LEARNING ENERGY-BASED GENERATIVE MODELS VIA COARSE-TO-FINE EXPANDING AND SAMPLING

Yang Zhao, Jianwen Xie, Ping Li

Cognitive Computing Lab
Baidu Research
10900 NE 8th St. Bellevue, WA 98004, USA
{yangzhao.eric, jianwen.kenny, pingli98}@gmail.com

ABSTRACT

Energy-based models (EBMs) parametrized by neural networks can be trained by Markov chain Monte Carlo (MCMC) sampling-based maximum likelihood estimation. Despite the recent significant success of EBMs in data generation, the current approaches to train EBMs can be unstable and sometimes may have difficulty synthesizing diverse and high-fidelity images. In this paper, we propose to train EBMs via a multistage coarse-to-fine expanding and sampling strategy, namely CF-EBM. To improve the learning procedure, we propose an effective net architecture and advocate applying smooth activations. The resulting approach is computationally efficient and achieves the best performance on image generation amongst EBMs and the spectral normalization GAN. Furthermore, we provide a recipe for being the first successful EBM to synthesize 512×512-pixel images and also improve out-of-distribution detection. In the end, we effortlessly generalize CF-EBM to the one-sided unsupervised image-to-image translation and beat baseline methods with the model size and the training budget largely reduced. In parallel, we present a gradient-based discriminative saliency method to interpret the translation dynamics which align with human behavior explicitly.

1 INTRODUCTION

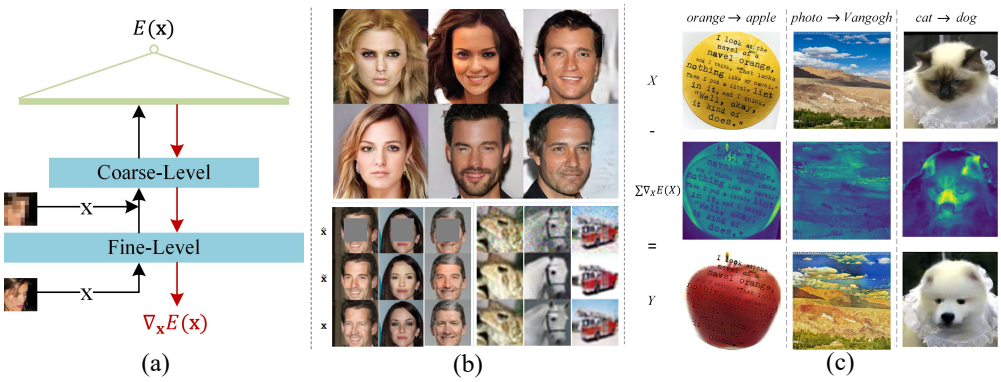


Figure 1: Highlights of our contributions. (a) Coarse-to-fine EBM with gradient-based MCMC sampling: $\mathbf{x} \leftarrow \mathbf{x} - \nabla_{\mathbf{x}} E(\mathbf{x})$; (b) Image generation (512^2), image denoising and image inpainting from top to bottom. ($\hat{x}$: corrupted image, $\tilde{x}$: restored image, x: ground truth); (c) Unsupervised image-to-image translation (256^2) interpreted by the discriminative saliency $\sum \nabla_X E(X)$.

Recently, energy-based models (EBMs) (Zhu et al., 1998; LeCun et al., 2006) parameterized by modern neural networks have drawn much attention from the deep learning communities. Successful applications with EBMs include generations of images (Xie et al., 2016; 2018b; Du & Mordatch, 2019), videos (Xie et al., 2017; 2019), 3D volumetric shapes (Xie et al., 2018c; 2020), unordered point clouds (Xie et al., 2021a), texts (Deng et al., 2020), molecules (Ingraham et al., 2018; Du

et al., 2019), etc., as well as image-to-image translation (Xie et al., 2021b;c), out-of-distribution detection (Liu et al., 2020) and inverse optimal control (Xu et al., 2019). EBMs are characterized by (i) Simplicity: The maximum likelihood learning of EBMs unifies representation and generation in a single model, and (ii) Explicitness: EBMs provide an explicit density distribution of data by training an energy function that assigns lower values to observed data and higher values to unobserved ones.

However, it is still difficult to train an EBM to synthesize diverse and high-fidelity images. The maximum likelihood estimation (MLE) of EBMs requires the Markov Chain Monte Carlo (MCMC) (Liu, 2008; Barbu & Zhu, 2020) to sample from the model and then updates the model parameters according to the difference between those samples and the observed data. Such an "analysis by synthesis" (Grenander et al., 2007) learning scheme is challenging because the sampling step is inefficient, unstable, and hard to traverse modes of the learned model, especially when the energy function is multimodal due to highly varied or high resolution training data. Fortunately, it is common knowledge that the manifold residing in a downsampled low-dimensional image space is smoother than that in the original high-dimensional counterpart. Thus, learning an EBM from low-dimensional data is much stabler and faster in terms of convergence (Odena et al., 2017; Gao et al., 2018).

Inspired by the above, we propose to train EBMs via a multistage coarse-to-fine expanding and sampling strategy (CF-EBM). As shown in Figure 1(a), the approach starts with learning a coarse-level EBM on low resolution images and then smoothly transits to learn the finer-level EBM by adding new layers that take into account the higher resolution information as the learning progresses. The gradient-based short-run MCMC (Nijkamp et al., 2019), *e.g.*, Langevin dynamics (Neal et al., 2011), is used for sampling. From the modeling aspect, the coarse-level training can be useful for exploring the global structure of image, while the fine-level training will then gradually refine the image details. Recent works have demonstrated the advantages of this incremental learning (Karras et al., 2018; Wang et al., 2018). However, there have been no works focusing on the incremental learning of EBMs that incorporates bottom-up modeling and top-down sampling in a single net. Besides, as shown in Figure 1(a), the top-down gradient information for synthesis flows through coarse-level layers towards fine-level layers. Thus, during the coarse-to-fine expanding, we can use the coarse-level synthesis to help the fine-level synthesis to stabilize the sampling. Such a coarse-to-fine expanding and sampling scheme is useful for high-fidelity synthesis in several vision tasks. See Figure 1(b).

Furthermore, we propose a one-sided energy-based unsupervised image-to-image translation method that can scale up to high resolution. The approach is immediately available with the FC-EBM by using iterative Langevin dynamic without an explicit cycle consistency (Zhu et al., 2017) or geometry constraints (Fu et al., 2019). Meanwhile, it works as a saliency method (Simonyan et al., 2014; Adebayo et al., 2018) by using gradients to explain essential aspects, namely the discriminative saliency, in translating from the source domain $\mathcal{X}$ to the target domain $\mathcal{Y}$ as shown in Figure 1(c). To accomplish this, EBM learns a scalar energy function $E : \mathcal{Y} \to \mathbb{R}$ and applies MCMC to sample $\mathcal{Y}$ starting from $\mathcal{X}$, instead of learning a block-box domain mapping network $F : \mathcal{X} \to \mathcal{Y}$ in prior works (Zhu et al., 2017; Huang et al., 2018; Park et al., 2020).

The contributions of our paper can be summarized as:

- To the best of our knowledge, this is the first work that trains EBMs under the "analysis by synthesis" scheme via a multistage coarse-to-fine expanding and sampling strategy. Besides, we propose eliminating the additive noise to the observed data while experimenting with several essential techniques for improving EBM, *e.g.* smooth activations. In particular, CF-EBM is the first pure EBM approach applied to model images as large as 512×512.
- We propose a novel energy-based unsupervised image translation approach, which is essentially different from all other existing GAN-based approaches. We demonstrate noticeable results in terms of both translation quality and efficiency of time and memory.
- We conduct extensive experiments to validate our approach, including image generation, denoising, inpainting, out-of-distribution detection and unsupervised image-to-image translation. Strong results are shown to outperform or be competitive with the prior art mostly.

The rest of the paper is organized as follows. Section 2 summarizes related works and how our proposal is differentiated from those. Section 3 introduces the proposed methodology in detail. Section 4 presents extensive experiments to test our method. Section 5 concludes our paper and discuss some future research directions.

2 RELATED WORK

2.1 ENERGY-BASED GENERATIVE MODELS

The main challenge to train EBMs via MLE lies in drawing fair samples from the model, especially when the energy function is parameterized by a highly non-linear ConvNet. The contrastive divergence (CD) (Hinton, 2002; Tieleman, 2008), with MCMC chains initialized from data distribution, is an efficient but biased way to train EBMs. Another direction is to adopt the idea of energy-based correction of a more tractable model to train EBMs. Noise contrastive estimation (NCE) (Gutmann & Hyvärinen, 2010; Gao et al., 2020) and introspective neural networks (INN) (Lazarow et al., 2017; Jin et al., 2017; Lee et al., 2018b) belong to this theme. Generative cooperative networks (CoopNets) (Xie et al., 2018b; 2021d;b) train an EBM with a generator or a variational auto-encoder (VAE) (Kingma & Welling, 2014) as amortized sampler by MCMC teaching (Xie et al., 2018a). Triangle divergence (Han et al., 2019) trains an EBM without MCMC by amortizing the MCMC via a VAE. However, these frameworks still struggle to scale up and model multimodal data. There have been several strategies to improve the EBM training. Gao et al. (2018) adopts a multi-grid method that trains multiple EBMs at different grids simultaneously, where the EBM at coarser grid is used to initialize the image generation by EBM at finer grid. However, optimizing and sampling from multiple EBMs will result in low efficiency of both time and memory. To stabilize the training, Nijkamp et al. (2019); Grathwohl et al. (2020) add the Gaussian white noise to the observed data, resulting in noisy synthesized images. In contrast, our paper proposes to train a single EBM via a coarse-to-fine growing strategy, along with some improved techniques. With smooth parameter training and image sampling, our model can preserve EBM's compatibility and synthesize high-fidelity images.

2.2 UNSUPERVISED IMAGE-TO-IMAGE TRANSLATION

Unsupervised image-to-image translation aims at learning two directions of mappings between two unpaired domains. Recent successes are all based on adversarial learning, *e.g.*, CycleGAN (Zhu et al., 2017), UNIT (Liu et al., 2017), MUNIT (Huang et al., 2018), DRIT (Lee et al., 2018a) and U-GAT-IT (Kim et al., 2020). These methods typically train two GANs with two levels of learning objectives: (i) Distribution level: Two adversarial losses are used to capture style discrepancy between source and target domain; (ii) Instance level: To tackle the difficulty of unpaired setting, they adopt a cycle consistency loss for content preservation. This loss enables an instance-level supervision to regularize the training of two mappings by enforcing them to be a bijective function between two domains. Except for works about two-sided translation, efforts on research about one-sided unsupervised image translation have also been made, *e.g.*, DistanceGAN (Benaim & Wolf, 2017), GcGAN (Fu et al., 2019) and CUT (Park et al., 2020), which apply geometric or contrastive constraints. We solve this problem from the prospective of EBM, which is different from GAN-based methods. The proposed concise EBM solution only relies on its built-in objective, which is a distribution-level statistics matching, to accomplish the one-sided image translation. It transfers the style and preserves the source content by MCMC without using the cycle-consistency loss. The model demonstrates better performances with less time and memory. Another distinction between our method and GAN-based methods is the natural interpretability of Langevin dynamics. It provides a gradient-based saliency map to visualize those key regions that make the two domains distinct, as illustrated in Figure 1(c).

3 METHOD

In this section, we first present the EBM learning framework, and then the proposed CF-EBM approach. After that, we generalize our model to the unsupervised image-to-image translation.

3.1 MCMC-BASED MAXIMUM LIKELIHOOD LEARNING OF ENERGY-BASED MODEL

Let $\mathbf{x} \in \mathbb{R}^D$ be the observed example, *e.g.* an image. An energy-based model is defined as follows:

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

where $E_\theta(\mathbf{x}) : \mathbb{R}^D \to \mathbb{R}$ is the energy function defined by a bottom-up ConvNet parameterized by θ. $Z(\theta) = \int \exp(-E_\theta(\mathbf{x}))d\mathbf{x}$ is the intractable normalizing constant or the partition function. Given N observed examples $\{\mathbf{x}_i\}_{i=1}^N \sim p_{\text{data}}(\mathbf{x})$, where $p_{\text{data}}(\mathbf{x})$ denotes the unknown data distribution, the model can be trained by maximizing the log-likelihood $L(\theta) = \frac{1}{N}\sum_{i=1}^N \log p_\theta(\mathbf{x}_i) \approx$

$\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})} \log(p_\theta(\mathbf{x}))$. The derivative of the negative log-likelihood is given by

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

where the second expectation term under $p_\theta(\mathbf{x})$ is intractable and can be approximated via MCMC. Given that, the EBM is updated by gradient descent. To sample $\tilde{\mathbf{x}} \sim p_\theta(\mathbf{x})$ via MCMC, we rely on gradient-based Langevin dynamics that recursively computes the following step

$$\tilde{\mathbf{x}}_{t+1} = \tilde{\mathbf{x}}_t - \frac{\eta_t}{2}\nabla_{\tilde{\mathbf{x}}} E_\theta(\tilde{\mathbf{x}}_t) + \sqrt{\eta_t}\boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I}), \tag{3}$$

where η_t is the step size of Langevin step and also the variance of Gaussian noise $\boldsymbol{\epsilon}_t$. Theoretically, to ensure convergence, the MCMC is typically performed with infinite steps and an infinitesimal stepsize (Welling & Teh, 2011). However, it is impractical for training EBMs. In this paper, we follow Nijkamp et al. (2019) to use short-run MCMC, which always starts from a fixed distribution and runs a fixed number T of Langevin steps in both training and testing stages. The training with a short-run MCMC might result in a biased estimation of EBM but the learned short-run MCMC is still a valid generator, which enables us to synthesize realistic images and efficiently train the model, as seen in most well-established EBM works (Nijkamp et al., 2019; Grathwohl et al., 2020; Pang et al., 2020). In this paper, we keep the step size constant and linearly decay the noise variance till 0.

3.2 CF-EBM: Coarse-to-fine Expanding and Sampling of Energy-Based Model

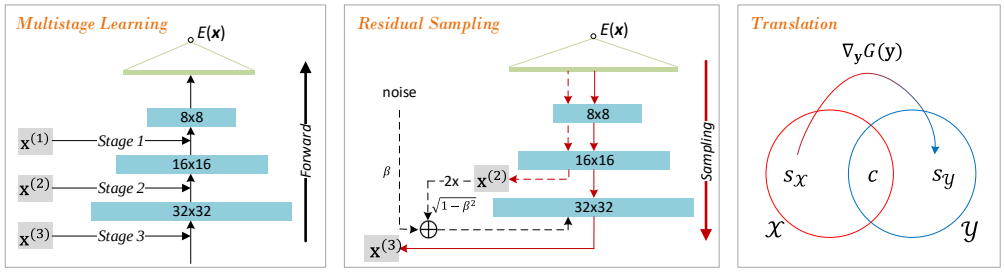
Our primary contribution is the multistage coarse-to-fine expanding and sampling methodology for training EBMs. The key idea is to incrementally grow the EBM from a low resolution (coarse model) to a high resolution (fine model) by gradually adding new layers to the energy function. In this way, both stability and time efficiency in training EBMs benefit. And eventually, we only keep the highest-resolution EBM for image generation using the short-run MCMC sampling.

Figure 2 (a) illustrates the proposed multistage coarse-to-fine EBM training strategy, which is accompanied by the pseudocode in Algorithm 1. Let S denote the total number of training stages and $(\mathbf{x}^{(s)},\ s = 1, ..., S)$ denote the multi-resolution versions of an image $\mathbf{x}$, with $\mathbf{x}^{(0)}$ being the minimal resolution version of $\mathbf{x}$, and $\mathbf{x}^{(S)} = \mathbf{x}$. For each $\mathbf{x}^{(s)}$, we can easily generate $\mathbf{x}^{(s-1)}$ by average pooling with a pool-

Algorithm 1: CF-EBM Training

Input: Multi-resolution data $\{\mathbf{x}_i^{(s)}, i = 1, ..., N; s = 1, ..., S\}$, numbers of Langevin steps $\{T^{(s)}\}_{s=1}^S$, and batch sizes $\{n^{(s)}\}_{s=1}^S$
Output: Energy function $E^{(S)}$

$E^{(0)} \leftarrow \varnothing$
for $s = 1 : S$ **do**
  $m = 0,\ \beta = 0$
  Expand model: $E^{(s)} \leftarrow E^{(s-1)} + \texttt{Expand}(\cdot)$
  while $(m \leq N)$ **do**
    # Draw real data and initial MCMC samples
    Draw observed $\{\mathbf{x}_i^{(s)}\}_{i=1}^{n^{(s)}} \sim p_{data}(\mathbf{x})$
    Draw initial $\{\tilde{\mathbf{x}}_i^{(s)}\}_{i=1}^{n^{(s)}} \sim \mathcal{U}^{(s)}(-1, 1)$
    # Smooth MCMC sampling
    if $(\beta < 1$ and $s > 1)$ **then**
      Draw initial $\{\tilde{\mathbf{x}}_i^{(s-1)}\}_{i=1}^{n^{(s)}} \sim \mathcal{U}^{(s-1)}(-1, 1)$
      Update $\{\tilde{\mathbf{x}}_i^{(s-1)}\}_{i=1}^{n^{(s)}}$ for $T^{(s-1)}(1-\beta)$ steps of Eq. (3) with $E^{(s-1)}$
      $\tilde{\mathbf{x}}_i^{(s)} = \sqrt{1-\beta^2}\texttt{Upsample}(\tilde{\mathbf{x}}_i^{(s-1)}) + \beta\mathcal{U}^{(s)}(-1, 1)$
    end
    Update $\{\tilde{\mathbf{x}}_i^{(s)}\}_{i=1}^{n^{(s)}}$ for $T^{(s)}$ steps of Eq. (3) with $E^{(s)}$
    # Optimize energy function
    Update $E^{(s)}$ with gradient descent in Eq. (2)
    $m \leftarrow m + n^{(s)},\ \beta = \max(1,\ 2N/n^{(s)})$
  end
end



(a) (b) (c)

Figure 2: The illustration of the proposed multistage CF-EBM.

ing size. At stage s, we associate the model with an energy function $E^{(s)}$ to learn from $\mathbf{x}^{(s)}$. The model starts from an initial minimal resolution at stage 1. When the training proceeds to the next stage $s+1$, we will add randomly initialized layers at the bottom to expand the resolution, and obtain a new energy function $E^{(s+1)}$. For example, in Figure 2 (a), when the training stage switches from stages 2 to 3, we expand $E^{(2)}$, which is an energy function of resolution 16×16, by adding a bottom block accounting for resolution 32×32 to get $E^{(3)}$. After that, we only need to train $E^{(3)}$ at stage 3.

Compared with PGAN (Karras et al., 2018) which proposes a progressive growing strategy to train GAN, our CF-EBM integrates both the learning (Eq. (2)) and the sampling (Eq. (3)) into a single neural network $E_\theta(\mathbf{x})$, which is essentially different from PGAN that applies a discriminator to guide the learning and a generator to produce samples. Therefore, training an EBM via a multistage growing strategy is more challenging and non-trivial. With the newly expanded energy function $E^{(s+1)}$ and the enlarged training examples $\mathbf{x}^{(s+1)}$, the stage transition must be taken care of to avoid instability issue. The CF-EBM deals with this issue from the following two aspects:

(i) Smooth EBM Learning: Let $\texttt{Expand}(\cdot)$ denote an expanding block. To expand the resolution of the EBM at each stage, we add an $\texttt{Expand}(\cdot)$ to the bottom of the energy function to double its resolution. We fade in the block smoothly to prevent sudden shocks to the well-trained coarse-level energy function. Specifically, $\texttt{Expand}(\cdot)$ is a composition of a primal block $\texttt{Primal}(\cdot)$ and a fading block $\texttt{Fade}(\cdot)$. We define $\texttt{Expand}(\mathbf{x}) = \beta\texttt{Primal}(\mathbf{x}) + (1-\beta)\texttt{Fade}(\mathbf{x})$, where β is a fading factor that gradually increases from 0 to 1 as more data are consumed in the training. As a result, the model will first rely on the well-trained coarse-level layers and gradually move the attention to the newly added fine-level layers. The auxiliary fading block will be abandoned when β reaches 1. In this paper, $\texttt{Primal}(\cdot)$ consists of two convolution layers and $\texttt{Fade}(\cdot)$ is a convolution layer followed by a 2×2 average pooling to link the previous coarse-level layers.

(ii) Smooth MCMC Sampling: The smooth sampling is realized both implicitly and explicitly.

(a) *Implicitly*: Considering the resolution transition from stage $s-1$ to stage s, we expand the energy function by $E^{(s)}(\mathbf{x}^{(s)}) = E^{(s-1)}(\texttt{Expand}(\mathbf{x}^{(s)}))$. The gradient $\nabla_{\mathbf{x}^{(s)}} E^{(s)}(\mathbf{x}^{(s)})$ for Langevin sampling in Eq. (3) can be unfolded as:

$$\nabla_{\texttt{Expand}(\mathbf{x}^{(s)})} E^{(s-1)}(\texttt{Expand}(\mathbf{x}^{(s)})) \times [\beta \nabla_{\mathbf{x}^{(s)}} \texttt{Primal}(\mathbf{x}^{(s)}) + (1-\beta)\nabla_{\mathbf{x}^{(s)}} \texttt{Fade}(\mathbf{x}^{(s)})].$$

Initially, because β is small, the newly added primal block $\texttt{Primal}(\cdot)$, whose weights are randomly initialized, would not largely affect the sampling immediately. Only the coarse-level function $E^{(s-1)}$ and the auxiliary fading block $\texttt{Fade}(\cdot)$ greatly contribute to the synthesis. As β increases, the $\texttt{Primal}(\mathbf{x}^{(s)})$ becomes increasingly well-trained. Meanwhile, the term $\nabla_{\mathbf{x}^{(s)}} \texttt{Fade}(\mathbf{x}^{(s)})$ fades away and $\nabla_{\mathbf{x}^{(s)}} \texttt{Primal}(\mathbf{x}^{(s)})$ gradually takes the lead in the sampling, thus the synthesized images would become sharper and sharper.

(b) *Explicitly*: To mitigate the impact of the sudden model expansion on sampling, we utilize the well-trained coarse model $E^{(s-1)}$ to initialize the MCMC sampling of the newly expanded model $E^{(s)}$. Specifically, at stage s, the algorithm first generates low-resolution samples from $E^{(s-1)}$ by running $T^{(s-1)}(1-\beta)$ Langevin steps. Those samples are then upsampled ($2\times$) by $\texttt{Upsample}(\cdot)$ and mixed with a uniform noise to initialize the MCMC sampling of $E^{(s)}$. As β increases, such an MCMC initialization assistance from the coarse model would fade away. Eventually, when β reaches 1, we can directly sample from $E^{(s)}$ with a purely noise-initialized MCMC. Figure 2 (b) illustrates the smooth sampling process at stage 3. We use different numbers of Langevin steps at different stages. For example, we only run MCMC with $T^{(1)} = 15$ steps at stage 1, and gradually increase the number of steps as the model grows. The maximal number of steps is 60 in our paper.

After training, we obtain the target EBM with $E^{(S)}$. The sampling procedure by the learned model is presented in Algorithm 2. The MCMC starts from a uniform distribution and runs $T^{(S)}$ steps of Langevin updates to generate samples. Note that, both the initial distribution and the number of Langevin steps are the same in both training and sampling. If we run MCMC with more steps than that used in training, the oversaturation occurs (Nijkamp et al., 2019).

Algorithm 2: CF-EBM Sampling

Input: model $E^{(S)}$, $T^{(S)}$ steps
Output: samples $\tilde{\mathbf{x}}^{(S)}$
Draw initial $\tilde{\mathbf{x}}^{(S)} \sim \mathcal{U}^{(S)}(-1, 1)$
for $t = 1 : T^{(S)}$ **do**
  Update $\tilde{\mathbf{x}}^{(S)}$ by using Eq. (3)
end

3.3 ENERGY-BASED UNSUPERVISED IMAGE-TO-IMAGE TRANSLATION

We then generalize the proposed approach to the task of unpaired image-to-image translation and scale it up to high resolution image datasets. Given two image domains, $\mathcal{X}$ and $\mathcal{Y}$, endowing two ground-truth distributions $p(\mathbf{x}) : \mathbf{x} \in \mathcal{X}$ and $p(\mathbf{y}) : \mathbf{y} \in \mathcal{Y}$, the goal is to learn two energy functions $F(\mathbf{x}) : \mathcal{X} \to \mathbb{R}$ and $G(\mathbf{y}) : \mathcal{Y} \to \mathbb{R}$. Here, we only explain how to learn G as the vice versa should be straightforward. The target distribution $p_G(\mathbf{y}) \propto \exp(-G(\mathbf{y}))$. Hence, the image translation process is given by:

$$\tilde{\mathbf{y}}_{t+1} = \tilde{\mathbf{y}}_t - \frac{\eta_t}{2}\nabla_{\tilde{\mathbf{y}}} G(\tilde{\mathbf{y}}_t) + \sqrt{\eta_t}\boldsymbol{\epsilon}_t,\ \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I}),\ \tilde{\mathbf{y}}_0 = \mathbf{x} \sim p_{\text{data}}(\mathbf{x}). \tag{4}$$

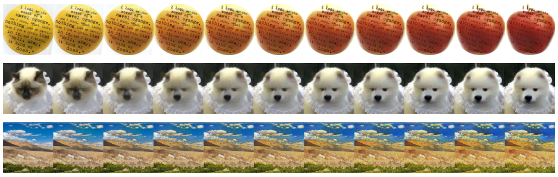


Figure 3: Discriminative saliency evolution.

The underlying assumption of applying EBM for image translation is that the two domains $(\mathcal{X}, \mathcal{Y})$ share the same ambient space where each data can be decomposed into a content code c and a domain specific style code $s_{\mathcal{X}}$ of domain $\mathcal{X}$ or $s_{\mathcal{Y}}$ of domain $\mathcal{Y}$ respectively. This assumption is somewhat similar to the partially shared latent space assumption in Huang et al. (2018); Choi et al. (2020); however, it is directly defined in the ambient space and is more efficient without resorting to complementary models. Consequently, as shown in Figure 2(c) and Figure 3, the above dynamic aims to evolve the style code $s_{\mathcal{X}}$ towards $s_{\mathcal{Y}}$, where the gradient descent steps into a lower energy area with content c unchanged as much as possible.

Discriminative Saliency Map: With the proposed method, we visualize the image translation dynamic in Figure 3. The three datasets are selected for demonstration because each contains different kinds of discriminative saliency, *e.g.* color and texture. But how are we supposed to know this? We adopt the gradient-based saliency maps (Simonyan et al., 2014) to highlight the discriminative aspect that leads to the translation, which we call *discriminative saliency map*. It is computed as the aggregated scores in Eq. (4), *i.e.*, $\sum_t \nabla_{\mathbf{y}_t} G(\mathbf{y}_t)$. It quantifies how much a change of each pixel would contribute to the image translation. From Figure 4, we can conclude all types of discriminative saliency in different datasets, *e.g.*, color and shape in *orange→apple*, and nose and ear in *cat→dog*.

(a) *cat → dog* (b) *photo → Vangogh* (c) *orange → apple*

Figure 4: Gradient-based discriminative saliency maps.

3.4 SMOOTH ACTIVATIONS IMPROVES TRAINING

We further claim that smooth activations potentially strengthen EBM learning. Given a simplest EBM model $E(\mathbf{x}) = Act(Block(\mathbf{x}))$ with a single $Block(\cdot)$ activated by $Act(\cdot)$, the score function in Eq. (3) is $\nabla_{\mathbf{x}} E(\mathbf{x}) = Act'(Block(\mathbf{x}))Block'(\mathbf{x})$. When the data $\mathbf{x}$ is continuous, the smoothness of $Act'(\cdot)$ will substantially affect the iterative sampling process, thus applying an activation with continuous gradients everywhere that enforces a smooth Langevin dynamics is potentially useful. In this regard, we consider common activations for demonstration, including ReLU, LeakyReLU, GELU, CELU, and Swish. See Appendix A.4 for their functional forms and visualizations. In Table 8, we observe that non-smooth activations ReLU and LeakyReLU tend to diverge, whereas others don't.

Note that most well-established EBM generative models add Gaussian noise to smooth data and make it easier for model convergence (Nijkamp et al., 2019; Grathwohl et al., 2020). Otherwise, the model will diverge quickly. Similarly, score-based models (Song & Ermon, 2019; 2020) inject multiple

level noises to improve score estimations. However, the additive noisy information will be leaked to the sampler and result in foggy synthesized samples (Grathwohl et al., 2020; Song & Ermon, 2020). In this papaer, we integrate smooth activations with the proposed model such that we can alleviate the instability issue effortlessly while preserving original data structures. See Appendix A.3.1 for more discussions and Figure 12 for a visualization of perturbed samples.

4 Experiments

In the main part, we showcase the results on various tasks, including image generation, image restoration (inpainting and denoising), out-of-distribution (OOD) detection and unpaired image-to-image translation. These are implemented on the same base neural architecture and TITAN X (12GB) platform mostly. Detailed experimental settings are provided in Appendix A.2. We adopt Fréchet Inception Distance (FID) (Heusel et al., 2017) and Kernel Inception Distance (KID) (Bińkowski et al., 2018) for quantitative evaluation , and Amazon Turker Platform (AMT) for perceptual human evaluation (detailed in Appendix A.1). KID score is proved more reliable when there are fewer testing images available, *e.g.* in the image translation. In Appendix A.4, we conduct a detailed ablation study, including activation functions, normalization strategies, and layer connections. Finally, in Appendix A.6, we scale the model up to 512^2-pixel images (e.g., CelebA-HQ) when all intermediate CF-EBMs learned in Algorithm 1 are incorporated in the learning and sampling process.

4.1 Image generation

Datasets (i) CIFAR-10 (Krizhevsky, 2009) is a dataset containing 60k images (32×32) of 10 classes; (ii) CelebA (Liu et al., 2015) is a celebrity facial dataset containing over 200k images. For fair comparisons with former EBM works, the 64×64 resolution is used for quantitative evaluation; (iii) CelebA-HQ 512×512 contains 30K high resolution facial images (Karras et al., 2018).

Approach	Models	FID
VAE	VAE (Kingma & Welling, 2014)	78.41
Autoregressive	PixelCNN Van den Oord et al. (2016)	65.93
	PixelIQN (Ostrovski et al., 2018)	49.46
GAN	WGAN-GP (Gulrajani et al., 2017)	36.40
	SN-GAN (Miyato et al., 2018)	21.70
	StyleGAN2-ADA (Karras et al., 2020)	**2.92**
Flow	Glow (Kingma & Dhariwal, 2018)	45.99
	Residual Flow (Chen et al., 2019a)	46.37
	Contrastive Flow (Gao et al., 2020)	37.30
Score matching	MDSM (Li et al., 2020)	30.93
	NCSN (Song & Ermon, 2019)	25.32
	NCK-SVGD (Chang et al., 2020)	21.95
EBM	Short-run EBM (Nijkamp et al., 2019)	44.50
	Multi-grid (Gao et al., 2018)	40.01
	EBM(ensemble) (Du & Mordatch, 2019)	38.20
	CoopNets (Xie et al., 2018b)	33.61
	CF-EBM	**16.71**

Table 1: FID on CIFAR-10.

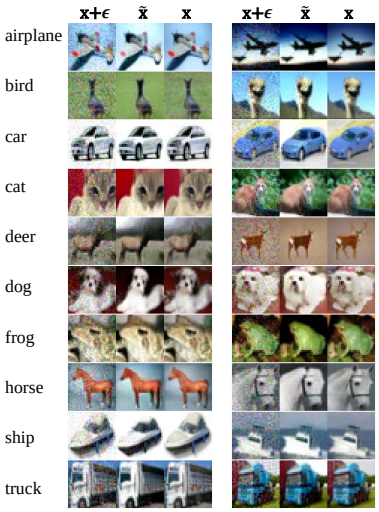


Figure 5: Image denoising.

Models	FID
VAE (Kingma & Welling, 2014)	79.09
DCGAN (Radford et al., 2016)	32.71
NCSNv2 (Song & Ermon, 2020)†	28.90
Short-run EBM (Nijkamp et al., 2019)	23.02
WGAN-GP (Gulrajani et al., 2017)†	21.40
mr-Langevin (Block et al., 2020)†	19.54
CF-EBM	**14.35**
CF-EBM†	**10.80**

Table 2: FID on CelebA.

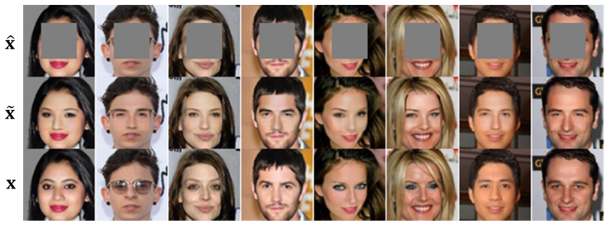


Figure 6: Image inpainting.

Quantitative Results: The quantitative unconditional image generation results on CIFAR-10 and CelebA are given in Table 1 and Table 2. In Table 2, † works are experimented on the center-

crop(140×140)-resized CelebA whereas others are learned on the resized CelebA. In all cases, the proposed CF-EBM is leading by a wide margin over other approaches. Du & Mordatch (2019) claims that large models are necessary to learn multimodal data. To examine this, in Table 3, we compare the number of parameters and relative evaluation time of representative open-sourced EBMs and score matching models since both kinds rely on Langevin dynamics for sampling. We use 50 steps for sampling (cross-validated) while all other three methods require either 60, 100 or even more steps. It is observed that our model consistently stands out in terms of efficiency while keeping the lowest FID on CIFAR-10.

Model Analysis: EBM generative models, once learned, can be directly applied to image restoration tasks. As in Du & Mordatch (2019), we conduct an experiment on the image restoration to demonstrate the model's ability of mode coverage and generalization to the unseen data. (i) *Image denoising*: We randomly pick up images from the testing dataset of CIFAR-10. Then, we pollute each image $\mathbf{x}$ with an additive Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 0.2)$. Figure 5 shows the restored result $\tilde{\mathbf{x}}$ on all ten classes of CIFAR-10, implying a good generalization ability. (ii) *Image inpainting*: it can be viewed as associative memory (Hopfield, 1982). As shown in Figure 6, the recovered images don't mimic the ground truth with 25% image areas occluded. We also provide diverse image inpaintings in Figure 11 in Appendix A.3. We demonstrate the EBM characterized by MLE indeed learns meaningful representations We additionally provide nearest neighbor search results in Figure 10 to validate that our model doesn't suffer from over-fitting.

Models	#Params(M)	Time
EBM(ensemble)	18.7	21.3
NCSN	7.6	14.2
Short-run EBM	2.8	2.2
CF-EBM	**2.7**	**1.0**

Table 3: Model comparison.

Qualitative Samples: We have presented qualitative samples of Celeba-HQ 256×256 in Figure 1. Please refer to Appendix and see Figure 8 for CIFAR-10 samples, Figure 15 for how the noise level affects the samples' visual effect in the evaluation and Figures 16, 17 for more high resolution samples.

4.2 Out-of-distribution detection

Out-of-distribution (OOD) detection is a binary classification problem. Given input examples, the model must produce a score that we expect to be higher for in-distribution examples and lower for out-distribution examples. Du & Mordatch (2019) has shown that the likelihoods from EBMs can be a reliable OOD detector. According to this principle, we construct the OOD detection experiments and compare our model with the other three generative models, Glow (Kingma & Dhariwal, 2018), conditional EBM (Du & Mordatch, 2019) and JEM (Grathwohl et al., 2020). All models are trained on CIFAR-10. We use SVHN (Netzer et al., 2011), a uniform distribution (Uniform), a constant distribution (Constant), CIFAR10 Interpolation (Interp) and CIFAR100 (Krizhevsky, 2009) as OOD distributions. We apply the area under the receiver-operating curve (AUROC) (Hendrycks & Gimpel, 2016a) as a metric for evaluation. As shown in Table 4, our model shows mostly better or competitive results, even though it doesn't incorporate any label information at the training stage (Grathwohl et al., 2020).

Models	SVHN	Uniform	Constant	Interp	CIFAR100	Average
Glow	.07	1.0	.00	.45	.51	.41
EBM	.63	1.0	.30	.70	.50	.63
JEM	**.67**	1.0	.51	.65	**.67**	.70
CF-EBM (Ours)	.65	1.0	**.78**	**.81**	.59	**.77**

Table 4: AUROC scores of OOD detection. The higher, the better. The other three models are all conditional generative models except for our CF-EBM.

4.3 Unsupervised Image-to-Image translation

We mainly use four unpaired image-to-image translation datasets for evaluation, including cat2dog, Yosemite summer2winter, vangogh2photo and apple2orange. All images are resized to 256×256 resolutions. More details are provided in Appendix A.7.1. We compare our approach against CycleGAN (Zhu et al., 2017), CUT (Park et al., 2020) and two latest state-of-the-art methods U-GAT-IT (Kim et al., 2020) and FQ-GAN (Zhao et al., 2020). CycleGAN is a popular GAN-based image translation framework and we use it to demonstrate a basic comparison between two kinds of

models, *GAN-based model with cycle constraint* and the proposed *non-GAN-based model without cycle constraint*. CUT (Park et al., 2020) is the state-of-the-art one-sided unpaired image translation model which applies the contrastive loss to preserve the content. U-GAT-IT and FQ-GAN are current leading models on cat2dog and vangogh2photo. We present the baselines' results mostly from the public results or checkpoints.

Quantitative Results: Results are shown in Table 5. In most cases, our model obtains better KID scores and preference scores in the human evaluation. Meanwhile, the training time and model size are incredibly low. We postulate that this efficiency comes from two aspects: *model design* and *optimization process*.

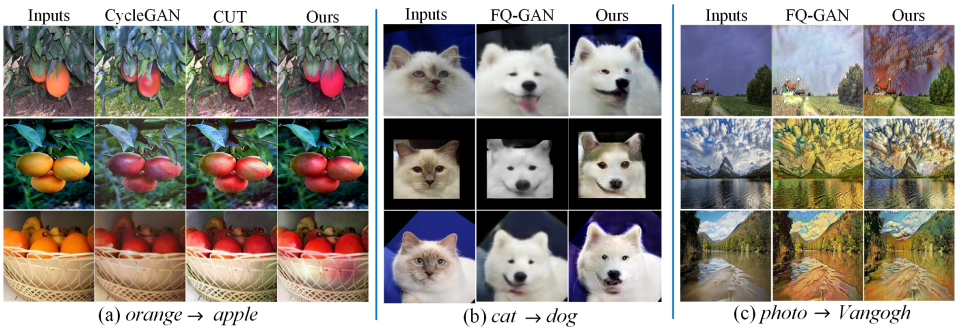
(i) *Model design*: The energy-based model is an image-to-scalar mapping instead of an image-to-image mapping as in GANs. It can also implicitly transfer styles and preserve contents without cycle constraint. Therefore, the model size will be smaller than GAN-based models where it applies two pairs of GANs.

(ii) *Optimization process*: GANs apply the adversarial loss to match styles and the cycle loss to preserve the content. In practice, the two losses are partially decoupled and optimized alternatively. In contrast, our approach only applies MLE on one neural network (Eq. (2)) and practically, we find that the training takes less iterations to uncover the discrepancy between two domains. Although we need to run additional Langevin sampling, the overall optimization cost is still low.

Models	$cat \to dog$		$dog \to cat$		$photo \to Vangogh$		$Vangogh \to photo$		**#Params**	**#Days**
	KID↓	AMT↑	KID↓	AMT↑	KID↓	AMT↑	KID↓	AMT↑		
CycleGAN	8.92	1.1	9.94	0.9	5.46	22.1	4.68	28.3	28.3M	-
U-GAT-IT	7.07	16.0	8.15	40.9	4.28	28.7	5.61	10.5	671M	6.0
FQ-GAN	6.44	31.3	8.90	35.3	6.54	15.9	5.21	20.4	671M	6.0
CF-EBM	**6.20**	**51.6**	9.21	22.9	**4.25**	**33.3**	**4.49**	**40.8**	**649K**	**0.6**

Table 5: Quantitative results for different image translation datasets.

Qualitative Results: As can be seen in Figure 7, our model can generate sharper images with superior visual quality. The source content is better preserved while focusing on evolving the source style to the target domain. We put the comparison between our model and one-sided translation model CUT (Park et al., 2020) in Appendix A.7.2. More uncurated results on $photo \leftrightarrow Vangogh$ and Yosemite $summer \leftrightarrow winter$ can be found in Appendix A.7.3.



(a) *orange* → *apple* (b) *cat* → *dog* (c) *photo* → *Vangogh*

Figure 7: Qualitative comparison with CycleGAN and FQ-GAN.

5 Conclusion

To tackle the training instability and the multimodal MCMC sampling difficulty of EBM generative models, we propose a coarse-to-fine expanding and sampling strategy, coined CF-EBM. Extensive experiments demonstrate the superior performance compared with various generative models, in sample quality and computational efficiency. This success is due to our newly proposed architecture, smoothing activations, and the flexibility of EBM. To the best of our knowledge, CF-EBM is the first pure EBM generative model that can produce large high-fidelity images and be competent in unsupervised image-to-image translation. The future work includes investigating EBM's intrinsic mechanics, experimenting on large architectures, and applications on other types of learning scenarios.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 9505–9515, 2018.

Adrian Barbu and Song-Chun Zhu. *Monte Carlo Methods*. Springer, 2020.

Jonathan T Barron. Continuously differentiable exponential linear units. *arXiv preprint arXiv:1704.07483*, 2017.

Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 752–762, 2017.

Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, 2018.

Adam Block, Youssef Mroueh, Alexander Rakhlin, and Jerret Ross. Fast mixing of multi-scale langevin dynamics underthe manifold hypothesis. *arXiv preprint arXiv:2006.11166*, 2020.

Wei-Cheng Chang, Chun-Liang Li, Youssef Mroueh, and Yiming Yang. Kernel stein generative modeling. *arXiv preprint arXiv:2007.03074*, 2020.

Ricky TQ Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 9916–9926, 2019a.

Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised GANs via auxiliary rotation loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12154–12163, 2019b.

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 (CVPR)*, pp. 8188–8197, 2020.

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 (ICLR)*, 2020.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3608–3618, 2019.

Yilun Du, Joshua Meier, Jerry Ma, Rob Fergus, and Alexander Rives. Energy-based models for atomic-resolution protein conformations. In *International Conference on Learning Representations (ICLR)*, 2019.

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 (CVPR)*, pp. 2427–2436, 2019.

Ruiqi Gao, Yang Lu, Junpei Zhou, Song-Chun Zhu, and Ying Nian Wu. Learning generative convnets via multi-grid modeling and sampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9155–9164, 2018.

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 (CVPR)*, pp. 7518–7528, 2020.

Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations (ICLR)*, 2020.

Ulf Grenander, Michael I Miller, Michael Miller, et al. *Pattern Theory: From Representation to Inference*. Oxford university press, 2007.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein GANs. In *Advances in Neural Information Processing systems (NIPS)*, pp. 5767–5777, 2017.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 297–304, 2010.

Tian Han, Erik Nijkamp, Xiaolin Fang, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Divergence triangle for joint training of generator model, energy-based model, and inferential model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8670–8679, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2016a.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016b.

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 (NIPS)*, pp. 6626–6637, 2017.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences (PNAS)*, 79(8):2554–2558, 1982.

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 (ECCV)*, pp. 172–189, 2018.

John Ingraham, Adam Riesselman, Chris Sander, and Debora Marks. Learning protein structure with a differentiable simulator. In *International Conference on Learning Representations (ICLR)*, 2018.

Long Jin, Justin Lazarow, and Zhuowen Tu. Introspective classification with convolutional nets. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 823–833, 2017.

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 (ICLR)*, 2018.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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 (ICLR)*, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 10215–10224, 2018.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Justin Lazarow, Long Jin, and Zhuowen Tu. Introspective neural networks for generative modeling. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2774–2783, 2017.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 1(0), 2006.

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 (ECCV)*, pp. 35–51, 2018a.

Kwonjoon Lee, Weijian Xu, Fan Fan, and Zhuowen Tu. Wasserstein introspective neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3702–3711, 2018b.

Zengyi Li, Yubei Chen, and Friedrich T. Sommer. Annealed denoising score matching: learning energy based model in high-dimensional spaces. *arXiv preprint arXiv:1910.07762*, 2020.

Jun S Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Science & Business Media, 2008.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 700–708, 2017.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015.

Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Internatinal Conference on Machine Learning (ICML)*, 2010.

Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

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 (NeurIPS)*, pp. 5232–5242, 2019.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *International Conference on Machine Learning (ICML)*, pp. 2642–2651, 2017.

Georg Ostrovski, Will Dabney, and Remi Munos. Autoregressive quantile networks for generative modeling. In *International Conference on Machine Learning (ICML)*, pp. 3936–3945, 2018.

Bo Pang, Tian Han, Erik Nijkamp, Song-Chun Zhu, and Ying Nian Wu. Learning latent space energy-based prior model. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020.

Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2016.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. In *International Conference on Learning Representations (ICLR) Workshops*, 2016.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations (ICLR) Workshops*, 2014.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 11918–11930, 2019.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *arXiv preprint arXiv:2006.09011*, 2020.

Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *International Conference on Machine Learning (ICML)*, pp. 1064–1071, 2008.

Ngoc-Trung Tran, Viet-Hung Tran, Bao-Ngoc Nguyen, Linxiao Yang, et al. Self-supervised GAN: Analysis and improvement with multi-class minimax game. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 13253–13264, 2019.

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 (NIPS)*, pp. 4790–4798, 2016.

Yifan Wang, Federico Perazzi, Brian McWilliams, Alexander Sorkine-Hornung, Olga Sorkine-Hornung, and Christopher Schroers. A fully progressive approach to single-image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 864–873, 2018.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning (ICML)*, pp. 681–688, 2011.

Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet. In *International Conference on Machine Learning (ICML)*, pp. 2635–2644, 2016.

Jianwen Xie, Song-Chun Zhu, and Ying Nian Wu. Synthesizing dynamic patterns by spatial-temporal generative convnet. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7093–7101, 2017.

Jianwen Xie, Yang Lu, Ruiqi Gao, and Ying Nian Wu. Cooperative learning of energy-based model and latent variable model via mcmc teaching. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 32, 2018a.

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 (TPAMI)*, 42(1):27–45, 2018b.

Jianwen Xie, Zilong Zheng, Ruiqi Gao, Wenguan Wang, Song-Chun Zhu, and Ying Nian Wu. Learning descriptor networks for 3D shape synthesis and analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8629–8638, 2018c.

Jianwen Xie, Song-Chun Zhu, and Ying Nian Wu. Learning energy-based spatial-temporal generative convnets for dynamic patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

Jianwen Xie, Zilong Zheng, Ruiqi Gao, Wenguan Wang, Song-Chun Zhu, and Ying Nian Wu. Generative VoxelNet: learning energy-based models for 3D shape synthesis and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

Jianwen Xie, Yifei Xu, Zilong Zheng, Song-Chun Zhu, and Ying Nian Wu. Generative PointNet: energy-based learning on unordered point sets for 3D generation, reconstruction and classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021a.

Jianwen Xie, Zilong Zheng, Xiaolin Fang, Song-Chun Zhu, and Ying Nian Wu. Cooperative training of fast thinking initializer and slow thinking solver for conditional learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021b.

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. In *Proceedings of The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, 2021c.

Jianwen Xie, Zilong Zheng, and Ping Li. Learning energy-based model with variational auto-encoder as amortized sampler. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, 2021d.

Yifei Xu, Jianwen Xie, Tianyang Zhao, Chris Baker, Yibiao Zhao, and Ying Nian Wu. Energy-based continuous inverse optimal control. *arXiv preprint arXiv:1904.05453*, 2019.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Yang Zhao, Chunyuan Li, Ping Yu, Jianfeng Gao, and Changyou Chen. Feature quantization improves gan training. In *International Conference on Machine Learning (ICML)*, pp. 11376–11386, 2020.

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 (ICCV)*, pp. 2223–2232, 2017.

Song-Chun Zhu, Yingnian Wu, and David Mumford. Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling. *International Journal of Computer Vision (IJCV)*, 27(2):107–126, 1998.

A APPENDIX

A.1 EVALUATION METRICS

We consider two commonly used metric for evaluating image synthesis quality. (i) FID (Heusel et al., 2017) compares the statistics (mean and variances of Gaussian distributions) between the generated samples and real samples. FID is consistent with increasing disturbances and human judgment. Lower scores indicate that the model can generate higher quality images. We follow the dataset split protocol in Zhang et al. (2020) to compute FID. (ii) KID (Bińkowski et al., 2018) improves FID as an unbiased estimator, making it more reliable when there are fewer available test images. 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. We follow Kim et al. (2020) to compute KID.

For AMT evaluation, we follow the setting in FQ-GAN. Each testing image is judged by six users, who are asked to select the best-translated image to target domain where two standards should be considered, visual image quality, and the extent of background information is kept. We inform the participants of the name of the target domain and six example images of the target domain as a visual illustration.

Design	**Output shape**
RGB image $x \in \mathbb{R}^{3\times256\times256}$	$3 \times 256 \times 256$
FromRGB $3 \to ch/2$	$ch/2 \times 256 \times 256$
BlockDown $ch/2 \to ch$	$ch \times 128 \times 128$
FromRGB $3 \to ch$	$ch \times 128 \times 128$
BlockDown $ch \to 2ch$	$2ch \times 64 \times 64$
BlockDown $2ch \to 4ch$	$4ch \times 32 \times 32$
BlockDown $4ch \to 4ch$	$4ch \times 16 \times 16$
BlockDown $4ch \to 4ch$	$4ch \times 8 \times 8$
BlockDown $4ch \to 4ch$	$4ch \times 4 \times 4$
Block $4ch \to 4ch$	$4ch \times 1 \times 1$
Flatten & Linear	1

Table 6: The neural network architecture which can input the maximum resolution 256×256. ch is the channel multiplier and the default is 32 across all experiments. When the model grows the resolution from 128 to 256, the gray-shadowed FromRGB is abandoned and two new blocks above that are added. And so on for other resolution transitions.

A.2 ARCHITECTURES AND SETTINGS

The architecture detail up to 256×256 resolution is given in Table 6. Each BlockDown contains the 2-layer convolution operations where the first layer won't change feature size whereas the second layer comes with an average pooling 2×2 for down-sampling. The kernel size is 3×3 with stride 1. The last Block (without downsampling) first uses a 3×3 kernel and then uses a 4×4 kernel. The residual connection used in BlockDown is a single convolution and an average pooling of size 2×2. We don't apply the residual connection in the last Block. The stride is set to be 1 in all convolutions. FromRGB is a block for transforming RGB images to feature maps. Thus, in Section 3, the $\texttt{Primal}(\cdot)$ consists of a FromRGB and a BlockDown while the $\texttt{Fade}(\cdot)$ consists of an average pooling operation and a FromRGB block.

We train the network using Adam optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. We set the learning rate schedule as $\alpha = \{8 : 0.001, 16 : 0.001, 32 : 0.001, 64 : 0.0012, 128 : 0.0015\}$ and the data feeding schedule as $N = \{8 : 50\text{K}, 16 : 75\text{K}, 32 : 100\text{K}, 64 : 125\text{K}, 128 : 150\text{K}\}$. In

terms of the sampling hyperparameters, we set a schedule regrading the number of steps $T = \{8 : 15,\ 16 : 30,\ 32 : 50,\ 64 : 50,\ 128 : 60\}$, step size 1.0 and the standard deviation of an additive noise $\eta_t^\epsilon = 2e^{-2} - 2e^{-2}/(T - t + 1)$ in most experiments. As most experiments were implemented on the less powerful TITAN X (12 GB), we don't sweep over these values and hyper-parameters are very likely not to be optimal due to limited computational resources. Nonetheless, the presentation of experimental results is adequate to validate the proposed approach.

CelebA experiment We preprocess CelebA to produce CelebA-64 in two settings. 1) Each image is firstly center-cropped into 140×140 and then resized to 64×64. In Table 2, three baselines NCSN (Song & Ermon, 2019), mr-Langevin (Block et al., 2020), and WGAN-GP (Lee et al., 2018b) applied this setting. The FID computation protocol is also based on NCSN where the score is calculated between 10k samples and all test images. 2) Each image is directly resized to 64×64. FID is computed on 40k generated samples. In Table 2, VAE, DCGAN, and short-run EBM exactly follow this setting.

A.3 More image synthesis results

Figure 8 shows the generated images of which the EBM is modeled on CIFAR-10 and ImageNet-1K (32×32). Table 7 describes the quantitative results of image generation on ImageNet-1k from which CF-EBM is observed to achieve competitive results with a much smaller model.

In Figure 9, we visualize the short-run MCMC initialized from uniform noises for CelebA. Figure 10 shows the generated samples and their nearest neighbors from the training data. As is shown, the synthesized images are not present in the training data. Figure 11 illustrates the diverse image completions on CelebA-64^2 test data. It demonstrates that our model can generalize well to unseen test data modes. It also learns meaningful representations which captures different facial details and expressions, *e.g.* eyes, smiles, eyebrows and beard, which allows to produce diverse images completions.

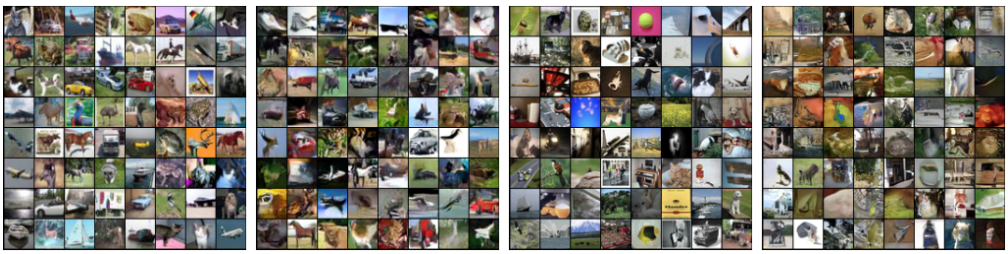


Figure 8: From Left to Right: ground truth of CIFAR-10, generated samples of CIFAR-10, ground truth of ImageNet-1K (32×32), and generated samples of ImageNet-1K (32×32).

Approach	Models	FID	#Params
Unconditional	PixelCNN (Van den Oord et al., 2016)	40.51	> 35M
	PixelIQN-small	37.62	>35M
	PixelIQN-big	26.56	>50M
	CF-EBM (ours)	26.31	5.82M
Conditional	PixelCNN	33.27	-
	PixelIQN (Ostrovski et al., 2018)	22.99	
	EBM (Du & Mordatch, 2019)	14.31	
Self-supervision	SS-GAN (Chen et al., 2019b)	17.10	-
	MS-GAN (Tran et al., 2019)	12.30	

Table 7: FID on ImageNet-1K (32×32). The channel multiplier is $ch = 48$.

Figure 9: Short-run MCMC generative sequences on CelebA (50 steps).

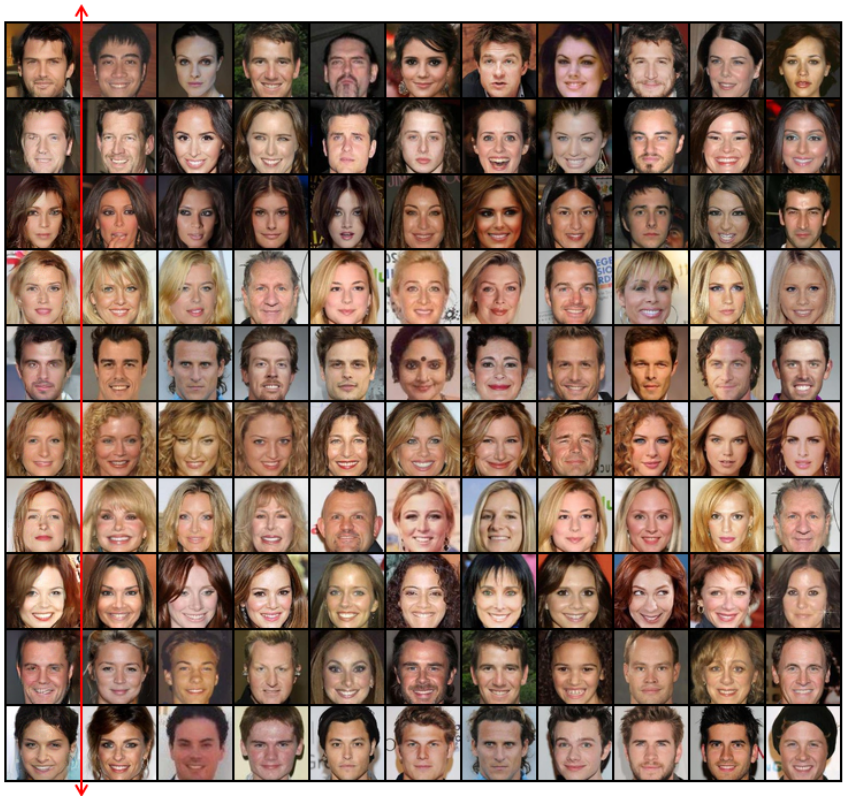


Figure 10: Nearest neighbors on CelebA (64×64). Generated samples are placed on the left side of red vertical line whereas 10 nearest neighbors are on the right side.

Figure 11: Illustration of diverse image inpaintings. The top row is the occluded images and the bottom row shows the original images. Our model can complete the occluded areas with different expressions, *e.g.* smile, and diverse face details, *e.g.* eyes, ears and beard.

A.3.1 THE EFFECT OF DATA PERTURBATION

As almost all EBM related generative models add Gaussian noise $\mathcal{N}(0, \sigma^2\mathbf{I})$ to perturb the data for the purpose of training stabilization, it is necessary to investigate the effect of such perturbation. Nijkamp et al. (2019) has already reported that a larger perturbation leads to lower FID score where the minimum noise standard deviation is $\sigma = 0.03$. Also, the work Song & Ermon (2019) applies a sequence of decayed perturbation from $\sigma_L = 1.0$ to $\sigma_1 = 0.01$. We claim that the model will memorize the leaked noisy information even if the final noise is small. As a result, the synthesized samples are noisy and foggy. Figure 12 shows the visualization effect at different levels of perturbation on the observed data. Note that, even the noise as small as $\sigma = 0.01$ can cause apparent visualization difference. Moreover, the noise term for sampling is usually not decayed to 0 such that each sample is noisy in the continuous image space. From Figure 2 in Nijkamp et al. (2019), Figure 1 in Song & Ermon (2020) and Figure 2 in Grathwohl et al. (2020), we observe the perturbation will be finally reflected in the generated samples. Therefore, the data perturbation should be taken as the last trick for generative modeling, or be discarded.

A.4 ABLATIONS

We mainly examine the activation functions, normalization and layer connections to see the model performance differences.

A.4.1 ACTIVATIONS

We show formulas of the following five activations and also give their derivatives:

- **Rectified Linear Unit** (ReLU) (Nair & Hinton, 2010): $f(x) = \max(0, x)$, and the derivative:

$$\frac{\partial f(x)}{\partial x} = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

- **Leaky Rectified Linear Unit** (LeakyReLU) (Maas et al., 2013): $f(x) = \max(0, x) + \alpha \cdot \min(0, x)$, and the derivative:

$$\frac{\partial f(x)}{\partial x} = \begin{cases} 1 & \text{if } x > 0, \\ \alpha & \text{otherwise.} \end{cases} \tag{6}$$

  where $\alpha = 0.2$ is the common setting.

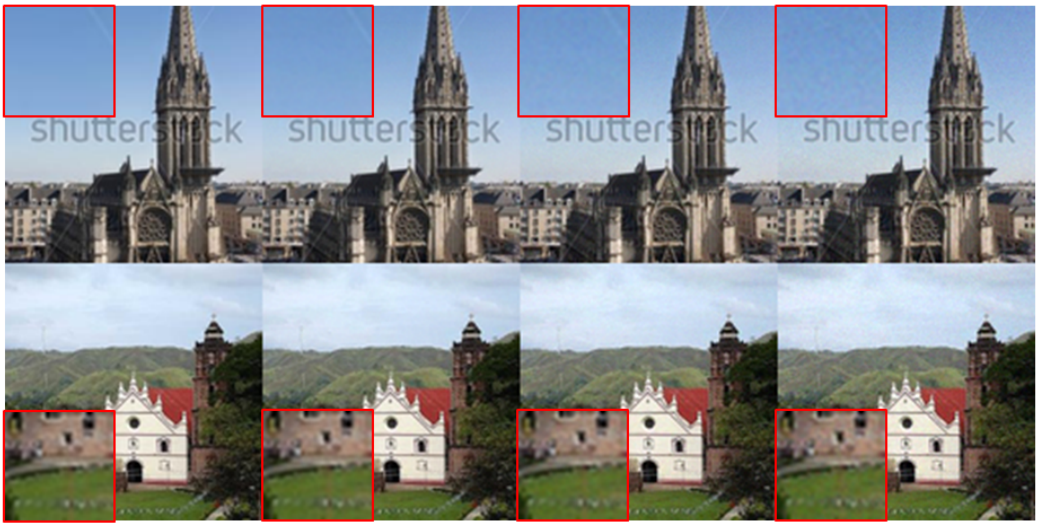


Figure 12: The visualization of different levels of perturbation, with red rectangle areas zoomed in. (LSUN Church 128×128, Yu et al. (2015))

- **Continuously Differential Exponential Linear Unit** (CELU) (Barron, 2017): $f(x) = \max(0, x) + \min(0, \alpha \cdot (\exp(x/\alpha) - 1)$, and the derivative:

$$\frac{\partial f(x)}{\partial x} = \begin{cases} 1 & \text{if } x > 0, \\ \alpha(\exp(x/\alpha) - 1) & \text{otherwise.} \end{cases} \tag{7}$$

- **Swish** (Ramachandran et al., 2016): $f(x) = x \cdot \text{Sigmoid}(x)$, and the derivative is:

$$\frac{\partial f(x)}{\partial x} = (2 - \text{Sigmoid}(x))\text{Sigmoid}(x) \tag{8}$$

- **Gaussian Error Linear Unit** (GELU) (Hendrycks & Gimpel, 2016b): $f(x) = x \cdot \Phi(x)$ where $\Phi(\cdot)$ is the cumulative distribution function for Gaussian distribution. It is successfully applied in natural language processing like BERT. The derivative can be approximated as:

$$\frac{\partial f(x)}{\partial x} = (2.702 - \text{Sigmoid}(1.702x))\text{Sigmoid}(1.702x) \tag{9}$$

Figure 13 visualizes above activations and their derivatives. Obviously, both ReLU and LeakyReLU are non-smooth around 0. Image synthesis results on CIFAR-10 are shown in Table 8. We find that the use of LeakyReLU and ReLU diverge at an early time. These phenomenons are also observed on CelebA experiment. Therefore, smooth activations indeed improve the training stability and sample quality of EBM.

Activation	FID
ReLU	N/A
LeakyReLU	78.21
GELU	28.71
CELU($\alpha = 1$)	21.50
Swish	16.71

Table 8: On the impact of activation functions.

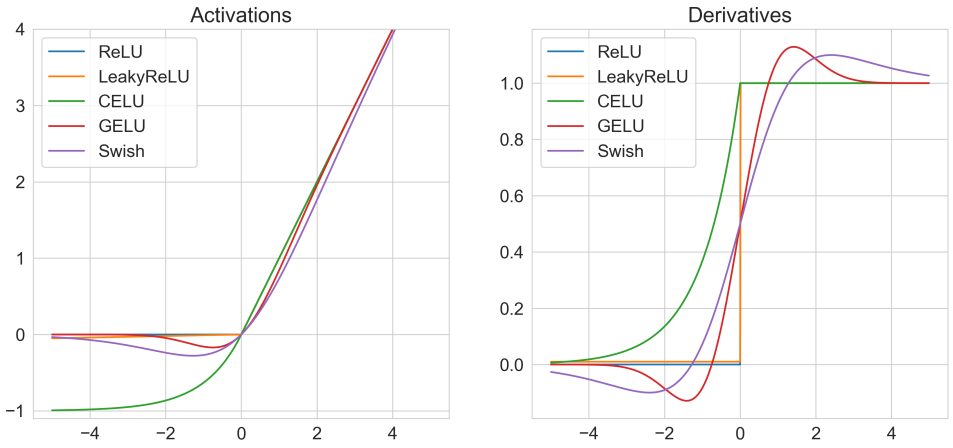


Figure 13: Activations and their derivatives

A.4.2 Normalization

Since the objective in Eq. (2) is similar to that in Wasserstein GAN, and the EBM acts as a discriminator, we thus naturally only consider spectral normalization (Miyato et al., 2018) to improve the performance and stability. SN constrains the Lipschitz constant of the learnable neural network parameters, which is widely used to stabilize the training of the discriminator network in GAN.

However, batch normalization is not appropriate here. Because each update of Langevin dynamics will rely on the running mean/variance of $\mathbf{x} \sim p_{data}$. Even if the statistic is not updated in Eq. (3), the distribution of $p_\theta(\mathbf{x}_t)$ is changing all the time along a chain. As a result, the pre-computed statistics of batch normalization don't fit each $\mathbf{x}_t$. Other normalizations, like instance normalization and layer normalization, are not appropriate in a network of which the target is to discriminate two distributions.

A.4.3 Ablate the components

We first examine how each component affects the model performances, including progressive growing, residual-connections, and spectral normalization. Results are shown in Table 9. From (a)-(c), we find that spectral normalization and residual connection improved the sample quality. (d) Besides, we also compare the time efficiency regarding the training with and without coarse-to-fine settings. It is observed that the model without coarse-to-fine learning requires approximately $\mathbf{3}\times$ more time than the counterpart. (e) When the coarse EBM is fixed during training higher resolution, the FID is worse than that of a single fine EBM.

Configuration	**FID**
(a) Basic CF-EBM	32.01
(b) + Spectral normalization	20.83
(c) + Residual connection	**16.71**
(d) – Coarse-to-fine	24.67
(e) Fix coarse EBM	21.25

Table 9: Ablation results on CIFAR10.

A.5 Evaluations about likelihood

We implement an experiment on the continuous MNIST following the setting in (Du & Mordatch, 2019). Its open source implementation is at `https://github.com/openai/ebm_code_release`. It applies Annealed Importance Sampling (AIS) (Neal, 2001) to obtain a bound of partition function. Figure 14 illustrates the generated samples on MNIST. As can be seen, our model can produce more realistic digits. Table 11 shows the log likelihood comparison across three models.

Models	FID
Glow (Kingma & Dhariwal, 2018)	68.93
SN-GAN (Miyato et al., 2018)	25.95
WGAN-GP (Gulrajani et al., 2017)	22.57
CR-GAN (Zhang et al., 2020)	**16.97**
CF-EBM (ours)	23.50

Table 10: FID for image generation on CelebA-HQ (128×128). Results of GAN-based approaches are taken from Zhang et al. (2020).

Models	Lower Bound	Upper Bound
NICE (Dinh et al., 2014)	1980.0	1980.0
EBM (Du & Mordatch, 2019)	1925.0	2218.3
CF-EBM (Ours)	**2197.7**	**2342.0**

Table 11: Log likelihood in Nats on Continuous MNIST.

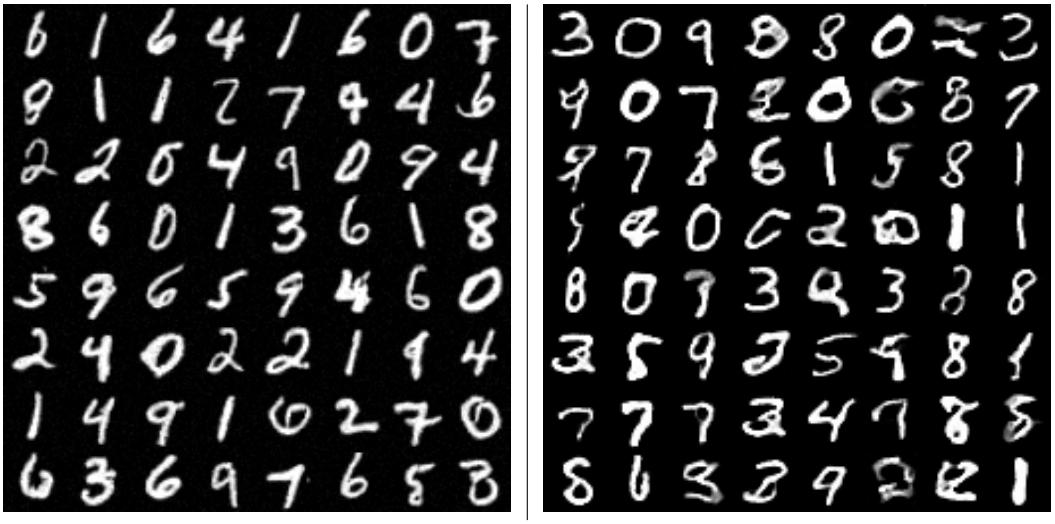


Figure 14: Generated samples on MNIST. (*Left*) Our model, (*Right*) EBM (Du & Mordatch, 2019)

A.6 SCALE UP CF-EBMS

To scale up the EBM generative model to be capable of sampling higher resolution images, we integrate intermediate CF-EBMs that have been learned in Algorithm 1. When the training enters into modeling higher resolution, we fix the coarse-level CF-EBM, which only produces initial points for fine-level CF-EBM samplers. In this scenario, the number of Langevin steps can be reduced to 15 to expedite sampling at each scale, meaning that generating a 256×256 images requires 90 steps. On a single TITAN X GPU (12GB), the total training time (200k iterations) of CelebA-HQ 128×128 and 256×256 are about 120 hours and 235 hours, respectively. On a single TITAN V100 GPU, these costs can be reduced to 55 hours and 100 hours (we tested it once).

In this experiment, we fix the initial noise variance $\eta_0^\epsilon = 0.03$. We observe that, in the testing phase, choosing a different value of η_0^ϵ largely affects the synthetic results, as shown in Figure 15. A larger η_0^ϵ tends to add details to look more realistic but have more noisy images. Figure 16 shows more generated samples on CelebA-HQ 256×256. In this experiment, we set the channel multiplier at 32 and train it for 5 days on a single GPU. Figure 17 shows generated samples on CelebA-HQ 512×512 where we set the channel multiplier at 48 and the model is trained for 10 days on a single GPU. Table 10 compares the FID of different models on CelebA-HQ 128×128. Note that our approach obtains competitive results without any regularization term.

Figure 15: Effect of changing the initial noise level η_0^ϵ at the testing phase.

Figure 16: CelebA-HQ 256×256 generated samples, with linearly cooling the "temperature" η_t^ϵ from 0.025 to 0.

A.7 Unpaired Image-to-Image translation

A.7.1 Dataset

In this paper, we experiment on five popular image-to-image translation datasets:

selfie2anime It is first introduced in Kim et al. (2020). The selfie and anime datasets each contains 3400 training images and 100 testing images.

photo2vangogh and apple2orange These datasets are used in Zhu et al. (2017). The training dataset size of each class: 6,287 (photo), 400 (vangogh), 995 (apple) and 1019 (navel orange). The test datasets consist of 751 (photo), 400 (vangogh), 266 (apple) and 248 (navel orange).

Yosemite summer2winter This dataset is used in Zhu et al. (2017). The training size of each class: 1273 (summer) and 854 (winter). The testing size of each class: 309 (summer) and 238 (winter).

cat2dog These datasets are used in DRIT (Lee et al., 2018a). The numbers of data for each class are 871 (cat) and 1,364 (dog). Follow (Kim et al., 2020), we use 100 (cat) and 100 (dog) as test data.

Figure 17: CelebA-HQ 512×512 generated samples, , with linearly cooling the "temperature" η_t^ϵ from 0.015 to 0.

Experimental settings For the unpaired image-to-image translation task, we only made two changes compared with the settings used in image generation: (i) the channel multiplier is $ch = 16$; (ii) the learning rate is multiplied by 10 only for $dog \rightarrow cat$.

A.7.2 Compare with CUT

We implement CUT based on its open-source code https://github.com/taesungp/contrastive-unpaired-translation. We then pick up the most realistic translation results from CUT on cat→dog. In Figure 18, we compare them with the proposed CF-EBM. As can be seen, CUT sometimes hallucinates a tongue, resulting in unfaithful translations. This observation is also mentioned in the original paper (Park et al., 2020). In contrast, CF-EBM returns better and more faithful translations mostly.

Figure 18: Qualitative comparison between our CF-EBM and CUT.

A.7.3 More Uncurated Image-to-Image Translation Examples

Figure 19: Translation dynamics: $orange \rightarrow apple$.

Figure 20: Translation dynamics: $cat \rightarrow dog$.

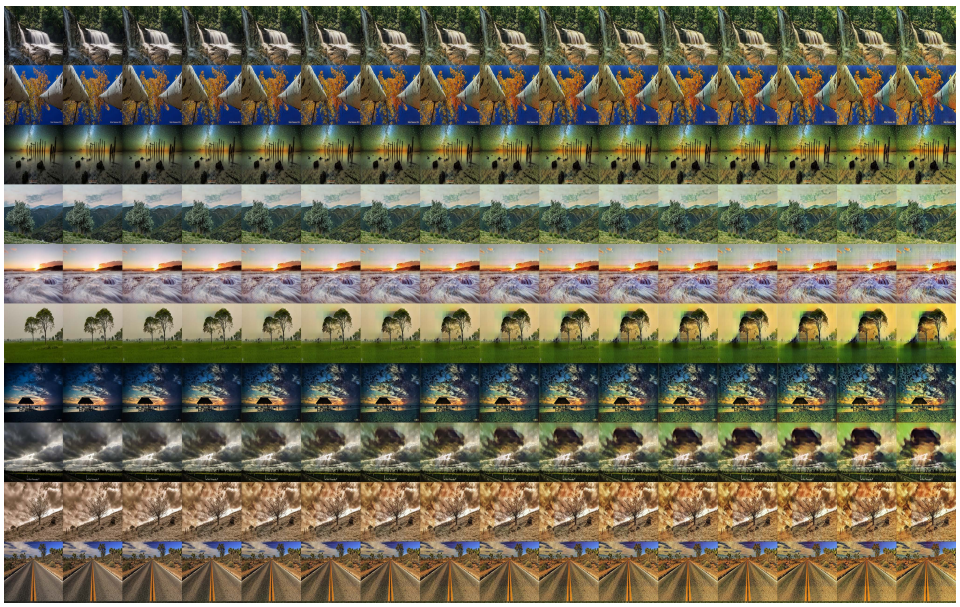


Figure 21: Translation dynamics: $photo \rightarrow vangogh$.

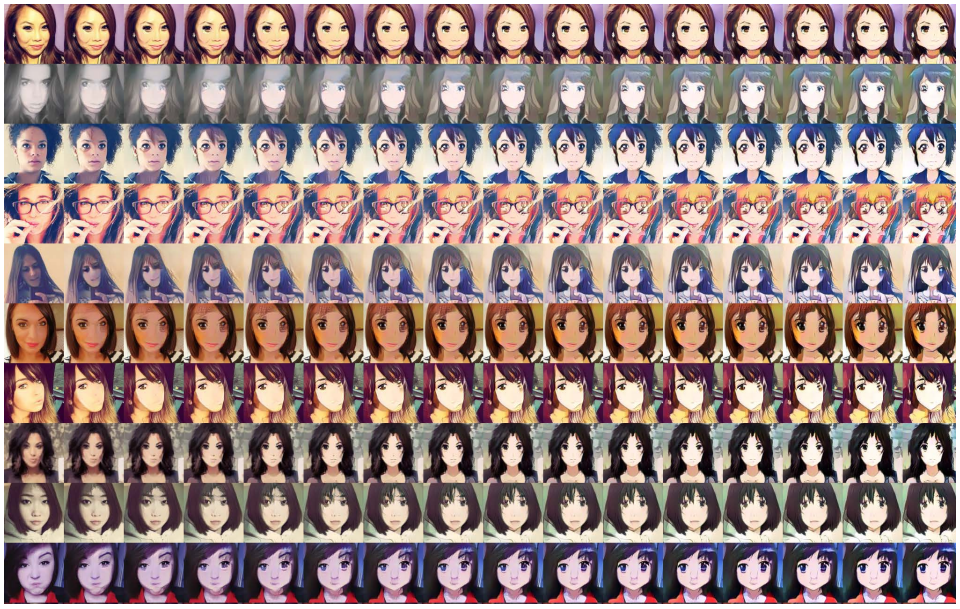


Figure 22: Translation dynamics: $selfie \rightarrow anime$.

Figure 23: Uncurated translation result: $photo \rightarrow Vangogh$

Figure 24: Uncurated translation result: $Vangogh \rightarrow photo$

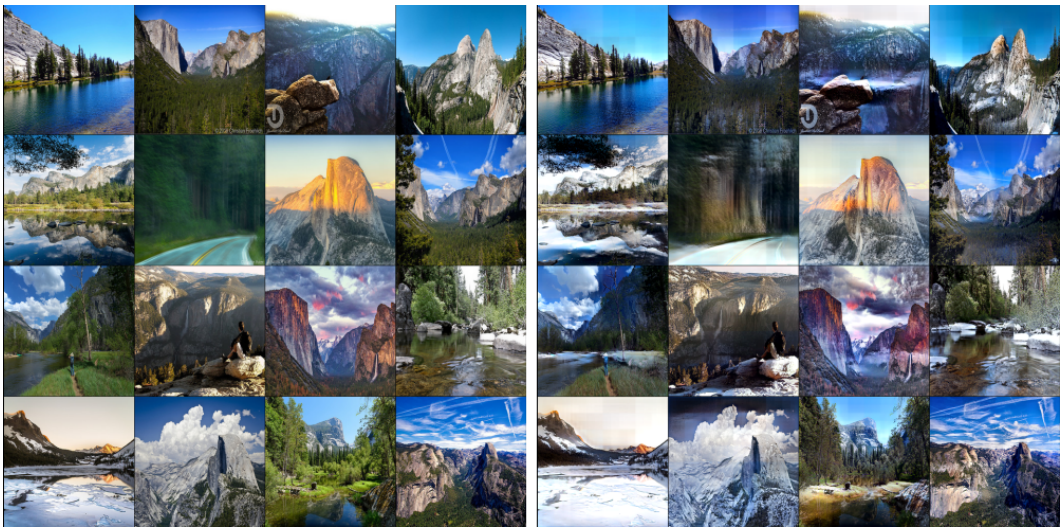


Figure 25: Uncurated translation result: Yosemite $summer \rightarrow winter$

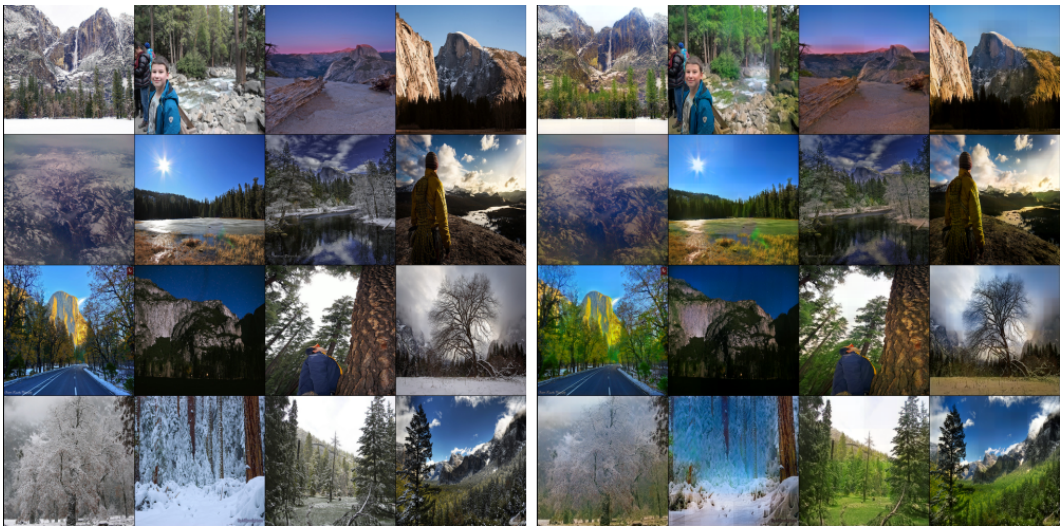


Figure 26: Uncurated translation result: Yosemite $winter \rightarrow summer$