Energy-Calibrated VAE with Test Time Free Lunch

Yihong Luo[1,2], Siya Qiu[1,2], Xingjian Tao[2], Yujun Cai[3], and Jing Tang[2,1⋆]

[1] The Hong Kong University of Science and Technology
[2] The Hong Kong University of Science and Technology (Guangzhou) [3] Meta
yluocg@connect.ust.hk, jingtang@ust.hk

Abstract. In this paper, we propose a novel generative model that utilizes a conditional Energy-Based Model (EBM) for enhancing Variational Autoencoder (VAE), termed Energy-Calibrated VAE (EC-VAE). Specifically, VAEs often suffer from blurry generated samples due to the lack of a tailored training on the samples generated in the generative direction. On the other hand, EBMs can generate high-quality samples but require expensive Markov Chain Monte Carlo (MCMC) sampling. To address these issues, we introduce a conditional EBM for calibrating the generative direction of VAE during training, without requiring it for the generation at test time. In particular, we train EC-VAE upon both the input data and the calibrated samples with adaptive weight to enhance efficacy while avoiding MCMC sampling at test time. Furthermore, we extend the calibration idea of EC-VAE to variational learning and normalizing flows, and apply EC-VAE to an additional application of zero-shot image restoration via neural transport prior and range-null theory. We evaluate the proposed method with two applications, including image generation and zero-shot image restoration, and the experimental results show that our method achieves competitive performance over single-step non-adversarial generation.

Keywords: Image Generation · VAEs · EBMs

1 Introduction

Deep generative models, including Generative Adversarial Nets (GANs) [14], Variational Autoencoders (VAEs) [28], flow-based generative models [8,9], Energy-Based Models (EBMs) [10, 54], and diffusion models [23], achieve excellent performance in a variety of applications. Compared to GANs, likelihood-based models, such as VAEs and EBMs, typically exhibit greater stability during training and more faithfully cover modes in the data. Moreover, unlike normalizing flows that suffer from architecture restrictions [30], VAEs and EBMs offer considerable potential for expressivity. Therefore, VAEs and EBMs have gained extensive attention recently.

⋆ Corresponding author: Jing Tang.

In particular, VAEs map the input data into a latent distribution and optimize the evidence lower bound (ELBO) on the data likelihood. However, VAEs do not explicitly optimize the generative direction. Specifically, VAEs assume that the prior distribution in latent space (*e.g.*, Gaussian distribution) matches the empirical distribution mapped from the input data. Unfortunately, there is often a gap between the two distributions in practice. As a result, VAEs struggle to generate high-quality images, producing blurry or corrupted samples. To tackle this issue, in addition to designing more flexible prior distribution [1,46,57], some work [21,41,47] involves GANs, yielding unstable adversarial games, while NVAE [48] proposes to adjust BatchNorm [22] statistics based on prior distribution that improves the generation slightly.

On the other hand, EBMs directly model the unnormalized density in data space by assigning low energy to high-probability areas. Unlike VAEs that usually assume a Gaussian prior, EBMs do not necessitate distribution assumptions in modeling. In fact, EBMs have shown comparable state-of-the-art performance in terms of generative results among non-adversarial methods [32]. However, a significant drawback of EBMs is the necessity for Markov Chain Monte Carlo (MCMC) sampling during the training and during the generation at test time, which suffers from slow convergence and is computationally expensive, particularly when the energy is parameterized by neural networks.

In this paper, we propose a novel generative model termed Energy-Calibrated VAE (EC-VAE) by involving a conditional EBM to calibrate the VAE for better generation while keeping high sampling efficiency. The VAE is trained by ELBO and the energy-based calibration. Specifically, to address the training of the generative side (*i.e.*, the generating process from prior to data) that is missing in the conventional VAE, we propose to incorporate the generated samples into training. That is, for the generated data $\hat{\mathbf{x}}$ by the generative model, we use a conditional EBM to sample data $\tilde{\mathbf{x}}$ initialized with $\hat{\mathbf{x}}$ that approximates the input real data. Then, integrating the minimization of the distance between $\hat{\mathbf{x}}$ and $\tilde{\mathbf{x}}$ into training can calibrate the decoder. We show that the proposed model can be jointly trained by adopting the primal-dual method in a constrained formulation. The conditional EBM is solely utilized for calibrating samples. Consequently, a short-run MCMC sampling is sufficient and does not suffer from slow convergence. Note that the EBM is involved in training (i.e., generation calibration) only, and is not required during test time sampling.

Moreover, we show that the idea of energy-based calibration can be extended to calibrate the variational inference and normalizing flows with large improvement. Take the former as an illustration while the latter is similar. We first sample $\mathbf{z}$ from variational posterior, and then calibrate $\mathbf{z}$ by running MCMC sampling on constructed conditional posterior $p(\tilde{\mathbf{z}}|\mathbf{z}, \mathbf{x})$ to obtain the calibrated $\tilde{\mathbf{z}}$. Finally, minimizing the distance between $\mathbf{z}$ and $\tilde{\mathbf{z}}$ can calibrate the encoder.

Experimental results show that the proposed EC-VAE outperforms previous EBMs and the state-of-the-art VAEs on image generation benchmarks in both low-resolution and high-resolution datasets by a large margin. In particular, EC-VAE achieves the strong performance with a single-step non-adversarial gener-

ation manner, competing with advanced GANs and diffusions across multiple datasets. We also propose and show how to apply EC-VAE to image restoration in a zero-shot way by constructing neural transport prior and leveraging range-null space theory with competitive performance.

Our main contributions are summarized as follows.

1. We propose a new generative model termed EC-VAE utilizing a conditional EBM to calibrate the VAE to generate sharper samples without incurring extra costs of MCMC sampling during the generation at test time.
2. We extend the energy-based calibration to enhance variational learning and normalizing flows, and apply EC-VAE to an additional application of zero-shot image restoration.
3. We demonstrate the strong empirical results of our proposed methods on various tasks, including image generation and image restoration.

2 Preliminaries

Notations. Denote by $\mathbf{x}$ the data and by $\mathbf{z}$ the latent variable. Let $\mathcal{X}$ be the data space and $\mathcal{Z}$ be the latent space. Let $p_d(\mathbf{x})$ be the data distribution. Denote by $f_\alpha \colon \mathcal{X} \to \mathcal{Z}$ the encoder parameterized by α, and by $g_\beta \colon \mathcal{Z} \to \mathcal{X}$ the decoder parameterized by β. Denote by $E_\omega \colon \mathcal{X} \to \mathbb{R}$ the energy function parameterized by ω.

Variational Autoencoders. VAE [29] adopts the encoder-decoder architecture with a prior distribution. To be more precise, VAE defines the joint distribution of $(\mathbf{x}, \mathbf{z})$ as $p_{\beta,\theta}(\mathbf{x}, \mathbf{z}) = p_\beta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})$. The model can be trained by maximizing the marginal log-likelihood $\mathcal{L} = \mathbb{E}_{p_d(\mathbf{x})}[\log p_{\beta,\theta}(\mathbf{x})]$. However, maximizing the log-likelihood needs sampling from intractable posterior $p_{\beta,\theta}(\mathbf{z}|\mathbf{x}) = \frac{p_{\beta,\theta}(\mathbf{x},\mathbf{z})}{p_{\beta,\theta}(\mathbf{x})}$.

Instead of sampling from intractable posterior $p_{\beta,\theta}(\mathbf{z}|\mathbf{x})$, VAEs propose using a variational inference $q_\alpha(\mathbf{z}|\mathbf{x})$ to approximate the posterior. In particular, VAEs optimize the evidence lower bound (ELBO) on $\log p_{\beta,\theta}(\mathbf{x})$ such that

$$\mathrm{ELBO}_\phi(\mathbf{x}) = \mathbb{E}_{q_\alpha(\mathbf{z}|\mathbf{x})}\big[\log p_\beta(\mathbf{x}|\mathbf{z})) - \mathrm{KL}(q_\alpha(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})\big], \tag{1}$$

where $\phi = \{\alpha, \beta, \theta\}$. The first term is the reconstruction loss and the second term is the KL divergence between the approximated posterior and the prior.

Sampling from VAE can be achieved by sampling $\mathbf{z}$ from prior $p_\theta(\mathbf{z})$ first, and then obtain generated samples $\mathbf{x}$ with probability $p_\beta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(g_\beta(\mathbf{z}), \sigma^2\mathbf{I})$. Equivalently, we denote samples $\mathbf{x} \sim p_{\beta,\theta}(\mathbf{x}) = \int p_\theta(\mathbf{z})p_{\beta,\theta}(\mathbf{x}|\mathbf{z})\mathrm{d}\mathbf{z}$. In practice, the samples are typically obtained directly by $g_\beta(\mathbf{z})$.

Energy-Based Models. A deep EBM assumes that $p_\omega(\mathbf{x})$ is a gibbs distribution with the form $p_\omega(\mathbf{x}) = \exp(-E_\omega(\mathbf{x}))/L_\omega$, where L_ω is the corresponding normalizing constant. EBM is trained by minimizing the negative log-likelihood (NLL) $\mathcal{L}(\omega)$ such that

$$\mathcal{L}(\omega) = -\mathbb{E}_{p_d(\mathbf{x})}[\log p_\omega(\mathbf{x})] = -\mathbb{E}_{p_d(\mathbf{x})}\big[\tfrac{\exp(-E_\omega(\mathbf{x}))}{L_\omega}\big].$$

The gradient of $\mathcal{L}(\omega)$ can be obtained as follows:

$$\nabla \mathcal{L}(\omega) = \mathbb{E}_{p_d(\mathbf{x})}[\nabla E_\omega(\mathbf{x})] - \mathbb{E}_{p_\omega(\mathbf{x})}[\nabla E_\omega(\mathbf{x})]. \tag{2}$$

In practice, sampling from $p_\omega(\mathbf{x})$ can be achieved by running MCMC sampling with K steps of Langevin dynamics, with initial samples $\mathbf{x}_0$ and step size s:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{s}{2}\nabla_{\mathbf{x}_k} E_\omega(\mathbf{x}_k) + \sqrt{s}\xi, \quad \text{where } \xi \sim \mathcal{N}(0, \mathbf{I}).$$

2.1 Additional Related Work

Our method shares some similarities with works that combine generative autoencoders and EBMs in different ways. VAEBM [51] learns EBM upon a pre-trained NVAE [48], and a recent work [15] jointly learns EBM and VAE by an adversarial game instead of by MCMC sampling, while our work is jointly trained without adversarial components. In addition, our learning algorithm bears some similarities to cooperative learning [19, 53, 55, 56], which also employs EBM to teach the base generative model, but only in small images. However, the base generative model in these approaches is purely trained upon *generated samples*, which can be quite biased. In contrast, our base generative model (i.e., VAE) is trained upon data and generated samples, with adaptive weights. Recently, dual MCMC [6] also consider incorporating real data into cooperative-like training. However, similar to other cooperative approaches [53, 55], they maximize the *marginal likelihood* of the generated samples by latent variable models. This requires extra effort in inferring latent variables, increasing the burden of the model and training cost. In contrast, we maximize the *conditional likelihood* of the generated samples via the decoder of VAE, which is easier to learn. Moreover, our method differs significantly from aforementioned works in the way that we discard MCMC during inference without adversarial components, providing a strong one-step generation that competes with diffusions and GANs.

3 The Design of Energy-Calibrated VAE

An issue with VAE is that the generative direction has not been explicitly trained during the training process, potentially leading to lower-quality output for the generated samples. To address this, we propose to incorporate the generated samples into the training. As the real data corresponding to the generated samples are unavailable, we propose utilizing a short-run MCMC, initialized with the generated samples, to approximate the corresponding real data.

As the aim is to calibrate the samples, we suggest constructing a **conditional EBM**, similar in [12], to model the conditional density, i.e.,

$$p_\omega(\tilde{\mathbf{x}}|\mathbf{x}) = \frac{1}{L_\omega(\mathbf{x})}\exp(-E_\omega(\tilde{\mathbf{x}}) - \frac{||\tilde{\mathbf{x}}-\mathbf{x}||_2^2}{2\sigma^2}),$$

where σ is a pre-defined hyper-parameter, $L_\omega(\mathbf{x}) = \int \exp(-E_\omega(\tilde{\mathbf{x}}) - \frac{||\tilde{\mathbf{x}}-\mathbf{x}||_2^2}{2\sigma^2})\mathrm{d}\tilde{\mathbf{x}}$ is the corresponding normalizing constant, $\mathbf{x}$ is the generated samples from VAE,

E_ω is an unconditional EBM. Compared to direct model $p_\omega(\tilde{\mathbf{x}})$, the extra distance term in $p_\omega(\tilde{\mathbf{x}}|\mathbf{x})$ constrains the high-density area localized around generated samples $\mathbf{x}$, making it easier to be learned by the base generative model (i.e., VAE). Sampling from $p_\omega(\tilde{\mathbf{x}}|\mathbf{x})$ can be achieved by MCMC sampling, i.e.,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{s}{2}\Big(\underbrace{\nabla_{\mathbf{x}_k} E_\omega(\mathbf{x}_k)}_{\text{direct to real}} + \underbrace{\nabla_{\mathbf{x}_k} \frac{||\mathbf{x}_k - \mathbf{x}||_2^2}{2\sigma^2}}_{\text{direct to origin}}\Big) + \sqrt{s}\xi, \quad \text{where } \xi \sim \mathcal{N}(0, \mathbf{I}). \tag{3}$$

For simplicity, we denote $\mathbf{x}^K = \text{MCMC}_\omega^K(\tilde{\mathbf{x}}|\mathbf{x})$. The learning of EBM can be achieved by minimizing $\mathcal{L}(\omega) \triangleq \mathbb{E}_{p_{\beta,\theta}(\mathbf{x})}[D_{\text{KL}}(p_d(\tilde{\mathbf{x}}|\mathbf{x})||p_\omega(\tilde{\mathbf{x}}|\mathbf{x}))]$:

$$\begin{aligned}
\nabla\mathcal{L}(\omega) &= \mathbb{E}_{p_{\beta,\theta}(\mathbf{x})}\big[\mathbb{E}_{p_d(\tilde{\mathbf{x}}|\mathbf{x})}[\nabla_\omega \log p_\omega(\tilde{\mathbf{x}}|\mathbf{x})] - \mathbb{E}_{p_\omega(\tilde{\mathbf{x}}|\mathbf{x})}[\nabla_\omega \log p_\omega(\tilde{\mathbf{x}}|\mathbf{x})]\big] \\
&= \mathbb{E}_{p_{\beta,\theta}(\mathbf{x})}[\mathbb{E}_{p_d(\tilde{\mathbf{x}}|\mathbf{x})}[\nabla_\omega E_\omega(\tilde{\mathbf{x}})]] - \mathbb{E}_{p_{\beta,\theta}(\mathbf{x})}[\mathbb{E}_{p_\omega(\tilde{\mathbf{x}}|\mathbf{x})}[\nabla_\omega E_\omega(\tilde{\mathbf{x}})]] \\
&= \mathbb{E}_{p_d(\tilde{\mathbf{x}})}[\nabla_\omega E_\omega(\tilde{\mathbf{x}})] - \mathbb{E}_{p_{\beta,\theta}(\mathbf{x})}[\mathbb{E}_{p_\omega(\tilde{\mathbf{x}}|\mathbf{x})}[\nabla_\omega E_\omega(\tilde{\mathbf{x}})]].
\end{aligned}$$

Since the distance term does not involve learnable parameters, hence the parameter ω can be learned in unconditional form. We only need to define $p_d(\tilde{\mathbf{x}}|\mathbf{x})$ to ensure the marginal distribution of $p_{\beta,\theta}(\mathbf{x})p_d(\tilde{\mathbf{x}}|\mathbf{x})$ being data distribution (e.g., the simplest case is $p_d(\tilde{\mathbf{x}}|\mathbf{x}) \triangleq p_d(\tilde{\mathbf{x}})$). It can be seen that the $\mathcal{L}(\omega)$ reaches minima at $\int p_{\beta,\theta}(\mathbf{x})p_\omega(\tilde{\mathbf{x}}|\mathbf{x})d\mathbf{x} = p_d(\tilde{\mathbf{x}})$. This is said that given rich enough $p_\omega(\tilde{\mathbf{x}}|\mathbf{x})$, we are able to calibrate samples from $p_{\beta,\theta}(\mathbf{x})$ to $p_d(\tilde{\mathbf{x}})$.

Then, we regard the $\tilde{\mathbf{x}}$ as calibrated samples, thus the generative direction can be calibrated by minimizing the distance between $\mathbf{x}$ and $\tilde{\mathbf{x}}$,

$$\mathcal{L}_{\text{calibration}} = \mathbb{E}_{p_{\beta,\theta}(\mathbf{x})}[\mathbb{E}_{p_\omega(\tilde{\mathbf{x}}|\mathbf{x})}[||\mathbf{x} - \tilde{\mathbf{x}}||_2^2]].$$

Note that the calibration loss can be considered as maximizing the likelihood of calibrated samples conditioned on generated samples under the form of normal distribution. This MCMC for sampling calibrated $\tilde{\mathbf{x}}$ can be regarded as a teacher, guiding the generative model to produce higher-quality generated samples. However MCMC sampling introduces some random noise into training, making it impossible to perfectly match the calibrated samples, thus we suggest using a constrained optimization form as follows:

Definition 1 (Energy-Calibrated VAE) *Given a fixed margin ϵ_1, the general optimization can be transformed into the following inequality-constrained optimization.*

$$\begin{aligned}
\min_{\phi,\omega} \quad & \mathcal{L}(\phi) + \mathcal{L}(\omega) \\
\text{s.t.} \quad & ||\tilde{\mathbf{x}} - \mathbf{x}||_2^2 < \epsilon_1, \ \forall \mathbf{x} \sim p_{\beta,\theta}(\mathbf{x}), \tilde{\mathbf{x}} = \text{MCMC}_\omega^K(\tilde{\mathbf{x}}|\mathbf{x}), \\
\text{where} \quad & \mathcal{L}(\omega) = \mathbb{E}_{p_{\beta,\theta}(\mathbf{x})} D_{\text{KL}}(p_d(\tilde{\mathbf{x}}|\mathbf{x})||p_\omega(\tilde{\mathbf{x}}|\mathbf{x})), \\
& \mathcal{L}(\phi) = -\mathbb{E}_{p_d(\mathbf{x})}\mathbb{E}_{q_\alpha(\mathbf{z}|\mathbf{x})}\big[\log p_\beta(\mathbf{x}|\mathbf{z}) - \tfrac{q_\alpha(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z})}\big].
\end{aligned}$$

We get a constrained optimization problem in Definition 1, which is hard to optimize directly, but we can consider it as a corresponding saddle-point problem as follows:

$$\max_{\lambda} \min_{\phi,\omega} \big\{\mathcal{L}(\phi) + \mathcal{L}(\omega) + \lambda \mathcal{L}_{\text{con}}(\phi)\big\}, \quad \lambda \geq 0$$

where the constraint-related loss $\mathcal{L}_{\text{con}}(\phi)$ is defined as:

$$\mathcal{L}_{\text{con}}(\phi) = \mathcal{L}_{\text{con}}(\beta) = \mathbb{E}_{p_{\beta,\theta}(\mathbf{x})}[\mathbb{E}_{p_\omega(\tilde{\mathbf{x}}|\mathbf{x})}[||\mathbf{x} - \tilde{\mathbf{x}}||_2^2 - \epsilon_1]].$$

Since the prior is typically less powerful, we suggest calibrating only the decoder, enabling the prior focus on maximizing likelihood related to the latent variable.

The final challenge is the lack of access to the ground truth of data distribution. To tackle this issue, we consider the corresponding empirical optimization problem as follows:

$$\begin{aligned}
&\max_{\lambda} \min_{\phi,\omega} \{\hat{\mathcal{L}}(\phi) + \hat{\mathcal{L}}(\omega) + \lambda \mathcal{L}_{\text{con}}(\phi)\} \\
&= \max_{\lambda} \min_{\phi,\omega} \Big\{ \sum_{i=1}^{n} \sum_{j=1}^{m} -\Big(\log p_\beta(\mathbf{x}_i|\mathbf{z}_{ij}) + \log \frac{q_\alpha(\mathbf{z}_{ij}|\mathbf{x}_i)}{p_\theta(\mathbf{z}_{ij})} \Big) + \sum_{i=1}^{n} -\log p_\omega(\mathbf{x}_i) \\
&\qquad + \lambda \sum_{i=1}^{n} [||\hat{\mathbf{x}}_i - \text{MCMC}_\omega^K(\tilde{\mathbf{x}}_i|\hat{\mathbf{x}}_i)||_2^2 - \epsilon_1] \Big\},
\end{aligned}$$

where $\mathbf{x}_i$ is sampled from data, $\hat{\mathbf{x}}_i$ is sampled from the VAE, $\mathbf{z}_{ij}$ is sampled from $q_\alpha(\mathbf{z}|\mathbf{x}_i)$. Notice that m is usually chosen to be one in practice that is efficient and effective enough to estimate the gradient. MCMC is not necessary during test time sampling.

Concrete Algorithm. To efficiently optimize the problem, we employ the primal-dual algorithm tailored for addressing the saddle-point problem corresponding to the constrained form. Specifically, in the *primal* step, the algorithm alternately optimizing parameters ω and $\phi = \{\beta, \alpha, \theta\}$ by minimizing the empirical Lagrangian under a given dual variable λ, i.e.,

$$\begin{aligned}
\omega_{t+1} &:= \arg\min_{\omega} \hat{\mathcal{L}}(\omega_t), \\
\beta_{t+1} &:= \arg\min_{\beta} \Big\{ \lambda \hat{\mathcal{L}}_{\text{con}}(\beta_t) - \sum_{i=1}^{n} \sum_{j=1}^{m} \log p_{\beta_t}(\mathbf{x}_i|\mathbf{z}_{ij}) \Big\}, \\
\alpha_{t+1} &:= \arg\min_{\alpha} \Big\{ \sum_{i=1}^{n} \sum_{j=1}^{m} \log \tfrac{q_{\alpha_t}(\mathbf{z}_{ij}|\mathbf{x}_i)}{p_{\theta_t}(\mathbf{z}_{ij})} - \log p_{\beta_t}(\mathbf{x}_i|\mathbf{z}_{ij}) \Big\}, \\
\theta_{t+1} &:= \arg\min_{\theta} \Big\{ \sum_{i=1}^{n} \sum_{j=1}^{m} -\log p_{\theta_t}(\mathbf{z}_{ij}) \Big\}.
\end{aligned}$$

In practice, we perform stochastic gradient descent that derives concrete update step for θ and ϕ. On the other hand, in the *dual* step, we update λ as follows,

$$\lambda_{t+1} := \max \Big\{ \lambda_t + \eta \cdot \big(\hat{\mathcal{L}}_{\text{con}} - \epsilon\big), 0 \Big\}, \tag{4}$$

where η is the learning rate of dual step.

Algorithm 1 in Appendix A gives the pseudo-code of our primal-dual algorithm for optimizing the VAE parameters ϕ, and EBMs parameters ω. Compared with using stochastic gradient descent directly in the constrained optimization problem in Definition (1), the primal-dual algorithm can dynamically tune λ to avoid introducing extra hyper-parameters (which may serve as an early-stopping condition, e.g., $\lambda = 0$), and can provide convergence guarantees with sufficiently long training using sufficiently small step size [3].

4 Extensions and Additional Application

In this section, we first show the calibration idea of EC-VAE can be extended to enhance variational learning and normalizing flows. Then we show how to apply our method in zero-shot image restoration.

4.1 Energy-Calibrated Variational Learning

Variational learning allows efficient training, but results in learning a lower bound for data likelihood. The gap is the KL divergence between variational posterior $q_\alpha(\mathbf{z}|\mathbf{x})$ and posterior $p_{\beta,\theta}(\mathbf{z}|\mathbf{x})$ which is not explicitly minimized in variational training, just like the training of the generative direction is missing in vanilla VAEs.

Similar to calibrating the generative direction as proposed in Sec. 3, we propose to incorporate the calibration of the variational posterior $q_\alpha(\mathbf{z}|\mathbf{x})$ into training. We first construct the conditional density $p_{\beta,\theta}(\tilde{\mathbf{z}}|\mathbf{z},\mathbf{x})$ as follows:

$$p_{\beta,\theta}(\tilde{\mathbf{z}}|\mathbf{z},\mathbf{x}) = p_{\beta,\theta}(\tilde{\mathbf{z}}|\mathbf{x}) \cdot \exp(-\tfrac{||\tilde{\mathbf{z}}-\mathbf{z})||_2^2}{2\sigma^2})/L_{\beta,\theta}(\mathbf{z}),$$

where σ is a pre-defined hyper-parameter, $p_{\beta,\theta}(\tilde{\mathbf{z}}|\mathbf{x})$ is the posterior and $L_{\beta,\theta}(\mathbf{z})$ is corresponding normalizing constant. The conditional density is constructed by adding the distance term to constrain the calibrated $\mathbf{z}$ to be close to $\mathbf{z}$. And the sampling can be achieved by MCMC with Langevin dynamics. Given a step size $s > 0$ and an initial value $\mathbf{z}_0$, the Langevin dynamics iterates:

$$\mathbf{z}_{k+1} = \mathbf{z}_k - \tfrac{s}{2}\nabla_{\mathbf{z}_k}(-\log p_{\beta,\theta}(\mathbf{z}_k|\mathbf{x}) + \tfrac{||\mathbf{z}_k-\mathbf{z}||_2^2}{2\sigma^2}) + \sqrt{s}\xi, \quad \xi \sim \mathcal{N}(0,\mathbf{I})$$

where the $\mathbf{z}_0$ is proposed to be sampled from variational posterior $q_\alpha(\mathbf{z}|\mathbf{x})$, thus the $\mathbf{z}_K$ can be considered as calibrated $\mathbf{z}$. For simplicity, we denote $\mathbf{z}^K = \text{MCMC}^K_{\beta,\theta}(\tilde{\mathbf{z}}|\mathbf{z})$. Note the $\nabla_{\mathbf{z}} \log p_{\beta,\theta}(\mathbf{z}|\mathbf{x})$ can be easily obtained by following form: $\nabla_{\mathbf{z}} \log p_{\beta,\theta}(\mathbf{z}|\mathbf{x}) = \nabla_{\mathbf{z}} \log p_{\beta,\theta}(\mathbf{x},\mathbf{z}) - \nabla_{\mathbf{z}} \log p_{\beta,\theta}(\mathbf{x}) = \nabla_{\mathbf{z}} \log p_{\beta,\theta}(\mathbf{x},\mathbf{z})$.

Once the MCMC Calibrated $\mathbf{z}_K$ is obtained, we can conduct the constrained learning formulation similar to Sec. 3. We directly give the corresponding saddle-point problem:

$$\mathcal{L}(\phi = \{\alpha,\beta,\theta\}) = -\mathbb{E}_{p_d(\mathbf{x})}[\text{ELBO}_\phi(\mathbf{x})] + \lambda_1 \cdot \mathcal{L}_{\text{con}}(\beta) + \lambda_2 \cdot \mathcal{L}_{\text{con}}(\alpha),$$

where $\lambda_1, \lambda_2 \geq 0$ and $\mathcal{L}_{\text{con}}(\alpha) = \mathbb{E}_{q_\alpha(\mathbf{z}|\mathbf{x})}\left[\mathbb{E}_{p_{\beta,\theta}(\tilde{\mathbf{z}}|\mathbf{z},\mathbf{x})}[||\mathbf{z}-\tilde{\mathbf{z}}||_2^2]\right]$. Given the saddle-point formulation, the learning and calibration of the encoder parameterized by α can be achieved with corresponding empirical optimization as follows:

$$\alpha_{t+1} := \arg\min_{\alpha} \left\{ \sum_{i=1}^{n}\sum_{j=1}^{m} \Big(\log \frac{q_{\alpha_t}(\mathbf{z}_{ij}|\mathbf{x}_i)}{p_{\theta_t}(\mathbf{z}_{ij})} - \log p_{\beta_t}(\mathbf{x}_i|\mathbf{z}_{ij}) + \lambda_2 \times ||\mathbf{z}_{ij} - \text{MCMC}^K_{\beta,\theta}(\tilde{\mathbf{z}}_{ij}|\mathbf{z}_{ij})||_2^2 \Big) \right\},$$

The update of the rest parameters $\{\beta, \theta, \omega\}$ remain the same, while the λ_i (i=1,2) is updated by Eq. (4).

4.2 Energy-Calibrated Normalizing Flow

The Energy-Calibrated Normalizing Flow (EC-Flow) can be easily formed, as normalizing flow h_ϕ directly models the density $p_\phi(\mathbf{x}) = p_z(\mathbf{z}) \times \left|\det\left(\frac{\partial h_\phi}{\partial \mathbf{z}}\right)\right|$, where h_ϕ is the invertible transformation, thus we just need to use the negative model log-likelihood $-\mathbb{E}_{p_d(\mathbf{x})}[\log p_\phi(\mathbf{x})]$ to serve as the $\mathcal{L}(\phi)$, and directly sampling from $p_\phi(\mathbf{x})$ to obtain samples, while other components keep the same.

4.3 Applying to Zero-Shot Image Restoration

In this section, we present the application of our method in zero-shot image restoration tasks, inspired by recent works [50] that employ the diffusion model and the Range-Null space theory for similar purposes.

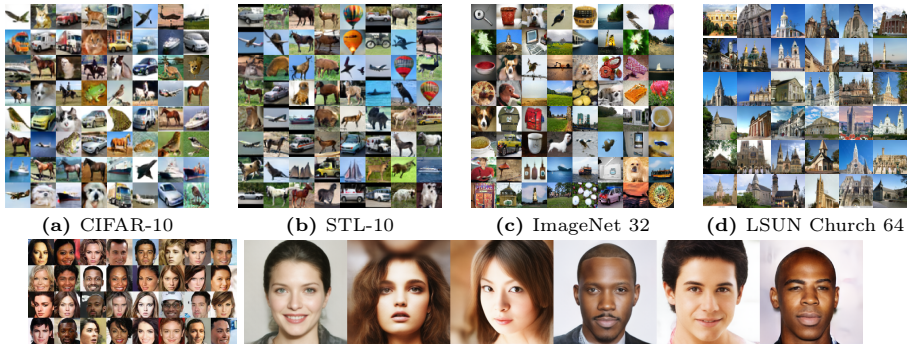
We start by briefly reviewing the necessary background. Given a known linear operator $\mathbf{A} \in \mathbb{R}^{D_1 \times D_2}$, there exists pseudo-inverse $\mathbf{A}^\dagger \in \mathbb{R}^{D_2 \times D_1}$ that satisfies $\mathbf{A}\mathbf{A}^\dagger\mathbf{A} = \mathbf{A}$. Considering degraded image: $\mathbf{y} = \mathbf{A}\mathbf{x}$. For any prediction $\hat{\mathbf{x}}_r$, we let $\hat{\mathbf{x}} = \mathbf{A}^\dagger\mathbf{y} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\hat{\mathbf{x}}_r$, then immediately gives: $\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}\mathbf{A}^\dagger\mathbf{y} + \mathbf{0} = \mathbf{A}\mathbf{x}$, which we make predicted images have the same degradation as original images. This can be regarded as $\mathbf{x}_r$ predicting the zero-space while remaining the original rang-space $\mathbf{A}\mathbf{x}$. For getting the $\hat{\mathbf{x}}_r$, we can employ $\mathbf{A}^\dagger\mathbf{y}$ which is the range-space part of $\mathbf{x}$ as biased ground truth and build a joint distribution

$$p_{\beta,\theta}(\mathbf{A}^\dagger\mathbf{y}, \mathbf{z}) = p_\theta(\mathbf{z})p(\mathbf{A}^\dagger\mathbf{y}|\mathbf{A}^\dagger\mathbf{A}g_\beta(\mathbf{z})). \tag{5}$$

However the $p_\theta(\mathbf{z})$ may not be powerful enough to handle image restoration. Recall that we have an EBM in data space, hence we can construct a more powerful prior via neural transport, i.e.,

$$p_{\theta,\omega}(\mathbf{z}) \propto \exp(-E_\omega(g_\beta(\mathbf{z})))p_\theta(\mathbf{z}) \propto \exp(-E_{\beta,\omega}(\mathbf{z}))p_\theta(\mathbf{z}). \tag{6}$$

This is achieved by transporting $\mathbf{z}$ to data space via decoder, then the EBM defined in data space is used, finally, the EBM defined in latent space for enhancing prior is well-constructed. By substituting the prior in Eq. (6) into Eq. (5), the



(a) CIFAR-10 **(b)** STL-10 **(c)** ImageNet 32 **(d)** LSUN Church 64

(e) CelebA 64 **(f)** CelebA-HQ-256

Fig. 1: Random generated samples from EC-VAE. For CelebA 64 and CelebA-HQ-256, we pick out samples for diversity.

new joint likelihood is obtained. Then we run MCMC on $\mathbf{z}$ to maximize the biased joint likelihood, i.e.,

$$g_\beta(\mathbf{z}_r^{k+1}) = g_\beta(\mathbf{z}_r^k) + \tfrac{s}{2}\nabla_{\mathbf{z}_r^k} \log p_{\theta,\beta,\omega}(\mathbf{A}^\dagger \mathbf{y}, \mathbf{z}_r^k) + \sqrt{s}\xi, \tag{7}$$

where $\xi \sim \mathcal{N}(0, \mathbf{I})$, $\mathbf{z}_r^0$ can be initialized by sampling from $p_\theta(\mathbf{z})$. In practice, the linear degraded operator $\mathbf{A}$ has various corresponding image restoration tasks, such as colorization, super-resolution, and inpainting. See Appendix B for concrete forms of $\mathbf{A}$ and $\mathbf{A}^\dagger$. It's worth noting that our proposed method doesn't need extra training for those tasks.

5 Experiment

In this section, we conduct comprehensive experiments to evaluate the proposed EC-VAE. We also evaluate the extension of Energy-based calibration to normalizing flow and variational inference. We use a simple ResNet [17] similar to used in VAEBM [51] or CLEL [32] as energy functions E_ω with 30 MCMC steps on CIFAR-10 and 15 MCMC steps on other datasets in all experiments. We adopt Fréchet Inception Distance (FID) [18] as quantitative metrics in most experiments. We apply Exponential Moving Average (EMA) with coefficient of 0.999 on Church-64 and 0.9999 on other datasets for VAE. Please note that by default, our proposed model does not incorporate the use of MCMC in test time, unless specifically indicated otherwise. The prior $p_\theta(\mathbf{z})$ is a simple Gaussian as default.

5.1 Image Generation

In this section, we evaluate EC-VAE on six datasets, including CIFAR-10 [58], STL-10 [5], ImageNet 32 [4,7], LSUN Church 64 [58], CelebA 64 [35], and CelebA-HQ-256 [35]. We use a flow as the prior in experiments on the CelebA-HQ-256.

Table 1: Generative performance on CIFAR-10. The † means without extra hyper-parameters tuning. The * means we evaluate the FID by officially released checkpoint.

	Model	NFE↓	FID↓	Times (s)↓
Score-based	NCSN [44]	1000	25.32	107.9
	Denoising Diffusion [20]	1000	<u>3.17</u>	80.5
	Consistency Models (LPIPS) [43]	1	8.70	-
GAN-based	SN-GAN [37]	1	21.7	-
	AutoGAN [13]	1	12.4	-
	StyleGAN2 w/o ADA [25]	1	9.9	0.04
VAEs+GANs	VAE+GAN [41]	1	39.8	-
	DC-VAE [41]	1	17.9	-
EBMs+GANs	FlowCE [11]	1	37.3	-
	Divergence Triangle [15]	1	30.1	-
Flow-based	Glow [27]	1	48.9	-
	SurVAE Flow [38]	1	49.03	-
NVAE-family	NVAE [48]	1	50.97	0.36
	NCP-VAE [1]	-	24.08	-
Energy-based	IGEBM [10]	60	40.58	-
	CoopFlow [56]	31	15.80 ($26.54^{\dagger *}$)	-
	CoopFlow w/o MCMC	1	79.45^{*}	-
	Hat EBM [19]	51	19.30	-
	CLEL-Large [32]	1200	8.61	-
	Dual MCMC [6]	31	9.26	-
	Diffusion EBM [11]	180	9.58	-
	VAEBM [51]	16	12.19	8.79
Ours	**EC-VAE**	1	**5.20**	**0.03**
	EC-Flow	1	31.12	0.37
Ablations	**EC-VAE w/ MCMC**	31	5.63	0.52
	EC-Flow w/ MCMC	31	22.74	0.86
	VAE	1	104.68	0.03
	Flow	1	91.33	0.37

See Appendix C for more experiment setting details. We show qualitative results in Fig. 1. See Appendix G for more qualitative results and qualitative comparison to other models. The quantitative results are reported in Tabs. 1 to 6, with the best results for GANs or Score-models highlighted by underlining, and the best results for other models in bold, respectively.

Our results are comparable to advanced GANs and Score-based Models, outperforming NVAE-family (including NVAE, NCP-VAE and VAEBM) and Glow, the most advanced Hierarchical VAE and flows, respectively, by a significant margin on all datasets, despite using much smaller latent space, and much fewer training resources. Our results also outperform existing EBMs, with a *single-step generation manner*. Additionally, Our results outperform the Consistency Models with Learned Perceptual Image Patch Similarity (LPIPS) [59] which is a strong model on single-step non-adversarial generation. Note our approach does not rely on LPIPS, which is known to potentially manipulate the FID metric [31]. Remarkably, we obtain a strong result of unconditional generation on ImageNet 32, even outperforming DDPM [23]. This indicates the strong potential and capability of our proposed model for learning highly diverse datasets.

Comparison with Other EBMs. The proposed method presents a crucial, clear difference from existing methods: we demonstrate that it is possible to

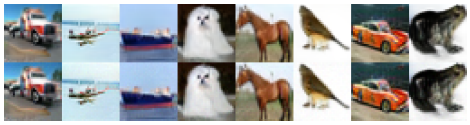


Fig. 2: Comparison of **EC-VAE** (Top) and **EC-VAE (w/ MCMC)** (Bottom) on CIFAR-10. Best viewed when zoomed in.

Table 2: Generative performance on STL-10.

Model	FID↓
ProbGAN [16]	46.7
SN-GAN [37]	40.1
Improv. MMD GAN [49]	37.6
AutoGAN [13]	31.0
DC-VAE [41]	41.91
EC-VAE (Ours)	**8.39**

Table 3: Generative performance on ImageNet 32.

Model	FID↓
DDPM [23]	6.99
Flow Matching [34]	5.02
PixelCNN [39]	40.51
CF-EBM [60]	26.31
CLEL-Large [32]	15.47
EC-VAE (Ours)	**5.76**

Table 4: Generative performance on Church 64.

Model	FID↓
NVAE [48]	41.3
GLOW [27]	59.35
Diffusion EBM [12]	7.02
Dual MCMC [6]	4.56
VAEBM [51]	13.51
EC-VAE (ours)	**4.28**

Table 5: Generative performance on CelebA 64.

Model	FID↓
NCSNv2 [45]	26.86
COCO-GAN [33]	4.0
QA-GAN [40]	6.42
Divergence Triangle [15]	24.7
Dual MCMC [6]	5.15
VAEBM [51]	5.31
NCP-VAE [1]	5.25
NVAE [48]	14.74
EC-VAE (Ours)	**2.71**

Table 6: Generative performance on CelebA-HQ-256.

Model	FID↓
ALAE [42]	19.21
DC-VAE [41]	15.81
PGGAN [24]	8.03
GLOW [27]	68.93
Dual MCMC [6]	15.89
VAEBM [51]	20.38
NCP-VAE [1]	24.79
NVAE [48]	45.11
EC-VAE (Ours)	**12.35**

drop MCMC steps during test time sampling without compromising the quality of generation. And the FID is even better without MCMC. This is likely because MCMC introduces noise at each step which cannot be perfectly denoised (See Fig. 2), resulting in noise still present in the image. In contrast, the decoder minimizes the distance $\mathbb{E}_{\mathbf{x}}||\tilde{\mathbf{x}}, \mathbf{x}||_2^2$, assuming $\tilde{\mathbf{x}} = \mathbf{y} + \xi$, where ξ is zero-mean random noise without any information. The minimum is ideally achieved at $\mathbf{x} = \mathbf{y}$, allowing the decoder to act as a filter and output sharp images without noise. Moreover, as observed, most previous EBMs [10, 51, 56] need to tune MCMC steps and step size carefully during inference to achieve high performance (e.g., CoopFlow† in Tab. 1), while our method does not need extra hyper-parameters tuning as even no MCMC required during inference.

Energy-Calibrated Normalizing Flow. We evaluate the EC-Flow on CIFAR-10. We use Glow as the flow architecture. As shown in Tab. 1, the EC-Flow outperforms existing flows in single-step generation, even outperforming the FlowCE

Fig. 3: Qualitative results of zero-shot image restoration (colorization, inpainting, 4× super-resolution).

Table 7: Quantitative results Calibrated posterior on CIFAR-10.

Model	FID↓	MSE↓	ELBO↑
VAE	104.68	0.0235	-953.43
+ **Calibrated posterior**	102.10	0.0198	-349.14
EC-VAE	**5.20**	0.0193	652.59
+ **Calibrated posterior**	5.81	**0.0170**	**1072.40**

Table 8: Sampling efficiency and training cost on CIFAR-10. Time is the seconds took to generate 50 images.

Model	FID↓	Latent dim↓	Time↓	GPU days ↓	GPU-Type
NCSN	25.32	3072	107.9	-	-
GLOW [38]	48.9	3072	-	60	-
SurVAE Flow [38]	49.03	1536	-	7	TITAN-X
NVAE [48]	50.97	153600	0.36	18.3	32G-V100
NCP-VAE [1]	24.08	153600	-	34.5	32G-V100
VAEBM [51]	12.19	153600	8.79	≥18.3	32G-V100
EC-VAE w/ MCMC (ours)	5.63	**128**	0.52	**3**	RTX-3090
EC-VAE (ours)	**5.20**	**128**	**0.03**	**3**	RTX-3090

which is trained by playing an adversarial game. However, we found that in EC-Flow, the role of MCMC remains crucial in enhancing generative performance. Specifically, EC-Flow w/ MCMC significantly improves the FID from 31.12 to 22.74. We believe that this is due to the expressivity limitations of invertible transformations in flows. Nevertheless, compared to CoopFlow which also combines flows and EBMs, our EC-Flow beat CoopFlow without MCMC by a large margin, demonstrating the effectiveness of the proposed EC-Flow. Note that the 15.8 FID achieved by CoopFlow needs extra hyper-parameter tuning, its result (26.54 FID) without extra tuning is worse than ours w/ MCMC (22.74 FID).

Energy-Calibrated Variational Learning. We also evaluate the Energy-Calibrated Variational Learning on CIFAR-10, with 5 MCMC steps for calibrating posterior. As shown in Tab. 7, the calibrated posterior consistently enhances the ELBO and MSE of related baseline VAE and EC-VAE. This indicates that the calibrated posterior can provide a more accurate posterior for better likelihood maximization. However, a slight FID deterioration is observed in the EC-VAE with Calibrated posterior, implying that the likelihood is not always consistent with generation quality. Interestingly, we found the EC-VAE also improves the ELBO and MSE compared to VAE, despite we only calibrate the generative side. This may be highly due to the energy-based calibration effectively improving the ability to map latent space to data space and reducing the gap between prior and aggregated posterior.

Table 9: Quantitative results of image restoration on CelebA-HQ.

Model	4× SR	Colorization
	PSNR↑ / FID↓	Cons↓ / FID↓
PULSE [36]	22.7 / 40.3	N/A
DDRM [26]	**31.6** / 31.0	456 / 31.2
DDNM [50]	**31.6** / **22.3**	26.2 / 26.4
EC-VAE (ours)	28.8 / 30.4	**0.004** / **13.3**

Table 10: Comparison for FID on CIFAR-10 between several related methods.

Model	FID↓
EC-VAE	5.20
EC-VAE w/ flow prior	4.85
EC-VAE w/ LPIPS	4.77
EC-VAE w/o primal-dual	5.72
VAE	104.68
EBM	45.52
EBM init w/ VAE samples	44.63
VAE + WGAN	33.78

5.2 Sampling Efficiency and Training Cost

Although the score-based model and NVAE's variants have shown outstanding sample quality, their sampling speed is limited by the necessity of expensive MCMC sampling steps. As shown in Tab. 8, the generation by EC-VAE, in contrast, takes just one pass, making it hundreds of and thousands of times faster than NCSN and VAEBM, respectively. Notably, even after performing MCMC steps, our method still only requires 0.52 seconds to generate 50 samples. This is because our energy network is lightweight and only requires 30 MCMC steps on the data space, while VAEBM runs MCMC in $(\mathbf{x}, \mathbf{z})$ space which requires backward through its heavy decoder at each step. Furthermore, due to the extremely large scale of latent variables used in NVAE-family and previous flows, yielding challenges in learning, they need at least 7 GPU days to be trained on CIFAR-10, while our model only needs 3 GPU days despite using costly MCMC steps in training.

5.3 Image Restoration

Here we show that well-trained EC-VAE is able to be zero-shot used in image restoration as described in Sec. 4.3. The Qualitative results are shown in Fig. 3. Our model can successfully restore those images with high quality and consistency. More qualitative results can be found in Appendix G. Following the setting as in [50], we compare our method with strong zero-shot baselines, using metrics FID, PSNR, and Consistency [50] (i.e., $||\mathbf{A}\hat{\mathbf{x}}-\mathbf{A}\mathbf{x}||_1$). As shown in Tab. 9, we outperform GAN-based PULSE [36] and compete with diffusion-based DDNM [50] and DDRM [26], which confirms our model provides competitive performance.

5.4 Ablation Study

All experiments here are performed on CIFAR-10 for faster training. See Appendix E for Experimental settings and additional ablations. See Appendix F for extra study on mode coverage.

Effect of Calibration. The proposed method is calibrating generative models by incorporating generated samples into training. This raises the question of

what is the performance of single generative model and train generative model and EBMs as independent components. We compare the proposed EC-VAE with three related baselines i.e., VAE, EBM, and EBM initialized with pre-trained VAE samples. As shown in Tab. 10, we significantly outperform these three variants by a huge margin. Note that there is almost no benefit of initializing EBM with pre-trained VAE samples, implying the effectiveness of calibration.

Energy-based Calibration vs. Adversarial-based Calibration. The gradient for updating EBMs E_ω is similar to the gradient updates of WGAN's discriminator [2]. The key difference is that WGAN updates the generator and discriminator by adversarial training, while we update EBMs by maximizing the likelihood and we update the VAE by maximizing adaptive weight between the lower bound of data likelihood and conditional likelihood on calibrated samples (i.e., $\log p(\tilde{\mathbf{x}}|\mathbf{x}) = -\frac{||\tilde{\mathbf{x}}-\mathbf{x}||_2^2}{2\sigma^2} + \text{Constant}$). Thus, a related variant of EC-VAE is updating the VAE by both ELBO and WGAN training objectives. We compare the adversarial variant with the proposed EC-VAE. As shown in Tab. 10, the adversarial variant achieves similar results (**33.78 FID**) to the VAEs+GANs (**39.8 FID**) in DC-VAE which is significantly worse than the proposed EC-VAE (**5.20 FID**), highlighting the advantage of our method compared to adversarial training.

Effect of Primal-Dual. We firstly emphasize that primal-dual has advantages with theoretical guarantees in constrained learning [3]. This section is for empirically showing the impact intuitively. As shown in Tab. 10, in comparison to primal-dual, the naive stochastic gradient descent (w/o dual variable λ) algorithm yields worse results.

Compatibility with Advanced Technique in VAEs. We employ MSE as the distance metric and a Gaussian as prior in CIFAR-10. Recognizing the efficacy of flexible priors and better distance metrics in improving vanilla VAEs [1, 52], we investigate the impact of integrating either flow prior or LPIPS into the EC-VAE. The superior performance of both variants (Tab. 10) suggests that EC-VAE is orthogonal to these advanced techniques to some extent. Additionally, we note that the improvement margin is not large, indicating that EC-VAE effectively addressed the prior hole issue and blurry generation issue in vanilla VAEs.

6 Conclusion

In this paper, we proposed EC-VAE, using conditional EBMs for calibrating the VAEs. Furthermore, the proposed energy-based calibration can enhance normalizing flows and variational posterior. We also propose and show that EC-VAE can effectively solve image restoration in a zero-shot manner. We show that MCMC sampling is not required once the VAE is calibrated while keeping high performance. In terms of efficiency, EC-VAE can be trained by a single GPU and is fast to converge, addressing the intensive computational resources consumption problem of previous state-of-the-art VAEs (i.e., NVAE). The proposed EC-VAE shows promising results over multiple datasets with high computational efficiency, significantly reducing the gap with GANs.

Acknowledgments

Jing Tang's work is partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. U22B2060, by National Key R&D Program of China under Grant No. 2023YFF0725100, by National Language Commission under Grant No. WT145-39, by The Department of Science and Technology of Guangdong Province under Grant No. 2023A1515110131, by Guangzhou Municipal Science and Technology Bureau under Grant No. 2023A03J0667 and 2024A04J4454, by Hong Kong Productivity Council (HKPC), and by Createlink Technology Co., Ltd.

References

1. Aneja, J., Schwing, A., Kautz, J., Vahdat, A.: A contrastive learning approach for training variational autoencoder priors. Advances in Neural Information Processing Systems **34** (2021) 2, 10, 11, 12, 14
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 214–223. PMLR (06–11 Aug 2017), https://proceedings.mlr.press/v70/arjovsky17a.html 14
3. Chamon, L.F., Paternain, S., Calvo-Fullana, M., Ribeiro, A.: Constrained learning with non-convex losses. arXiv:2103.05134 (2021) 7, 14
4. Chrabaszcz, P., Loshchilov, I., Hutter, F.: A downsampled variant of imagenet as an alternative to the cifar datasets. ArXiv **abs/1707.08819** (2017), https://api.semanticscholar.org/CorpusID:7304542 9
5. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: AISTATS. pp. 215–223 (2011) 9
6. Cui, J., Han, T.: Learning energy-based model via dual-mcmc teaching. Advances in Neural Information Processing Systems **36** (2024) 4, 10, 11
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (Jun 2009). https://doi.org/10.1109/cvpr.2009.5206848, http://dx.doi.org/10.1109/cvpr.2009.5206848 9
8. Dinh, L., Krueger, D., Bengio, Y.: NICE: non-linear independent components estimation. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings (2015), http://arxiv.org/abs/1410.8516 1
9. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real nvp. arXiv preprint arXiv:1605.08803 (2016) 1
10. Du, Y., Mordatch, I.: Implicit generation and modeling with energy based models. In: Advances in Neural Information Processing Systems. pp. 3608–3618 (2019) 1, 10, 11
11. Gao, R., Nijkamp, E., Kingma, D.P., Xu, Z., Dai, A.M., Wu, Y.N.: Flow contrastive estimation of energy-based models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7518–7528 (2020) 10
12. Gao, R., Song, Y., Poole, B., Wu, Y.N., Kingma, D.P.: Learning energy-based models by diffusion recovery likelihood. arXiv preprint arXiv:2012.08125 (2020) 4, 11

13. Gong, X., Chang, S., Jiang, Y., Wang, Z.: Autogan: Neural architecture search for generative adversarial networks. In: ICCV (2019) 10, 11
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014) 1
15. Han, T., Nijkamp, E., Zhou, L., Pang, B., Zhu, S.C., Wu, Y.N.: Joint training of variational auto-encoder and latent energy-based model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7978–7987 (2020) 4, 10, 11
16. He, H., Wang, H., Lee, G.H., Tian, Y.: Probgan: Towards probabilistic gan with theoretical guarantees. In: ICLR (2019) 11
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016) 9
18. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems. pp. 6626–6637 (2017) 9
19. Hill, M., Nijkamp, E., Mitchell, J.C., Pang, B., Zhu, S.C.: Learning probabilistic models from generator latent spaces with hat EBM. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022), https://openreview.net/forum?id=AluQNIIb_Zy 4, 10
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. arXiv preprint arXiv:2006.11239 (2020) 10
21. Huang, H., He, R., Sun, Z., Tan, T., et al.: Introvae: Introspective variational autoencoders for photographic image synthesis. Advances in neural information processing systems **31** (2018) 2
22. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. CoRR **abs/1502.03167** (2015), http://arxiv.org/abs/1502.03167 2
23. Jonathan, H., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020) 1, 10, 11
24. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2018) 11
25. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. arXiv preprint arXiv:2006.06676 (2020) 10
26. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. In: Advances in Neural Information Processing Systems (2022) 13
27. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. arXiv preprint arXiv:1807.03039 (2018) 10, 11
28. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: The International Conference on Learning Representations (ICLR) (2014) 1
29. Kingma, D.P., Welling, M.: Stochastic gradient vb and the variational auto-encoder. In: Second International Conference on Learning Representations, ICLR. vol. 19, p. 121 (2014) 3
30. Kong, Z., Chaudhuri, K.: The expressive power of a class of normalizing flow models. International Conference on Artificial Intelligence and Statistics,International Conference on Artificial Intelligence and Statistics (May 2020) 1

31. Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., Lehtinen, J.: The role of imagenet classes in fréchet inception distance. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=4oXTQ6m_ws8 10
32. Lee, H., Jeong, J., Park, S., Shin, J.: Guiding energy-based models via contrastive latent variables. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=CZmHHj9MgkP 2, 9, 10, 11
33. Lin, C.H., Chang, C.C., Chen, Y.S., Juan, D.C., Wei, W., Chen, H.T.: Coco-gan: generation by parts via conditional coordinating. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4512–4521 (2019) 11
34. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=PqvMRDCJT9t 11
35. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015) 9
36. Menon, S., Damian, A., Hu, S., Ravi, N., Rudin, C.: Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 2437–2445 (2020) 13
37. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: ICLR (2018) 10, 11
38. Nielsen, D., Jaini, P., Hoogeboom, E., Winther, O., Welling, M.: Survae flows: Surjections to bridge the gap between vaes and flows. In: NeurIPS (2020) 10, 12
39. van den Oord, A., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A., Kavukcuoglu, K.: Conditional image generation with pixelcnn decoders (2016) 11
40. Parimala, K., Channappayya, S.: Quality aware generative adversarial networks. In: Advances in Neural Information Processing Systems. pp. 2948–2958 (2019) 11
41. Parmar, G., Li, D., Lee, K., Tu, Z.: Dual contradistinctive generative autoencoder. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 823–832 (June 2021) 2, 10, 11
42. Pidhorskyi, S., Adjeroh, D.A., Doretto, G.: Adversarial latent autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14104–14113 (2020) 11
43. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023) 10
44. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: Advances in Neural Information Processing Systems. pp. 11918–11930 (2019) 10
45. Song, Y., Ermon, S.: Improved techniques for training score-based generative models. arXiv preprint arXiv:2006.09011 (2020) 11
46. Tomczak, J.M., Welling, M.: VAE with a vampprior. In: Storkey, A.J., Pérez-Cruz, F. (eds.) International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain. Proceedings of Machine Learning Research, vol. 84, pp. 1214–1223. PMLR (2018), http://proceedings.mlr.press/v84/tomczak18a.html 2
47. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7167–7176 (2017) 2
48. Vahdat, A., Kautz, J.: NVAE: A deep hierarchical variational autoencoder. In: Neural Information Processing Systems (NeurIPS) (2020) 2, 4, 10, 11, 12

49. Wang, W., Sun, Y., Halgamuge, S.: Improving mmd-gan training with repulsive loss function. In: ICLR (2019) 11
50. Wang, Y., Yu, J., Zhang, J.: Zero-shot image restoration using denoising diffusion null-space model. In: International Conference on Learning Representations (2023), https://openreview.net/forum?id=mRieQgMtNTQ 8, 13
51. Xiao, Z., Kreis, K., Kautz, J., Vahdat, A.: Vaebm: A symbiosis between variational autoencoders and energy-based models. In: International Conference on Learning Representations (2021) 4, 9, 10, 11, 12
52. Xiao, Z., Yan, Q., Chen, Y., Amit, Y.: Generative latent flow: A framework for non-adversarial image generation. CoRR **abs/1905.10485** (2019), http://arxiv.org/abs/1905.10485 14
53. Xie, J., Lu, Y., Gao, R., Wu, Y.N.: Cooperative learning of energy-based model and latent variable model via mcmc teaching. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018) 4
54. Xie, J., Lu, Y., Zhu, S.C., Wu, Y.: A theory of generative convnet. In: International Conference on Machine Learning. pp. 2635–2644 (2016) 1
55. Xie, J., Zheng, Z., Li, P.: Learning energy-based model with variational auto-encoder as amortized sampler. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 10441–10451 (2021) 4
56. Xie, J., Zhu, Y., Li, J., Li, P.: A tale of two flows: Cooperative learning of langevin flow and normalizing flow toward energy-based model. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=31d5RLCUuXC 4, 10, 11
57. Yang, G., Huang, X., Hao, Z., Liu, M.Y., Belongie, S., Hariharan, B.: Pointflow: 3d point cloud generation with continuous normalizing flows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4541–4550 (2019) 2
58. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015) 9
59. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 10
60. Zhao, Y., Xie, J., Li, P.: Learning energy-based generative models via coarse-to-fine expanding and sampling. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=aD1_5zowqV 11