FRAME Revisited: An Interpretation View Based on Particle Evolution

Xu Cai,[1†] **Yang Wu,**[1†] **Guanbin Li,**[1] **Ziliang Chen,**[1] **Liang Lin**[1,2*]
[1]School of Data and Computer Science, Sun Yat-Sen University, China
[2]Dark Matter AI Inc.
caitree@foxmail.com, wuyang36@mail2.sysu.edu.cn,
liguanbin@mail.sysu.edu.cn, c.ziliang@yahoo.com, linliang@ieee.org

Abstract

FRAME (Filters, Random fields, And Maximum Entropy) is an energy-based descriptive model that synthesizes visual realism by capturing mutual patterns from structural input signals. The maximum likelihood estimation (MLE) is applied by default, yet conventionally causes the unstable training energy that wrecks the generated structures, which remains unexplained. In this paper, we provide a new theoretical insight to analyze FRAME, from a perspective of particle physics ascribing the weird phenomenon to KL-vanishing issue. In order to stabilize the energy dissipation, we propose an alternative Wasserstein distance in discrete time based on the conclusion that the Jordan-Kinderlehrer-Otto (JKO) discrete flow approximates KL discrete flow when the time step size tends to 0. Besides, this metric can still maintain the model's statistical consistency. Quantitative and qualitative experiments have been respectively conducted on several widely used datasets. The empirical studies have evidenced the effectiveness and superiority of our method.

Introduction

FRAME (Filters, Random fields, And Maximum Entropy) (Zhu, Wu, and Mumford 1997) is a model built on Markov random field that can be applied to approximate various types of data distributions, such as images, videos, audios and 3D shapes (Lu, Zhu, and Wu 2015; Xie, Zhu, and Wu 2017; Xie et al. 2018). It is an energy-based descriptive model in the sense that besides its parameters are estimated, samples can be synthesized from the probability distribution the model specifies. Such distribution is derived from maximum entropy principle (MEP), which is consistent with the statistical properties of the observed filter responses. FRAME can be trained via an information theoretical divergence between real data distribution $\mathbb{P}_r$ and model distribution P_θ. Primitive efforts model it as KL-divergence by default, which also leads to the same results of MLE.

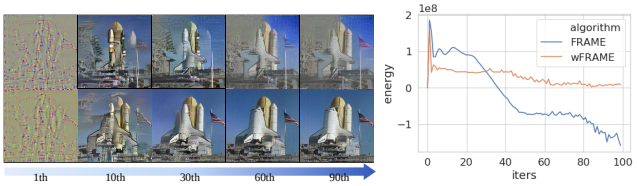

Figure 1: Visual and numerical results of FRAME and wFRAME. Left: the generating steps and selected typical results of "spaceship" from two algorithms. The first and the second-row images are respectively from FRAME and wFRAME. wFRAME achieves higher quality images compared with FRAME, which collapses at the very beginning of the sampling iteration. Right: the observed model energy of both algorithms. The instability of the energy curve is the signal of the model collapse. The detailed discussion can be found in the experiment section.

A large number of experimental results reveal that FRAME tends to generate inferior synthesized images and is often arduous to converge during training. For instance, displayed in Fig. 1, the synthesized images of FRAME seriously deteriorates along with the model energy. This phenomenon is caused by KL-vanishing in the stepwise parameters estimation of the model due to the existence of the great filter responses disparity between P_θ and $\mathbb{P}_r$. Specifically, the MLE-based learning algorithm attempts to optimize a transformation from the high dimensional support of P_θ to the non-existing support of $\mathbb{P}_r$, i.e., it starts from an initialization of a Gaussian noise covering the whole support of P_θ and $\mathbb{P}_r$, then gradually updates θ by calculating the KL discrete flow step-wisely. Therefore in the discrete time setting of the actual iterative training process, the dissipation of the model energy may become considerably unstable, and the stepwise minimization scheme may suffer serious KL-vanishing issue during the communicative parameters estimation.

To tackle the above shortcomings, we first investigate this model from a particle perspective by regarding all the observed signals as Brownian particles (pre-condition of KL

*Xu Cai and Yang Wu contribute equally to this work and share first-authorship. The corresponding author is Liang Lin (Email: linliang@ieee.org). This work was supported in part by the National Key Research and Development Program of China under Grant No.2018YFC0830103, in part by the NSFC-Shenzhen Robotics Projects (U1613211), in part by the National Natural Science Foundation of China under Grant No.61702565, No.61622214 and No.61836012 and in part by National High Level Talents Special Support Plan (Ten Thousand Talents Program).

discrete flow), which helps explore the reasons for the collapses of the FRAME model. This is inspired by the fact that the empirical measure of a set of Brownian particles generated by P_θ satisfies Large Deviation Principle (LDP) with rate functional coincides exactly with the KL discrete flow (see Lemma 1). We then delve into the model in discrete time state and translate its learning mechanism from KL discrete flow into the Jordan-Kinderlehrer-Otto (JKO) (Jordan, Kinderlehrer, and Otto 1998) discrete flow, which is a procedure for finding time-discrete approximations to solutions of diffusion equations in Wasserstein space. By resorting to the geometric distance between P_θ and $\mathbb{P}_r$ through optimal transport (OT) (Villani 2003) and replacing the KL-divergence with Wasserstein distance (a.k.a. the earth mover's distance (Rubner, Tomasi, and Guibas 2000)), this method manages to stabilize the energy dissipation scheme in FRAME and maintain its statistical consistency. The whole theoretical contribution can be summed up as the following deduction process:

- We deduce the learning process of data density in FRAME model from a view of particle evolution and confirm that it can be approximated by a discrete flow model with gradually decreasing energy driven by the minimization of the KL divergence.
- We further propose Wasserstein perspective of FRAME (wFRAME) by reformulating the FRAME's learning mechanism from KL discrete flow into the JKO discrete flow, of which the former theoretically explains the cause of the vanishing problem, while the latter overcomes the drawbacks, including the instability of sample generation and the failure of model convergence during training.

Qualitative and quantitative experiments demonstrate that the proposed wFRAME greatly ameliorates the vanishing issue of FRAME and can generate more visually promising results, especially for structurally complex training data. Moreover, to our knowledge, this method can be applied to most sampling processes which aim at abridging the KL-divergence between real data distribution and the generated data distribution by time sequence.

Related Work

Descriptive Model for Generation. The descriptive models originated from statistical physics have an explicit probability distribution of the signal, where they are ordinarily called the Gibbs distributions (Landau and Lifshitz 2013). With the massive developments of Convolutional Neural Networks (CNN) (Krizhevsky, Sutskever, and Hinton 2012) which has been proven to be a powerful discriminator, recently, increasing researches on the generative perspective of this model have drawn a lot of attention. (Dai, Lu, and Wu 2014) first introduces a generative gradient for pre-training discriminative ConvNet by a non-parametric importance sampling scheme and (Lu, Zhu, and Wu 2015) proposes to learn FRAME using pre-learned filters of modern CNN. (Xie et al. 2016b) further studies the theory of generative ConvNet intensively and show that the model has a representational structure which can be viewed as a hierarchical version of the FRAME model.

Implicit Model for Generation. Apart from the descriptive models, another popular branch of deep generative models is black-box models which map the latent variables to signals via a top-down CNN, such as the Generative Adversarial Network (GAN) (Goodfellow et al. 2014) and its variants. These models have gained remarkable success in generating realistic images and learn the generator network with an assistant discriminator network.

Relationship. Unlike the majority of implicit generative models, which use an auxiliary network to guide the training of the generator, descriptive models maintain a single model which simultaneously serves as a descriptor and generator, though FRAME can be served as an auxiliary and be combined with GAN to facilitate each other (Xie et al. 2016a). They factually generate samples directly from the input set, rather than from the latent space, which to a certain extent ensures that the model can be efficiently trained and produce stable synthesized results with relatively less model structure complexity. In this paper, FRAME and its variants as described above share the same MLE based learning mechanism, which follows an analysis-by-synthesis scheme and works by first generating synthesized samples from the current model using Langevin dynamics and then learn the parameters through observed-synthesized samples' distance.

Preliminaries

Let $\mathcal{P}$ denote the space of Borel probability measures on any given subset of space $\mathcal{X}$, where $\forall \boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{x} \in \mathbb{R}^d$. Given some sufficient statistics $\phi : \mathcal{X} \to \mathbb{R}$, scalar $\alpha \in \mathbb{R}$ and base measure q, the space of distributions satisfying linear constraint is defined as $\mathcal{P}_\alpha^{lin} = \left\{p, f \in \mathcal{P} : p = fq, f \geq 0, \int p dx = 1, E_p[\phi(x)] = \alpha\right\}$. The Wasserstein space of order $r \in [1, \infty]$ is defined as $\mathcal{P}_r = \left\{p \in \mathcal{P} : \int |x|^r dp < \infty\right\}$, where $|\cdot|^r$ denotes the r-norm on $\mathcal{X}$. $|\mathcal{X}|$ is the number of elements in domain $\mathcal{X}$. ∇ denotes gradient and $\boldsymbol{\nabla}\cdot$ denotes the divergence operator.

Markov Random Fields (MRF). MRF belongs to the family of undirected graphical models, which can be written in the Gibbs form as

$$P(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left\{\sum_{k=1}^{K} \theta_k f_k(\boldsymbol{x})\right\}, \tag{1}$$

where K stands for the number of features $\{f_k\}_{k=1}^K$ and $Z(\cdot)$ is the partition function (Koller and Friedman 2009). Its MLE learning process follows the iteration of the following two steps:

I. Update model parameter $\boldsymbol{\theta}$ by ascending the gradient of the log likelihood

$$\frac{\partial}{\partial \theta_k} \frac{1}{N} \log P(\boldsymbol{x}; \boldsymbol{\theta}) = \mathbb{E}_{\mathbb{P}_r}[f_k(\boldsymbol{x})] - \mathbb{E}_{P(\boldsymbol{x};\boldsymbol{\theta})}[f_k(\boldsymbol{x})], \tag{2}$$

where $\mathbb{E}_{\mathbb{P}_r}[f_k(\boldsymbol{x})]$ and $\mathbb{E}_{P(\boldsymbol{x};\boldsymbol{\theta})}[f_k(\boldsymbol{x})]$ is respectively the feature response over real data distribution $\mathbb{P}_r$ and current model distribution $P(\boldsymbol{x}; \boldsymbol{\theta})$.

II. Sample from the current model by parallel MCMC chains. The sampling process, according to (Younes 1989), does not necessarily converge at each $\boldsymbol{\theta}_t$, thus we only establish one persistent sampler that converges globally in order to reduce calculus.

FRAME Model. Based on an energy function, FRAME is defined on the exponential tilting of a reference distribution q, which is a reformulation of MRF and can be written as (Lu, Zhu, and Wu 2015):

$$P(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left\{ \sum_{k=1}^{K} \sum_{x\in\mathcal{X}} \theta_k h\left(\langle \boldsymbol{x}, \boldsymbol{w}\rangle + \boldsymbol{b}\right)_k \right\} q(\boldsymbol{x}), \tag{3}$$

where $h(\boldsymbol{x}) = \max(0, \boldsymbol{x})$ is the nonlinear activation function, $\langle \boldsymbol{x}, \boldsymbol{w}\rangle$ is the filtered image or feature map and $q(\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{|\mathcal{X}|/2}} \exp\left[-\frac{1}{2\sigma^2}\|\boldsymbol{x}\|^2\right]$ denotes the Gaussian white noise model with mean 0 and variance σ^2.

KL Discrete Flow. This flow is related to discrete probability distributions (evolutions discretized in time) with finite dimensional problems. More precisely, it indicates the system of n independent Brownian particles $\{\boldsymbol{x}^i\}_{i=1}^n \in \mathbb{R}^d$ whose position in $\mathbb{R}^d$ is given by a Wiener process satisfies the following stochastic differential equation (SDE)

$$d\boldsymbol{x}_t = \mu(\boldsymbol{x}_t)dt + \varepsilon(\boldsymbol{x}_t)d\boldsymbol{B}_t. \tag{4}$$

μ is the drift term, ε stands for the diffusion term, $\boldsymbol{B}$ denotes the Wiener process and subscript t denotes time point. The empirical measure of those particles is proved to approximate Eq. 3 by an implicit descent step $\rho^* = \operatorname{argmin}_\rho \mathcal{I}_t$, where $\mathcal{I}_t$ is the so called KL discrete flow consists of KL divergence and energy function $\Phi : \mathbb{R}^d \to \mathbb{R}$.

$$\mathcal{I}_t = \mathcal{K}(\rho \mid \rho_t) + \int \Phi d\rho. \tag{5}$$

Particle Perspective of FRAME Model

Although there is a traditional statistical perspective to interpret the FRAME theory (Xie et al. 2016b), we still need a more stable sampling process to avoid this frequent generation failure. We revisit the FRAME model from a completely new particle perspective and prove that its parameter update mechanism is actually equivalent to the reformulation of KL discrete flow. Its further transformation, a mechanism in JKO discrete flow manner which we will next prove the equivalence on condition of enough sampling time steps, has ameliorated this unpredictably vanishing phenomenon. All the proofs in detail are added to Appendix A.

Discrete Flow Driven by KL-divergence

Herein we first introduce FRAME in discrete flow manner. If we regard the observed signals $\{\boldsymbol{x}_t^i\}_{i=1}^n$ with the generating function of Markov property as Brownian particles, then Theorem 1 points out that Langevin dynamics can be deduced from KL discrete flow sufficiently and necessarily through Lemma 1.

Lemma 1. *For i.i.d. particles $\{\boldsymbol{x}_t^i\}_{i=1}^n$ with common generating function $\mathbb{E}[e^{\Phi(\boldsymbol{x};\boldsymbol{\theta})}]$ which has Markov property, the empirical measure $\rho_t = \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{x}_t^i}$ satisfies LDP with rate functional in the form of $\mathcal{I}_t$.*

Theorem 1. *Given a base measure q, a clique potential Φ, the density of FRAME in Eq. 3 can be obtained **sufficiently** and **necessarily** by solving the following constrained optimization.*

$$\begin{aligned} &\rho_{t+1} = \operatorname*{argmin}_{\rho} \mathcal{K}(\rho \mid \rho_t), \\ &s.t. \int \Phi d\rho = \int \Phi d\mathbb{P}_r, \quad \rho_0 = q, \quad \forall \rho \in \mathcal{P}_\alpha^{lin}. \end{aligned} \tag{6}$$

Let $\boldsymbol{\theta}$ be the Lagrange multiplier integrated in $\Phi(\boldsymbol{x};\boldsymbol{\theta})$ and ensure $\mathbb{E}[e^{\Phi(\boldsymbol{x};\boldsymbol{\theta})}] < \infty$, the optimizing objective can be reformulated as

$$\mathcal{I}_t^l = \min_{\rho}\max_{\boldsymbol{\theta}} \Big\{ \mathcal{K}(\rho \mid \rho_t) + \int \Phi(\boldsymbol{x};\boldsymbol{\theta}) d\rho - \int \Phi(\boldsymbol{x};\boldsymbol{\theta}) d\mathbb{P}_r \Big\}. \tag{7}$$

Since $\nabla_{\boldsymbol{x}} \log P(\boldsymbol{x};\boldsymbol{\theta}) = \nabla_{\boldsymbol{x}} \Phi(\boldsymbol{x};\boldsymbol{\theta})$, the SDE iteration of $\boldsymbol{x}_t$ in Eq. 4 can be expressed in the Langevin form as

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \nabla_{\boldsymbol{x}} \log P(\boldsymbol{x}_t;\boldsymbol{\theta}) + \sqrt{2}\boldsymbol{\xi}_t. \tag{8}$$

By Lemma 1, if we fix $\boldsymbol{\theta}$, the sampling scheme in Eq. 8 approaches the KL discrete flow $\mathcal{I}_t^l$, the flow will fluctuate in case $\boldsymbol{\theta}$ varies. $\boldsymbol{\theta}$ is updated by calculating $\nabla_\theta \mathcal{I}_t^l$, which implies $\boldsymbol{\theta}$ can dynamically transform the transition map into desired. The sampling process of FRAME can be summed up as

$$\begin{cases} \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \left(\dfrac{\boldsymbol{x}_t}{\sigma^2} - \nabla_{\boldsymbol{x}} \Phi(\boldsymbol{x}_t;\boldsymbol{\theta}) \right) + \sqrt{2}\boldsymbol{\xi}_t \\ \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \nabla_{\boldsymbol{\theta}} \mathbb{E}_{\rho_t}[\Phi(\boldsymbol{x};\boldsymbol{\theta})] - \nabla_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}_r}[\Phi(\boldsymbol{x};\boldsymbol{\theta})], \end{cases} \tag{9}$$

where $-\boldsymbol{x}_t/\sigma^2$ is the derivative of initial Gaussian noise q. If we take a close look at the objective function, there is an adversarial mechanism while updating $\boldsymbol{x}_t$ and $\boldsymbol{\theta}_t$. Regardless of fixing $\boldsymbol{\theta}$ updating $\boldsymbol{x}$, or fixing $\boldsymbol{x}$ updating $\boldsymbol{\theta}$, the correct direction cannot be insured to the optimal of minimizing $\mathcal{K}(P(\boldsymbol{x};\theta) \mid \mathbb{P}_r)$.

Discrete Flow Driven by Wasserstein Metric

Although KL approach is relatively rational in the methodology of FRAME, there exists a risk of the KL-vanishing problem as we have discussed, since the parameter updating mechanism of MLE may not converge. To avoid this problem, we introduce the Wasserstein metric to discrete flow, according to the statement of (Montavon, Müller, and Cuturi 2016) that P_θ can be closer from a KL method given empirical measure ρ_t, but further from the same measure in the Wasserstein distance. And (Arjovsky, Chintala, and Bottou 2017) also claims that a better convergence and approximated results can be obtained since Wasserstein metric defines a weaker topology. The conclusion that $\mathcal{I}_t \approx \mathcal{J}_t$ when time step size $\tau \to 0$ rationalizes the proposed method. The proof of this conclusion in the one-dimensional situation has shown in (Adams et al. 2011) and in higher-dimensional has been proved by (Duong, Laschos, and Renger 2013; Erbar et al. 2015). Here we first provide some background knowledge about the transformation then we briefly show the derivation process.

Fokker-Planck Equation. Under the influence of drifts and random diffusions, this equation describes the evolution of the probability density of the particle velocity. Let F be an integral function and $\delta F/\delta\rho$ denote its Euler-Lagrange first variation, the equations are

$$\begin{cases} \partial_t \rho + \boldsymbol{\nabla} \cdot (\rho \nu) = 0 & \text{(Continuity equation)} \\ \nu = -\nabla \dfrac{\delta F}{\delta \rho} & \text{(Variational condition)} \\ \rho(\cdot, 0) = \rho_0 & \rho_0 \in L^1(\mathbb{R}^d), \rho_0 \geq 0. \end{cases} \tag{10}$$

Wasserstein Metric. The Benamou-Brenier form of this metric (Benamou and Brenier 2000) of order r involves solving a smoothy OT problem over any probabilities μ_1 and μ_2 in $\mathcal{P}_r$ using the continuity equation showed in Eq. 10 as follows, where ν belongs to the tangent space of the manifold governed by some potential and associated with curve ρ_t.

$$\begin{aligned} \mathcal{W}^r(\mu_1, \mu_2) := \min_{\rho_t \in \mathcal{P}_r} \{ \int_0^1 \int_{\mathbb{R}^d} |\nu_t|^r d\rho_t dt : \partial_t \rho_t \\ + \boldsymbol{\nabla} \cdot (\rho_t \cdot \nu_t) = 0 \mid \rho_0 = \mu_1, \rho_1 = \mu_2 \}. \end{aligned} \tag{11}$$

JKO Discrete Flow. Following the initial work (Jordan, Kinderlehrer, and Otto 1998), which shows how to recover Fokker-Planck diffusions of distributions in Eq. 10 when minimizing entropy functionals according to Wasserstein metric $\mathcal{W}^2$, the JKO discrete flow is applied by our method to replace the initial KL divergence with the entropic Wasserstein distance $\mathcal{W}^2 - H(\rho)$. The function of the flow is

$$\mathcal{J}_t = \frac{1}{2}\mathcal{W}^2(\rho, \rho_t) + \int \log \rho d\rho + \int \Phi d\rho. \tag{12}$$

Remark 1. *The initial Gaussian term q is left out for convenience to facilitate the derivation, otherwise, the entropy $-H(\rho) = \int \log \rho d\rho$ in Eq. 12 should be written as the relative entropy $\mathcal{K}(\rho \mid q)$.*

By Theorem 1, $\mathcal{J}_t$ instead of $\mathcal{I}_t$ can be calculated in approximation and its steady state will approach Eq. 3. Applying $\mathcal{J}_t$ in the manner of dissipation mechanism as a substitute of $\mathcal{I}_t$ allows regarding the diffusion Eq. 4 as the steepest descent of clique energy Φ and entropy $-H(P)$ w.r.t. Wasserstein metric. Solving such optimization problem using $\mathcal{W}$ is identical to solve the Monge-Kantorovich mass transference problem.

With Second Mean Value theorem for definite integrals, we can approximately recover the integral $\mathcal{W}^2$ by two randomly interpolated rectangles

$$\begin{aligned} &\mathcal{W}^2(\rho_{t_0}, \rho_{t_1}) := \inf_{\rho_t} \int_{t_0}^{t_1} \int_{\mathbb{R}^d} |\nabla \Phi|^2 d\rho_t dt \\ &\approx (\zeta - t_0) \int_{\mathbb{R}^d} |\nabla \Phi|^2 d\rho_{t_0} + (t_1 - \zeta) \int_{\mathbb{R}^d} |\nabla \Phi|^2 d\rho_{t_1} \\ &= -\beta \left((1-\gamma) \int_{\mathbb{R}^d} |\nabla \Phi|^2 d\rho_{t_0} + \gamma \int_{\mathbb{R}^d} |\nabla \Phi|^2 d\rho_{t_1} \right). \end{aligned} \tag{13}$$

where $\beta = t_1 - t_0$ parameterizes the time piece and $\gamma = \zeta/\beta$ $(0 \leq \gamma \leq 1)$ represents random interpolated parameter since ζ is random. With Eq. 13, the functional derivative of $\mathcal{W}^2(\rho_{t_0}, \rho_{t_1})$ w.r.t. ρ_{t_1} is then proportional to

$$\frac{\delta \mathcal{W}^2(\rho_{t_0}, \rho_{t_1})}{\delta \rho_{t_1}} \propto |\nabla \Phi|^2, \tag{14}$$

which is exactly the result of Proposition 8.5.6 in (Ambrosio, Gigli, and Savaré 2008). Assume Φ be at least twice differentiable and treat Eq. 14 as the variational condition in Eq. 10, then plug Eq. 14 into the continuity equation of Eq. 10, which turns into a modified Wasserstein gradient flow in Fokker-Planck form as follows

$$\partial_t \rho = \Delta \rho - \boldsymbol{\nabla} \cdot \big(\rho (\nabla \Phi - \nabla |\nabla \Phi(\boldsymbol{x})|^2) \big). \tag{15}$$

Then the corresponding SDE can be written in Euler-Maruyama form as

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \nabla \Phi(\boldsymbol{x}_t) - \nabla |\nabla \Phi(\boldsymbol{x}_t)|^2 + \sqrt{2}\boldsymbol{\xi}_t. \tag{16}$$

By Remark 1, if we reconsider the initial Gaussian term, the discrete flow of $\boldsymbol{x}_{t+1}$ in Eq. 16 should be added with $-\boldsymbol{x}_t/\sigma^2$.

Remark 2. *If Φ is the energy function defined in Eq. 3, then $\nabla |\nabla \Phi(\boldsymbol{x})|^2 = 0$.*

It's a direct result since $\Phi(\boldsymbol{x}, \boldsymbol{\theta})$ defined in FRAME only involves inner-product, ReLu (piecewise linear) and other linear operations, the second derivative is obviously 0. Therefore, both the time evolution of density ρ_t in Eq. 15 and sample $\boldsymbol{x}_t$ in Eq. 16 will respectively degenerate to Eq. 10 and Eq. 8. Thus the SDE of $\boldsymbol{x}_t$ remains default, i.e. Langevin form while the gradients of the model parameter $\boldsymbol{\theta}_t$ doesn't degenerate.

Alike to the parameterized KL flow $\mathcal{I}_t^l$ defined in Eq. 7, we propose a similar form in JKO manner. With Eq. 13 and Eq. 14, the **final optimization objective function** $\mathcal{J}_t^l$ can be formulated as

$$\begin{aligned} \mathcal{J}_t^l = &\min_{\rho} \max_{\boldsymbol{\theta}} \Big\{ -\frac{\beta}{2}(1-\gamma) \int_{\mathbb{R}^d} |\nabla_{\boldsymbol{x}} \Phi(\boldsymbol{x}; \boldsymbol{\theta})|^2 d\rho_t \\ &- \frac{\beta}{2}\gamma \int_{\mathbb{R}^d} |\nabla_{\boldsymbol{x}} \Phi(\boldsymbol{x}; \boldsymbol{\theta})|^2 d\rho + \int log \rho d\rho \\ &+ \int \Phi(\boldsymbol{x}; \boldsymbol{\theta}) d\rho - \int \Phi(\boldsymbol{x}; \boldsymbol{\theta}) d\mathbb{P}_r \Big\}. \end{aligned} \tag{17}$$

With all discussed above, the learning progress of wFRAME can be constructed by ascending the gradient of θ, i.e. $\nabla_{\boldsymbol{\theta}} \mathcal{J}_t^l$. The calculating steps in formulation are summarized in Eq. 18.

$$\begin{cases} \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \Big(\dfrac{\boldsymbol{x}_t}{\sigma^2} - \nabla_{\boldsymbol{x}} \Phi(\boldsymbol{x}_t; \boldsymbol{\theta}) \Big) + \sqrt{2}\boldsymbol{\xi}_t \\ \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \nabla_{\boldsymbol{\theta}} \mathbb{E}_{\rho_t}[\Phi(\boldsymbol{x}; \boldsymbol{\theta})] - \nabla_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}_r}[\Phi(\boldsymbol{x}; \boldsymbol{\theta})] \\ \qquad - \dfrac{\beta}{2}(1-\gamma) \nabla_{\boldsymbol{\theta}} \mathbb{E}_{\rho_{t-1}}[|\nabla_{\boldsymbol{x}} \Phi(\boldsymbol{x}; \boldsymbol{\theta})|^2] \\ \qquad - \dfrac{\beta}{2}\gamma \nabla_{\boldsymbol{\theta}} \mathbb{E}_{\rho_t}[|\nabla_{\boldsymbol{x}} \Phi(\boldsymbol{x}; \boldsymbol{\theta})|^2]. \end{cases} \tag{18}$$

The equation above indicates that the gradient of $\boldsymbol{\theta}$ in Wasserstein manner is being added with some soft gradient norm constraints between the last two iterations. Such

Algorithm 1 Persistent Learning and Synthesizing in Wasserstein FRAME

Input: Training data $\{\boldsymbol{y}^i, i = 1, ..., N\}$
Output: Synthesized data $\{\boldsymbol{x}^i, i = 1, ..., M\}$

Initialize $\boldsymbol{x}_0^i \leftarrow 0$
for $t = 1$ **to** T **do**
  $H^{obs} \leftarrow \frac{1}{N}\sum_i^N \nabla_{\boldsymbol{\theta}_t}\Phi(\boldsymbol{y}^i)$
  for $j = 1$ **to** L **do**
    $\mathcal{G} \leftarrow \nabla_{\boldsymbol{x}_{t\times L+j-1}}\Phi(\boldsymbol{x}_{t\times L+j-1})$
    $\mathcal{S} \leftarrow \frac{\boldsymbol{x}_{t\times L+j-1}}{\sigma^2}$
    Sample $\boldsymbol{\Sigma} \leftarrow \mathcal{N}(0, \sigma^2 \cdot \mathbf{I}_d)$
    $\boldsymbol{x}_{t\times L+j} \leftarrow \boldsymbol{x}_{t\times L+j-1} + \frac{\delta^2}{2}(\mathcal{G} - \mathcal{S}) + \delta\boldsymbol{\Sigma}$
  end for
  $H^{syn} \leftarrow \frac{1}{M}\sum_i^M \nabla_{\boldsymbol{\theta}_t}\Phi(\boldsymbol{x}^i_{(t+1)\times L})$
  $\mathcal{P}_t \leftarrow \frac{1}{M}\sum_i^M \nabla_{\boldsymbol{\theta}_t}|\nabla_{\boldsymbol{x}_{(t+1)\times L}}\Phi(\boldsymbol{x}^i_{(t+1)\times L})|^2$
  $\mathcal{P}_{t-1} \leftarrow \frac{1}{M}\sum_i^M \nabla_{\boldsymbol{\theta}_t}|\nabla_{\boldsymbol{x}_{t\times L}}\Phi(\boldsymbol{x}^i_{t\times L})|^2$
  Sample $\gamma \sim U[0, 1]$
  Update $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \lambda \cdot (H^{obs} - H^{syn}) - \frac{\beta}{2}\left((1-\gamma)\,\mathcal{P}_{t-1} + \gamma\mathcal{P}_t\right)$
end for

gradient norm has the following **advantages** compared with the original iteration process (Eq. 9).

First the norm serves as the constant speed geodesic connecting ρ_t with ρ_{t+1} in the manifold spanned by $P_{\boldsymbol{\theta}}$ and $\mathbb{P}_r$, which may provide a speedup on converge. Next, it can be interpreted as the soft anti-force against the original gradient and prevent the whole learning process from vanishing. Moreover, in experiments, we find it can preserve data inner structural information. The new learning and synthesizing process of wFRAME is summarized in Algorithm 1 in detail.

Experiments

In this section, we intensively compare our proposed method with FRAME from two aspects, one is the confirmatory experiment of model collapse under varied settings with respect to the baseline, the other is the quantitative and qualitative comparison of generated results on extensively used datasets. In the first stage, as expected, the proposed wFRAME is verified to be more robust in training and the synthesized images are of higher quality and fidelity in most circumstances; In the second stage, we evaluate both models on the whole datasets. We propose a new metric, response distance R, which measures the gap between the generated data distribution and the real data distribution.

Confirmation of Model Collapse

We recognize that under some circumstances FRAME will suffer serious model collapse issue. Due to MEP, the expected well-learned FRAME model $P^*_{\boldsymbol{\theta}}$ should achieve minimum $\mathcal{K}(P^*_{\boldsymbol{\theta}} \mid q)$, i.e. the minimum amount of transformations to the reference measure. But such minimization of KL divergence might be the unpredictable cause of the energy to 0, namely the learned model will degenerate to produce initial noise instead of the desired minimum modification. So, in case $\Phi(\boldsymbol{x}, \boldsymbol{\theta}) \leq 0$, the learned model intends to degenerate, the images synthesized from FRAME driven by KL divergence will collapse immediately and the quality may barely restore. Consequently, the best curve of Φ is slowly asymptotic to and slightly above 0.

To manifest the superiority of our method over FRAME compared with the baseline settings, we conduct the validation experiments on a subset of SUN dataset (Xiao et al. 2010) under different circumstances. Intuitively, a simple trick to the model collapse issue is to restrict $\boldsymbol{\theta}$ in a safe range, a.k.a. weight clipping. The experimental settings include respectively altering λ and δ to an insecure range, turning on or off the weight clipping and varying the inputs dimensions. The results are presented in Fig. 3, which shows the property of a more robust generation compared with the original strategy or FRAME with weight clipping trick.

Empirical Setup on Common Datasets

We apply wFRAME on several widely used datasets in the field of generative modeling. As for default experimental settings, $\sigma = 0.01$, $\beta = 60$, the number of learning iterations is set to $T = 100$, the step number L of Langevin sampling within each learning iteration is 50 and the batch size is $N = M = 9$. The implementation of $\Phi(x)$ in our method is the first 4 convolutional layers of a pre-learned VGG-16 (Simonyan and Zisserman 2014). Input shape varies by datasets and is specified following. The hyper-parameters appear in Algorithm 1 differs on each dataset in order to achieve the best results. As for FRAME we use default settings in (Lu, Zhu, and Wu 2015).

CelebA (Liu et al. 2015) and **LSUN-Bedroom** (Yu et al. 2015) images are cropped and resized to 64×64. we set $\lambda = 1e^{-3}$ in both datasets, $\delta = 0.2$ in CelebA and $\delta = 0.15$ in LSUN-Bedroom. The visualizations of two methods are exhibited in Fig. 2.

CIFAR-10 (Krizhevsky and Hinton 2009) includes various categories and we learn both algorithms conditioned

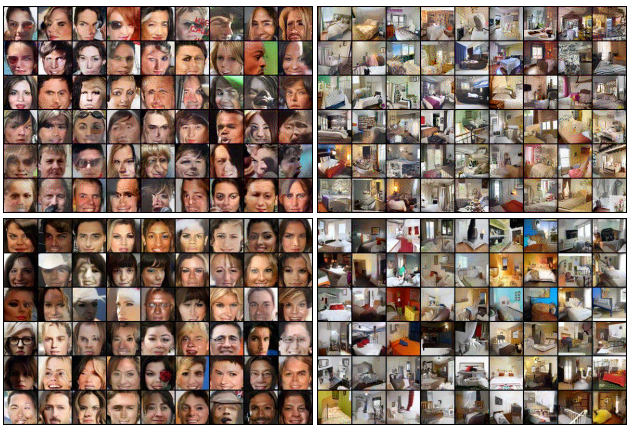

Figure 2: Comparison on LSUN-Bedroom and CelebA, where the first row is synthesized from FRAME the second is from wFRAME. More visual results have been added to the Appendix B.

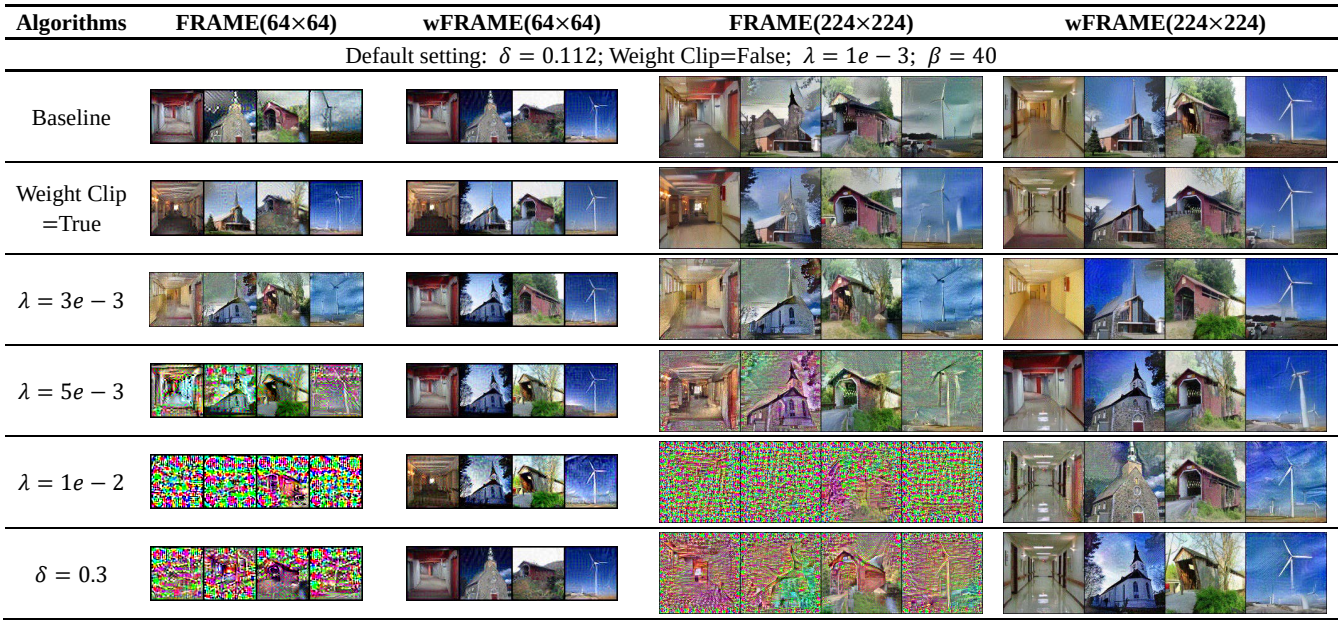

Figure 3: The synthesized results under different circumstances.

on the class label. In this experiment, we set $\delta = 0.15$, $\lambda = 2e^{-3}$ and images' size are of 32×32. Numerically and visually in Fig. 4, 5 and Table 1, the results show great improvement.

For a fair comparison, two metrics are utilized to evaluate FRAME and wFRAME. We offer a new metric response distance to measure the disparity between two distributions according to the results sampled out, while the Inception score is a widely used standard in measuring samples diversity.

Response distance R is defined as

$$R = \frac{1}{K}\sum_{k=1}^{K}\left|\frac{1}{N}\sum_{i=1}^{N}F_k(\boldsymbol{x}^i) - \frac{1}{M}\sum_{i=1}^{M}F_k(\boldsymbol{y}^i)\right|$$

where F_k denotes the kth filter. The smaller the R is, the better the generated results will be, since $R \propto max_\theta \mathbb{E}_r[F(\boldsymbol{y}^i)] - \mathbb{E}_{P_\theta}[F(\boldsymbol{x}^i)]$, which implies that R provides an approximation of the divergence between the target data distribution and the generated data distribution. Furthermore, by Eq. 2, the faster R falls the better $\boldsymbol{\theta}$ converges.

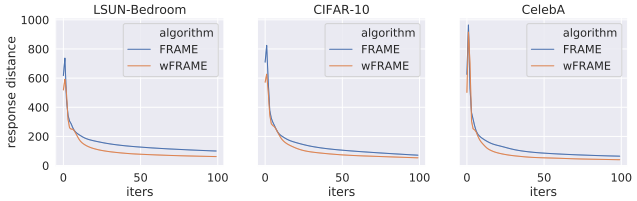

Figure 4: The averaged learning curves of response distance R on CelebA, LSUN-Bedroom and CIFAR-10.

Inception score (IS) is the most widely adopted metric of generative models, which estimates the diversity of the generated samples. It uses a network Inception v2 (Szegedy et al. 2016) pre-trained on ImageNet (Deng et al. 2009) to capture the classifiable properties of samples. This method has the drawbacks of neglecting the visual quality of the generated results and prefers models who generate objects rather than realistic scene images, but it can still provide essential diversity information of synthesized samples in evaluating generative models.

Model Type	Name	Inception Score
	Real Images	11.24±0.11
Implicit Models	DCGAN	**6.16±0.07**
	Improved GAN	4.36±0.05
	ALI	5.34±0.05
Descriptive Models	WINN-5CNNs	5.58±0.05
	FRAME (wl)	4.95±0.05
	FRAME	4.28±0.05
	wFRAME (ours,wl)	**6.05±0.13**
	wFRAME (ours)	5.52±0.13

Table 1: Inception score on datasets CIFAR-10 where 'wl' means training with labels. The IS result of ALI is reported in (Warde-Farley and Bengio 2016). IS of DCGAN is reported in (Wang and Liu 2016), and the result of Improved GAN(wl) is reported in (Salimans et al. 2016). WINN's is reported in (Lee et al. 2018). In the Descriptive Model plate, wFRAME outperforms the most methods.

Comparison with GANs

We compare FRAME and wFRAME with GAN models implemented on CIFAR-10 via the Inception score in Table 1. Most GAN-family models achieve pretty high on this score, however, our method is a descriptive model instead of an

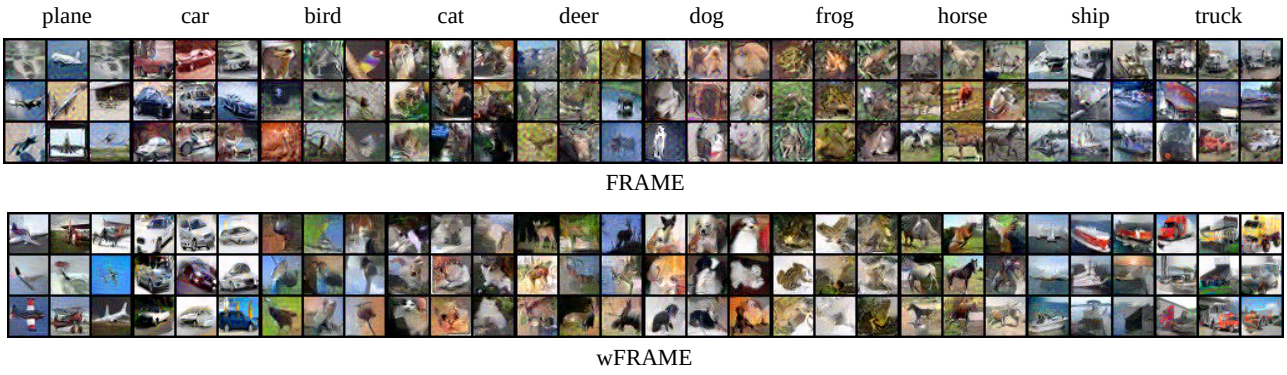

Figure 5: Images generated by two algorithms conditioned on labels in CIFAR-10, every three columns are of one class, the first group is from FRAME and the second is from wFRAME.

implicit model. GANs with high scores perform badly in descriptive situations, for example, the image reconstruction task or training on a small amount of data. FRAME can handle most of these situations properly. The performance of DCGAN in modeling mere few images is presented in Fig. 6 where for equal comparison, we duplicate the input images several times to the total amount of 10000 to adopt the training environment of DCGAN. The compared wFRAME is trained in our own method. The DCGAN's training procedure is ceased as it converges but still remains collapsed results.

Figure 6: The first left row is the selected input images from the SUN dataset, the right first row is the random outputs of DCGAN, the right last row is the outputs of our method.

Comparison of FRAME and wFRAME

From two aspects, we analyze FRAME and wFRAME as a summary of the whole experiments conducted above. As expected, our algorithm is more suitable for synthesizing complex and varied scene images and the resulting images are apparently more authentic compared with FRAME.

Quality of Generation Improvement. According to our performances on response distance R, the quality of the image synthesis is improved. This measurement is corresponding with the iteration learning process of both FRAME and wFRAME. The learning curves presented in Fig. 4 are the observations of the overall datasets synthesis. From the curves can we draw the conclusion that wFRAME converges better than FRAME. The results of generation on CelebA, LSUN-Bedroom and CIFAR-10 in Fig. 2 and 5 shows that even if the training images are relatively aligned with conspicuous structural information, or with only simple categorical context information, the images produced by FRAME are still abundant with motley noise and twisted texture, while ours are more reasonably mixed, more sensible structured and bright-colored with less distortion.

Training Steadiness Improvement. Compared with FRAME as shown in Fig. 1 which illustrates the typical evolution of generated samples, we found an improvement in the training steadiness. The generated images are almost identical at the beginning, however, images produced by our algorithm are able to be back on track after 30 iterations while FRAME's deteriorate. Quantitatively in Fig. 4, the curves are calculated by averaging across the whole dataset. wFRAME reaches lower cost on response distance, namely the direct L_1 critic of filter banks between synthesized samples and target samples is smaller and decreases more steadily. To be more specific, our algorithm has mostly solved the model collapse problem of FRAME for it not only ensures the closeness between the generated samples and "ground-truth" samples but also stabilizes the learning phase of the model parameter $\boldsymbol{\theta}$. The three plots clearly show the quantitative measures are well correlated with qualitative visualizations of generated samples. In the absence of collapsing, we attain comparable or even better results over FRAME.

Conclusion

In this paper, we re-derivatively track the origin of FRAME from the viewpoint of particle evolution and have discovered the potential factors that may lead to the deterioration of sample generation and the instability of model training, i.e, the inherent vanishing problem existing in the minimization of KL divergence. Based on this discovery, we propose wFRAME by reformulating the KL discrete flow in the FRAME to the JKO scheme, and prove through empirical examination that it can overcome the above-mentioned deficiencies. The experiments are carried out to demonstrate the superiority of the proposed wFRAME model and comparable results have shown that it can greatly ameliorate the vanishing issue of FRAME and can produce more visually

promising results.

References

Adams, S.; Dirr, N.; Peletier, M. A.; and Zimmer, J. 2011. From a large-deviations principle to the wasserstein gradient flow: a new micro-macro passage. *Communications in Mathematical Physics* 307(3):791–815.

Ambrosio, L.; Gigli, N.; and Savaré, G. 2008. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 214–223.

Benamou, J.-D., and Brenier, Y. 2000. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik* 84(3):375–393.

Dai, J.; Lu, Y.; and Wu, Y.-N. 2014. Generative modeling of convolutional neural networks. *arXiv preprint arXiv:1412.6296*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR), 2009. IEEE Conference on*, 248–255. IEEE.

Duong, M. H.; Laschos, V.; and Renger, M. 2013. Wasserstein gradient flows from large deviations of many-particle limits. *ESAIM: Control, Optimisation and Calculus of Variations* 19(4):1166–1188.

Erbar, Matthias anErbar, M.; Maas, J.; Renger, M.; et al. 2015. From large deviations to wasserstein gradient flows in multiple dimensions. *Electronic Communications in Probability* 20.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2672–2680.

Jordan, R.; Kinderlehrer, D.; and Otto, F. 1998. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis* 29(1):1–17.

Koller, D., and Friedman, N. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

Krizhevsky, A., and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical report, Citeseer.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Landau, L. D., and Lifshitz, E. M. 2013. *Course of theoretical physics*. Elsevier.

Lee, K.; Xu, W.; Fan, F.; and Tu, Z. 2018. Wasserstein introspective neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, 3730–3738.

Lu, Y.; Zhu, S.-C.; and Wu, Y. N. 2015. Learning frame models using cnn filters. *arXiv preprint arXiv:1509.08379*.

Montavon, G.; Müller, K.-R.; and Cuturi, M. 2016. Wasserstein training of restricted boltzmann machines. In *Advances in Neural Information Processing Systems*, 3718–3726.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision* 40(2):99–121.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2234–2242.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Villani, C. 2003. *Topics in optimal transportation*. Number 58. American Mathematical Soc.

Wang, D., and Liu, Q. 2016. Learning to draw samples: With application to amortized mle for generative adversarial learning. *arXiv preprint arXiv:1611.01722*.

Warde-Farley, D., and Bengio, Y. 2016. Improving generative adversarial networks with denoising feature matching.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, 3485–3492. IEEE.

Xie, J.; Lu, Y.; Zhu, S.-C.; and Wu, Y. N. 2016a. Cooperative training of descriptor and generator networks. *arXiv preprint arXiv:1609.09408*.

Xie, J.; Lu, Y.; Zhu, S.-C.; and Wu, Y. 2016b. A theory of generative convnet. In *International Conference on Machine Learning*, 2635–2644.

Xie, J.; Zheng, Z.; Gao, R.; Wang, W.; Zhu, S.-C.; and Wu, Y. N. 2018. Learning descriptor networks for 3d shape synthesis and analysis. *arXiv preprint arXiv:1804.00586*.

Xie, J.; Zhu, S.-C.; and Wu, Y. N. 2017. Synthesizing dynamic patterns by spatial-temporal generative convnet. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7093–7101.

Younes, L. 1989. Parametric inference for imperfectly observed gibbsian fields. *Probability Theory and Related Fields* 82(4):625–645.

Yu, F.; Zhang, Y.; Song, S.; Seff, A.; and Xiao, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Zhu, S. C.; Wu, Y. N.; and Mumford, D. 1997. Minimax entropy principle and its application to texture modeling. *Neural Computation* 9(8):1627–1660.