Pretrained Reversible Generation as Unsupervised Visual Representation Learning

ICCV 2025

Rongkun Xue[1*] Jinouwen Zhang[2*] Yazhe Niu[2,4] Dazhong Shen[5]

Bingqi Ma[4] Yu Liu[4] Jing Yang[1]

[1] Xi'an Jiaotong University [2]Shanghai AI Laboratory [3]SenseTime

[4]The Chinese University of Hong Kong

[5]Nanjing University of Aeronautics and Astronautics

Code Paper Wechat

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Pretrained Reversible Generation as Unsupervised Visual Representation Learning

ICCV OCT 19-23, 2025 HONOLULU HAWAII

SDE $dx = f(t, x_t)dt + g(t)dw_t$ ⟷

ODE $$\frac{dx_t}{dt} = v(x_t) = v(x_t, t)x_t - \frac{1}{2}g^2(t)\nabla_{x_t} \log p(x_t)$$

$$\frac{d\mathbf{a}(t)}{dt} = -\mathbf{a}(t)^\top \frac{\partial f(\mathbf{z}(t), t, \theta)}{\partial \mathbf{z}}$$

$$\frac{dL}{d\theta} = -\int_{t_1}^{t_0} \mathbf{a}(t)^\top \frac{\partial f(\mathbf{z}(t), t, \theta)}{\partial \theta} dt$$

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \varepsilon$$

$$x = \alpha_t \mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}_t = \mathbf{x}] + \sigma_t \mathbb{E}[\varepsilon \mid \mathbf{x}_t = \mathbf{x}]$$

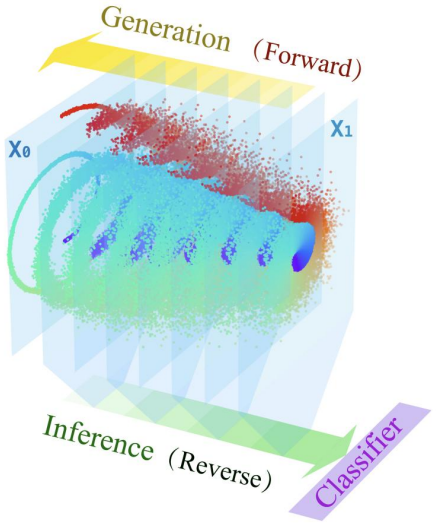
$$\mathbf{v}(\mathbf{x}, t) = \dot{\alpha}_t \mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}_t = \mathbf{x}] + \dot{\sigma}_t \mathbb{E}[\varepsilon \mid \mathbf{x}_t = \mathbf{x}]$$

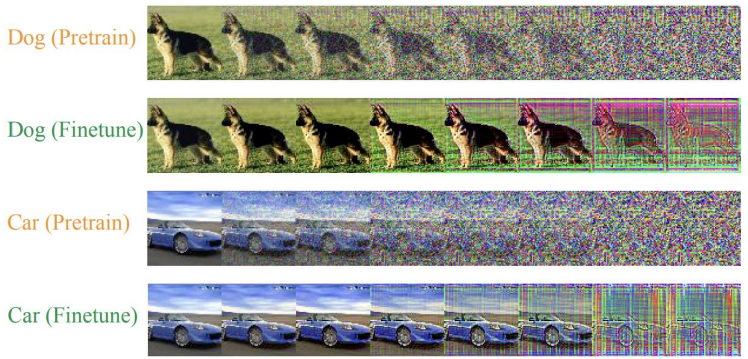
$$\mathbf{s}(\mathbf{x}, t) = -\sigma_t^{-1} \mathbb{E}[\varepsilon \mid \mathbf{x}_t = \mathbf{x}]$$

$$\mathbf{s}(\mathbf{x}, t) = \sigma_t^{-1} \frac{\alpha_t \mathbf{v}(\mathbf{x}, t) - \dot{\alpha}_t \mathbf{x}}{\dot{\alpha}_t \sigma_t - \alpha_t \dot{\sigma}_t} \longrightarrow \frac{\partial p(x_t)}{\partial t} + \nabla_x \cdot (p(x_t) v(x_t)) = 0$$

$$\mathcal{L}_s(\theta) = \int_0^T \mathbb{E}\left[\|\mathbf{v}_\theta(\mathbf{x}_t, t) - \dot{\alpha}_t \mathbf{x}_0 - \dot{\sigma}_t \varepsilon\|^2\right] \mathrm{d}t$$

$$\mathcal{L}_v(\theta) = \int_0^T \mathbb{E}\left[\|\mathbf{v}_\theta(\mathbf{x}_t, t) - (x_1 - x_0)\|^2\right] \mathrm{d}t$$



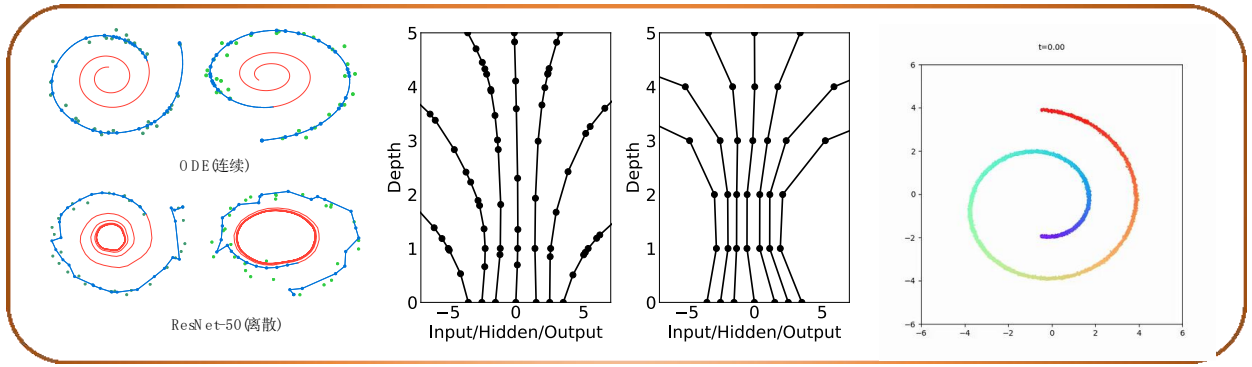


Motivation:

- As Richard Feynman famously said, "What I cannot create, I do not understand." Guided by this, PRG effectively reuses generative models, leveraging their high capacity to serve as robust and generalizable feature extractors for downstream tasks. This framework enables the flexible selection of feature hierarchies tailored to specific downstream tasks.

Advantage of PRG:

- Model-agnostic, infinitely expressive, and highly generalizable—enabling the extraction of strong understanding abilities from powerful community generative models.



Mechanism:

Flow matching raises the lower bound between the input and its optimal representation hidden within the generative model. The reverse trajectory preserves complete image information, while fine-tuning prunes irrelevant features and amplifies discriminative ones.

$$\theta^* = \arg\max_\theta \mathcal{J}(X,Z) = \arg\max_\theta \mathbb{E}_{p(z,x)}[\log p(x \mid z)]$$

$$\mathbb{E}_{p(z,x)}[\log p(x \mid z)] \geq \mathbb{E}_{p(z,x)}[\log p_{\theta'}(x \mid z)]$$

$$= -\mathbb{D}_{\mathrm{KL}}(p(x)\|p_{\theta'}(x \mid z = F_\theta(x))) - \mathcal{H}(X)$$

$$-\mathbb{D}_{\mathrm{KL}}(p(x_0)\|p_{\theta'}(x_0)) = -\mathbb{D}_{\mathrm{KL}}(p(x_1)\|p_{\theta'}(x_1)) - \mathcal{L}_{\mathrm{ODE}}$$

$$\geq \underbrace{-\mathbb{D}_{\mathrm{KL}}(p(x_1)\|p_{\theta'}(x_1))}_{=0} - \sqrt{\mathcal{L}_{\mathrm{SM}}}\sqrt{\mathcal{L}_{\mathrm{Fisher}}}$$

$$\mathcal{L}_{\mathrm{SM}} = \frac{1}{2}\mathbb{E}_{p(x_t)}\left[g^2(t)\|s_\theta(x) - \nabla \log p(x_t)\|^2\right]\mathrm{d}t$$

$$\mathcal{L}_{\mathrm{Fisher}} = \frac{1}{2}\mathbb{E}_{p(x_t)}\left[g^2(t)\|\nabla \log p_{\theta'}(x_t) - \nabla \log p(x_t)\|^2\right]\mathrm{d}t$$

Method	Param. (M)	Acc. (%)
Discriminative methods		
ViT-L/16 (384^2) [17]	307	76.5
ResNet-152 (224^2) [23]	60	77.8
Swin-B (224^2) [40]	88	83.5
Generative methods		
HybViT (32^2) [70]	43	53.5
DMSZC-DiTXL2 (256^2) [35]	338	77.5
iGPT-L (48^2) [10]	1362	72.6
Our methods		
PRG-GVP-onlyPretrain (64^2)	122	20.18
PRG-GVP-XL (64^2)	122	77.84
PRG-ICFM-XL (64^2)	122	78.12
PRG-OTCFM-XL (64^2)	122	78.13

Tiny-ImageNet-C			
Method	**Clean**	**Average**	**Corruption-5**
Adversarial Training			
PGD [44]	51.08	33.46 ↓ 17.62	24.00 ↓ 27.08
PLAT [32]	51.29	37.92 ↓ 13.37	29.05 ↓ 22.24
Noise Injection			
RSE [38]	53.74	27.99 ↓ 25.75	18.92 ↓ 34.82
ENResNet [66]	49.26	25.83 ↓ 23.43	19.01 ↓ 30.25
Data Augment			
AugMix [25]	52.82	37.74 ↓ 15.08	28.66 ↓ 24.16
AutoAug [15]	52.63	35.14 ↓ 17.49	25.36 ↓ 27.27
Generative Methods			
PDE+ [71]	53.72	39.41 ↓ 14.31	30.32 ↓ 23.40
PRG-ICFM-S (ours)	56.85	46.93 ↓ 9.92	33.32 ↓ 23.53

Method	Param. (M)	Acc. (%)
Discriminative methods		
WideResNet-28-10 [73]	36	69.3
Generative methods		
HybViT [70]	43	56.7
DDAE [68]	40	69.4
Our methods		
PRG-GVP-onlyPretrain	42	15.34
PRG-GVP-S	42	70.98
PRG-ICFM-S	42	71.12
PRG-OTCFM-S	42	71.33

Method	Param. (M)	Acc. (%)
Discriminative methods		
WideResNet-28-10 [73]	36	96.3
ResNeXt-29-16 * 64d [69]	68	96.4
Generative methods		
GLOW [19]	N/A	84.0
Energy model [18]	N/A	92.9
SBGC [77]	N/A	95.0
HybViT [70]	43	96.0
DDAE [68]	36	97.2
Our methods		
PRG-GVP-onlyPretrain	42	54.10
PRG-GVP-S	42	97.35
PRG-ICFM-S	42	97.59
PRG-OTCFM-S	42	97.65

Method	Param. (M)	Cifar Acc.	Tiny Acc.
Classic self-supervised learning methods			
MAE (ViT-B/16) [41]	86	96.5	76.5
SimCLR Res-50-4x [12]	375	98.6	N/A
Generative methods			
DDAE-DiT-XL2 [68]	338	98.4	77.8
iGPT-L [10]	1362	99.0	N/A
Our methods			
PRG-SiT-XL2	338	98.72	78.33

Dataset	CIFAR-10	Tiny-ImageNet	ImageNet
GPU	4×A100	8×A100	32×A100
Optimizer	AdamW	AdamW	AdamW
Eps	1e-8	1e-8	1e-8
Betas	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)
LR base	1.25e-4	1.25e-4	1.25e-4
Weight Decay	0.05	0.05	0.05
Scheduler	CosineLR	CosineLR	CosineLR
Warmup LR base	1.25e-7	1.25e-7	1.25e-7
Min LR base	1.25e-6	1.25e-6	1.25e-6
Epochs	200	200	200
Warmup Epochs	5	10	10
Image Size	32	64	64
Batch Size	256	128	128
T Span	20	32	64
Solver	Euler	Euler	Euler

Model	CIFAR-10	CIFAR-10-C	
	Clean	Corr Severity All	w/ Noise
Adversarial Training			
PGD *[44]	93.91	83.08 ↓ 10.83	82.10 ↓ 11.81
PLAT * [66]	94.73	88.28 ↓ 6.45	88.56↓ 6.17
Noise Injection			
RSE *[38]	95.59	77.86 ↓ 17.73	N/A
ENResNet *[66]	83.33	74.34 ↓ 8.99	N/A
Data Augment			
AugMix § [25]	95.83	89.09 ↓ 6.74	88.71 ↓ 7.12
AutoAug *[15]	95.61	85.37 ↓ 10.24	76.47↓ 19.14
Generative Methods			
SBGC [77]	95.04	76.24 ↓ 18.80	75.38↓ 19.66
PDE+ [71]	95.59	89.11 ↓ 6.48	85.59↓ 10.00
Our methods			
PRG-GVP-S	97.35	91.21 ↓ 6.14	92.17 ↓ 5.18
PRG-ICFM-S	97.59	92.13 ↓ 5.46	93.07 ↓ 4.52
PRG-OTCFM-S	97.65	92.26 ↓ 5.39	93.10 ↓ 4.55