Microsoft Research
微软亚洲研究院

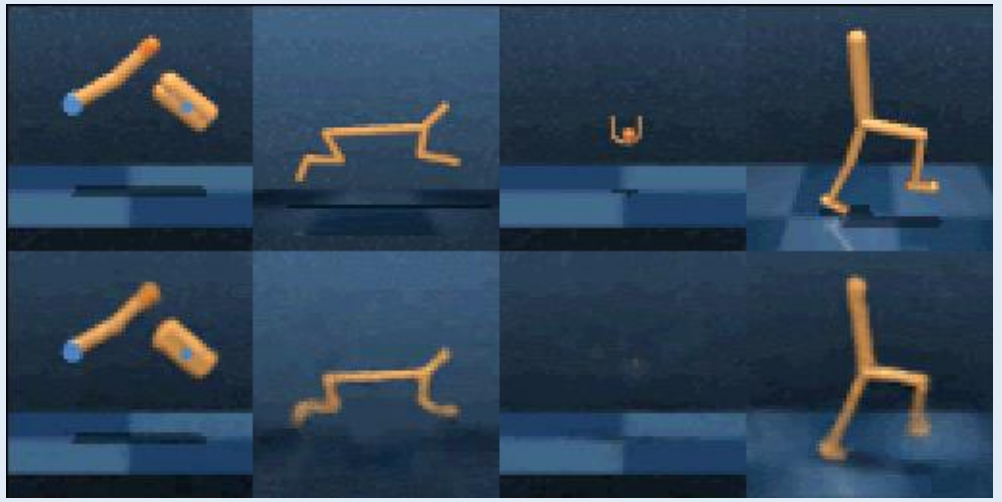
PlayVirtual: Augmenting Cycle-Consistent Virtual Trajectories for Reinforcement Learning

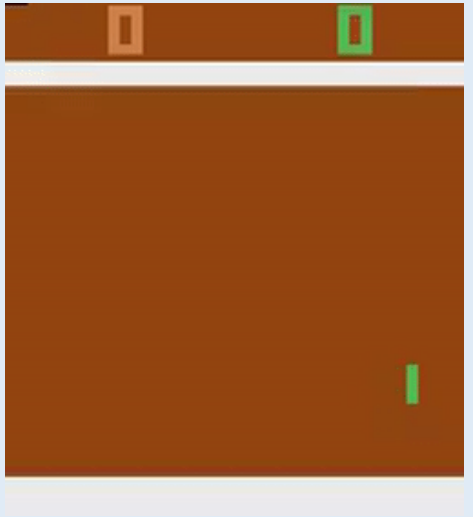
TAO YU[1], *CUILING LAN*[2], *WENJUN ZENG*[2], *MINGXIAO FENG*[1], *ZHIZHENG ZHANG*[2], *ZHIBO CHEN*[1]

[1]*UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA*, [2]*MICROSOFT RESEARCH ASIA*

NeurIPS 2021

Reinforcement Learning from Pixels





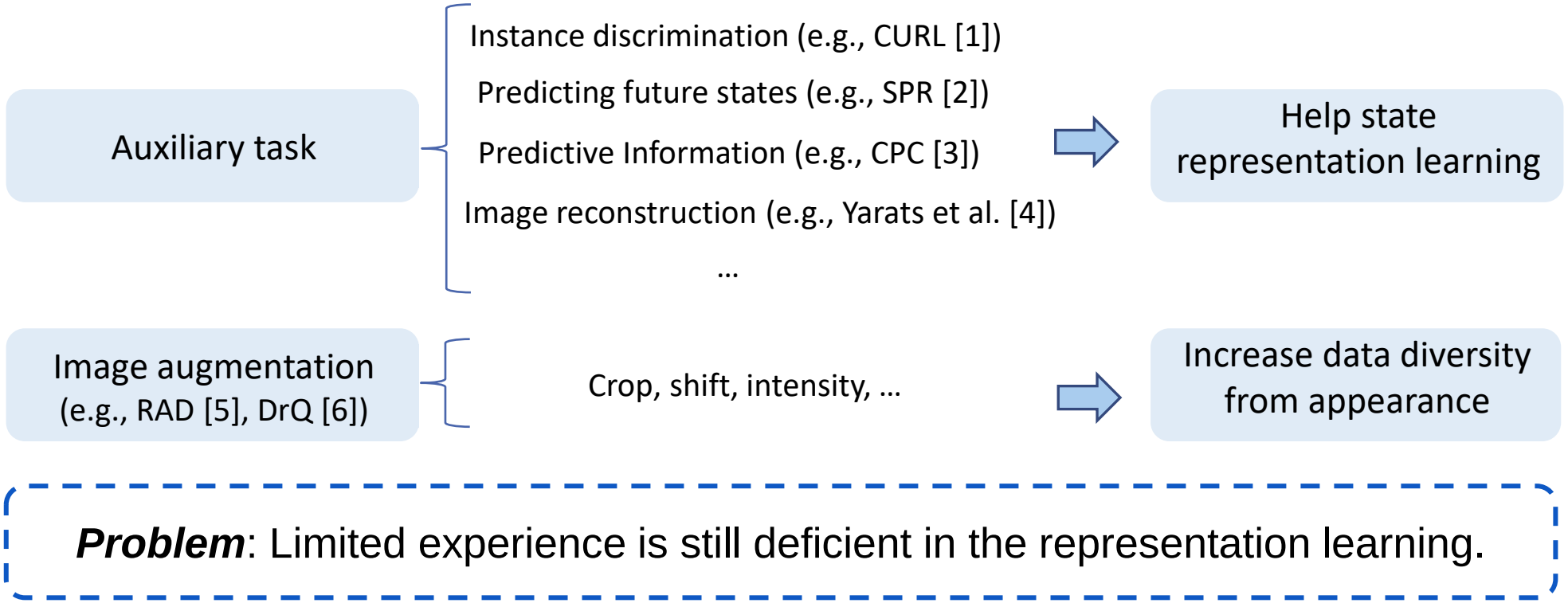
Challenge:

High dimensional observation

Limited interaction with the environment

→ Lower sample efficiency

Previous Sample-Efficient RL Methods

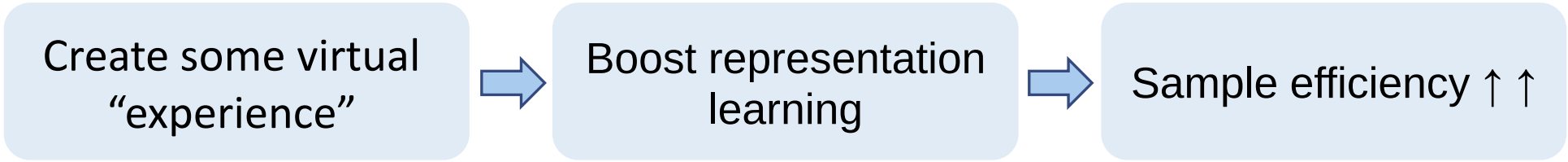


Problem: Limited experience is still deficient in the representation learning.

[1] *Laskin M, Srinivas A, Abbeel P. Curl: Contrastive unsupervised representations for reinforcement learning. ICML 2020.*
[2] *Schwarzer M, Anand A, Goel R, et al. Data-efficient reinforcement learning with self-predictive representations. ICLR 2021.*
[3] *Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv:1807.03748, 2018.*
[4] *Kostrikov D Y A Z I, Fergus B A J P R. Improving Sample Efficiency in Model-Free Reinforcement Learning from Images[J]. AAAI 2021.*
[5] *Laskin M, Lee K, Stooke A, et al. Reinforcement learning with augmented data. NeurIPS2020.*
[6] *Yarats D, Kostrikov I, Fergus R. Image Augmentation Is All You Need: Regularizing Deep Reinforcement Learning from Pixels. ICLR 2021.*

Our Idea

- Problem: limited experience
- Our idea:



Proposed Method

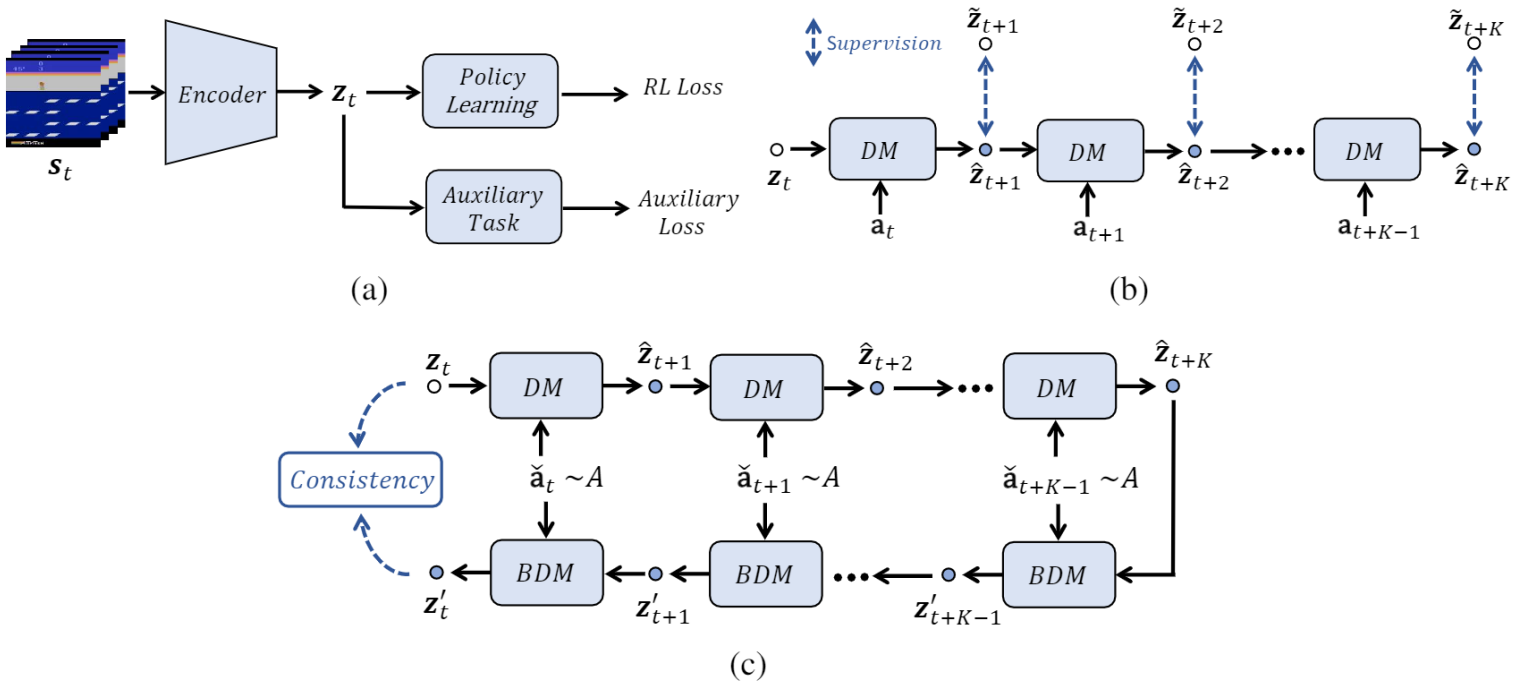


Illustration of the proposed ***PlayVirtual***.

Forward-backward with random sampled actions

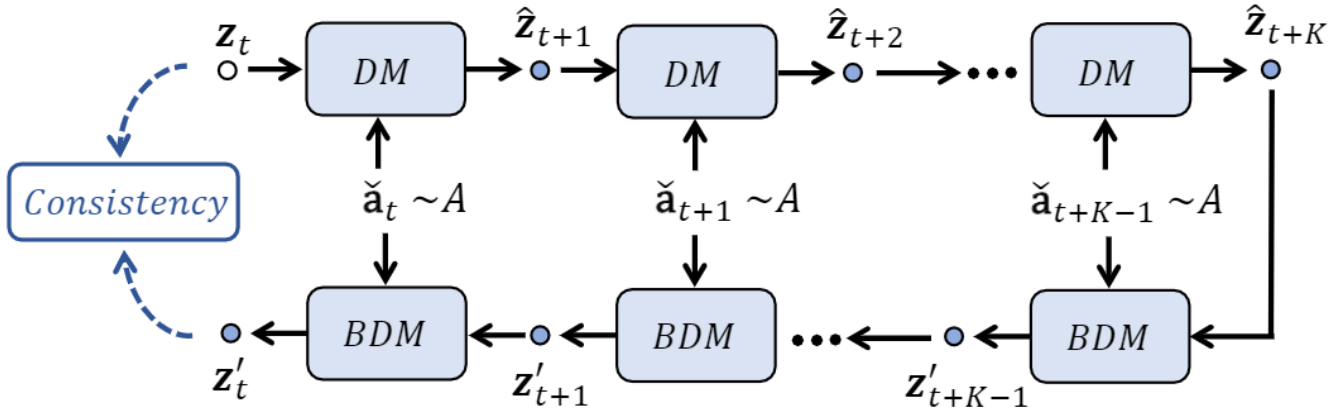
↓

Generate virtual trajectories

↓

Enrich "experience" & boost representation learning

Proposed Method



Our Cycle Consistency Definition: $\mathbb{E}_{\tau^c}[d_{\mathcal{M}}(\mathbf{z}'_t, \mathbf{z}_t)] = 0$

$$forward: \quad \hat{\mathbf{z}}_t = \mathbf{z}_t,\ \hat{\mathbf{z}}_{t+k+1} = h(\hat{\mathbf{z}}_{t+k}, \mathbf{a}_{t+k}),\ \ for\ k = 0, 1, \cdots, K-1,$$

$$backward: \quad \mathbf{z}'_{t+K} = \hat{\mathbf{z}}_{t+K},\ \mathbf{z}'_{t+k} = b(\mathbf{z}'_{t+k+1}, \mathbf{a}_{t+k}),\ \ for\ k = K-1, K-2, \cdots, 0.$$

Cycle loss: $\mathcal{L}_{cyc} = \frac{1}{M} \sum_{m=1}^{M} d_{\mathcal{M}}(\mathbf{z}'_t, \mathbf{z}_t)$

Proposed Method

Overall Objective:

$$\mathcal{L}_{total} = \mathcal{L}_{rl} + \lambda_{pred}\mathcal{L}_{pred} + \lambda_{cyc}\mathcal{L}_{cyc}$$

- **RL loss:** The loss of Rainbow [7] or SAC [8].
- **Prediction loss:** The loss of forward prediction using real trajectories like SPR [2].
- **Cycle loss:** Our proposed cycle loss using the virtual trajectories.

[2] *Schwarzer M, Anand A, Goel R, et al. Data-efficient reinforcement learning with self-predictive representations. ICLR 2021.*
[7] Hessel M, Modayil J, Van Haasselt H, et al. Rainbow: Combining improvements in deep reinforcement learning. AAAI 2018.
[8] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. ICML 2018.

Ablation Study

Table 1: Effectiveness of PlayVirtual.

Model	Atari-100k	DMControl-100k
Baseline w/o Pred	33.4	680.0
Baseline	37.1	728.0
Baseline+BDM	38.4	741.0
PlayVirtual	**47.2**	**797.0**

Comparison with the State-of-the-arts

Table 2: Scores achieved by different methods on discrete-control benchmark Atari-100k.

Game	Human	Random	SimPLe[61]	DER[45]	OTR[23]	CURL[29]	DrQ[50]	SPR[40]	PlayVirtual
Alien	7127.7	227.8	616.9	739.9	824.7	558.2	771.2	801.5	**947.8**
Amidar	1719.5	5.8	88.0	**188.6**	82.8	142.1	102.8	176.3	165.3
Assault	742.0	222.4	527.2	431.2	351.9	600.6	452.4	571.0	**702.3**
Asterix	8503.3	210.0	**1128.3**	470.8	628.5	734.5	603.5	977.8	933.3
Bank Heist	753.1	14.2	34.2	51.0	182.1	131.6	168.9	**380.9**	245.9
Battle Zone	37187.5	2360.0	5184.4	10124.6	4060.6	14870.0	12954.0	**16651.0**	13260.0
Boxing	12.1	0.1	9.1	0.2	2.5	1.2	6.0	35.8	**38.3**
Breakout	30.5	1.7	16.4	1.9	9.8	4.9	16.1	17.1	**20.6**
Chopper Command	7387.8	811.0	**1246.9**	861.8	1033.3	1058.5	780.3	974.8	922.4
Crazy Climber	35829.4	10780.5	**62583.6**	16185.3	21327.8	12146.5	20516.5	42923.6	23176.7
Demon Attack	1971.0	152.1	208.1	508.0	711.8	817.6	1113.4	545.2	**1131.7**
Freeway	29.6	0.0	20.3	**27.9**	25.0	26.7	9.8	24.4	16.1
Frostbite	4334.7	65.2	254.7	866.8	231.6	1181.3	331.1	1821.5	**1984.7**
Gopher	2412.5	257.6	771.0	349.5	**778.0**	669.3	636.3	715.2	684.3
Hero	30826.4	1027.0	2656.6	6857.0	6458.8	6279.3	3736.3	7019.2	**8597.5**
Jamesbond	302.8	29.0	125.3	301.6	112.3	**471.0**	236.0	365.4	394.7
Kangaroo	3035.0	52.0	323.1	779.3	605.4	872.5	940.6	**3276.4**	2384.7
Krull	2665.5	1598.0	**4539.9**	2851.5	3277.9	4229.6	4018.1	3688.9	3880.7
Kung Fu Master	22736.3	258.5	**17257.2**	14346.1	5722.2	14307.8	9111.0	13192.7	14259.0
Ms Pacman	6951.6	307.3	**1480.0**	1204.1	941.9	1465.5	960.5	1313.2	1335.4
Pong	14.6	-20.7	**12.8**	-19.3	1.3	-16.5	-8.5	-5.9	-3.0
Private Eye	69571.3	24.9	58.3	97.8	100.0	**218.4**	-13.6	124.0	93.9
Qbert	13455.0	163.9	1288.8	1152.9	509.3	1042.4	854.4	669.1	**3620.1**
Road Runner	7845.0	11.5	5640.6	9600.0	2696.7	5661.0	8895.1	**14220.5**	13534.0
Seaquest	42054.7	68.4	**683.3**	354.1	286.9	384.5	301.2	583.1	527.7
Up N Down	11693.2	533.4	3350.3	2877.4	2847.6	2955.2	3180.8	**28138.5**	10225.2
Median HNS (%)	100	0	14.4	16.1	20.4	17.5	26.8	41.5	**47.2**

Comparison with the State-of-the-arts

Table 3: Scores achieved by different methods on continuous-control benchmark DMControl.

100k Step Scores	**PlaNet**[16]	**Dreamer**[17]	**SAC+AE**[49]	**SLAC**[30]	**CURL**[29]	**DrQ** [50]	**SPR**†[40]	**PlayVirtual**
Finger, spin	136 ± 216	341 ± 70	740 ± 64	693 ± 141	767 ± 56	901 ± 104	868 ± 143	$\mathbf{915 \pm 49}$
Cartpole, swingup	297 ± 39	326 ± 27	311 ± 11	-	582 ± 146	759 ± 92	799 ± 42	$\mathbf{816 \pm 36}$
Reacher, easy	20 ± 50	314 ± 155	274 ± 14	-	538 ± 233	601 ± 213	638 ± 269	$\mathbf{785 \pm 142}$
Cheetah, run	138 ± 88	235 ± 137	267 ± 24	319 ± 56	299 ± 48	344 ± 67	467 ± 36	$\mathbf{474 \pm 50}$
Walker, walk	224 ± 48	277 ± 12	394 ± 22	361 ± 73	403 ± 24	$\mathbf{612 \pm 164}$	398 ± 165	460 ± 173
Ball in cup, catch	0 ± 0	246 ± 174	391 ± 82	512 ± 110	769 ± 43	913 ± 53	861 ± 233	$\mathbf{926 \pm 31}$
Median Score	137.0	295.5	351.0	436.5	560.0	685.5	719.0	**800.5**
500k Step Scores								
Finger, spin	561 ± 284	796 ± 183	884 ± 128	673 ± 92	926 ± 45	938 ± 103	924 ± 132	$\mathbf{963 \pm 40}$
Cartpole, swingup	475 ± 71	762 ± 27	735 ± 63	-	841 ± 45	868 ± 10	$\mathbf{870 \pm 12}$	865 ± 11
Reacher, easy	210 ± 390	793 ± 164	627 ± 58	-	929 ± 44	942 ± 71	925 ± 79	$\mathbf{942 \pm 66}$
Cheetah, run	305 ± 131	570 ± 253	550 ± 34	640 ± 19	518 ± 28	660 ± 96	716 ± 47	$\mathbf{719 \pm 51}$
Walker, walk	351 ± 58	897 ± 49	847 ± 48	842 ± 51	902 ± 43	921 ± 45	916 ± 75	$\mathbf{928 \pm 30}$
Ball in cup, catch	460 ± 380	879 ± 87	794 ± 58	852 ± 71	959 ± 27	963 ± 9	963 ± 8	$\mathbf{967 \pm 5}$
Median Score	405.5	794.5	764.5	757.5	914.0	929.5	920.0	**935.0**

THANKS!