Unpaired Image-to-Image Translation via Latent Energy Transport

Yang Zhao Changyou Chen
University at Buffalo, SUNY
{yzhao63, changyou}@buffalo.edu

Abstract

Image-to-image translation aims to preserve source contents while translating to discriminative target styles between two visual domains. Most works apply adversarial learning in the ambient image space, which could be computationally expensive and challenging to train. In this paper, we propose to deploy an energy-based model (EBM) in the latent space of a pretrained autoencoder for this task. The pretrained autoencoder serves as both a latent code extractor and an image reconstruction worker. Our model, ***LETIT***[1]*, is based on the assumption that two domains share the same latent space, where latent representation is implicitly decomposed as a content code and a domain-specific style code. Instead of explicitly extracting the two codes and applying adaptive instance normalization to combine them, our latent EBM can implicitly learn to transport the source style code to the target style code while preserving the content code, an advantage over existing image translation methods. This simplified solution is also more efficient in the one-sided unpaired image translation setting. Qualitative and quantitative comparisons demonstrate superior translation quality and faithfulness for content preservation. Our model is the first to be applicable to 1024×1024-resolution unpaired image translation to the best of our knowledge. Code is available at* https://github.com/YangNaruto/latent-energy-transport.

1. Introduction

The unpaired image-to-image translation aims to learn pairwise domain mappings without being aware of any paired-image information. Suppose a task of translating between two domains of male and female, denoted as $\mathcal{X}$ and $\mathcal{Y}$ and illustrated in Figure 1. Ideally, one should be able to retain the *shared contents*, *e.g.*, the irrelevant background and the rough facial skeleton, and only focus on transferring *discriminative styles*, *e.g.*, hair and beard. Most existing models adopt generative adversarial nets (GANs) [12, 17] to enforce the translated style of source instances to be indistinguishable from that of the target domain, which typically relies on an explicit cycle consistency regularizer [42, 46, 22] to maintain the content. However, the enforced cycle consistency is often restrictive, and the learning of two roughly invertible mappings, to some extent, can hinder model optimization efficiency. CUT [31] resorts to contrastive learning as an alternative in the one-sided translation setting. Still, GAN-based approaches need to learn at least one set of an encoder-decoder structured generator and an encoder-based discriminator, which is usually computationally expensive to train [45].

Apart from GAN-based solutions, CF-EBM [45], one of the most recent works, applies the energy-based model (EBM) to realize implicit image translation by direct maximum likelihood estimation (MLE). However, EBM learning leverages the Langevin dynamics for Markov Chian Monte Carlo (MCMC) sampling in the ambient data space, which usually distorts image pixels and is challenging to scale up. Besides, it is unclear whether EBM can learn a disentangled representation of the content and style for better image translation [25].

To overcome the above issues, we propose a plug-and-play EBM-based model in the latent space. Specifically, we first pretrain an autoencoder (or use an existing one) and then plug the EBM into the latent space to manipulate the extracted latent code to realize image translation. Our latent EBM models an explicit density distribution of latent variables by training a bottom-up latent energy function, which always assigns lower energy to latent variables of the target domain data and higher energy to those of the source domain data. Hence, when sampling from the EBM via Langevin dynamics, the energy gradient (score function) describes a path to transport the latent codes from source to target domains. Most interestingly, we demonstrate that the score function can implicitly and automatically separate both the content and style codes from the whole latent embedding. Thus image translation corresponds to simply evolving the style code. As can be seen in Figure 1, only the style appearance is translated while the content information

[1] **L**atent **E**nergy **T**ransport for **I**mage **T**ranslation

Figure 1. Unpaired image-to-image translation on 1024×1024-resolution images. (*Left*) female to male; (*Right*) male to female.

(the background) is preserved.

Extensive experiments and analysis show that our contributions can be summarized in the following five aspects:

- **Universality:** Our plug-and-play of latent EBM is a universal framework, which can be applied to most autoencoders without introducing any auxiliary networks and engineered loss functions.

- **Efficiency:** Our latent EBM learning is extraordinarily efficient. For example, after pretraining an autoencoder, a very light-design choice of multiple layer perceptron for the latent EBM is adequate to outperform state-of-the-art methods, with only less than one thousand iterations for learning.

- **Transferability:** A pretrained autoencoder can be reused for multiple image translation datasets, even when pretrained on the human facial dataset and aiming to translate between apples and oranges.

- **Scalability:** Without being restricted by the U-shape design of GANs, we can effortlessly scale the model up for translating images of 1024×1024 resolution, as demonstrated in Figure 1.

- **Faithfulness:** Our model can learn faithful translation mappings, where a translated image not only is stylized but also preserves the original content. Take the third column of Figure 1 for instance, the facial color and background are perfectly preserved when translating from female to male.

The rest of paper is organized as follows. Section 2 reviews and differentiates related works. Section 3 outlines the EBM learning preliminaries. Section 4 describes our proposed method in detail. Section 5 presents extensive experiments on various kinds of autoencoders to validate our approach. Section 6 concludes our work and highlights some future research directions.

2. Related Work

2.1. Image-to-Image Translation

The solution to unpaired image-to-image translation is usually decomposed into two cooperative paths, the domain-level distribution style matching and the instance-level content preservation. Current notable research efforts can be broadly categorized into two types: GAN-based model and energy-based model.

GAN-based Models MUNIT [16], DRIT [25], U-GAT-IT [21] and StarGAN [4, 5]. However, most of these works need to leverage cycle consistency to constrain the domain mapping and enforce the content to be unchanged. The cycle consistency regularizes the training by reconstructing an original image from its backward translated image at the instance level. Thus, these schemes generally require a combination of generators, discriminators or style encoders [25, 4, 5, 44], and complicate engineered loss designs. Efforts for one-sided unpaired image-to-image translation also have been made, *e.g.*, DistanceGAN [1], GcGAN [10] and CUT [31]. These models apply geometry constraints or contrastive constraints to improve cycle consistency.

Energy-based Model Recently, energy-based generative models have drawn significant attention [40, 39, 9, 29, 7, 11, 43, 41]. A recent work CF-EBM [45] demonstrates EBM as a powerful tool to simplify the conditional image generation problem for image translation. In CF-EBM, the source data is taken as the condition, which moves along the energy decay direction via Langevin dynamics [24, 37]. The aforementioned two-level matching is implicitly integrated through MLE-based EBM learning.

Our proposed approach differs from all the above methods. We summarize the learning schemes in Table 1. Though our practice falls into the EBM learning scheme, it is defined in the high-level lower-dimensional latent space rather than the low-level high-dimensional data space. In this way, the learned latent EBM is better at capturing

domain discrepancy to facilitate conditional learning with much fewer iterations because the MCMC sampling is much more efficient.

Approaches	Distribution	Instance	Transferability
CycleGAN [46]	adversarial	cycle	✗
DRIT [25]	adversarial	cycle	✗
(M)UNIT [26, 16]	adversarial	cycle	✗
StarGAN [4, 5]	adversarial	cycle	✗
Distance [1]	adversarial	cycle+distance	✗
CUT [31]	adversarial	contrastive	✗
CF-EBM [45]	mle	implicit	✗
Ours	mle (latent)	implicit	✓

Table 1. Feature-by-feature comparisons of unpaired image-to-image translation models.

2.2. Autoencoders

To extract the latent code, we adopt the autoencoder, where an encoder is used to encode an image into a latent code, and a decoder then reconstructs it back to the image space. The recent focus, on the one hand, has been on providing a probabilistic manner to predict the posterior distribution over the latent variables such that the autoencoder is turned into a competitive generative model, *e.g.* Variational AE (VAE) [23], Vector Quantized VAE (VQ-VAE) [36, 33] and NVAE [34]. On the other hand, the unsupervised disentanglement representation learning is trending, *e.g.* β-VAE [14, 15], Factor-VAE [20], Guided-VAE [8], TCVAE [20] and Adversarial Latent AE [32]. The goal is to learn factorized and interpretable latent representations that can encode different generative factors, *e.g.*, hair, gender, and age in the human faces dataset. We refer readers to an excellent repository for more details on this topic[2]. This paper validated our approach on different AE variations, including a vanilla VQ-VAE-2 [33] without any generative capability and disentangled generative AEs with other objectives, *e.g.*, β-VAE and ALAE.

3. MCMC-based Maximum Likelihood Learning of EBM

Given an observed image $x \in \mathbb{R}^D$ sampled from a data distribution p_{data}, an energy-based model follows:

$$p_\theta(x) = \frac{1}{Z(\theta)} \exp(-E_\theta(x)), \tag{1}$$

where $E_\theta(x)$: $\mathbb{R}^D \rightarrow \mathbb{R}$ is the scalar energy function parameterized by θ and $Z(\theta) = \int \exp(-E_\theta(x))q(x)dx$ is the intractable partition function. Given N observed data points $\{x_i\}_{i=1}^N$ from the data distribution, the model can be trained by maximizing the log-likelihood $L(\theta) = \frac{1}{N}\sum_{i=1}^N \log p_\theta(x_i) \approx \mathbb{E}_{x\sim p_{data}} \log(p_\theta(x))$. The derivative of the negative log-likelihood is

$$-\frac{\partial L(\theta)}{\partial \theta} = \mathbb{E}_{x\sim p_{data}}[\frac{\partial}{\partial\theta}E_\theta(x)] - \mathbb{E}_{\tilde{x}\sim p_\theta}[\frac{\partial}{\partial\theta}E_\theta(\tilde{x})], \tag{2}$$

where the second expectation term under p_θ is intractable. We will approximate it via MCMC such that the EBM can be updated by gradient descent. To sample $\tilde{x} \sim p_\theta$ via MCMC, we rely on Langevin dynamics that recursively computes the following step

$$\tilde{x}^{t+1} = \tilde{x}^t - \frac{\eta_t}{2}\frac{\partial}{\partial\tilde{x}^t}E_\theta(\tilde{x}^t) + \sqrt{\eta^t}\epsilon^t,\ \epsilon^t \sim \mathcal{N}(0, \mathbf{I}), \tag{3}$$

where η^t is the step size typically with polynomially decay to ensure convergence [37]; ϵ^t is a Gaussian sample to capture the data uncertainty and ensure sample convergence.

4. The Proposed Framework

4.1. Model

From a probabilistic perspective, given two image domains $\mathcal{X}$ and $\mathcal{Y}$, our goal of unpaired image translation is to infer a joint distribution based on the marginal distributions $P_\mathcal{X}(x)$ and $P_\mathcal{Y}(y)$ without awareness of paired instances of the two domains. We impose some desired properties on the joint distribution space.

Assumption: Suppose we want to translate from $\mathcal{X}$ to $\mathcal{Y}$. We can achieve this by performing image-to-image translation in both the ambient data space and latent space:
(i) In the data space, the goal can be achieved in a straightforward way to learn an image-to-image mapping F: $\mathcal{X} \rightarrow \mathcal{Y}$. The mapping F should satisfy two conditions: content preservation and style transfer.
(ii) Let the associated two latent spaces corresponding to $\mathcal{X}$ and $\mathcal{Y}$ be $\mathcal{Z}_\mathcal{X}$ and $\mathcal{Z}_\mathcal{Y}$ respectively. For our purpose, a latent code should contain both the latent content and latent style information. We thus formulate the problem as learning a mapping G: $\mathcal{Z}_\mathcal{X} \rightarrow \mathcal{Z}_\mathcal{Y}$ such that it satisfies two conditions: latent content preservation and latent style transfer.

As we claimed before, most existing works are based on the first framework, which is computationally expensive in training and requires complicated loss design to satisfy both conditions. Our approach is based on the second framework. As the latent space is an abstract-level and compact representation of the data, it is reasonable to assume that two domains $\mathcal{X}$ and $\mathcal{Y}$ share the same latent space $\mathcal{Z}$, with each code z decomposed into a content code c and a domain-specific code $s_\mathcal{X}$ for $\mathcal{X}$ or $s_\mathcal{Y}$ for $\mathcal{Y}$, *i.e.*, $z = [c, s_\mathcal{X}]$ or $z = [c, s_\mathcal{Y}]$. For a source instance x with a latent code $z_x = [c, s_x]$, the translation aims at only transforming the source style code s_x to the target style code space $s_\mathcal{Y}$. We

[2] https://github.com/matthewvowels1/Awesome-VAEs

will demonstrate emperically that our model achieve automatically learn to do this without explicitly specifying out the content code and domain-specific code.

Note our setting is essentially different from the purpose of the partially shared latent space assumption in MUNIT [16], which aims to ease the learning of a pair of underlying encoder and decoder. In our model, our assumption enables one to directly learn the latent code mapping between two domains to realize image translation.

Method: As described above, we aim to transport latent codes via EBM directly and postulate (but verify empirically) that the latent EBM can transport the source style code s_x into the target style space s_y while preserving source content codes c. To achieve this, one needs an encoder to extract the latent code and a decoder to reconstruct the target image. The model is illustrated in Figure 2.

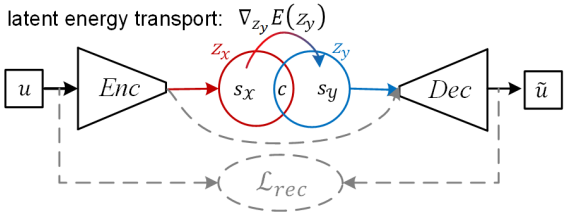

Figure 2. The proposed model: data flow and energy transport. The dashed data flow only exists in the autoencoder pretraining stage; u could be either x or y.

Specifically, we consider a pretrained autoencoder, including an encoder $Enc(\cdot)$ and a decoder $Dec(\cdot)$. The pretraining procedure follows:

$$\begin{aligned} &\text{Encoding}: z = Enc(u), u \sim P_{\mathcal{X}} \cup P_{\mathcal{Y}} \\ &\text{Decoding}: \tilde{u} = Dec(z) \\ &\text{Reconstruction loss}: \mathcal{L}_{rec} = \mathbb{E}_{u \sim P_{\mathcal{X}} \cup P_{\mathcal{Y}}} ||u - \tilde{u}||_2 \end{aligned} \tag{4}$$

Suppose the task is to adapt $P_{\mathcal{X}}$ to $P_{\mathcal{Y}}$, we aim to learn an EBM $E_{x \to y}$ satisfying:

$$p_\theta(z_y) = \frac{1}{Z(\theta)} \exp(-E_{x \to y}(z_y)), z_y = Enc(y). \tag{5}$$

The learning process of $E_{x \to y}$ is very simple by adopting (3). To sample from the EBM $E_{x \to y}$, we modify (3) as:

$$\tilde{z}_y^{t+1} = \tilde{z}_y^t - \frac{\eta^t}{2} \frac{\partial}{\partial \tilde{z}_y^t} E_{x \to y}(\tilde{z}_y^t) + \sqrt{\eta^t} \epsilon^t, \tag{6}$$

where $\tilde{z}_y^0 = z_x = Enc(x)$ and $x \sim P_{\mathcal{X}}$. The optimization of $E_{x \to y}$ exactly follows (2). After T Langevin steps, the reconstructed $Dec(z_y^T)$ will serve as the translation of x. Note the above learning only requires optimization of the EBM; thus, it is very computationally efficient.

In practice, it might happen that reconstructions from the decoder exhibit a blurry effect. Existing works introduce an additional EBM to refine the reconstructed output. This approach has proved successful in the recent proposed VAEBM [38]. [30] also provides a promising method to learn a generator and the latent EBM cooperatively. By contrast, our model does not consider this extra effort, as we want to focus on the simple plug-and-play setting, and our results already demonstrate excellent image quality.

4.2. Proof-of-Concept Verification

We conduct a proof-of-concept verification on our assumption with two synthetic domains, the *red pie* and the *blue pie* as shown in Figure 3. Following the proposed approach, we firstly pretrain until convergence a one-layer autoencoder and set the latent dimension to 8. A two-layer EBM is then learned to transport the latent code from one domain to the other using a 10-step MCMC. In Figure 3, we visualize the process of sample transportation from the black dots of the *red pie* to the *blue pie* in the data space. We observe that the proposed approach indeed can translate between the domains. To illustrate that our model can also automatically learn to distinguish the content and style codes, we calculate the **aggregated absolute latent code shift**, which is defined as the sum of absolute gradients during MCMC in (6), *i.e.*, $\sum_t \|\nabla_{z_y^t} E(z_y^t)\|$. As shown in the heatmap of Figure 3, the model does not learn a uniform shift among all latent dimensions. Instead, the 5-th latent dimension of most samples exhibits almost zero shift, meaning that this dimension can be considered as the shared content code.

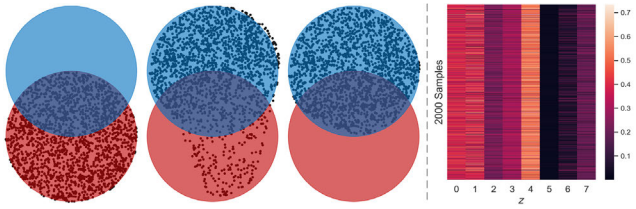

Figure 3. Domain evolving process from *red pie* to *blue pie* on synthetic data. (*Left*) Latent code transition corresponding to epochs 0, 5 and 50. (*Right*) The latent code shift after 10 Langevin steps.

5. Experiments

We evaluate our approach across several image-to-image translation datasets with three autoencoder structures.

Autoencoders: We adopt three autoencoder structures, with key characterizations listed in Table 2.

- β**-VAE** [15]: It is a well-known unsupervised disentangle representation learning model. It modifies the

objective – the evidence lower bound (ELBO), of variational autodencoder [23] and enforces the latent bottleneck (weighted Kullback–Leibler (KL) divergence) to encourage a *factorized representation*.

- **ALAE** [32]: It leverages the adversarial data distribution learning and latent distribution learning in AE. It learns a *descent disentangle representation* and demonstrates *comparable sample quality* with GAN. Thus, it is reasonable to compare it with state-of-the-art image translation models directly. And it is also beneficial for the plug-and-play of much higher-resolution models.

- **VQ-VAE-2** [33]: We treat it as a vanilla autoencoder without generative ability if it is trained only with the reconstruction loss and the second-stage PixelCNN prior training is discarded [35]. Hence, the decoder can only remember specific latent codes but not learn a distribution mapping like that of a generator. The reasons why we consider it is: (*i*) the latent space is built on several feature maps which are more challenging than 1-d vector; (*ii*) the reconstruction performance is in the lead; (*iii*) most importantly, we use VQ-VAE-2 to demonstrate the strong *transferability* of image translation via latent energy transport after it is pretrained on a large-scale or essentially irrelevant dataset.

Models	Data	Latent	Generative?	Disentangle?
β-VAE	ELBO*	KL	✓	✓
ALAE	Adversarial	MSE	✓	✓
VQ-VAE-2	MSE	Quantization	✗	✗

Table 2. Comparison of the three autoencoder models.

Datasets: For β-VAE, we apply the CelebFaces Attributes Dataset (CelebA) [27] with over 200K celebrity facial images. The dataset is also widely used in disentangled representation learning. We resize the images into 64×64 resolution and divide the dataset into *male* and *female* according to the gender attribute for translation.

To compare our model with StarGAN v2, where we directly adopt the two high-fidelity image-to-image translation datasets that are used in its experiment: CelebA-HQ [18] and Animal Faces (AFHQ) [5]. (*i*) CelebA-HQ: it contains 30k celebrity facial images, which are manually split into 17,943 female faces and 10,057 male faces for training [5]. The rest 2000 images are evenly divided as testing data. We conduct experiments on both 256×256 and 1024×1024 resolutions. (*ii*) AFHQ (256×256): it includes 15k animal faces and is evenly distributed into three challenging domains, cat, dog and wildlife. Each domain uses 500 images for testing and the rest for training.

To investigate the generalization capability through the combination of the latent energy transport and VQ-VAE-2, we conduct experiments across a couple of datasets used in CycleGAN [46]: *apple2orange* and two painting style transfer datasets, *vangogh2photo* and *ukiyoe2photo*. More details are given in the Appendix.

For the large-scale pretraining, we use the following two datasets: (*i*) The Flickr-Faces-HQ (FFHQ) dataset [19] which consists of 70k high-quality images (1024× 1024) with more variations than CelebA-HQ in terms of accessories, age, ethnicity and image background; (*ii*) ImageNet-1k [6] dataset containing over 1 million images of 1000 distinct classes. The images are resized to 256×256.

Evaluation Metrics: We consider two commonly used metrics for evaluating image synthesis quality and human perceptual study for visual quality evaluation.

(*i*) FID [13] compares the statistics (mean and variances of Gaussian distributions) between generated samples and real samples. FID is consistent with increasing disturbances and human judgment. Lower scores indicate that a model can create higher quality images.

(*ii*) KID [2] improves FID as an unbiased estimator, making it more reliable when fewer test images are available. We use generated images translated from all test images in the source domain vs. test images in the target domain to compute KID. Lower KID values indicate that images are better translated.

(*iii*) We use the Amazon Turker (AMT) for human percpetual study.

5.1. Analysis on β-VAE

β-VAE is learned by maximizing a lower bound:

$$\mathsf{ELBO} = -\mathcal{L}_{rec} - \beta \mathsf{KL}(q(z|u)||p(u)) \tag{7}$$

where $p(z)$ is the Gaussian prior and $q(z|u)$ is the approximate posterior; β is the adjustable hyperparameter and when $\beta = 1$, it recovers the original VAE [23]. Generally, choosing a different value of β leads to a trade-off between the reconstruction quality and the disentangled latent space.

Implementation Since we focus on the latent energy transport aspect for the unpaired image translation task, we simply choose a random $\beta = 10$ for evaluation. We set $dim(z) = 32$ and construct the EBM with a 1-layer MLP. More detailed experimental settings are given in the Appendix. Note that after the pretraining of β-VAE[3], the latent EBM only requires around 200 additional training iterations (takes 1 minute on TITAN XP) to find a descent translation path on the latent space. Figure 4 shows the results of the two-direction image translation on CelebA. It

[3] https://github.com/1Konny/Beta-VAE

is observed that the discriminative features, like hair, beard and facial-skeleton, can be effectively uncovered for image translation via latent energy transport. Note we do not expect this model to generate high-quality images as β-VAE itself is not a good model for this purpose. The other two models will achieve high-quality images.

Figure 4. Uncurated image translation results based on β-VAE in CelebA. x: the input, $\tilde{x}$: the reconstruction, y: the translated output. (*Top*) female to male. (*Bottom*) male to female.

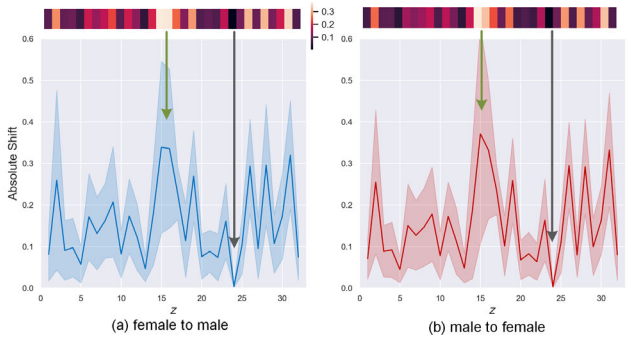

Figure 5. Visualization of the latent code shift via MCMC, which is acquired from 1,000 test images.

Behavior of latent space We use this model to visualize the aggregated absolute latent code shift via MCMC, plotted in Figure 5. We find translations of the two directions reveal nearly identical responses to the same latent dimensions, meaning the models have been likely to learn two almost *mutually inverse* mappings of the two domains, although two EBMs are learned separately without explicitly enforced cycle consistency. Specifically, dimensions 15, 16 (green arrow) are activated, representing style codes, while dimension 24 (gray arrow) stays inactivated, representing content code. Moreover, it is observed that image translation through manipulating the encoded latent code is done by the cooperation of various latent dimensions instead of one single dimension. This experiment validates our assumption and the effectiveness of latent energy transport.

5.2. Analysis on ALAE

Implementations We follow all the experiment settings in [32] and pretrain ALAE[4] on CelebA-HQ, AFHQ and FFHQ datasets. We test the unpaired image-to-image translation on CelebA-HQ and AFHQ datasets. The latent dimension of ALAE is 512. The latent EBM has one hidden layer (512-2048-1) activated by LeakyReLU. We find that the performance is not very sensitive to the architecture. To optimize the EBM, we apply the stochastic gradient descent (SGD) with a learning rate 0.1. We set the batch size to 32 and train the EBM for 2000 iterations. We run 20 Langevin steps with a step size of 1.0.

Qualitative results: StarGAN v2 [5] is the current state-of-the-art image-to-image translation approach on AFHQ and CelebA-HQ. We show the qualitative comparison between our method and StarGAN v2 in Figure 6 and Figure 7. For StarGAN v2, we feed all the reconstructed images from ALAE for translation and pick the best image from its diverse results. It's fair for us to use the reconstructed image from ALAE as inputs because we are manipulating the latent space with the proposed approach and the change will be reflected in the reconstruction. Figure 6 presents two variants of outcomes from our approach: one ALAE is pretrained on CelebA-HQ and the other on FFHQ. We observe that both results give faithful and high-fidelity translations. In particular, compared with StarGAN v2, our approach obviously works better at preserving the facial color and backgrounds while translating between males and females. Similar observations are also seen in AFHQ as shown in Figure 7.

Models	CelebA-HQ	AFHQ
MUNIT [16]	31.4	41.5
DRIT [25]	52.1	95.6
MSGAN [28]	33.1	61.4
StarGAN v2 [5]	**13.7**	**16.2**
LETIT (Ours)	**12.5**	**15.9**

Table 3. Quantitative comparison with baselines. All numbers except for our approach are from StarGAN v2 [5].

Quantitative results: We report the FID scores on several models in Table 3. On both CelebA-HQ and AFHQ datasets, our approach consistently performs better than the baselines. For the AMT perceptual study, each image is judged by six users, who are asked to select the best qualitative translated image with two standards considered: the visual **Quality** and the **Faith**ful translation measured by the

[4] https://github.com/podgorskiy/ALAE

Figure 6. Faithful unpaired image-to-image translation on CelebA-HQ. The third row (**Ours**) presents the results from the model pretrained on CelebA-HQ whereas the fourth row (**Ours-T**) gives the transferred translations from the model that is pretrained on FFHQ.

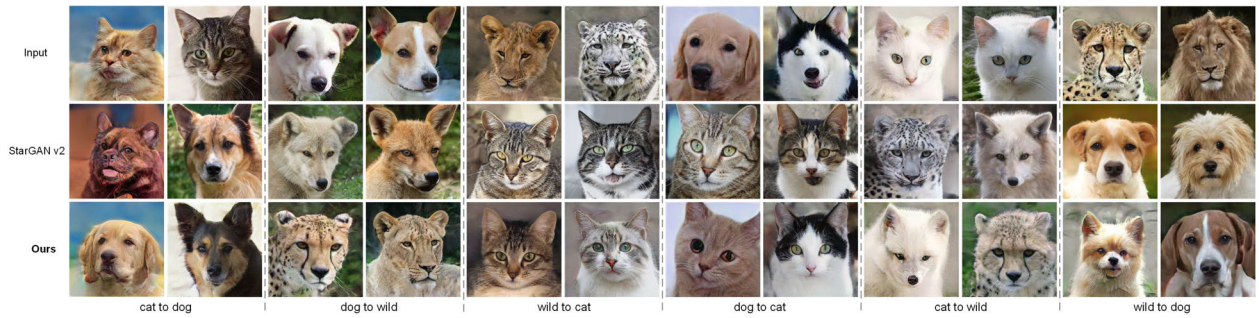

Figure 7. Faithful unpaired image-to-image translation on AFHQ.

Models	CelebA-HQ		AFHQ	
	Quality	Faith	Quality	Faith
StarGAN v2	25.6	2.2	56.3	38.4
LETIT (Ours)	35.4	**51.7**	**43.7**	**61.6**
LETIT (Ours-T)	**39.0**	46.1	-	-

Table 4. Human perceptual study regarding translation quality and faithfulness.

extent of source content preservation, including background and expression. We inform the participants of the name of the target domain and six example images of the target domain as a visual illustration. The result is shown in Table 4, where our models again outperform StarGAN v2.

High-resolution translation We conduct the image-to-image translation on CelebA-HQ with a resolution of 1024×1024. To the best of our knowledge, this is the first trial on such a high-resolution setting. The autoencoder is pretrained on FFHQ-1024^2. Other implementation details are the same as CelebA-HQ, except that the batch size is made to 16 to fit the GPU memory. Since the model only deals with the latent space, the cost is similar as translating 256×256 images. Some visualized results are shown in Figure 1. We additionally visualize the smooth translation evolution in the data space as the latent code evolves in the latent space. Some example images are shown in Figure 8, from which we can see how images are smoothly transported to the target domains. More results are included in the Appendix.

5.3. Analysis on VQ-VAE-2

Implementations: We use the original VQ-VAE-2 implementation [33] with source code. To accelerate the pretraining and facilitate the latent EBM learning, we make several modifications to the original architecture [33], including (i) We adopt two-level latent maps, bottom and top, such that the EBM is defined on the latent space just before the entrance to the bottom decoder. (ii) We inject layer-wise random Gaussian noise to the decoder for more flexibility. (iii) We adjust the codebook dimension and codebook size. Note that both bottom and top codebooks use the same configuration. The impact on reconstruction quality regarding the Mean Squared Error (MSE) is given in Table 8. We set the codebook dimension to 32 and codebook size to 256 such that the latent dimension is $64 \times 64 \times 64$ [33]. The detailed latent EBM architecture and training settings are given in Appendix. We compare our model with two one-sided translation models, CF-EBM and CUT.

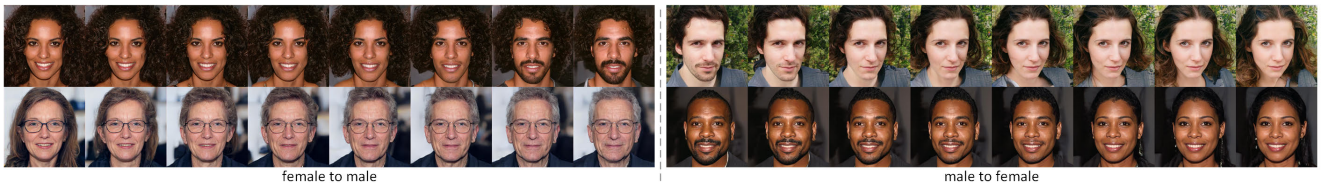

Figure 8. Smooth unpaired image translation dynamics driven by MCMC, on 1024×1024-resolution images.

Qualitative results: Figure 9 visualizes some results from different approaches. In the top row of Figure 9, we show the outcomes from different pretrained autoencoders on photo→vangogh. We observe, although the autoencoder is pretrained on irrelevant datasets, that our model can still generate reasonable translations. However, the most competitive results come from the pretrained autoencoders with the same dataset, which has better color controllability. The bottom row in Figure 9 compares CUT with our model on AFHQ cat→dog. Additional results are given in Appendix.

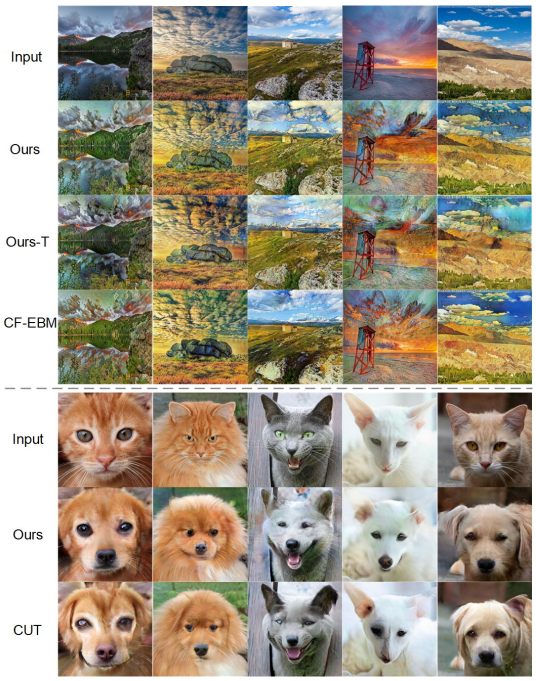

Figure 9. Qualitative comparisons based on VQ-VAE-2. Ours-T means this autoencoder is pretrained on AFHQ.

Quantitative results: We firstly compare our approach with the recent CF-EBM mode [45] on photo→vangogh. As shown in Table 5, our approach requires much less time than CF-EBM while achieving comparable KID. We postulate that: (i) A slightly higher KID is due to the blurry output from the decoder; (ii) The latent code is a high-level compact representation compared with the raw data so that our latent EBM requires less iterations to train. CUT is the current stat-of-the-art one-sided image translation approach on AFHQ cat→dog. We compare the performance including the sample quality and memory requirement. As shown in Table 5, our method achieves greatly better FID score while requiring much less GPU memory. The GPU memory is measured when both methods set the batch size to 1 on GTX 1080Ti.

Models	photo → vangogh		cat → dog	
	KID↓	Days	FID↓	Mem(GB)
CF-EBM [45]	**4.25**	0.6	55.1	2.50
CUT [31]	4.81	0.7	76.2	3.03
LETIT (Ours)	4.42	**0.1**	**45.2**	**1.24**

Table 5. VQ-VAE-2-based model comparison with CF-EBM and CUT. We only report available numbers from the original papers.

5.4. Training Cost

We compare the training cost between our approach and the two strong baselines, StarGAN v2 [5] and CUT [31]. As seen in Table 5, the time budget for training the latent EBM is incredibly low. Our model is still much more efficient even if the autoencoder training cost counts. Therefore, our approach is a light-weight and practical choice in scenarios where the computational resources are limited.

Models	#Param(M)	Time(days)	GPU
StarGAN v2 [5]	79	3.0	TITAN V100
CUT [31]	15	2.0	GTX 1080Ti
EBM-VQ-VAE-2 (ours)	17 +1.4	0.6 +0.3	GTX 1080Ti
EBM-ALAE (ours)	54 +1.6	1.1 +0.05	TITAN XP

Table 6. Comparison of computational cost on AFHQ. a+b denotes the time for pretraining is a, and b for training the EBM.

6. Conclusion

We propose an efficient and readily plug-and-play approach for one-sided unpaired image-to-image translation, via latent energy transport in a pretrained latent space. The introduced latent EBM learns to implicitly and simultaneously transfer styles and preserve content, without relying on a complicated cycle constraint. Extensive experiments on various autoencoder structures and datasets have demonstrated the strong performance, remarkable efficiency, practical faithfulness and scalability of the proposed universal approach.

Acknowledgments This work was supported in part by the Verizon Media FREP program.

References

[1] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. In *Advances in neural information processing systems*, pages 752–762, 2017.

[2] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[4] 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.

[5] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. Residual energy-based models for text generation. In *International Conference on Learning Representations*, 2020.

[8] Zheng Ding, Yifan Xu, Weijian Xu, Gaurav Parmar, Yang Yang, Max Welling, and Zhuowen Tu. Guided variational autoencoder for disentanglement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7920–7929, 2020.

[9] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In *Advances in Neural Information Processing Systems*, pages 3608–3618, 2019.

[10] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2427–2436, 2019.

[11] Ruiqi Gao, Erik Nijkamp, Diederik P Kingma, Zhen Xu, Andrew M Dai, and Ying Nian Wu. Flow contrastive estimation of energy-based models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7518–7528, 2020.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[14] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2016.

[15] I. Higgins, Loïc Matthey, A. Pal, C. Burgess, Xavier Glorot, M. Botvinick, S. Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

[16] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision*, pages 172–189, 2018.

[17] 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*, pages 1125–1134, 2017.

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[20] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658, 2018.

[21] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwang Hee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *International Conference on Learning Representations*, 2020.

[22] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[24] Paul Langevin. On the theory of brownian motion. *Comptes rendus*, 1908.

[25] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision*, pages 35–51, 2018.

[26] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017.

[27] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[28] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, pages 1429–1437, 2019.

[29] Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run mcmc toward energy-based model. In *Advances in Neural Information Processing Systems*, pages 5232–5242, 2019.

[30] Bo Pang, Tian Han, Erik Nijkamp, Song-Chun Zhu, and Ying Nian Wu. Learning latent space energy-based prior model. *arXiv preprint arXiv:2006.08205*, 2020.

[31] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, 2020.

[32] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[33] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in Neural Information Processing Systems*, pages 14866–14876, 2019.

[34] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *arXiv preprint arXiv:2007.03898*, 2020.

[35] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016.

[36] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NIPS*, pages 6309–6318, 2017.

[37] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning*, pages 681–688, 2011.

[38] Zhisheng Xiao, Karsten Kreis, Jan Kautz, and Arash Vahdat. Vaebm: A symbiosis between variational autoencoders and energy-based models. *arXiv preprint arXiv:2010.00654*, 2020.

[39] Jianwen Xie, Yang Lu, Ruiqi Gao, Song-Chun Zhu, and Ying Nian Wu. Cooperative training of descriptor and generator networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(1):27–45, 2018.

[40] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet. In *International Conference on Machine Learning*, pages 2635–2644, 2016.

[41] Jianwen Xie, Zilong Zheng, Xiaolin Fang, Song-Chun Zhu, and Ying Nian Wu. Learning cycle-consistent cooperative networks via alternating mcmc teaching for unsupervised cross-domain translation. 2021.

[42] 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.

[43] Lantao Yu, Yang Song, Jiaming Song, and Stefano Ermon. Training deep energy-based models with f-divergence minimization. In *International Conference on Machine Learning*, 2020.

[44] Yang Zhao, Chunyuan Li, Ping Yu, Jianfeng Gao, and Changyou Chen. Feature quantization improves gan training. In *International Conference on Machine Learning*, pages 11376–11386. PMLR, 2020.

[45] Yang Zhao, Jianwen Xie, and Ping Li. Learning energy-based generative models via coarse-to-fine expanding and sampling. In *International Conference on Learning Representations*, 2021.

[46] 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.