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

Anonymous authors
Paper under double-blind review

ABSTRACT

Energy-based models (EBMs) for generative modeling parametrize a single net and can be directly trained by maximum likelihood estimation. Despite the simplicity and tractability, current approaches are either unstable to learn or difficult to synthesize diverse and high-fidelity images. We propose to train EBM via a multistage coarse-to-fine expanding and sampling strategy, namely CF-EBM. To improve the learning procedure, we construct 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 256×256-pixel images. 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 reduced by $1000\times$ and the training budget by $9\times$. In parallel, we present a gradient-based discriminative saliency method to interpret the translation dynamics which align with human behavior explicitly.

1 INTRODUCTION

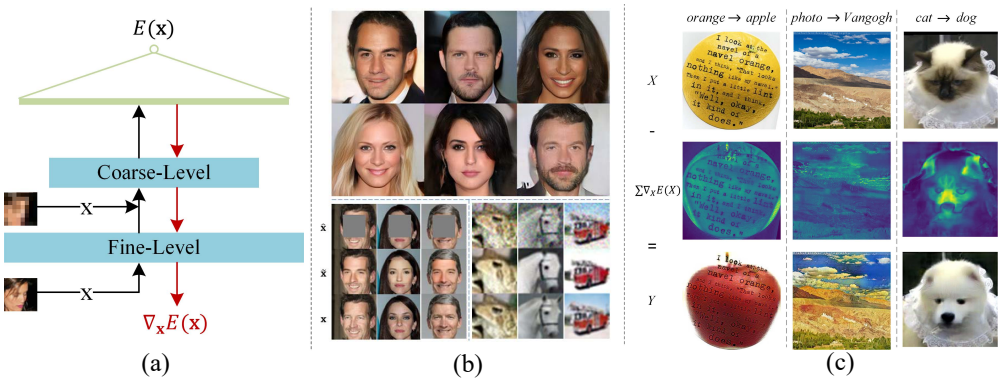


Figure 1: Highlights of our contributions. (a) Coarse-to-fine EBM with gradient-based MCMC sampling: $\mathbf{x} \leftarrow \mathbf{x} - \nabla_{\mathbf{x}} E(\mathbf{x})$; (b) Image generation (256^2), image denoising and image inpainting from top to bottom. ($\hat{x}$: corrupted image, $\tilde{x}$: 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 such as convolutional network (ConvNet) have drawn much attention from the deep learning communities (Xie et al., 2016; 2018; Du & Mordatch, 2019; Nijkamp et al., 2019; Deng et al., 2020; Gao et al., 2020; Yu et al., 2020). EBMs are characterized by simplicity and explicitness. (i) The maximum likelihood learning of EBM unifies the representation learning and generative modeling in a single net. (ii) EBMs provide an explicit density distribution of the data by training a bottom-up energy function $E(\cdot)$ that assigns lower values to observed data and higher values to unobserved ones.

However, it is not easy to train an EBM to synthesize diverse and high-fidelity images. The learning process requires the Markov Chain Monte Carlo (MCMC) to sample from the current model and then

updates the model parameters given the difference between those samples and the observed data. Such an "analysis by synthesis" learning scheme is challenging because the sampling step is inefficient and unstable to traverse different 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 a manifold residing in a downsampled low-dimensional ambient space is more straightforward than that in the original high-dimensional counterpart in the case of images. Therefore, modeling and sampling low-dimensional data is much stabler and faster to converge (Odena et al., 2017).

Inspired by the above, we advocate a multistage coarse-to-fine energy-based modeling and sampling methodology (CF-EBM) in this paper. As presented in Figure 1(a), the approach starts with learning a coarse-level EBM on easier 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 (Welling & Teh, 2011), is used for efficient sampling. At a high level of the **modeling aspect**, the coarse-level training can easily explore the global data structure, and finer-level training will then gradually refine the detail. 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 both the bottom-up modeling and the top-down sampling in a single net where the "entrance" is also the "exit."* In particular, as to the **sampling aspect**, we are more interested in investigating how the multistage learning strategy helps alleviate the difficulty of MCMC synthesis. We demonstrate that the top-down gradient information $\nabla_{\mathbf{x}} E(\mathbf{x})$ always flows through coarse-level layers towards finer-level layers, which can be treated as mapping a less multi-modal low-dimensional latent "prior" to a high-dimensional target by adding more delicate details, as illustrated in Figure 1(a). The proposed methodology is thus less sensitive to the multi-modal manifold structure embedded in high-dimensional ambient space, as given in Figure 1(b).

Furthermore, we present a one-sided unsupervised image-to-image translation method that can scale up to high resolution. The approach is directly feasible by iterative gradient descents without explicit cycle or geometry constraints (Zhu et al., 2017; 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 the source domain $\mathcal{X}$ and 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.

Contributions can be summarized as:

- To the best of our knowledge, it is the first work that approaches the multistage coarse-to-fine EBM learning based on a single neural network. 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 256×256.
- The proposed image translation approach is essentially different from all other well known 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 to image generation, denoising, inpainting, and unsupervised image-to-image translation. Results are shown to outperform or be competitive with the prior art mostly.

The rest of paper is organized as follows. Section 2 summarizes related works and how our proposal is differentiated from those. Section 3 describes our contribution in detail. Section 4 presents extensive experiments. Section 5 concludes our paper and highlights future research directions.

2 Related Work

2.1 Energy-based Generative Models

The main challenge to train the EBMs via MLE lies in drawing fair samples from the model, especially when the energy function is parameterized by a highly non-linear structural ConvNet. The contrastive divergence (CD) (Hinton, 2002; Tieleman, 2008) framework, with MCMC chains initialized from the 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.

However, these techniques still struggle to scale up and model multimodal data. There have been several attempts to improve the training of the EBMs. Gao et al. (2018) adopts a multi-grid method, which needs to train multiple models at different grids at each iteration and samples in a coarse-to-fine manner, where synthesized example at each subsequent grid is obtained by a finite-step MCMC initialized from the synthesized example generated at the previous coarser scale. Both time and memory efficiency are pretty low in this framework. To stabilize the training, Nijkamp et al. (2019); Du & Mordatch (2019) added the Gaussian white noise to the observed data, resulting in both a high-variance learning and noisy synthesized samples. 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 data sampling, the learned model can preserve EBM's compatibility and synthesize high-fidelity images.

2.2 Unsupervised image-to-image translation

GAN-based solution: Image-to-image translation aims to learn two mappings $F : \mathcal{X} \rightarrow \mathcal{Y}$ and $G : \mathcal{Y} \rightarrow \mathcal{X}$. 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 schemes generally adopt two sets of GANs and two-level learning objectives, a *distribution level* and an *instance level*. (i) At the *distribution-level*, two adversarial losses capture discrepancy between source and target domain; (ii) To tackle the difficulty of unpaired settings, they also need to leverage a cycle or reconstruction consistency. The cycle consistency can enforce an *instance-level* supervision, which aims to regularize the training by reconstructing an original image from its backward translated image.

Limitations: A significant limitation of this line of works is the mutual correlation between F and G since they are optimized simultaneously towards being roughly invertible of each other. To some extent, it may intensify the long-standing instability issues of GAN and increase training time cost. They also require an elegant design, *e.g.* instance normalization (Ulyanov et al., 2016), positional normalization (Li et al., 2019) and attention mechanism (Kim et al., 2020), and a large-scale architecture to ensure the model to explore the inherent discrepancy between two domains and preserve source contents in translation.

Our EBM Solution: Our formulation of solving the unpaired image-to-image translation is a different story from GANs. The proposed concise EBM solution can only rely on a *distribution-level* matching. It can achieve better performances using less time and memory, although there is no cycle constraint or fancy architecture. Another distinction between our method and GAN-based methods is the interpretability of image translation. The Langevin dynamics provide a gradient-based saliency to highlight relevant attributes that make the two domains distinct, as illustrated in Figure 1(c).

3 Method

In this section, we describe the EBM learning landscape, our proposed CF-EBM approach and the generalization 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 $E_\theta(\mathbf{x}) : \mathbb{R}^D \rightarrow \mathbb{R}$ is the energy function defined by a bottom-up ConvNet whose parameters are denoted by θ. $Z(\theta) = \int \exp(-E_\theta(\mathbf{x}))q(\mathbf{x})d\mathbf{x}$ is the intractable normalizing constant or the partition function. Given N observed data points $\{\mathbf{x}_i\}_{i=1}^N \sim p_{data}$, where p_{data} is the unknown data distribution, the model can be trained by maximizing the log-likelihood

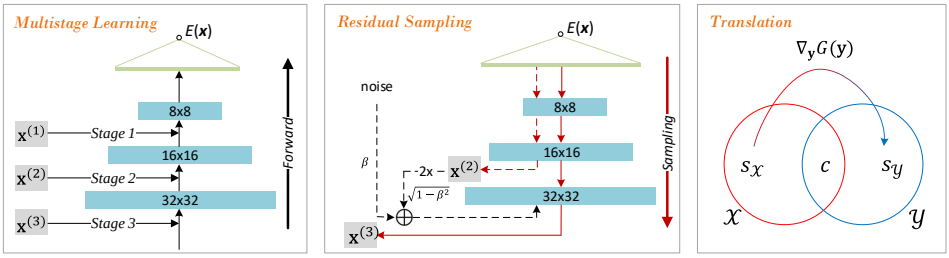
$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log p_\theta(\mathbf{x}_i) \approx \mathbb{E}_{\mathbf{x}\sim p_{data}} \log(p_\theta(\mathbf{x}))$. The derivative of the negative log-likelihood is

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

where the second expectation term under p_θ is intractable and can be approximated via MCMC. Given that, the EBM is updated by gradient descent. To sample $\tilde{\mathbf{x}} \sim p_\theta$ via MCMC, we rely on Langevin dynamics (Langevin, 1908; Zhu & Mumford, 1998) that recursively computes the following step

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

where η_t is the step size. To ensure convergence, typically, η_t is polynomially decayed (Welling & Teh, 2011). However, in practice, EBM often uses another version of Langevin dynamics where the step size $\eta_t/2$ and standard deviation $\sqrt{\eta_t}$ of ϵ_t are chosen separately. This kind of modified sampler allows for faster training (Grathwohl et al., 2020). In this paper, we keep the step size constant and linearly decay the noise variance η_t^ϵ till 0.



(a) (b) (c)

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

3.2 CF-EBM: Coarse-to-fine EBM Learning

Our primary contribution is the coarse-to-fine training methodology for 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 during a multistage training. In this way, both stability and time efficiency in training EBMs are benefited. Figure 2 (a)-(b) illustrate the proposed EBM learning and sampling strategy, which is accompanied by the pseudocode in Algorithm 1.

Algorithm 1: CF-EBM Expanding and Sampling

Input: Multi-resolution data $\{\mathbf{x}_i^{(s)}, i = 1, ..., N; s = 1, ..., S\}$, Langevin steps $\{T^{(s)}\}_{s=1}^S$, and batch size $\{n^{(s)}\}_{s=1}^S$, $\theta^{(0)} \leftarrow \emptyset$
Output: model parameters $\theta^{(S)}$

for $s = 1 : S$ **do**
  $m = 0,\ \beta = 0$
  Expand model $\theta^{(s)} \leftarrow \cup(\theta^{(s-1)}, \theta_{new}^{(s-1)})$
  while $(m \leq N)$ **do**
    Draw observed $\{\mathbf{x}_i^{(s)}\}_{i=1}^{n^{(s)}} \sim p_{data}(\mathbf{x})$
    # Sample from $p_\theta^{(s)}$
    Draw initial $\{\tilde{\mathbf{x}}_i^{(s)}\}_{i=1}^{n^{(s)}} \sim \mathcal{U}^{(s)}(-1, 1)$
    if $(\beta \neq 1$ and $s \neq 1)$ **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 $\theta^{(s)}$
    Update $\theta^{(s)}$ with gradient descent in Eq. (2)
    $m \leftarrow m + n^{(s)},\ \beta = \max(1,\ 2N/n^{(s)})$
  end
end

Assuming the total number of stages is S and the training starts from the minimal resolution 8×8. Let $(\mathbf{x}^{(s)},\ s = 1, ..., S)$ denote the multi-resolution versions of an image $\mathbf{x}$, then at stage s, the model will be associated with an energy function $E_{\theta^{(s)}}^{(s)}$ parameterized by $\theta^{(s)}$ to learn $\mathbf{x}^{(s)}$ of which the resolution is $2^{s+2} \times 2^{s+2}$. In what follows, be aware of the only difference between $\theta^{(s)}$ and $\theta^{(s+1)}$ is a new block $\theta_{new}^{(s)}$, we will suppress $\theta^{(s)}$ to avoid clutter.

At the stage transition, we add randomly initialized layers to the model, with which we simultaneously apply MCMC to learn the energy function. A natural question is how to smooth the transition and avoid the instability issues resulting from the energy function's sudden change. We solve it from the following two aspects:

Smooth Energy Function Expanding: This is in the same spirit of Progressive GAN (Karras et al., 2018). We define a basic expanding block $\texttt{Expand}(\cdot)$ that is used to double the model resolution. We fade in the newly added block smoothly to prevent sudden shocks to the already 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. In this sense, the model will first rely on the well trained coarse-level layers and gradually move attention to the newly added fine-level layers. The auxiliary fading block will be abandoned eventually when β reaches 1. In our paper, $\texttt{Primal}(\cdot)$ consists of two convolution layers, while $\texttt{Fade}(\cdot)$ is a convolution layer followed by an 2×2 average pooling layer to link the previous coarse-level layers.

Smooth MCMC Sampling: The smoothness of sampling is realized both implicitly and explicitly.

(i) *Implicitly*: Considering the resolution transition from stage s to stage $s+1$, we expand the energy function by $E^{(s+1)}(\mathbf{x}^{(s+1)}) = E^{(s)}(\texttt{Expand}(\mathbf{x}^{(s+1)}))$. The computation of the gradient $\nabla_{\mathbf{x}^{(s+1)}} E^{(s+1)}(\mathbf{x}^{(s+1)})$ in Langevin sampling can be decomposed into

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

Initially, via the fading block $\texttt{Fade}(\cdot)$, the coarse-level layers contribute greatly to the score function, thus resulting in a blurry synthesis output. Meanwhile, the primal block $\texttt{Primal}(\cdot)$ is randomly initialized such that the score function will lead to a noisy update. As the term $\nabla_{\mathbf{x}^{(s+1)}} \texttt{Fade}(\mathbf{x}^{(s+1)})$ fades away, $\nabla_{\mathbf{x}^{(s+1)}} \texttt{Primal}(\mathbf{x}^{(s+1)})$ takes the lead gradually to synthesize a sharper output.

(ii) *Explicitly*: At each stage transition, to adapt to the randomly initialized layer, the MCMC under the current fine model is initialized from the samples of the previous well-trained coarse model. Specifically, at stage s, the algorithm first generates low-resolution samples with the well-trained $E^{(s-1)}$ by running $T(1-\beta)$ Langevin steps. Through spherical interpolation, those samples are then upsampled to $2\times$ resolution by $\texttt{Upsample}(\cdot)$ and mixed with the uniform noise to initialize the MCMC chain induced by $E^{(s)}$. As β increases, the MCMC initialization effect of the coarse model fades away. Figure 2(b) illustrates the initialization strategy for Langevin dynamics from stage 2 to stage 3.

3.3 Unsupervised Discriminative Saliency Translation

We then generalize the proposed approach on the unpaired image-to-image translation and scale it up to high resolution image datasets. Given two 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:

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

The underlying assumption of applying EBM for image translation is that the two domains $(\mathcal{X}, \mathcal{Y})$ share the same content ambient space c, and each of them is characterized by a different style ambient space $(s_{\mathcal{X}}, s_{\mathcal{Y}})$ 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 explicitly steps into a lower energy area with the content code c being unchanged as much as possible. We name the style that can differentiate different domains *discriminative saliency*.

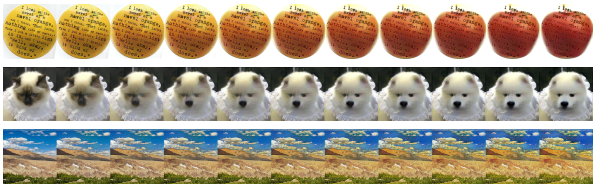


Figure 3: Visualization of disciminative saliency evolving process. Top row: $orange \to apple$, color saliency. Middle row: $cat \to dog$, texture saliency. Bottom row: $photo \to Vangogh$: color and texture saliency.

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? A major benefit of applying the EBM generative model is *interpretability*. We adopt the gradient-based saliency maps (Simonyan et al., 2014) to highlight the discriminative aspect that leads to the translation. In this paper, the accumulative values of the score function in Eq. (4) is computed as the gradient explanation $\sum_t \nabla_{\mathbf{y}_t} G(\mathbf{y}_t)$. It quantifies how much a change of each pixel would contribute to the image translation. From the saliency map in Figure 4, we can conclude all types of the discriminative saliency of different datasets.

(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 7, we observe that non-smooth activations ReLU and LeakyReLU tend to diverge, whereas others don't.

Note that, almost all current EBM generative models tend to add Gaussian noise to smooth but violate the ground-truth data (Nijkamp et al., 2019; Grathwohl et al., 2020). Otherwise, the model will diverge quickly. Similarly, Song & Ermon (2020) inject multiple level noises, resulting in foggy generated results. The effect of noise on the observed data can produce small jumps between adjacent pixels such that they may not be continuous anymore. With smooth activations integrated into the proposed model, we alleviate the instability issue effortlessly while preserving original data structures. See Appendix A.3.1 for more discussions and Figure 11 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), 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.5, we scale the model up to 256×256-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. To be fairly comparable with former EBM works, the 64×64 resolution is used for quantitative evaluation; (iii) CelebA-HQ 256×256 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	DCGAN (Radford et al., 2016)	37.11
	WGAN-GP (Gulrajani et al., 2017)	36.40
	SN-GAN (Miyato et al., 2018)	**21.70**
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., 2018)	33.61
	CF-EBM	**16.71**

Table 1: FID on CIFAR-10.

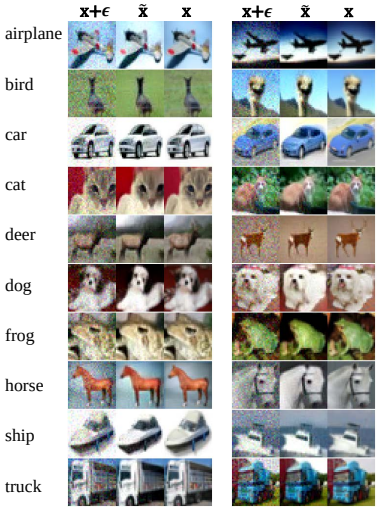


Figure 5: Image denoising.

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.

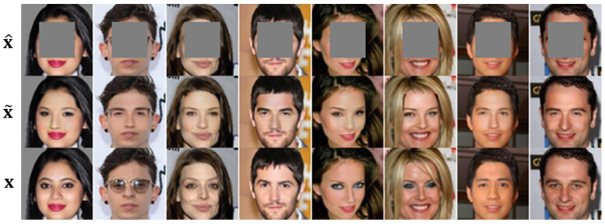


Figure 6: Image inpainting.

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 requires 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.

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.

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 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.

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 13 for how the noise level affects the samples' visual effect in the evaluation and Figure 14 for more high resolution samples.

4.2 UNSUPERVISED IMAGE-TO-IMAGE TRANSLATION

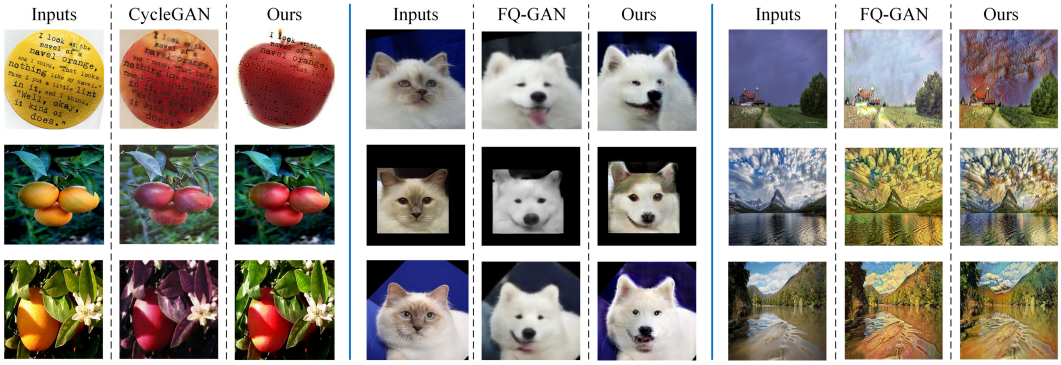
We mainly use four unpaired image-to-image translation datasets for evaluation, including cat2dog, Yosemite summer2winter, vangogh2photo and apple2orange. All images are resized to 256×256 resolutions. More details are provided in Appendix A.6.1.

Quantitative Results: We compare our approach against the well known CycleGAN (Zhu et al., 2017) and two latest state-of-the-art methods U-GAT-IT (Kim et al., 2020) and FQ-GAN (Zhao et al., 2020). Results are shown in Table 4. In most cases, our model obtains better KID scores and preference scores in the human evaluation while being significantly more efficient.

Models	$cat \rightarrow dog$		$dog \rightarrow cat$		$photo \rightarrow Vangogh$		$Vangogh \rightarrow photo$		**#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 4: Quantitative results for different image translation datasets.

Qualitative Results: We also compare the qualitative translation results with CycleGAN and FQ-GAN. Both baselines were implemented based on their provided checkpoints. Generated images from CycleGAN are selected as relatively successful results. We randomly pick up samples from FQ-GAN and our method. In comparison, our model can generate sharp images with superior visual quality while preserving the source domain's particular semantics by focusing on evolving the discriminative saliency map of the target domain. In the most straightforward case of translating orange to apple, CycleGAN struggles to discriminate between the orange and other insignificant semantics, leading to inferior translation quality. We conjecture that this is because of the entanglement of the adversarial objective and the cycle constraint. More uncurated results on $photo \leftrightarrow Vangogh$ and Yosemite $summer \leftrightarrow winter$ can be found in Appendix A.6.2.



(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. We owe the success to the newly proposed network architecture, smoothing activations, the powerful interpretability of EBM, and the layer-wise expanding process from coarse to fine stage by stage. 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*, pp. 9505–9515, 2018.

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

Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1lUOzWCW.

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*, 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*, 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*, 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*, 2020. URL https://openreview.net/forum?id=B1l4SgHKDH.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In *Advances in Neural Information Processing Systems*, pp. 3608–3618, 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*, 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*, 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*, 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*, 2020. URL https://openreview.net/forum?id=Hkxzx0NtDB.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing systems*, 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*, pp. 297–304, 2010.

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

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*, 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*, 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*, pp. 172–189, 2018.

Long Jin, Justin Lazarow, and Zhuowen Tu. Introspective classification with convolutional nets. In *Advances in Neural Information Processing Systems*, 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*, 2018. URL https://openreview.net/forum?id=Hk99zCeAb.

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. URL https://openreview.net/forum?id=BJlZ5ySKPH.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. URL https://openreview.net/forum?id=33X9fd2-9FyZd.

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

A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009.

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

Justin Lazarow, Long Jin, and Zhuowen Tu. Introspective neural networks for generative modeling. In *Proceedings of the IEEE International Conference on Computer Vision*, 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*, 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*, pp. 3702–3711, 2018b.

Boyi Li, Felix Wu, Kilian Q Weinberger, and Serge Belongie. Positional normalization. In *Advances in Neural Information Processing Systems*, pp. 1622–1634, 2019.

Zengyi Li, Yubei Chen, and Friedrich T. Sommer. Annealed denoising score matching: learning energy based model in high-dimensional spaces, 2020. URL https://openreview.net/forum?id=HJeFmkBtvB.

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

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*, 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*. Citeseer, 2013.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1QRgziT-.

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

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*, pp. 5232–5242, 2019.

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

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

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

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. In *International Conference on Learning Representations Workshops*, 2016. URL https://openreview.net/forum?id=Hkuq2EkPf.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations Workshops*, 2014. URL https://openreview.net/forum?id=cO4ycnpqxKcS9.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 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*, 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*, pp. 13253–13264, 2019.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

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*, 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 Workshops*, pp. 864–873, 2018.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning*, 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*, pp. 2635–2644, 2016.

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.

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.

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.

Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1lxKlSKPH.

Yang Zhao, Chunyuan Li, Ping Yu, Jianfeng Gao, and Changyou Chen. Feature quantization improves GAN training. In *International Conference on Machine Learning*, 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*, pp. 2223–2232, 2017.

Song-Chun Zhu and David Mumford. Grade: Gibbs reaction and diffusion equations. In *Proceedings of the International Conference on Computer Vision*, pp. 847–854, 1998.

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*, 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. To evaluate FID, we adopt the open-source code `https://github.com/mseitzer/pytorch-fid`. Specifically, 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 use the same code as in Kim et al. (2020) to compute KID, which is open-sourced at `https://github.com/taki0112/GAN_Metrics-Tensorflow`.

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 $3 \rightarrow ch/2$	$ch/2 \times 256 \times 256$
BlockDown $ch/2 \rightarrow ch$	$ch \times 128 \times 128$
FromRGB $3 \rightarrow ch$	$ch \times 128 \times 128$
BlockDown $ch \rightarrow 2ch$	$2ch \times 64 \times 64$
BlockDown $2ch \rightarrow 4ch$	$4ch \times 32 \times 32$
BlockDown $4ch \rightarrow 4ch$	$4ch \times 16 \times 16$
BlockDown $4ch \rightarrow 4ch$	$4ch \times 8 \times 8$
BlockDown $4ch \rightarrow 4ch$	$4ch \times 4 \times 4$
Block $4ch \rightarrow 4ch$	$4ch \times 1 \times 1$
Flatten & Linear	1

Table 5: 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 5. 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.

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 6 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.

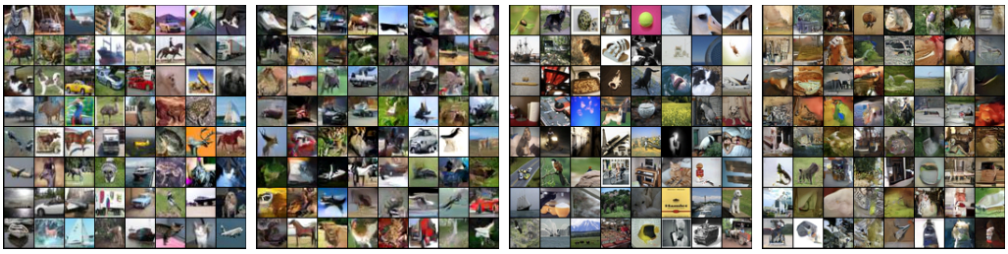


Figure 8: From Left to Right: ground truth of CIFAR-10, generated samples of CIFAR-10, ground truth of ImageNet-1K (32×32), and generated samples of ImageNet-1K (32×32).

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

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 (2020) applies a sequence of decayed perturbation from 1 to 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 11 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

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

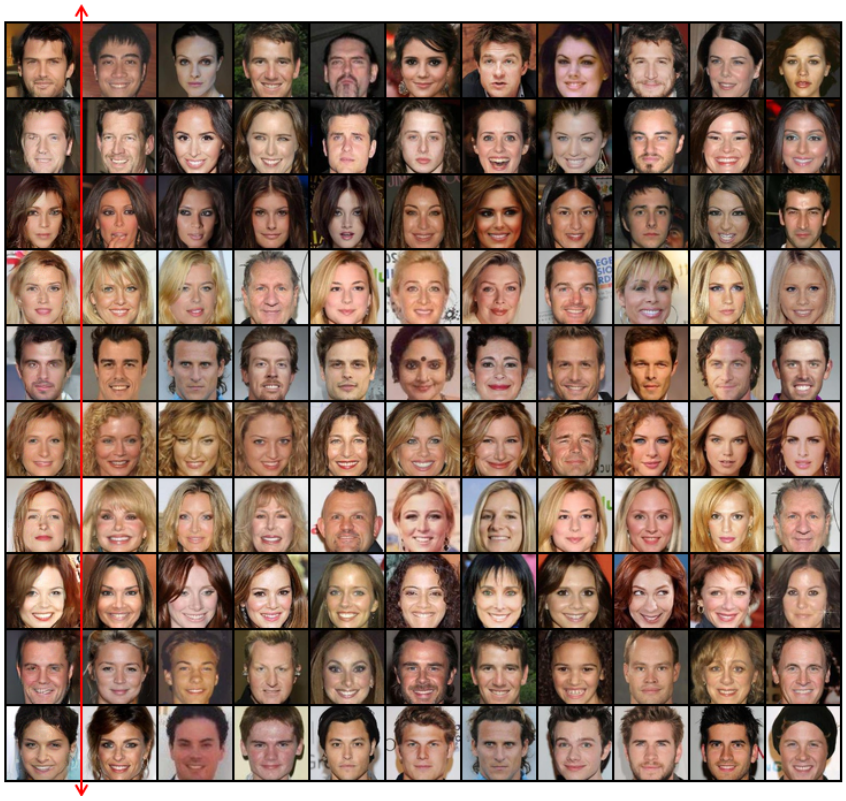


Figure 10: Nearest neighbors on CelebA (64×64). Generated samples are placed on the left side of red vertical line whereas 10 nearest neighbors are on the right side.

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.

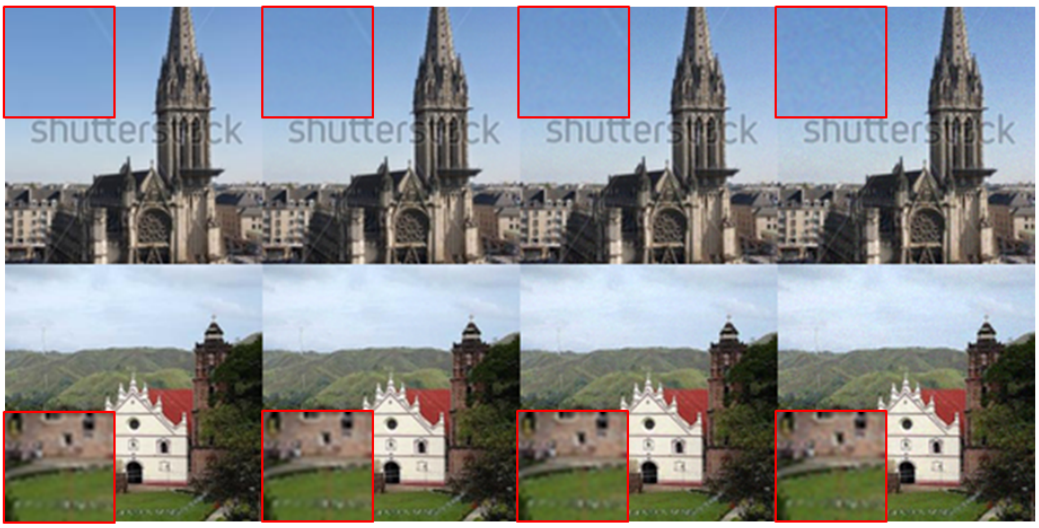


Figure 11: The visualization of different levels of perturbation, with red rectangle areas zoomed in. (LSUN Church 128×128, Yu et al. (2015))

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.

- **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, 2016): $f(x) = x \cdot \Phi(x)$ 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 12 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 7. 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 7: On the impact of activation functions.

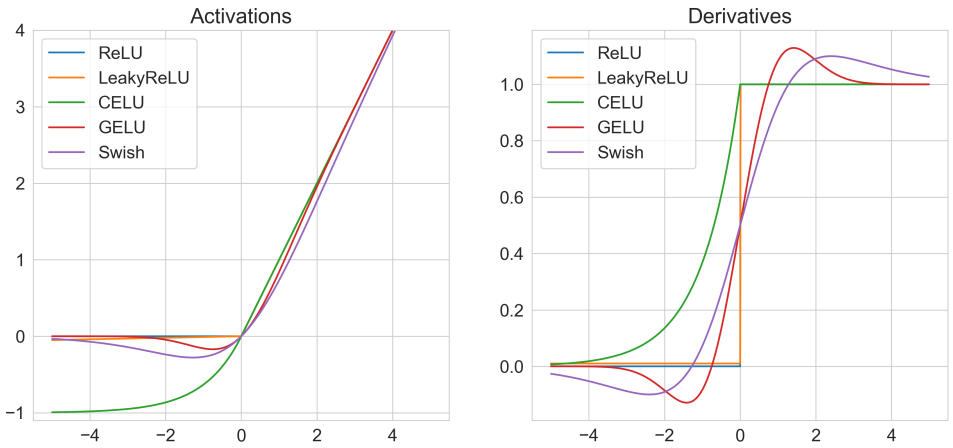


Figure 12: Activations and their derivatives

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 8. 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 8: Ablation results on CIFAR10.

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

A.5 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 13. A larger η_0^ϵ tends to add details to look more realistic but have more noisy images. Figure 14 shows more generated samples on CelebA-HQ 256×256. Table 9 compares the FID of different models on CelebA-HQ 128×128. Note that our approach obtains competitive results without any regularization term.

A.6 Unpaired Image-to-Image translation

A.6.1 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).

Figure 13: Effect of changing the initial noise level η_0^ϵ at the testing phase.

Figure 14: CelebA-HQ 256×256 generated samples, with linearly cooling the "temperature" η_t^ϵ from 0.025 to 0.

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.

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.6.2 More Uncurated Image-to-Image Translation Examples

Figure 15: Translation dynamics: $orange \rightarrow apple$.

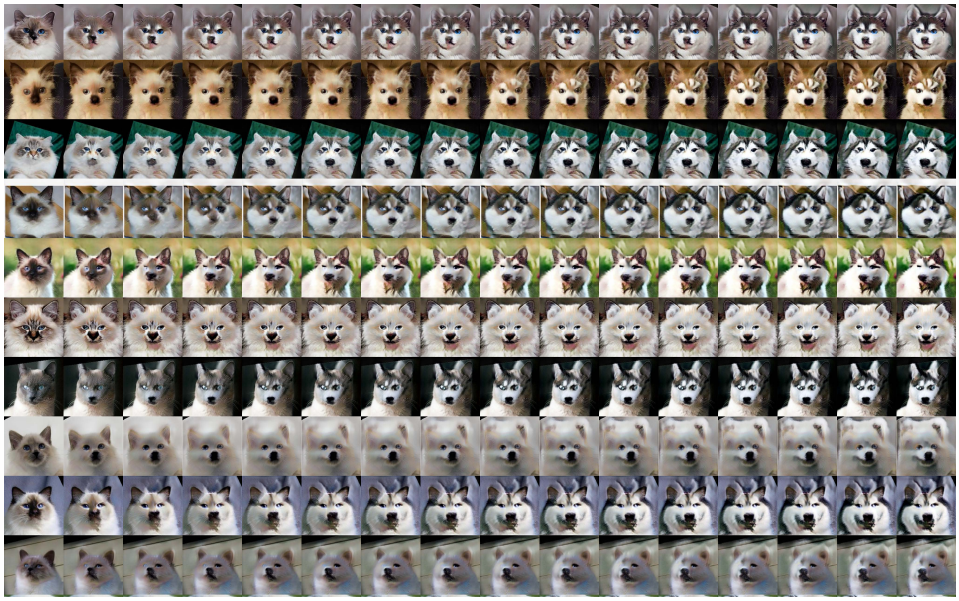


Figure 16: Translation dynamics: $cat \rightarrow dog$.

Figure 17: Translation dynamics: $photo \rightarrow vangogh$.

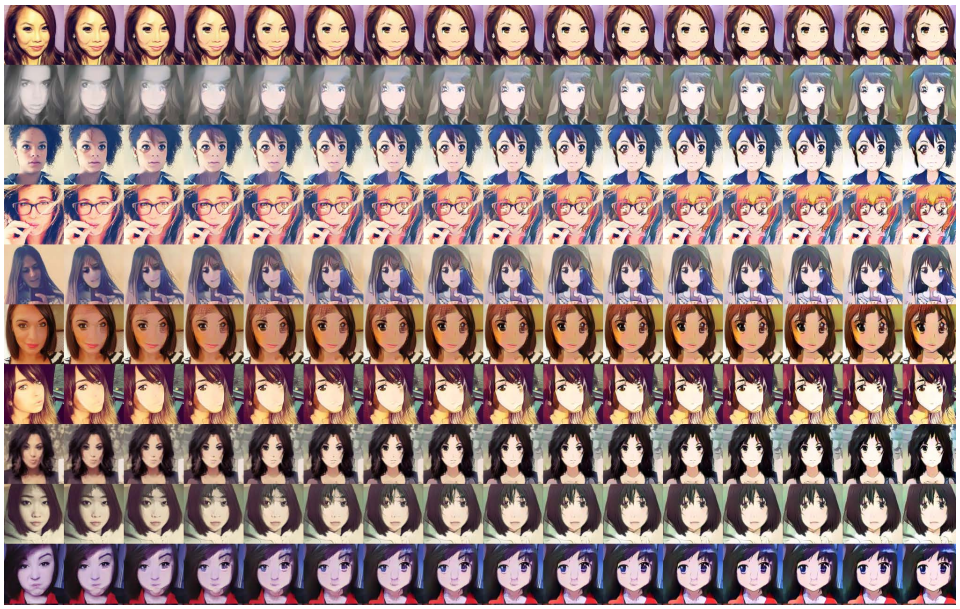


Figure 18: Translation dynamics: $selfie \rightarrow anime$.

Figure 19: Uncurated translation result: $photo \rightarrow Vangogh$

Figure 20: Uncurated translation result: $Vangogh \rightarrow photo$

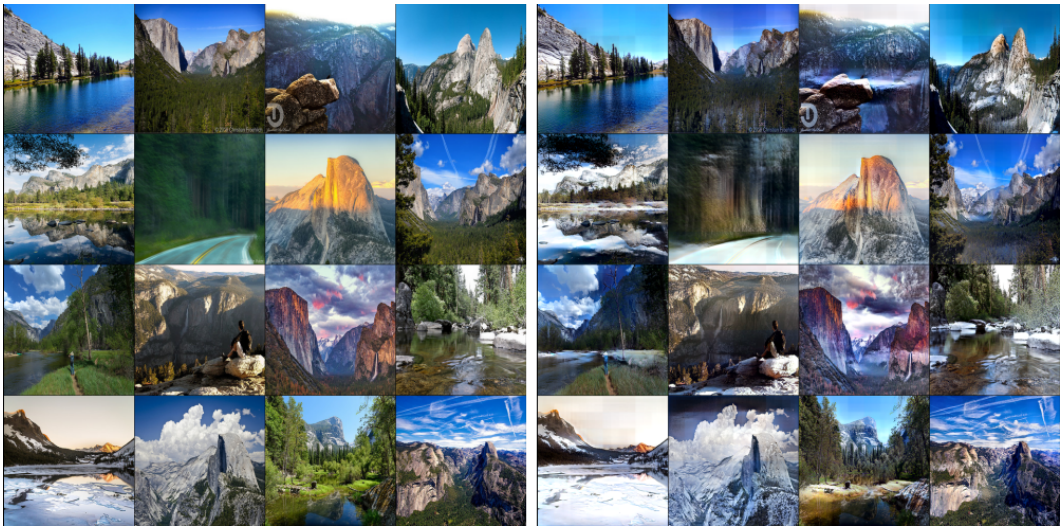


Figure 21: Uncurated translation result: Yosemite $summer \rightarrow winter$

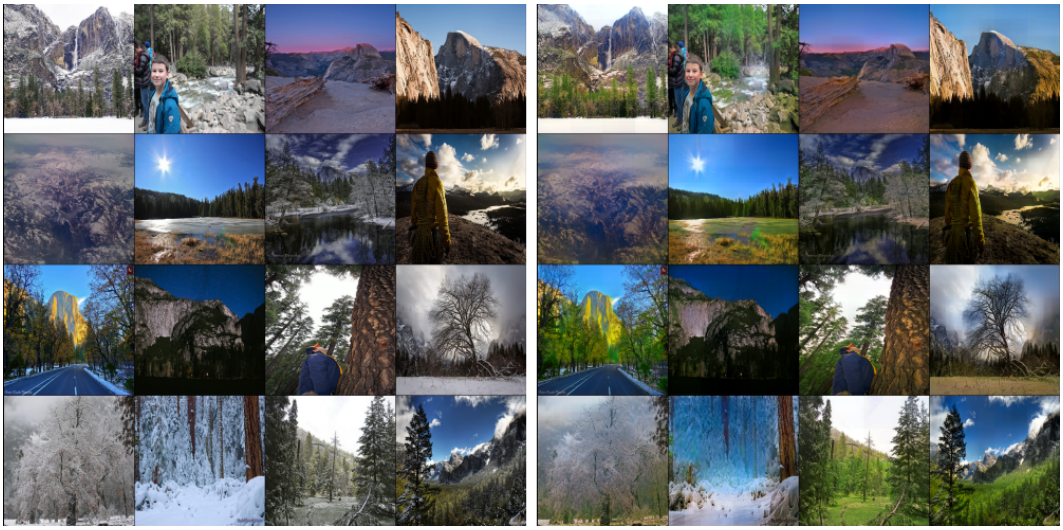


Figure 22: Uncurated translation result: Yosemite $winter \rightarrow summer$