Learning Energy-Based Models by Cooperative Diffusion Recovery Likelihood

Yaxuan Zhu[1], Jianwen Xie[2], Ying Nian Wu[1], Ruiqi Gao[3]

[1] Department of Statistics University of California (UCLA),
[2] Akool Research,
[3] Google DeepMind

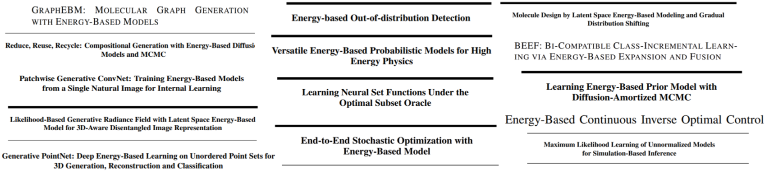
February 29, 2024

Introduction: Energy Based Model

- Energy-based models (EBMs) provide a powerful formulation for estimating the distribution of random variable $\mathbf{x}$

$$p_\theta(\mathbf{x}) = \frac{1}{Z_\theta} \exp(f_\theta(\mathbf{x})),$$

- EBMs have exhibited their flexibility and practicality in a variety of application scenarios.



GraphEBM: Molecular Graph Generation with Energy-Based Models

Reduce, Reuse, Recycle: Compositional Generation with Energy-Based Diffusion Models and MCMC

Patchwise Generative ConvNet: Training Energy-Based Models from a Single Natural Image for Internal Learning

Likelihood-Based Generative Radiance Field with Latent Space Energy-Based Model for 3D-Aware Disentangled Image Representation

Generative PointNet: Deep Energy-Based Learning on Unordered Point Sets for 3D Generation, Reconstruction and Classification

Energy-based Out-of-distribution Detection

Versatile Energy-Based Probabilistic Models for High Energy Physics

Learning Neural Set Functions Under the Optimal Subset Oracle

End-to-End Stochastic Optimization with Energy-Based Model

Molecule Design by Latent Space Energy-Based Modeling and Gradual Distribution Shifting

BEEF: Bi-Compatible Class-Incremental Learning via Energy-Based Expansion and Fusion

Learning Energy-Based Prior Model with Diffusion-Amortized MCMC

Energy-Based Continuous Inverse Optimal Control

Maximum Likelihood Learning of Unnormalized Models for Simulation-Based Inference

....

And many other interesting works

Introduction: Energy Based Model

Training energy-based models (EBMs) on high-dimensional data can be both challenging and time-consuming. There exists a noticeable gap in sample quality between EBMs and other generative frameworks like GANs and diffusion models.

Models	FID ↓
EBM based method	
NT-EBM (Nijkamp et al., 2022)	78.12
LP-EBM (Pang et al., 2020)	70.15
Adaptive CE (Xiao & Han, 2022)	65.01
EBM-SR (Nijkamp et al., 2019)	44.50
JEM (Grathwohl et al., 2020)	38.40
EBM-IG (Du & Mordatch, 2019)	38.20
EBM-FCE (Gao et al., 2020)	37.30
CoopVAEBM (Xie et al., 2021b)	36.20
CoopNets (Xie et al., 2018a)	33.61
Divergence Triangle (Han et al., 2020)	30.10
VARA (Grathwohl et al., 2021b)	27.50
EBM-CD (Du et al., 2021)	25.10
GEBM (Arbel et al., 2021)	19.31
HAT-EBM (Hill et al., 2022)	19.30
CF-EBM (Zhao et al., 2021)	16.71
CoopFlow (Xie et al., 2022)	15.80
CLEL-base (Lee et al., 2023)	15.27
VAEBM (Xiao et al., 2021)	12.16
DRL (Gao et al., 2021)	9.58
CLEL-large (Lee et al., 2023)	8.61
EGC (Unsupervised) (Guo et al., 2023)	5.36

Models	FID ↓
GAN based method	
WGAN-GP(Gulrajani et al., 2017)	36.40
SN-GAN (Miyato et al., 2018)	21.70
BigGAN (Brock et al., 2019)	14.80
StyleGAN2-DiffAugment (Zhao et al., 2020)	5.79
Diffusion-GAN (Xiao et al., 2022)	3.75
StyleGAN2-ADA (Karras et al., 2020)	2.92
Score based and Diffusion method	
NCSN (Song & Ermon, 2019)	25.32
NCSN-v2 (Song & Ermon, 2020)	10.87
NCSN++ (Song et al., 2021)	2.20
DDPM Distillation (Luhman & Luhman, 2021)	9.36
DDPM++(VP, NLL) (Kim et al., 2021)	3.45
DDPM (Ho et al., 2020)	3.17
DDPM++(VP, FID) (Kim et al., 2021)	2.47

FID Score for CIFAR-10 unconditional Generation

Challenges of Training EBM

- The training of EBM requires sampling from the current estimated distribution

$$\mathcal{L}'(\theta) = \mathbb{E}_{p_{\text{data}}}\left[\frac{\partial}{\partial\theta} f_\theta(\mathbf{x})\right] - \mathbb{E}_{p_\theta}\left[\frac{\partial}{\partial\theta} f_\theta(\mathbf{x})\right],$$

- Sampling is often achieved through MCMC, which can be challenging with high-dimensional data;
- Previous work [7] proposed to estimate a sequence of EBMs defined on increasingly noisy versions of the data and jointly estimate them by maximizing recovery likelihood.

Diffusion Recovery Likelihood (DRL)

- Assume a sequence of noisy training example perturbed by a Gaussian diffusion process: $\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_T$.
- $\mathbf{x}_{t+1} = \alpha_{t+1}\mathbf{x}_t + \sigma_{t+1}\boldsymbol{\epsilon}$, $\mathbf{x}_0 \sim p_{\text{data}}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$.
- Denote $\mathbf{y}_t = \alpha_{t+1}\mathbf{x}_t$ and fit $p_\theta(\mathbf{y}_t) = \frac{1}{Z_{\theta,t}} \exp(f_\theta(\mathbf{y}_t; t))$ for marginal distribution;
- The conditional distribution is given by

$$p_\theta(\mathbf{y}_t|\mathbf{x}_{t+1}) = \frac{1}{\tilde{Z}_{\theta,t}(\mathbf{x}_{t+1})} \exp\left(f_\theta(\mathbf{y}_t; t) - \frac{1}{2\sigma_{t+1}^2}\|\mathbf{y}_t - \mathbf{x}_{t+1}\|^2\right)$$

- Sample iteratively with Langevin dynamics:

$$\tilde{\mathbf{y}}_t^{\tau+1} = \tilde{\mathbf{y}}_t^{\tau} + \frac{s_t^2}{2}\left(\nabla_{\mathbf{y}} f_\theta(\tilde{\mathbf{y}}_t^{\tau}; t) - \frac{1}{\sigma_{t+1}^2}(\tilde{\mathbf{y}}_t^{\tau} - \mathbf{x}_{t+1})\right) + s_t\boldsymbol{\epsilon}^{\tau}$$

Cooperative Diffusion Recovery Likelihood (CDRL)

- Although DRL makes sampling easier by sampling from $p_\theta(\mathbf{y}_t|\mathbf{x}_{t+1})$, the initialization of MCMC sampling $\mathbf{x}_{t+1}$, may still be far from the data manifold of $\mathbf{y}_t$.
- We propose to learn an extra initializer model to further close this gap $q_\phi(\mathbf{y}_t|\mathbf{x}_{t+1}) \sim \mathcal{N}(\boldsymbol{g}_\phi(\mathbf{x}_{t+1}; t), \tilde{\sigma}_t^2 \boldsymbol{I})$.
- Cooperative Training:
  - Given observation $\mathbf{x}_{t+1}$, initializer q_ϕ makes the initial guess $\hat{\mathbf{y}}_y$.
  - $\hat{\mathbf{y}}_t$ is then modified by learned recovery-likelihood through Langevin sampling to get $\tilde{\mathbf{y}}_t$.
  - Update initialized using modified samples as target:

$$\mathcal{L}_t(\phi) = \frac{1}{n}\sum_{i=1}^{n}\left[-\frac{1}{2\tilde{\sigma}_t^2}\|\tilde{\mathbf{y}}_{t,i} - \boldsymbol{g}_\phi(\mathbf{x}_{t+1,i}; t)\|^2\right]$$

  - Update energy models using modified samples:

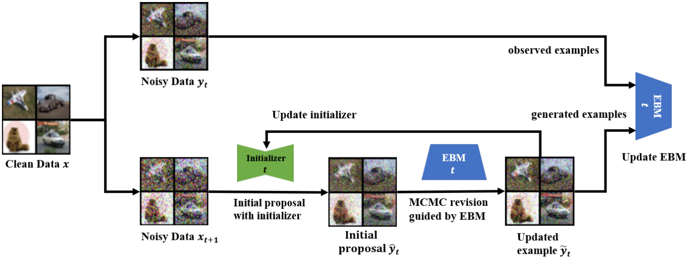
$$\nabla_\theta \mathcal{J}_t(\theta) = \nabla_\theta\left[\frac{1}{n}\sum_{i=1}^{n} f_\theta(\mathbf{y}_{t,i}; t) - \frac{1}{n}\sum_{i=1}^{n} f_\theta(\tilde{\mathbf{y}}_{t,i}; t)\right]$$

Training of CDRL

Algorithm 1 CDRL Training

Input: (1) observed data $\mathbf{x}_0 \sim p_{\text{data}}(\mathbf{x})$; (2) Number of noise levels T; (3) Number of Langevin sampling steps K per noise level; (4) Langevin step size at each noise level s_t; (5) Learning rate η_θ for EBM f_θ; (6) Learning rate η_ϕ for initializer g_ϕ;
Output: Parameters θ, ϕ

Randomly initialize θ and ϕ.
repeat
  Sample noise level t from $\{0, 1, ..., T-1\}$.
  Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$. Let $\mathbf{x}_{t+1} = \bar{\alpha}_{t+1}\mathbf{x}_0 + \bar{\sigma}_{t+1}\boldsymbol{\epsilon}$, $\mathbf{y}_t = \alpha_{t+1}(\bar{\alpha}_t\mathbf{x}_0 + \bar{\sigma}_t\boldsymbol{\epsilon})$.
  Generate the initial sample $\hat{\mathbf{y}}_t$.
  Generate the refined sample $\mathbf{y}_t$ by running K steps of Langevin dynamics starting from $\hat{\mathbf{y}}_t$.
  Update EBM parameter θ.
  Update initializer parameter ϕ.
until converged

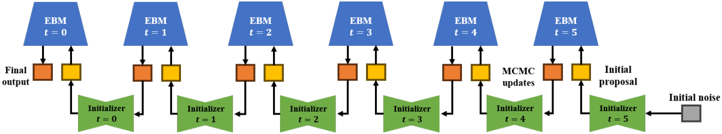


Sampling of CDRL

Algorithm 2 CDRL Sampling

Input: (1) Number of noise levels T; (2) Number of Langevin sampling steps K at each noise level; (3) Langevin step size at each noise level δ_t; (4) Trained EBM f_θ; (5) Trained initializer g_ϕ;
Output: Samples $\tilde{\mathbf{x}}_0$

Randomly initialize $\mathbf{x}_T \sim \mathcal{N}(0, \boldsymbol{I})$.
for $t = T - 1$ **to** 0 **do**
  Generate initial proposal $\hat{\mathbf{y}}_t$.
  Update $\hat{\mathbf{y}}_t$ to $\tilde{\mathbf{y}}_t$ by K iterations of Langevin Sampling.
  Let $\tilde{\mathbf{x}}_t = \tilde{\mathbf{y}}_t / \alpha_{t+1}$.
end for

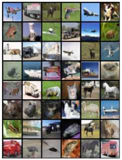


Unconditional Generation

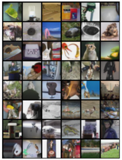
Models	FID ↓
EBM based method	
NT-EBM (Nijkamp et al., 2022)	78.12
LP-EBM (Pang et al., 2020)	70.15
Adaptive CE (Xiao & Han, 2022)	65.01
EBM-SR (Nijkamp et al., 2019)	44.50
JEM (Grathwohl et al., 2020)	38.40
EBM-IG (Du & Mordatch, 2019)	38.20
EBM-FCE (Gao et al., 2020)	37.30
CoopVAEBM (Xie et al., 2021b)	36.20
CoopNets (Xie et al., 2018a)	33.61
Divergence Triangle (Han et al., 2020)	30.10
VARA (Grathwohl et al., 2021b)	27.50
EBM-CD (Du et al., 2021)	25.10
GEBM (Arbel et al., 2021)	19.31
HAT-EBM (Hill et al., 2022)	19.30
CF-EBM (Zhao et al., 2021)	16.71
CoopFlow (Xie et al., 2022)	15.80
CLEL-base (Lee et al., 2023)	15.27
VAEBM (Xiao et al., 2021)	12.16
DRL (Gao et al., 2021)	9.58
CLEL-large (Lee et al., 2023)	8.61
EGC (Unsupervised) (Guo et al., 2023)	5.36
CDRL (Ours)	4.31
CDRL-large (Ours)	3.68

Models	FID ↓
Other likelihood based method	
VAE (Kingma & Welling, 2014)	78.41
PixelCNN (Salimans et al., 2017)	65.93
PixelIQN (Ostrovski et al., 2018)	49.46
Residual Flow (Chen et al., 2019)	47.37
Glow (Kingma & Dhariwal, 2018)	45.99
DC-VAE (Parmar et al., 2021)	17.90
GAN based method	
WGAN-GP(Gulrajani et al., 2017)	36.40
SN-GAN (Miyato et al., 2018)	21.70
BigGAN (Brock et al., 2019)	14.80
StyleGAN2-DiffAugment (Zhao et al., 2020)	5.79
Diffusion-GAN (Xiao et al., 2022)	3.75
StyleGAN2-ADA (Karras et al., 2020)	2.92
Score based and Diffusion method	
NCSN (Song & Ermon, 2019)	25.32
NCSN-v2 (Song & Ermon, 2020)	10.87
NCSN++ (Song et al., 2021)	2.20
DDPM Distillation (Luhman & Luhman, 2021)	9.36
DDPM++(VP, NLL) (Kim et al., 2021)	3.45
DDPM (Ho et al., 2020)	3.17
DDPM++(VP, FID) (Kim et al., 2021)	2.47

FID scores for CIFAR-10 Unconditional Generation



CIFAR-10 Samples



ImageNet (32 × 32) Samples

Models	FID ↓
EBM-IG (Du & Mordatch, 2019)	60.23
PixelCNN (Salimans et al., 2017)	40.51
EBM-CD (Du et al., 2021)	32.48
CF-EBM (Zhao et al., 2021)	26.31
CLEL-base (Lee et al., 2023)	22.16
DRL (Gao et al., 2021)	- (not converge)
DDPM++(VP, NLL) (Kim et al., 2021)	8.42
CDRL (Ours)	9.35

FID scores for ImageNet (32 × 32) Unconditional Generation

Sampling Acceleration

We applied post-training techniques to accelerate sampling.

- Langevin sampling step:

$$\tilde{\mathbf{y}}_t^{\tau+1} = \tilde{\mathbf{y}}_t^{\tau} + \frac{s_t^2}{2}\left(\nabla_{\mathbf{y}} f_\theta(\tilde{\mathbf{y}}_t^{\tau}; t) - \frac{1}{\sigma_{t+1}^2}(\tilde{\mathbf{y}}_t^{\tau} - \mathbf{x}_{t+1})\right) + s_t \boldsymbol{\epsilon}^{\tau},$$

- We decrease the number of sampling steps, and meanwhile adjust the MCMC sampling step size to be inversely proportional to the square root of the number of sampling steps.

Models	**Number of noise level $\times$ Number of MCMC steps**	**FID $\downarrow$**
DRL (Gao et al, 2021)	$6 \times 30 = 180$	9.58
CDRL	$6 \times 15 = 90$	4.31
CDRL (step 8)	$6 \times 8 = 48$	4.58
CDRL (step 5)	$6 \times 5 = 30$	5.37
CDRL (step 3)	$6 \times 3 = 18$	9.67

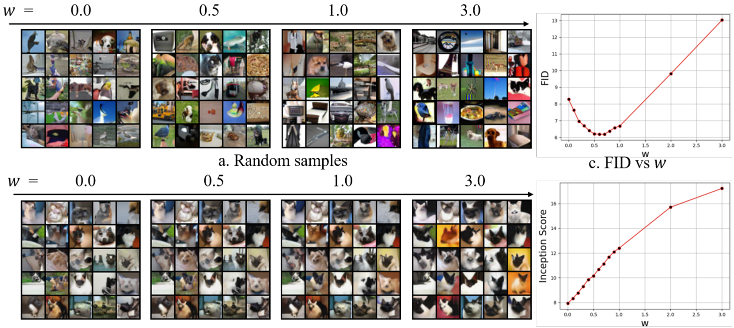
FID scores for CIFAR-10 with Sampling Adjustment

Conditional Generation

Following [15], we can apply classifier-free guidance to CDRL by estimating both conditional and unconditional distributions.

$$\log \tilde{p}_\theta(\mathbf{y}_t|c) = (w+1)f_\theta(\mathbf{y}_t; c, t) - wf_\theta(\mathbf{y}_t; t) + \text{const}.$$

$$\tilde{q}_\phi(\mathbf{y}_t|c, \mathbf{x}_{t+1}) \sim \mathcal{N}\left((w+1)\boldsymbol{g}_\phi(\mathbf{x}_{t+1}; c, t) - w\boldsymbol{g}_\phi(\mathbf{x}_{t+1}; t), \tilde{\sigma}_t^2\boldsymbol{I}\right).$$



a. Random samples

c. FID vs w

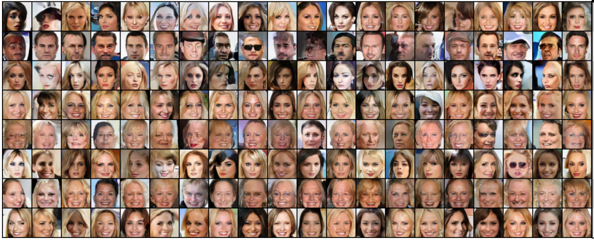
b. Samples with fixed noise

d. Inception Score vs w

Attribution-Compositional Generation

Similar to [5][20], given two energy functions trained on two conditionally independent concepts c_1 and c_2, we can estimate the energy conditioning on both concepts using:

$$\log p_\theta(\mathbf{x}|c_1, c_2) = \log p_\theta(\mathbf{x}|c_1) + \log p_\theta(\mathbf{x}|c_2) - \log p_\theta(\mathbf{x}) + \text{const.}$$



Male	Smile	Young
×	-	-
√	-	-
×	×	-
×	√	-
×	-	×
×	-	√
×	√	×
×	√	√

Out-Of-Distribution Detection

- The fitted energy function in CDRL naturally forms a score reflecting the log-likelihood of the data.
- We employ the model trained on CIFAR-10 as a detector and use the energy at the lowest noise level to serve as the OOD prediction score.

	Cifar-10 interpolation	Cifar-100	CelebA
PixelCNN (Salimans et al., 2017)	0.71	0.63	-
GLOW (Kingma & Dhariwal, 2018)	0.51	0.55	0.57
NVAE (Vahdat & Kautz, 2020)	0.64	0.56	0.68
EBM-IG (Du & Mordatch, 2019)	0.70	0.50	0.70
VAEBM (Xiao et al., 2021)	0.70	0.62	0.77
EBM-CD (Du et al., 2021)	0.65	0.83	-
CLEL-Base (Lee et al., 2023)	0.72	0.72	0.77
CDRL (ours)	0.75	0.78	0.84

AUROC scores in OOD detection using CDRL and other explicit density models on CIFAR-10

Conclusion and our contributions

This paper tries to push forward the progress of development of Energy Based Models.

- We propose cooperative diffusion recovery likelihood (CDRL) that tractably and efficiently learns and samples from a sequence of EBMs and MCMC initializers;
- Empirically we demonstrate that CDRL achieves significant improvements on sample quality compared to existing EBM approaches;
- We show that CDRL has great potential to enable more efficient sampling with sampling adjustment techniques;
- We demonstrate CDRL's ability in various tasks like unconditional generation, conditional generation, compositional generation, out-of-distribution (OOD) detection, etc.

Thank you for listening!

References I

[1] Michael Arbel, Liang Zhou, and Arthur Gretton. Generalized energy based models. In *The Ninth International Conference on Learning Representations, ICLR*, 2021.

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *The Seventh International Conference on Learning Representations, ICLR*, 2019.

[3] Ricky TQ Chen, Jens Behrmann, David Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. In *Advances in Neural Information Processing Systems*, 2019.

[4] Yilun Du, Shuang Li, Joshua B. Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy-based models. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, 2021.

[5] Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*, 2019.

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

[7] Ruiqi Gao, Yang Song, Ben Poole, Ying Nian Wu, and Diederik P. Kingma. Learning energy-based models by diffusion recovery likelihood. In *The Ninth International Conference on Learning Representations, ICLR*, 2021.

[8] Will Grathwohl, Kuan-Chieh Wang, Jörn-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 *The Eighth International Conference on Learning Representations, ICLR*, 2020.

[9] Will Sussman Grathwohl, Jacob Jin Kelly, Milad Hashemi, Mohammad Norouzi, Kevin Swersky, and David Duvenaud. No MCMC for me: Amortized sampling for fast and stable training of energy-based models. In *The Ninth International Conference on Learning Representations, ICLR*, 2021.

[10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *The Sixth International Conference on Learning Representations, ICLR*, 2017.

References II

[11] Qiushan Guo, Chuofan Ma, Yi Jiang, Zehuan Yuan, Yizhou Yu, and Ping Luo. Egc: Image generation and classification via a diffusion energy-based model. *arXiv preprint arXiv:2304.02012*, 2023.

[12] Tian Han, Erik Nijkamp, Linqi Zhou, Bo Pang, Song-Chun Zhu, and Ying Nian Wu. Joint training of variational auto-encoder and latent energy-based model. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, 2020.

[13] Mitch Hill, Erik Nijkamp, Jonathan Mitchell, Bo Pang, and Song-Chun Zhu. Learning probabilistic models from generator latent spaces with hat ebm. In *NeurIPS*, 2022.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

[17] Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation. In *International Conference on Machine Learning, ICML*, 2021.

[18] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, 2018.

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

[20] Hankook Lee, Jongheon Jeong, Sejun Park, and Jinwoo Shin. Guiding energy-based models via contrastive latent variables. In *The Eleventh International Conference on Learning Representations, ICLR*, 2023.

References III

[21] Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021.

[22] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *The Sixth International Conference on Learning Representations, ICLR*, 2018.

[23] Erik Nijkamp, Ruiqi Gao, Pavel Sountsov, Srinivas Vasudevan, Bo Pang, Song-Chun Zhu, and Ying Nian Wu. Learning energy-based model with flow-based backbone by neural transport mcmc. In *The Tenth International Conference on Learning Representations, ICLR*, 2022.

[24] 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*, 2019.

[25] Georg Ostrovski, Will Dabney, and Rémi Munos. Autoregressive quantile networks for generative modeling. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, 2018.

[26] Bo Pang, Tian Han, Erik Nijkamp, Song-Chun Zhu, and Ying Nian Wu. Learning latent space energy-based prior model. In *Advances in Neural Information Processing Systems*, 2020.

[27] Gaurav Parmar, Dacheng Li, Kwonjoon Lee, and Zhuowen Tu. Dual contradistinctive generative autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2021.

[28] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. In *The Fifth International Conference on Learning Representations, ICLR*, 2017.

[29] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 2019.

[30] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems*, 2020.

References IV

[31] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *The Ninth International Conference on Learning Representations, ICLR 2021*, 2021.

[32] Zhisheng Xiao and Tian Han. Adaptive multi-stage density ratio estimation for learning latent space energy-based model. In *NeurIPS*, 2022.

[33] Zhisheng Xiao, Karsten Kreis, Jan Kautz, and Arash Vahdat. Vaebm: A symbiosis between variational autoencoders and energy-based models. In *The Ninth International Conference on Learning Representations, ICLR*, 2021.

[34] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. In *The Tenth International Conference on Learning Representations, ICLR*, 2022.

[35] 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.

[36] Jianwen Xie, Zilong Zheng, and Ping Li. Learning energy-based model with variational auto-encoder as amortized sampler. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 10441–10451, 2021.

[37] Jianwen Xie, Yaxuan Zhu, Jun Li, and Ping Li. A tale of two flows: Cooperative learning of langevin flow and normalizing flow toward energy-based model. In *The Tenth International Conference on Learning Representations, ICLR*, 2022.

[38] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient GAN training. In *Advances in Neural Information Processing Systems*, 2020.

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