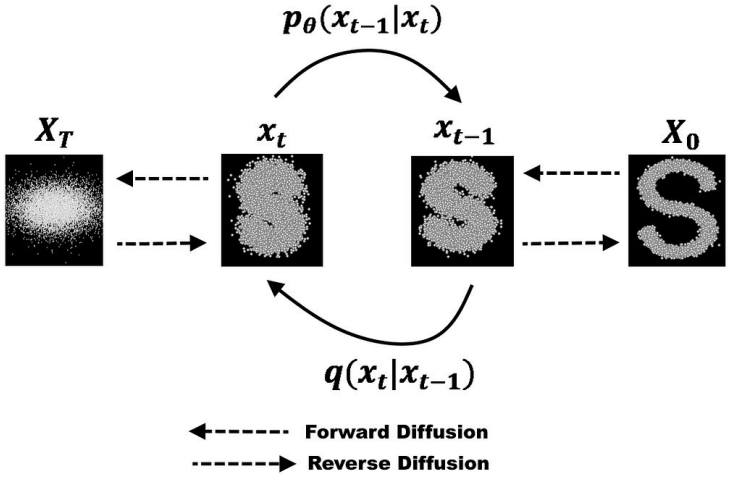
Consistency Models

Сорокин Дмитрий
БПМИ203

Что такое диффузионные модели?

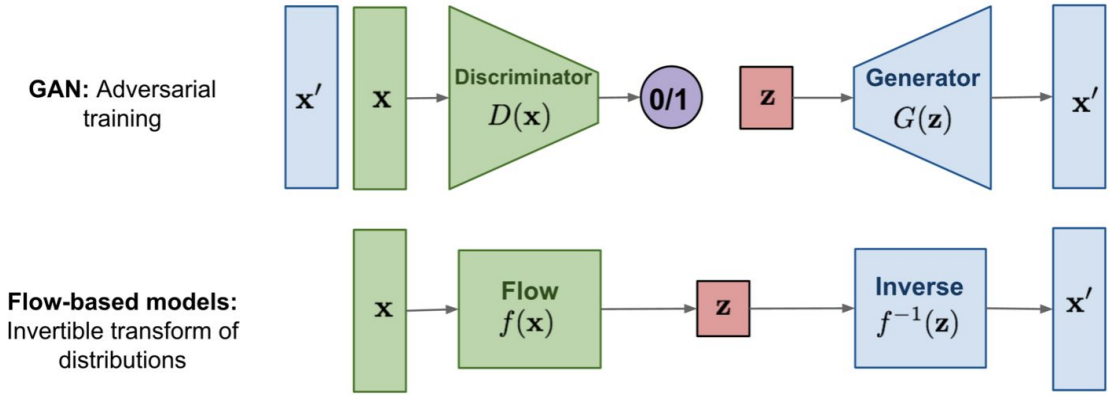
- Итеративно искажаем изображение, всего T итераций
- Получаем понятное распределение
- Обучаем модель удалять шум: получаем изображение шага t-1 из t
- Модель умеет определять шум добавленный на этапе (t-1) -> t

В итоге умеем превращать шум в изображение

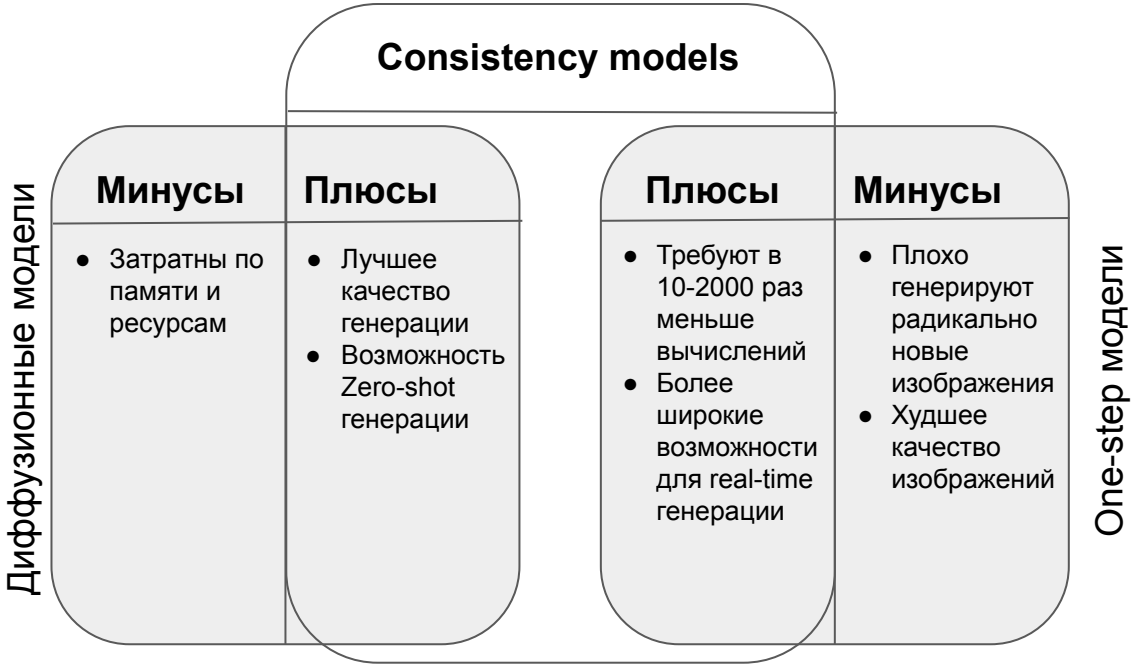


One-step генеративные модели

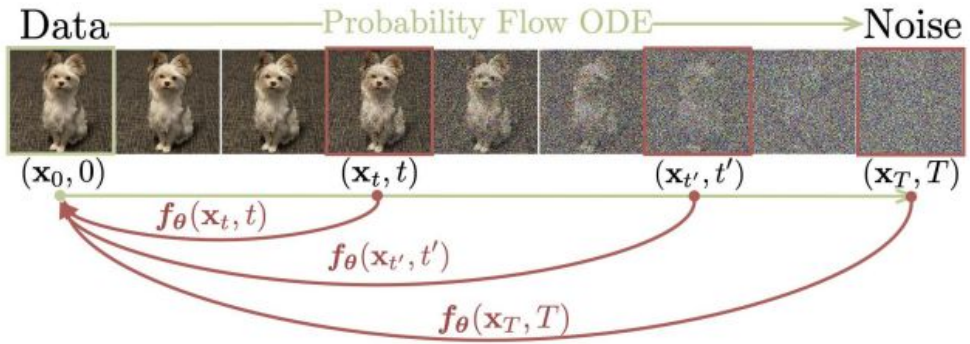
- Сразу обучаются находить исходное распределение (condition-sample)
- Экономим ресурсы, но жертвуем качеством



Consistency models мотивация



Важная особенность



Метод должен быть самосогласованным (self-consistent)

Аппроксимируем исходное распределение

Pretrained diffusion + ODE

$$\mathrm{d}\mathbf{x}_t = \boldsymbol{\mu}(\mathbf{x}_t, t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}\mathbf{w}_t,$$

$$\mathrm{d}\mathbf{x}_t = \left[\boldsymbol{\mu}(\mathbf{x}_t, t) - \frac{1}{2}\sigma(t)^2 \nabla \log p_t(\mathbf{x}_t)\right]\mathrm{d}t.$$

$$\boldsymbol{\mu}(\mathbf{x}, t) = \mathbf{0} \text{ and } \sigma(t) = \sqrt{2t}.$$

$$\boldsymbol{s}_{\boldsymbol{\phi}}(\mathbf{x}, t) \approx \nabla \log p_t(\mathbf{x})$$

$$\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = -t\boldsymbol{s}_{\boldsymbol{\phi}}(\mathbf{x}_t, t).$$

Идейно:

- С помощью SDE генерируем шум для какого то временного отрезка
- Начинаем решать ODE и аппроксимируем неизвестное распределение скор-функцией
- Вспоминаем, что решаем задачу с гауссовым распределением и сокращаем уравнение
- Генерируем шум и получаем траекторию data-noise
- Решаем ODE в обратном порядке и получаем аппроксимацию нашего распределения

Consistency model

- Определение

$\{\mathbf{x}_t\}_{t\in[\epsilon,T]}$ $\quad \boldsymbol{f} : (\mathbf{x}_t, t) \mapsto \mathbf{x}_\epsilon.$ $\qquad \boldsymbol{f}(\mathbf{x}_t, t) = \boldsymbol{f}(\mathbf{x}_{t'}, t') \text{ for all } t, t' \in [\epsilon, T].$

- Параметризация

1) $\boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x}, t) = \begin{cases} \mathbf{x} & t = \epsilon \\ F_{\boldsymbol{\theta}}(\mathbf{x}, t) & t \in (\epsilon, T] \end{cases}.$

2) $\boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x}, t) = c_{\text{skip}}(t)\mathbf{x} + c_{\text{out}}(t)F_{\boldsymbol{\theta}}(\mathbf{x}, t),$
*c - просто индикаторные функции

- Семплирование

Теперь можем сгенерировать случайный шум нормальным распределением N(0, T*T*I), применить функцию f(X, T) и получить желаемый результат

Consistency model

Такой подход позволяет применить “процедуру замены” подобно тому как это устроено для диффузионных моделей и решать zero-shot задачи!

Algorithm 1 Multistep Consistency Sampling

Input: Consistency model $\boldsymbol{f_\theta}(\cdot,\cdot)$, sequence of time points $\tau_1 > \tau_2 > \cdots > \tau_{N-1}$, initial noise $\hat{\mathbf{x}}_T$

$\mathbf{x} \leftarrow \boldsymbol{f_\theta}(\hat{\mathbf{x}}_T, T)$

for $n = 1$ **to** $N-1$ **do**

  Sample $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$

  $\hat{\mathbf{x}}_{\tau_n} \leftarrow \mathbf{x} + \sqrt{\tau_n^2 - \epsilon^2}\mathbf{z}$

  $\mathbf{x} \leftarrow \boldsymbol{f_\theta}(\hat{\mathbf{x}}_{\tau_n}, \tau_n)$

end for

Output: $\mathbf{x}$

Distillation / Isolation

Дистилляция:

- Берем предобученную скор функцию и следуем обычному пайплайну
- Аппроксимируем x’ на текущем шаге из предыдущих шагов
- Считаем x на текущем шаге
- Считаем лосс как dst(x, x’)
- Обновляем градиент
- Повторяем до сходимости

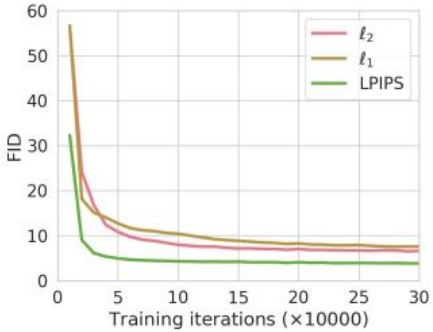
Изоляция:

- Можем оценить скор функцию следующей по одной из математических лемм

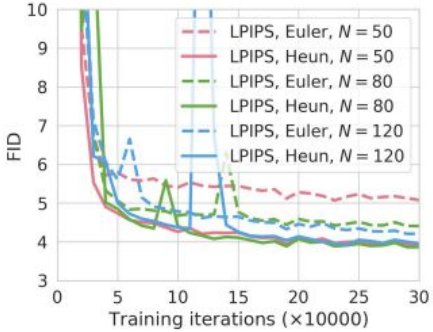
$$\nabla \log p_t(\mathbf{x}_t) = -\mathbb{E}\left[\frac{\mathbf{x}_t - \mathbf{x}}{t^2} \,\middle|\, \mathbf{x}_t\right],$$

- Теперь можем ничего не аппроксимировать а считать x’ напрямую, используя лишь консистентую функцию предыдущего шага
- Дальше все как при дистилляции
- Повторяем до сходимости

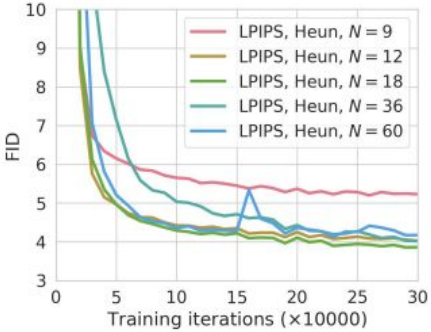
Эксперименты



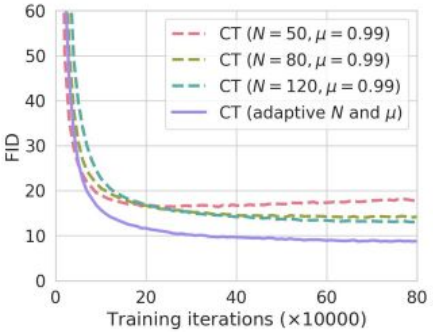
(a) Metric functions in CD.



(b) Solvers and N in CD.



(c) N with Heun solver in CD.



(d) Adaptive N and μ in CT.

Table 1: Sample quality on CIFAR-10. *Methods that require synthetic data construction for distillation.

METHOD	NFE (↓)	FID (↓)	IS (↑)
Diffusion + Samplers			
DDIM (Song et al., 2020)	50	4.67	
DDIM (Song et al., 2020)	20	6.84	
DDIM (Song et al., 2020)	10	8.23	
DPM-solver-2 (Lu et al., 2022)	10	5.94	
DPM-solver-fast (Lu et al., 2022)	10	4.70	
3-DEIS (Zhang & Chen, 2022)	10	**4.17**	
Diffusion + Distillation			
Knowledge Distillation* (Luhman & Luhman, 2021)	1	9.36	
DFNO* (Zheng et al., 2022)	1	4.12	
1-Rectified Flow (+distill)* (Liu et al., 2022)	1	6.18	9.08
2-Rectified Flow (+distill)* (Liu et al., 2022)	1	4.85	9.01
3-Rectified Flow (+distill)* (Liu et al., 2022)	1	5.21	8.79
PD (Salimans & Ho, 2022)	1	8.34	8.69
CD	1	**3.55**	**9.48**
PD (Salimans & Ho, 2022)	2	5.58	9.05
CD	2	**2.93**	**9.75**
Direct Generation			
BigGAN (Brock et al., 2019)	1	14.7	9.22
Diffusion GAN (Xiao et al., 2022)	1	14.6	8.93
AutoGAN (Gong et al., 2019)	1	12.4	8.55
E2GAN (Tian et al., 2020)	1	11.3	8.51
ViTGAN (Lee et al., 2021)	1	6.66	9.30
TransGAN (Jiang et al., 2021)	1	9.26	9.05
StyleGAN2-ADA (Karras et al., 2020)	1	2.92	**9.83**
StyleGAN-XL (Sauer et al., 2022)	1	**1.85**	
Score SDE (Song et al., 2021)	2000	2.20	**9.89**
DDPM (Ho et al., 2020)	1000	3.17	9.46
LSGM (Vahdat et al., 2021)	147	2.10	
PFGM (Xu et al., 2022)	110	2.35	9.68
EDM (Karras et al., 2022)	35	**2.04**	9.84
1-Rectified Flow (Liu et al., 2022)	1	378	1.13
Glow (Kingma & Dhariwal, 2018)	1	48.9	3.92
Residual Flow (Chen et al., 2019)	1	46.4	
GLFlow (Xiao et al., 2019)	1	44.6	
DenseFlow (Grcić et al., 2021)	1	34.9	
DC-VAE (Parmar et al., 2021)	1	17.9	8.20
CT	1	**8.70**	**8.49**
CT	2	**5.83**	**8.85**

Table 2: Sample quality on ImageNet 64 × 64, and LSUN Bedroom & Cat 256 × 256. †Distillation techniques.

METHOD	NFE (↓)	FID (↓)	Prec. (↑)	Rec. (↑)
ImageNet 64 × 64				
PD† (Salimans & Ho, 2022)	1	15.39	0.59	0.62
DFNO† (Zheng et al., 2022)	1	8.35		
CD†	1	6.20	0.68	0.63
PD† (Salimans & Ho, 2022)	2	8.95	0.63	**0.65**
CD†	2	**4.70**	**0.69**	0.64
ADM (Dhariwal & Nichol, 2021)	250	**2.07**	0.74	0.63
EDM (Karras et al., 2022)	79	2.44	0.71	**0.67**
BigGAN-deep (Brock et al., 2019)	1	4.06	**0.79**	0.48
CT	1	13.0	0.71	0.47
CT	2	11.1	0.69	0.56
LSUN Bedroom 256 × 256				
PD† (Salimans & Ho, 2022)	1	16.92	0.47	0.27
PD† (Salimans & Ho, 2022)	2	8.47	0.56	**0.39**
CD†	1	7.80	0.66	0.34
CD†	2	**5.22**	**0.68**	**0.39**
DDPM (Ho et al., 2020)	1000	4.89	0.60	0.45
ADM (Dhariwal & Nichol, 2021)	1000	**1.90**	0.66	**0.51**
EDM (Karras et al., 2022)	79	3.57	0.66	0.45
PGGAN (Karras et al., 2018)	1	8.34		
PG-SWGAN (Wu et al., 2019)	1	8.0		
TDPM (GAN) (Zheng et al., 2023)	1	5.24		
StyleGAN2 (Karras et al., 2020)	1	2.35	0.59	0.48
CT	1	16.0	0.60	0.17
CT	2	7.85	**0.68**	0.33
LSUN Cat 256 × 256				
PD† (Salimans & Ho, 2022)	1	29.6	0.51	0.25
PD† (Salimans & Ho, 2022)	2	15.5	0.59	0.36
CD†	1	11.0	0.65	0.36
CD†	2	**8.84**	**0.66**	**0.40**
DDPM (Ho et al., 2020)	1000	17.1	0.53	0.48
ADM (Dhariwal & Nichol, 2021)	1000	**5.57**	0.63	**0.52**
EDM (Karras et al., 2022)	79	6.69	**0.70**	0.43
PGGAN (Karras et al., 2018)	1	37.5		
StyleGAN2 (Karras et al., 2020)	1	7.25	0.58	0.43
CT	1	20.7	0.56	0.23
CT	2	11.7	0.63	0.36

- Среди дистилляций над диффузиями показывает лучший результат
- В основном лучше one-step методов, но проигрывает некоторым ганам
- Лучше дистилляций, но как отдельный метод отстает от конкурентов
- Результаты схожи с предыдущим датасетом
- Такая же картина

Zero-shot результаты

(a) *Left*: The gray-scale image. *Middle*: Colorized images. *Right*: The ground-truth image.

(b) *Left*: The downsampled image (32 × 32). *Middle*: Full resolution images (256 × 256). *Right*: The ground-truth image (256 × 256).

(c) *Left*: A stroke input provided by users. *Right*: Stroke-guided image generation.

Вывод

- Неплохо справляется с широким спектром задач
- Превосходит аналогичные методы
- Хуже генерирует изображения, чем ГАНы
- Может справляться с zero-shot задачами, но едва ли лучше чем диффузии