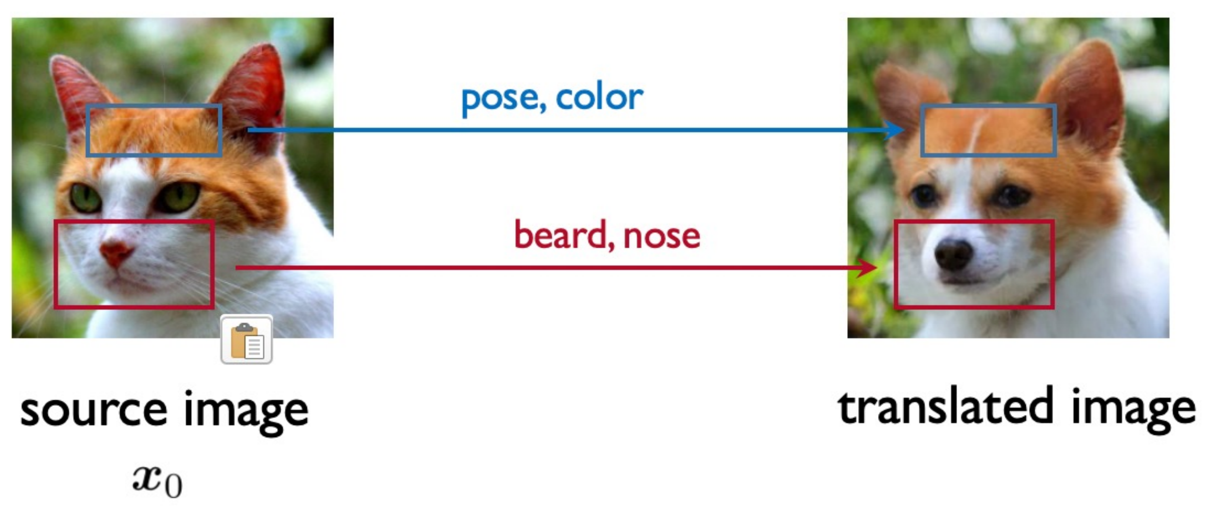
EGSDE: Unpaired Image-to-Image Translation via Energy-Guided Stochastic Differential Equations

https://github.com/ML-GSAI/EGSDE NeurIPS 2022

Tsinghua University, Renmin University of China

Min Zhao, Fan Bao, Chongxuan Li, Jun Zhu

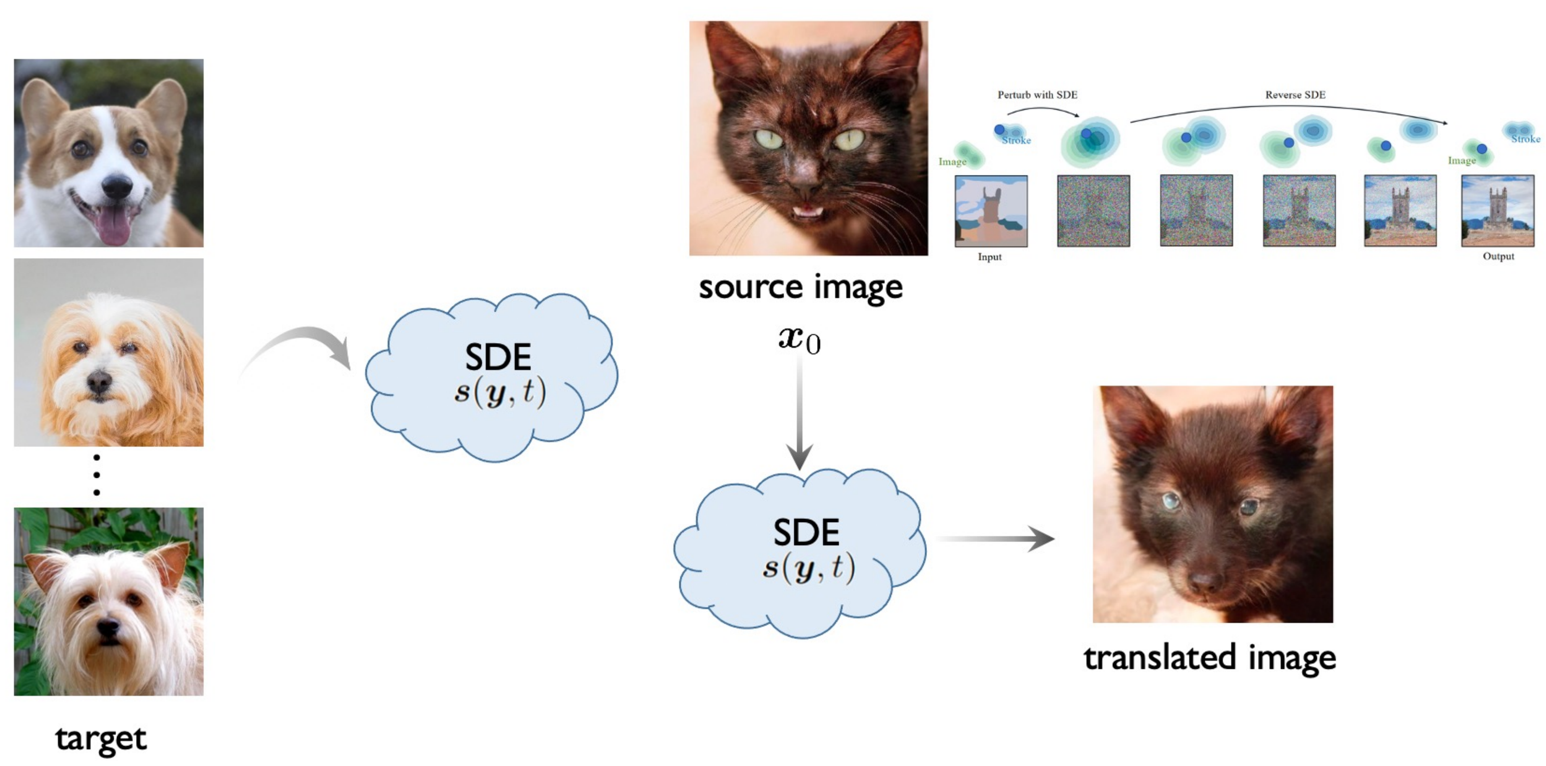
Motivation



➢ Be ***realistic*** for the target domain by changing the domain-specific features

➢ Be ***faithful*** for the source image by preserving the domain-independent features

- The goal of unpaired image-to-image translation (I2I)



- Existing methods trained a diffusion model solely on the target domain and exploited the test source image during inference. They ***did not leverage the training data in the source domain at all***.

➢ Recall the goal of I2I:

EGSDE as product of experts

Transition kernel

$$\tilde{p}(\boldsymbol{y}_t|\boldsymbol{x}_0) = \frac{p_{r1}(\boldsymbol{y}_t|\boldsymbol{x}_0)p_{r2}(\boldsymbol{y}_t|\boldsymbol{x}_0)p_f(\boldsymbol{y}_t|\boldsymbol{x}_0)}{Z_t} \longrightarrow \tilde{p}(\boldsymbol{y}_t|\boldsymbol{y}_s)$$

where $p_{r1}(\boldsymbol{y}_t|\boldsymbol{x}_0)$ is the marginal distribution defined by SDE conditioned on $\boldsymbol{x}_0$, $p_{r2}(\boldsymbol{y}_t|\boldsymbol{x}_0) \propto \exp(-\lambda_s \mathcal{E}_s(\boldsymbol{y}_t, \boldsymbol{x}_0, t))$, $p_f(\boldsymbol{y}_t|\boldsymbol{x}_0) \propto \exp(-\lambda_i \mathcal{E}_i(\boldsymbol{y}_t, \boldsymbol{x}_0, t))$.

Equivalent

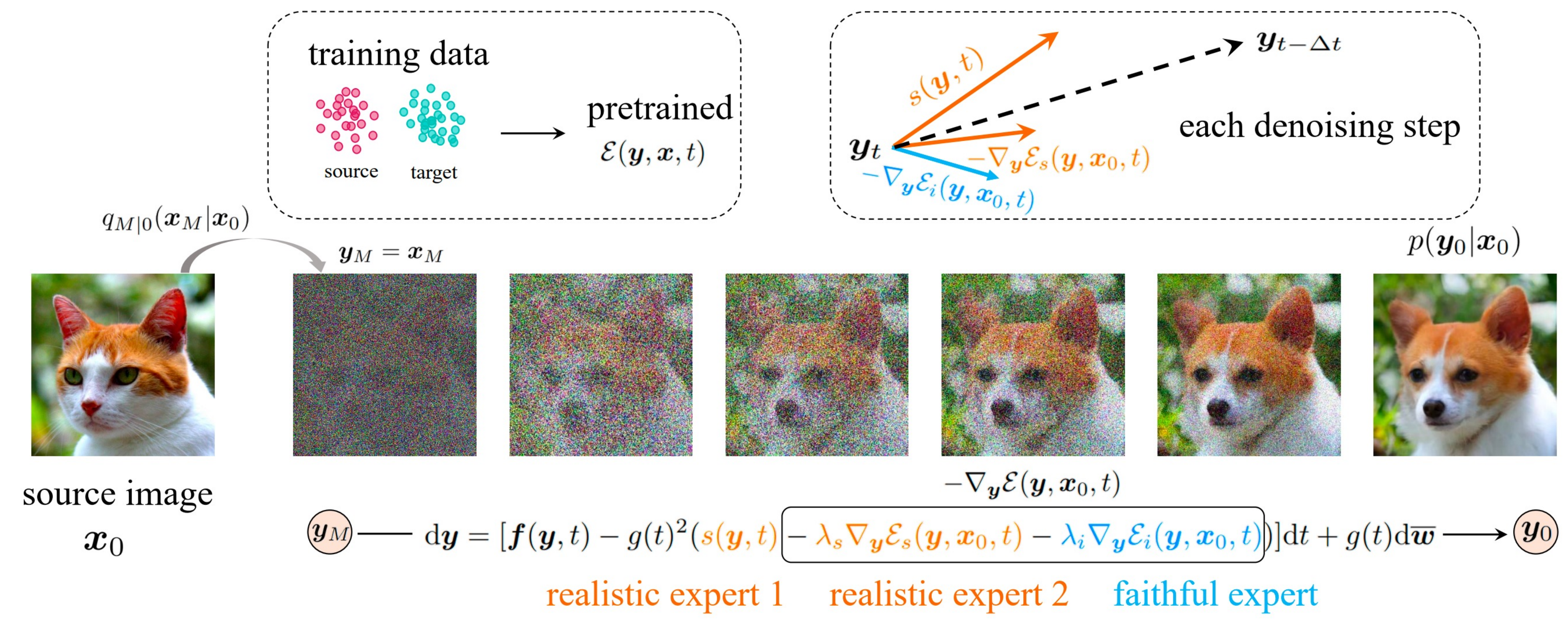
EGSDE

Euler-Maruyama solver

$$\mathrm{d}\boldsymbol{y} = [\boldsymbol{f}(\boldsymbol{y},t) - g(t)^2(s(\boldsymbol{y},t) - \lambda_s \nabla_{\boldsymbol{y}} \mathcal{E}_s(\boldsymbol{y}, \boldsymbol{x}_0, t) - \lambda_i \nabla_{\boldsymbol{y}} \mathcal{E}_i(\boldsymbol{y}, \boldsymbol{x}_0, t))]\mathrm{d}t + g(t)\mathrm{d}\overline{\boldsymbol{w}} \longrightarrow p(\boldsymbol{y}_t|\boldsymbol{y}_s)$$

Methods

- We propose energy-guided stochastic differential equations (EGSDE) that employs an energy function pretrained across the two domains to guide the inference process of a pretrained SDE for realistic and faithful unpaired I2I.



- Choice of Energy:

➢ Recall the goal of I2I:

Be ***realistic*** for the target domain by changing the domain-specific features
Be ***faithful*** for the source image by preserving the domain-independent features

➢ Decompose the energy function $\mathcal{E}(y, x, t)$ as the sum of two log potential functions:

$$\mathcal{E}(y,x,t) = \boxed{\lambda_s\ \mathcal{E}_s(y,x,t)} + \boxed{\lambda_i\ \mathcal{E}_i(y,x,t)}$$

$$= \lambda_s\ \mathbb{E}_{q_{t|0}(x_t|x)} S_s(y, x_t, t) - \lambda_i\ \mathbb{E}_{q_{t|0}(x_t|x)} S_i(y, x_t, t),$$

where $q_{t|0}(x_t|x)$ is the perturbation kernel from time 0 to time t in the forward SDE. $S_s(\cdot,\cdot,\cdot)$ and $S_i(\cdot,\cdot,\cdot)$ are two functions measuring the similarity between the sample and perturbed source image.

➢ Suppose $E_s(\cdot,\cdot) \in R^{C \times H \times W}$ is a domain-specific feature extractor, $S_s(\cdot,\cdot,\cdot)$ is defined as the cosine similarity between the features extracted from the generated sample and the source image :

$$S_s(y, x_t, t) = \cos(E_s(y,t), E_s(x_t, t))$$

➢ Suppose $E_i(\cdot,\cdot) \in R^{C \times H \times W}$ is a domain-independent feature extractor, $S_i(\cdot,\cdot,\cdot)$ is defined as the negative squared L2 distance between the features extracted from the generated sample and the source image :

$$S_i(y, x_t, t) = -\|E_i(y,t) - E_i(x_t, t)\|_2^2$$

Reference

[1] *Meng et al. Sdedit: Guided image synthesis and editing with stochastic differential equations*

[2] *Song et al. Score-based generative modeling through SDEs.*

Experiments

Table 1: Quantitative comparison

Model	FID ↓	L2 ↓	PSNR ↑	SSIM ↑	AMT ↑
		Cat → Dog			
CycleGAN* [55]	85.9	-	-	-	-
MUNIT* [17]	104.4	-	-	-	-
DRIT* [25]	123.4	-	-	-	-
Distance* [3]	155.3	-	-	-	-
SelfDistance* [3]	144.4	-	-	-	-
GCGAN* [10]	96.6	-	-	-	-
LSeSim* [53]	72.8	-	-	-	-
ITTR (CUT)* [54]	68.6	-	-	-	-
StarGAN v2 [8]	54.88 ± 1.01	133.65 ± 1.54	10.63 ± 0.10	0.27 ± 0.003	-
CUT* [35]	76.21	59.78	17.48	**0.601**	79.6%
ILVR [7]	74.37 ± 1.55	56.95 ± 0.14	17.77 ± 0.02	0.363 ± 0.001	75.4%
SDEdit [31]	74.17 ± 1.01	47.88 ± 0.06	19.19 ± 0.01	**0.423 ± 0.001**	65.2%
EGSDE	**65.82 ± 0.77**	**47.22 ± 0.08**	**19.31 ± 0.02**	0.415 ± 0.001	-
EGSDE†	**51.04 ± 0.37**	62.06 ± 0.10	17.17 ± 0.02	0.361 ± 0.001	-
		Wild → Dog			
CUT [35]	92.94	62.21	17.2	**0.592**	82.4%
ILVR [7]	75.33 ± 1.22	63.40 ± 0.15	16.85 ± 0.02	0.287 ± 0.001	73.4%
SDEdit [31]	68.51 ± 0.65	55.36 ± 0.05	17.98 ± 0.01	**0.343 ± 0.001**	57.2%
EGSDE	**59.75 ± 0.62**	**54.34 ± 0.08**	**18.14 ± 0.01**	**0.343 ± 0.001**	-
EGSDE†	**50.43± 0.52**	66.52± 0.09	16.40± 0.01	0.300± 0.001	-
		Male → Female			
CUT [35]	**31.94**	46.61	19.87	**0.74**	58.6%
ILVR [7]	46.12 ± 0.33	52.17 ± 0.10	18.59 ± 0.02	0.510 ± 0.001	88.2%
SDEdit [31]	49.43 ± 0.47	43.70 ± 0.03	20.03 ± 0.01	0.572 ± 0.000	74.4%
EGSDE	**41.93 ± 0.11**	**42.04 ± 0.03**	**20.35 ± 0.01**	**0.574 ± 0.000**	-
EGSDE†	**30.61 ± 0.19**	53.44 ± 0.09	18.32 ± 0.02	0.510 ± 0.001	-

(a)

(b)

(a) Ablation studies of energy function. (b) Rrepresentative translation results on three unpaired I2I tasks.