the discrepancy computation, the mask ends up

Table 3: Ablation study for different modifications to our baseline approach on Wild→Dog. In each test other components are set to default unless ablated for.

Setting	KID ↓	LPIPS ↓
constant mask $M_{0.5T}$	33.3	0.309
constant mask M_0	13.1	0.533
constant δ_t	10.5	0.698
w/o batching (N=1)	16.2	0.617
w/o RePaint	28.0	0.4373

Table 4: Ablation study for different values of the threshold coefficient hyperparameter λ on Wild→Dog. We observe a tradeoff between realism and faithfulness.

λ	FID ↓	KID ↓	SSIM↑	LPIPS↓
1.0	54.12	17.1	0.453	0.459
1.2	55.64	17.2	0.479	0.445
1.4	58.34	19.4	0.497	0.442
1.6	61.35	21.4	0.521	0.413
1.8	62.89	22.6	0.531	0.406

covering too much of the image and lowers faithfulness to the reference image. Finally demonstrate results of our method without utilizing the Repaint inpainting technique. Although the LPIPS score in this setting is similar to our default setting, the KID of the resulting images is low. This shows that applying RePaint is highly beneficial for realism while not significantly hindering faithfulness.

6 Conclusion

In this paper, we present a novel source-agnostic framework for unsupervised image-to-image translation. By leveraging the internal dynamics of denoising diffusion models, our method overcomes the limitations of static masking and the requirement for source-domain supervision. We achieve robust dynamic masking by introducing a time-dependent, statistical thresholding mechanism. By dynamically isolating domain-specific discrepancies while preserving invariant structural features, our method ensures faithfulness to the original content without compromising the realism of the target transformation. Extensive experiments on the AFHQ and Celeba-HQ datasets demonstrate that our method consistently outperforms existing state-of-the-art unsupervised methods in both realism and faithfulness while not integrating knowledge about the source domain. Because our method requires only a pretrained model of the target domain and no additional training or source-domain data, it offers a highly flexible solution for image translation. However, a current limitation of our approach is the underlying assumption that the source and target domains are of the same modality. Future work could explore implementing our method using more sophisticated discrepancy metrics that could be adaptable to cross-domain translation.

Acknowledgement

This work was supported by the Korea government (MSIT): IITP-RS-2021-II211341, Artificial Intelligence Graduate School, Chung-Ang University and NRF-RS-2025-25462275.

References

1. Benaim, S., Wolf, L.: One-sided unsupervised domain mapping. Advances in neural information processing systems **30** (2017)
2. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
3. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22560–22570 (2023)
4. Choi, J., Chen, Y., Choi, J.: Improving neural optimal transport via displacement interpolation. In: The Thirteenth International Conference on Learning Representations (2025)
5. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938 (2021)
6. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8188–8197 (2020)
7. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022)
8. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
9. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Zhang, K., Tao, D.: Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2427–2436 (2019)
10. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 5111–5120 (2024)
11. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 5111–5120 (2024)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
13. Gushchin, N., Selikhanovych, D., Kholkin, S., Burnaev, E., Korotin, A.: Adversarial schrödinger bridge matching. Advances in Neural Information Processing Systems **37**, 89612–89651 (2024)
14. Han, J., Shoeiby, M., Petersson, L., Armin, M.A.: Dual contrastive learning for unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 746–755 (2021)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
17. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV). pp. 172–189 (2018)

18. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
19. Ju, X., Zeng, A., Bian, Y., Liu, S., Xu, Q.: Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In: International Conference on Learning Representations. vol. 2024, pp. 23395–23422 (2024)
20. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. International Conference on Learning Representations (2018)
21. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
22. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6007–6017 (2023)
23. Kim, B., Kwon, G., Kim, K., Ye, J.C.: Unpaired image-to-image translation via neural schr\" odinger bridge. arXiv preprint arXiv:2305.15086 (2023)
24. Kim, J., Kim, M., Kang, H., Lee, K.: U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. arXiv preprint arXiv:1907.10830 (2019)
25. Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: International conference on machine learning. pp. 1857–1865. Pmlr (2017)
26. Korotin, A., Selikhanovych, D., Burnaev, E.: Neural optimal transport. arXiv preprint arXiv:2201.12220 (2022)
27. Li, B., Xue, K., Liu, B., Lai, Y.K.: Bbdm: Image-to-image translation with brownian bridge diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition. pp. 1952–1961 (2023)
28. Li, M., Huang, H., Ma, L., Liu, W., Zhang, T., Jiang, Y.: Unsupervised image-to-image translation with stacked cycle-consistent adversarial networks. In: Proceedings of the European conference on computer vision (ECCV). pp. 184–199 (2018)
29. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. Advances in neural information processing systems **30** (2017)
30. Liu, S., Chen, W., Tang, Y., He, Z.: Latentedit: Adaptive latent control for consistent semantic editing. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 91–105. Springer (2025)
31. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11461–11471 (2022)
32. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
33. Mokrov, P., Korotin, A., Kolesov, A., Gushchin, N., Burnaev, E.: Energy-guided entropic neural optimal transport. arXiv preprint arXiv:2304.06094 (2023)
34. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)

35. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European conference on computer vision. pp. 319–345. Springer (2020)
36. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: International Conference on Learning Representations. vol. 2024, pp. 1862–1874 (2024)
37. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
38. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
39. Sasaki, H., Willcocks, C.G., Breckon, T.P.: Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. arXiv preprint arXiv:2104.05358 (2021)
40. Shi, Y., De Bortoli, V., Campbell, A., Doucet, A.: Diffusion schrödinger bridge matching. Advances in neural information processing systems **36**, 62183–62223 (2023)
41. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
42. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. International Conference on Learning Representations (2021)
43. Sun, S., Wei, L., Xing, J., Jia, J., Tian, Q.: Sddm: Score-decomposed diffusion models on manifolds for unpaired image-to-image translation. In: International Conference on Machine Learning. pp. 33115–33134. PMLR (2023)
44. Torbunov, D., Huang, Y., Yu, H., Huang, J., Yoo, S., Lin, M., Viren, B., Ren, Y.: Uvcgan: Unet vision transformer cycle-consistent gan for unpaired image-to-image translation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 702–712 (2023)
45. Tumanyan, N., Bar-Tal, O., Bagon, S., Dekel, T.: Splicing vit features for semantic appearance transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10748–10757 (2022)
46. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1921–1930 (2023)
47. Wang, G., Shi, H., Chen, Y., Wu, B.: Unsupervised image-to-image translation via long-short cycle-consistent adversarial networks. Applied Intelligence **53**(14), 17243–17259 (2023)
48. Wang, J., Liu, P., Xu, W.: Unified diffusion-based rigid and non-rigid editing with text and image guidance. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2024)
49. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
50. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)

51. Wolleb, J., Bieder, F., Sandkühler, R., Cattin, P.C.: Diffusion models for medical anomaly detection. In: International Conference on Medical image computing and computer-assisted intervention. pp. 35–45. Springer (2022)
52. Wu, C.H., De la Torre, F.: A latent space of stochastic diffusion models for zero-shot image editing and guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7378–7387 (2023)
53. Wu, C., Huang, L., Zhang, Q., Li, B., Ji, L., Yang, F., Sapiro, G., Duan, N.: Godiva: Generating open-domain videos from natural descriptions. arXiv preprint arXiv:2104.14806 (2021)
54. Xie, S., Xu, Y., Gong, M., Zhang, K.: Unpaired image-to-image translation with shortest path regularization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10177–10187 (2023)
55. Yao, H., Liu, M., Yin, Z., Yan, Z., Hong, X., Zuo, W.: Glad: Towards better reconstruction with global and local adaptive diffusion models for unsupervised anomaly detection. In: European Conference on Computer Vision. pp. 1–17. Springer (2024)
56. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: Proceedings of the IEEE international conference on computer vision. pp. 2849–2857 (2017)
57. Zhang, H., Wang, Z., Zeng, D., Wu, Z., Jiang, Y.G.: Diffusionad: Norm-guided one-step denoising diffusion for anomaly detection. IEEE transactions on pattern analysis and machine intelligence (2025)
58. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
59. Zhao, M., Bao, F., Li, C., Zhu, J.: Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. Advances in Neural Information Processing Systems **35**, 3609–3623 (2022)
60. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
61. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. Advances in neural information processing systems **30** (2017)