Cross-Modal Domain Adaptation for Cost-Efficient Visual Reinforcement Learning

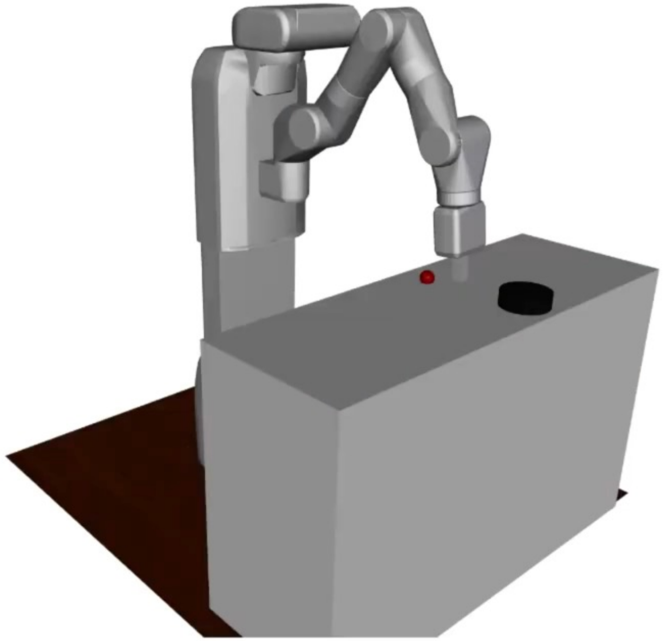
Xiong-Hui Chen*, Shengyi Jiang*, Feng Xu, Zongzhang Zhang†, Yang Yu
National Key Laboratory of Novel Software Technology
Nanjing University, Nanjing 210023, China
Pazhou Lab, Guangzhou 510330, China
{chenxh, jiangsy, xufeng}@lamda.nju.edu.cn, {zzzhang, yuy}@nju.edu.cn

南京大學 NANJING UNIVERSITY

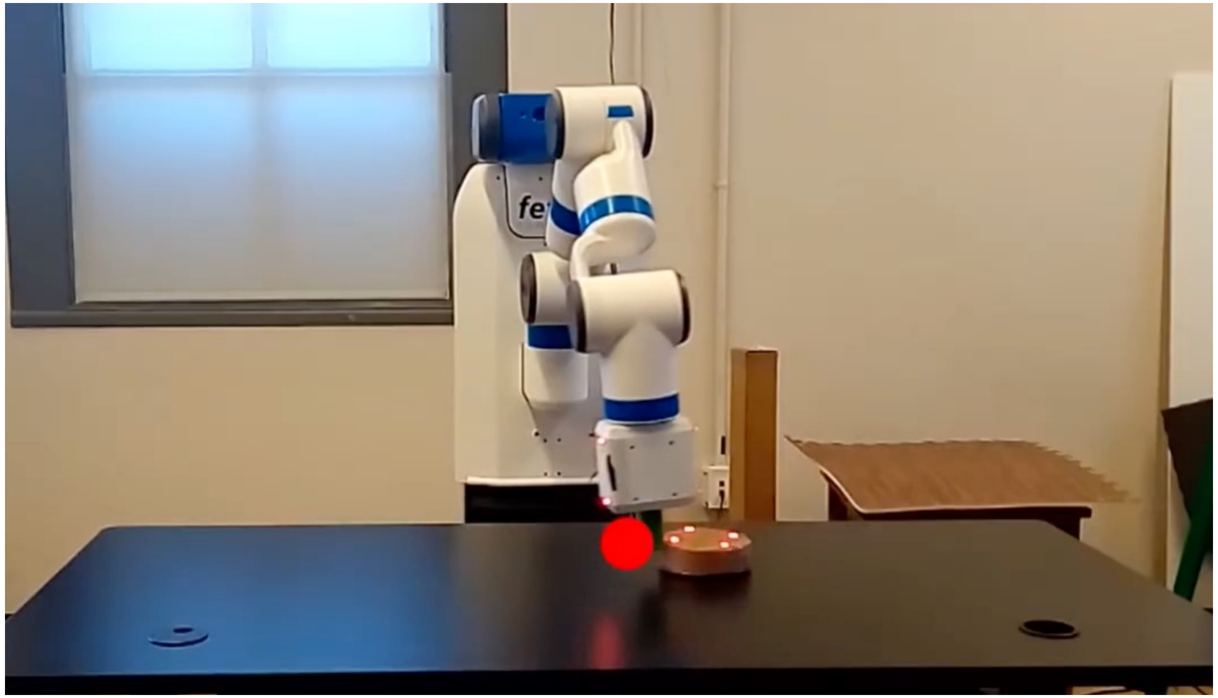
Learning And Mining from DatA

Table of Contents

An example of Sim2Real Reinforcement Learning



Simulation



Real world

An example of reality-gap which is the core challenge
in Sim2Real RL

The Framework of Unsupervised Domain Adaptation in Sim2Real RL

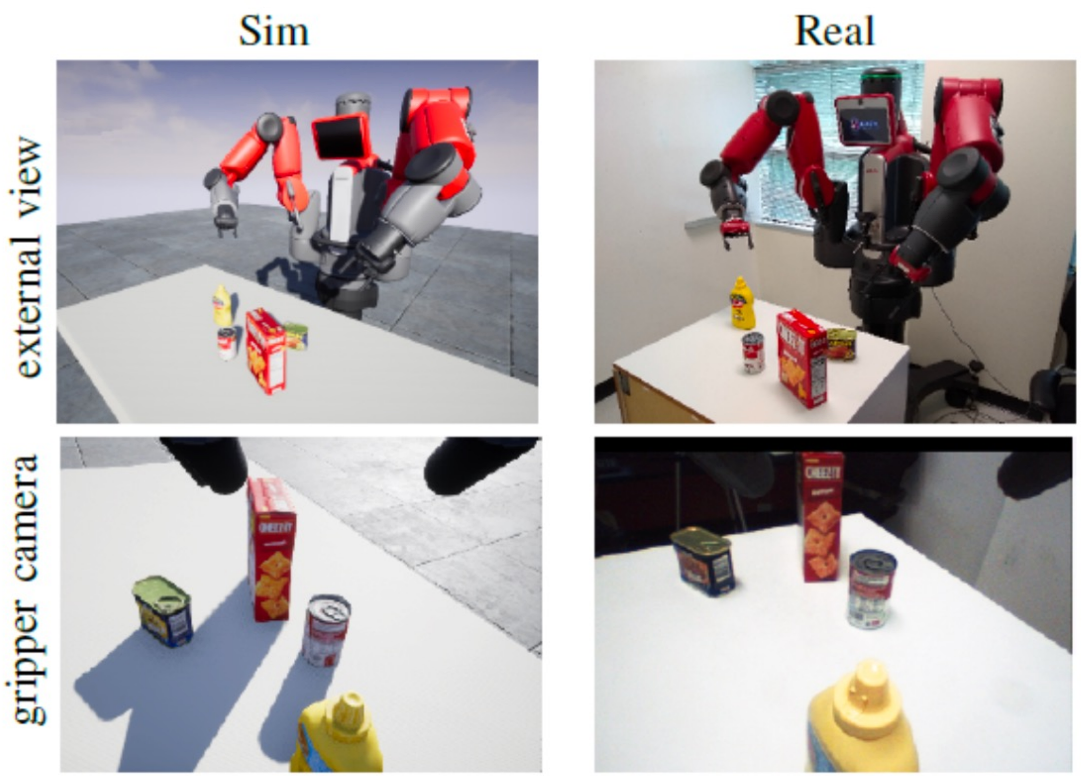
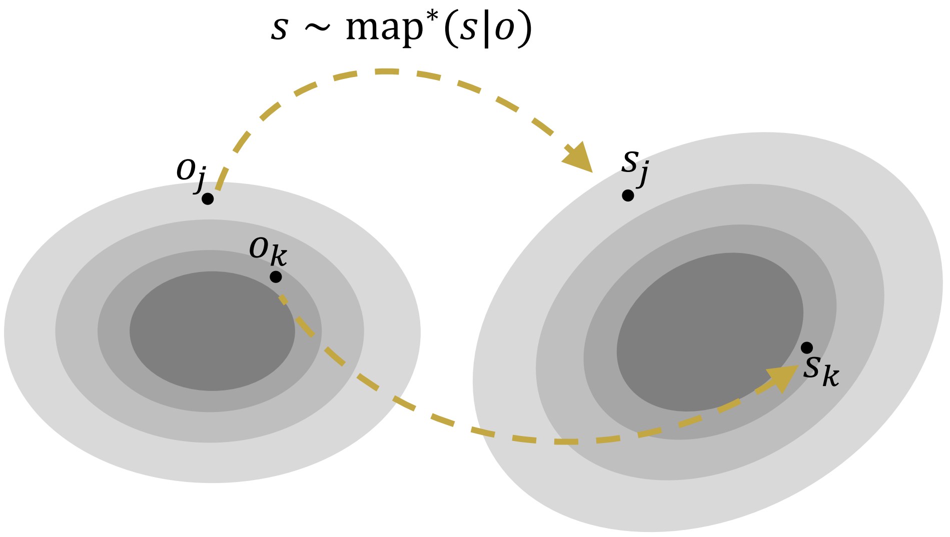


Fig. An example of reality-gaps in observation-space in Sim2Real RL [1]



$$\mathrm{map}^* = \min_{\mathrm{map}} \mathrm{Dist}_{JS}(\mathrm{SimData}||\mathrm{map}(\mathrm{RealData}))$$

-> trained via GAN-style algorithms

Fig. The framework of Unsupervised Domain Adaptation for Sim2Real RL

Unsupervised domain adaptation (UDA) learns a mapping function to align the data distribution of the source and the target domain to handle the challenge of reality-gap on observation-space in Sim2Real RL

Cross-Modal UDA: A cost-efficient Framework for Sim2Real RL

Cross-Modal UDA: A cost-efficient Framework for Sim2Real RL

Image-to-image UDA introduce three extra cost, which is ignored in discussion in previous work.

1. human labor of building a visual simulator
2. huge computation resource required by running the simulator
3. inferior policy training on visual simulator.

Cross-Modal UDA: A cost-efficient Framework for Sim2Real RL

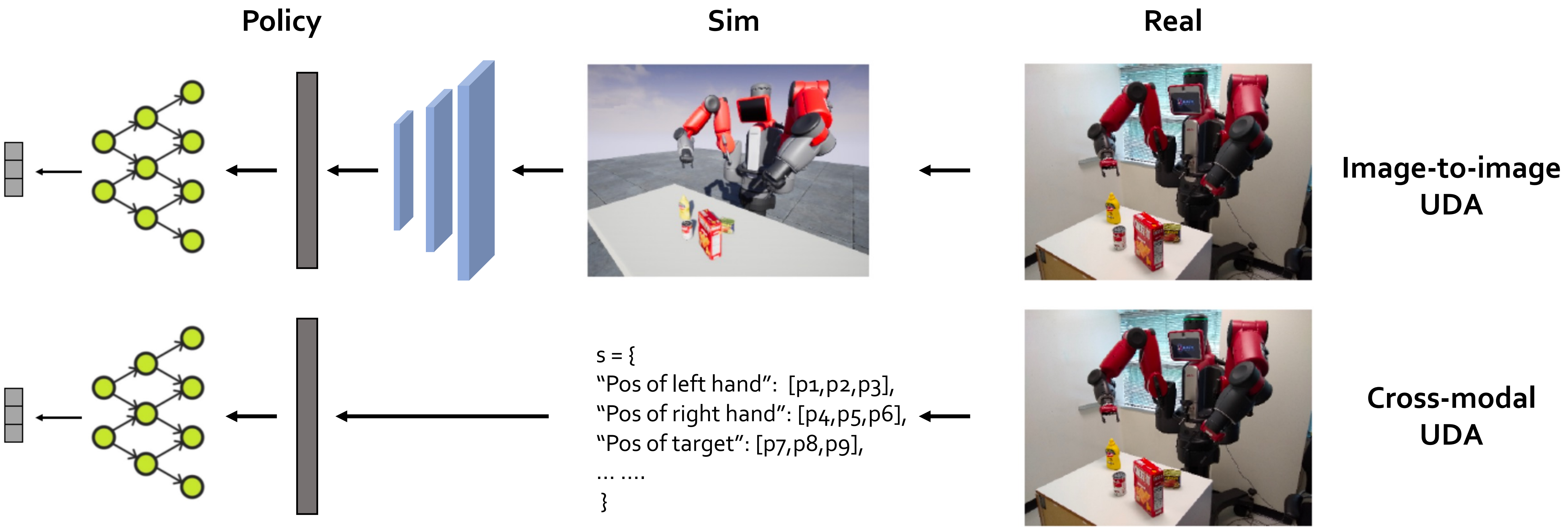
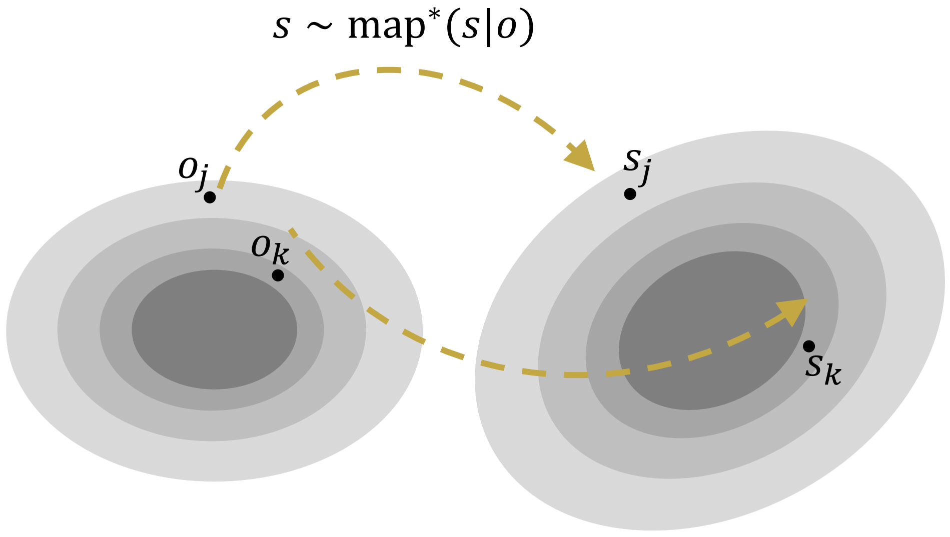


Image-to-image UDA introduce three extra cost, which is ignored in discussion in previous work.

1. human labor of building a visual simulator
2. huge computation resource required by running the simulator
3. inferior policy training on visual simulator.

⇨ **Can be solved in Cross-modal UDA**

Ill-posedness of the raw objective in current UDA solutions



$$\text{map}^* = \min_{\text{map}} \text{Dist}_{JS}(\text{SimData}||\text{map}(\text{RealData}))$$

-> trained via GAN-style algorithms

We cannot adopt the previous Unsupervised Domain Adaptation to Cross-Modal UDA setting.

Ill-posedness of the raw objective in current UDA solutions

$$\text{map}^* = \min_{\text{map}} \text{Dist}_{JS}(\text{SimData}||\text{map}(\text{RealData}))$$

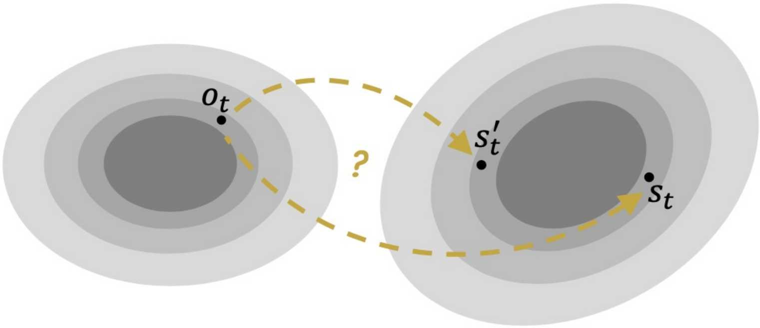


Fig. An example of ill-posedness of the distribution minimization objective in current UDA algorithms

Since s_t and s_t' have similar probabilities, mapping an instance o_t to anywhere of a similar probability in the source domain is "reasonable" if we only consider distribution matching.

In image-to-image UDA, current methods rely on additional constraints on modality consistency to handle the problem implicitly

- special model structure [2], e.g. U-Net, Cycle-GAN;
- auxiliary losses, e.g. geometry consistency [3];

However, these constraints cannot hold anymore in Cross-modal UDA setting.

Ill-posedness of the raw objective in current UDA solutions

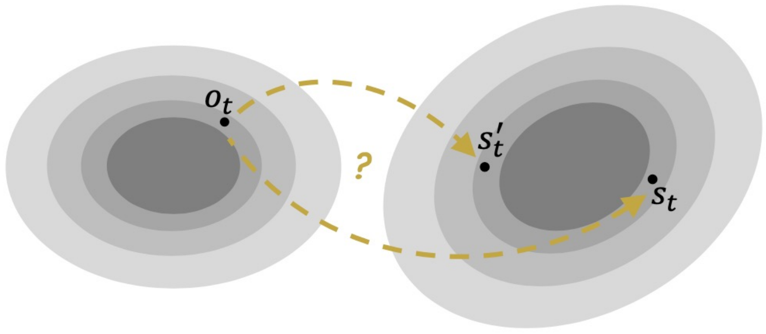


Fig. An example of ill-posedness of the distribution minimization objective in current UDA algorithms

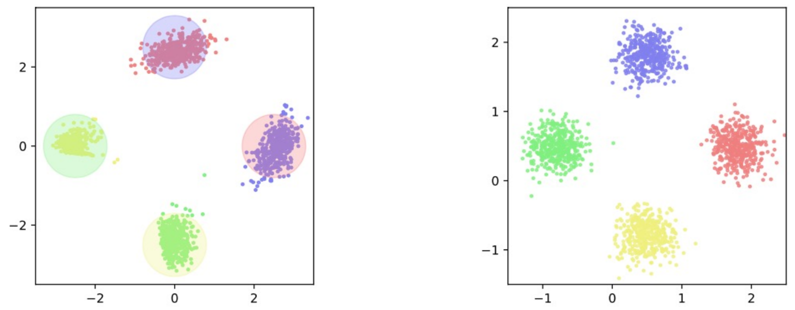


Fig. A toy example of ill-posed UDA

Since s_t and $s_t{}'$ have similar probabilities, mapping an instance o_t to anywhere of a similar probability in the source domain is "reasonable" if we only consider distribution matching.

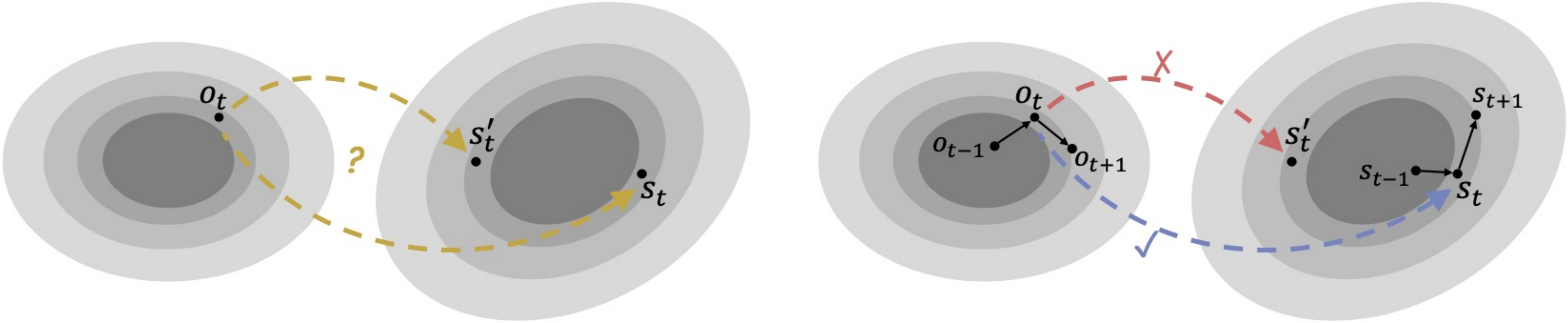
Our research question: Can we handle the ill-posedness of the objective directly?

NANJING UNIVERSITY

Learning And Mining from DatA

Table of Contents

Any other potential way to handle the ill-posedness of the objective?



(a) Mapping only considering state-distribution matching

(b) Mapping with sequential structure

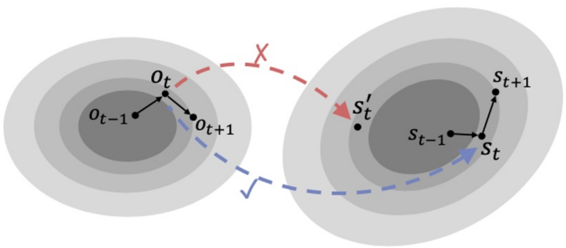
If we can make use of **the sequential structure in the Markov Decision Processes**,
the historical information will give us the ability to identify the difference between $s_t{}'$ and s_t,

then the proposed ill-posedness of the objective will be fixed.

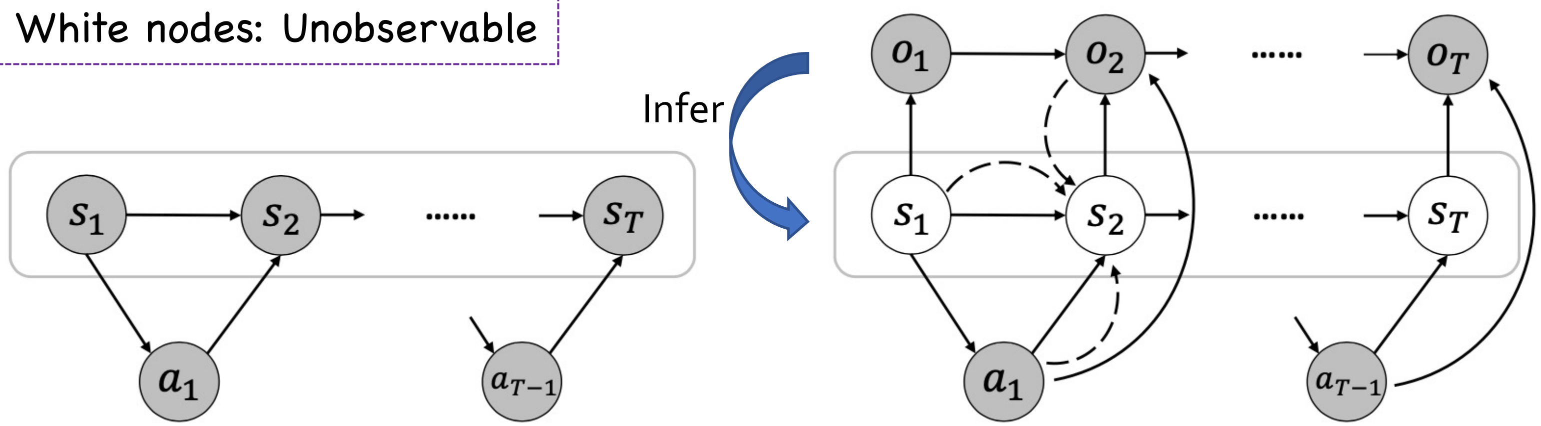
南京大學 NANJING UNIVERSITY

LAMDA Learning And Mining from DatA

Reformulate the objective of UDA in RL based on the framework of variational inference



Shaded nodes: Observable
White nodes: Unobservable

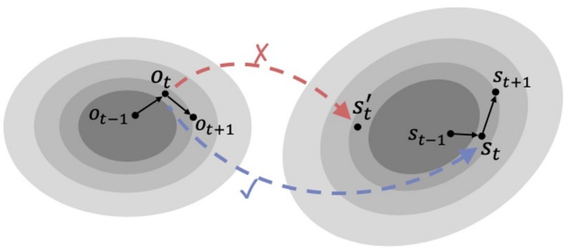


(a) Generation process in the source domain

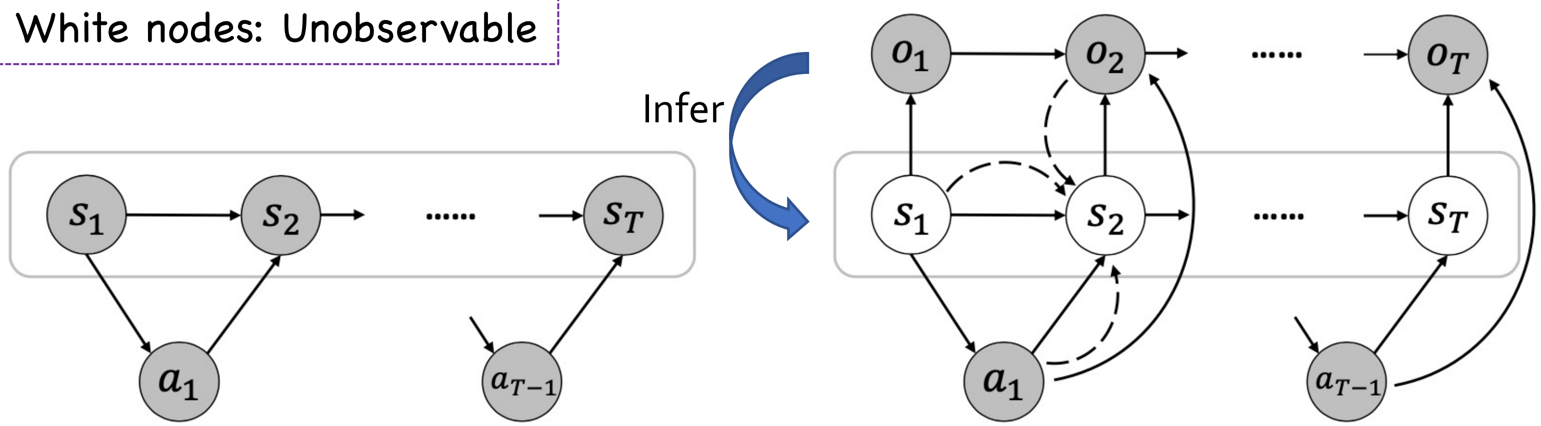
(b) Generation and inference process in the target domain

Fig. The generation and inference process of UDA based on the framework of variational inference

Reformulate the objective of UDA in RL based on the framework of variational inference



Shaded nodes: Observable
White nodes: Unobservable



(a) Generation process in the source domain (b) Generation and inference process in the target domain

Fig. The generation and inference process of UDA based on the framework of variational inference

$$\min_{\phi} \mathbb{E}_{\tau^o}\left[D_{\mathrm{KL}}\left[q_\phi(\tau^s \mid \tau^o) \mid\mid p(\tau^s \mid \tau^o)\right]\right]$$

$$\Downarrow (ELBO)$$

$$\max_{\phi,\theta} \mathbb{E}_{\tau^o}\left[\underbrace{\mathbb{E}_{\hat{\tau}^s \sim q_\phi(\tau^s \mid \tau^o)}\left[\log p_\theta\left(\tau^o \mid \hat{\tau}^s\right)\right]}_{\text{Reconstruction Error}} - \underbrace{D_{\mathrm{KL}}\left[q_\phi\left(\tau^s \mid \tau^o\right) \mid\mid p\left(\tau^s\right)\right]}_{\text{Divergence between prior and inferred state distributions}}\right]$$

D_{KL}: KL Divergence
τ: trajectory
o: observation
s: state
ϕ: parameters to learn

南京大學 NANJING UNIVERSITY

LAMDA Learning And Mining from DatA

Reformulate the objective of UDA in RL based on the framework of variational inference

$$\min_{\phi} \mathbb{E}_{\tau^o}\left[D_{\mathrm{KL}}\left[q_\phi(\tau^s \mid \tau^o) \,||\, p(\tau^s \mid \tau^o)\right]\right]$$

$$\Downarrow (ELBO)$$

$$\max_{\phi,\theta} \mathbb{E}_{\tau^o}\Big[\underbrace{\mathbb{E}_{\hat{\tau}^s\sim q_\phi(\tau^s|\tau^o)}\left[\log p_\theta\left(\tau^o \mid \hat{\tau}^s\right)\right]}_{\text{Reconstruction Error}} - \underbrace{D_{\mathrm{KL}}\left[q_\phi\left(\tau^s \mid \tau^o\right) \,||\, p\left(\tau^s\right)\right]}_{\text{Divergence between prior and inferred state distributions}}\Big]$$

D_{KL}: KL Divergence
τ: trajectory
o: observation
s: state
ϕ: parameters to learn

$$\Downarrow$$

Reconstruction loss

$$\max_{\phi,\theta} \mathbb{E}_{\tau^o\sim\mathcal{D}^o}\Big[\sum_{t=1}^{T} \mathbb{E}_{\hat{s}_t\sim q_\phi(s_t|\hat{s}_{t-1},a_{t-1},o_t)}\Big[\log p_\theta(o_t \mid \hat{s}_t, o_{t-1}, a_{t-1})$$

Trajectory-distribution mismatch loss

$$-\lambda_D \log\left(1 - D_{\omega^*}(\hat{s}_t, a_t, h_{t-1})\right)\Big]\Big],$$

Discriminator loss

$$\text{s.t. } \omega^* = \arg\max_{\omega} \mathbb{E}_{\tau^s\sim\mathcal{D}^s}\Big[\sum_{t=1}^{T} \log D_\omega(s_t, a_t, h_{t-1})\Big]$$

$$+\mathbb{E}_{\tau^o\sim\mathcal{D}^o}\Big[\sum_{t=1}^{T} \mathbb{E}_{\hat{s}_t\sim q_\phi(s_t|\hat{s}_{t-1},a_{t-1},o_t)} \log\left(1 - D_\omega(\hat{s}_t, a_t, h_{t-1})\right)\Big].$$

Embedded Dynamics Model for Stable Training

Inference Function only outputs a small Δs_t. The main part is from **Embedded DM** (p_ϕ).

$$\hat{s}_t = p_\varphi(s_{t-1}, a_{t-1}) + \alpha \Delta s_t,$$
$$\text{where} \Delta s_t \sim q_\phi(\Delta s \mid s_{t-1}, a_{t-1}, o_t)$$

The parameters of **Embedded DM** are copied from a DM trained by:

$$\min_\varphi \mathbb{E}_{(s,a,s') \sim D^s \cup D^{\hat{s}}} [(p_\varphi(s, a) - s')^2]$$

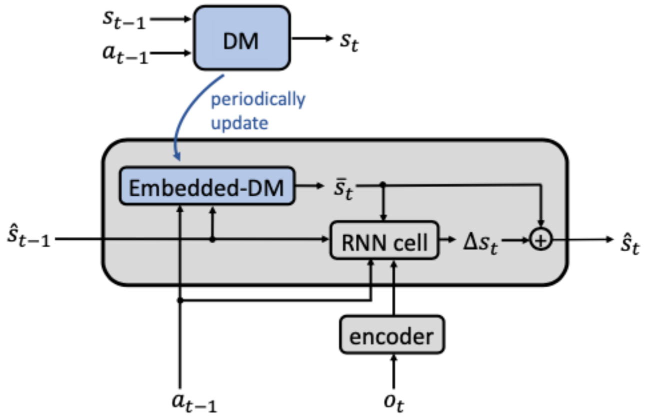
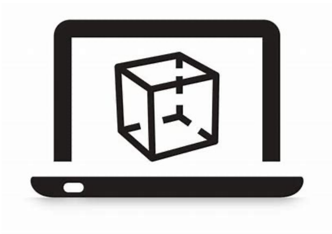


Figure: Detailed Structure of the Inference Function with Embedded DM

Overall Model Structure

(s_{t-1}, a_{t-1}, s_t)



A simulator of the source domain

(o_{t-1}, a_{t-1}, o_t)

Pre-collected dataset in
the target domain

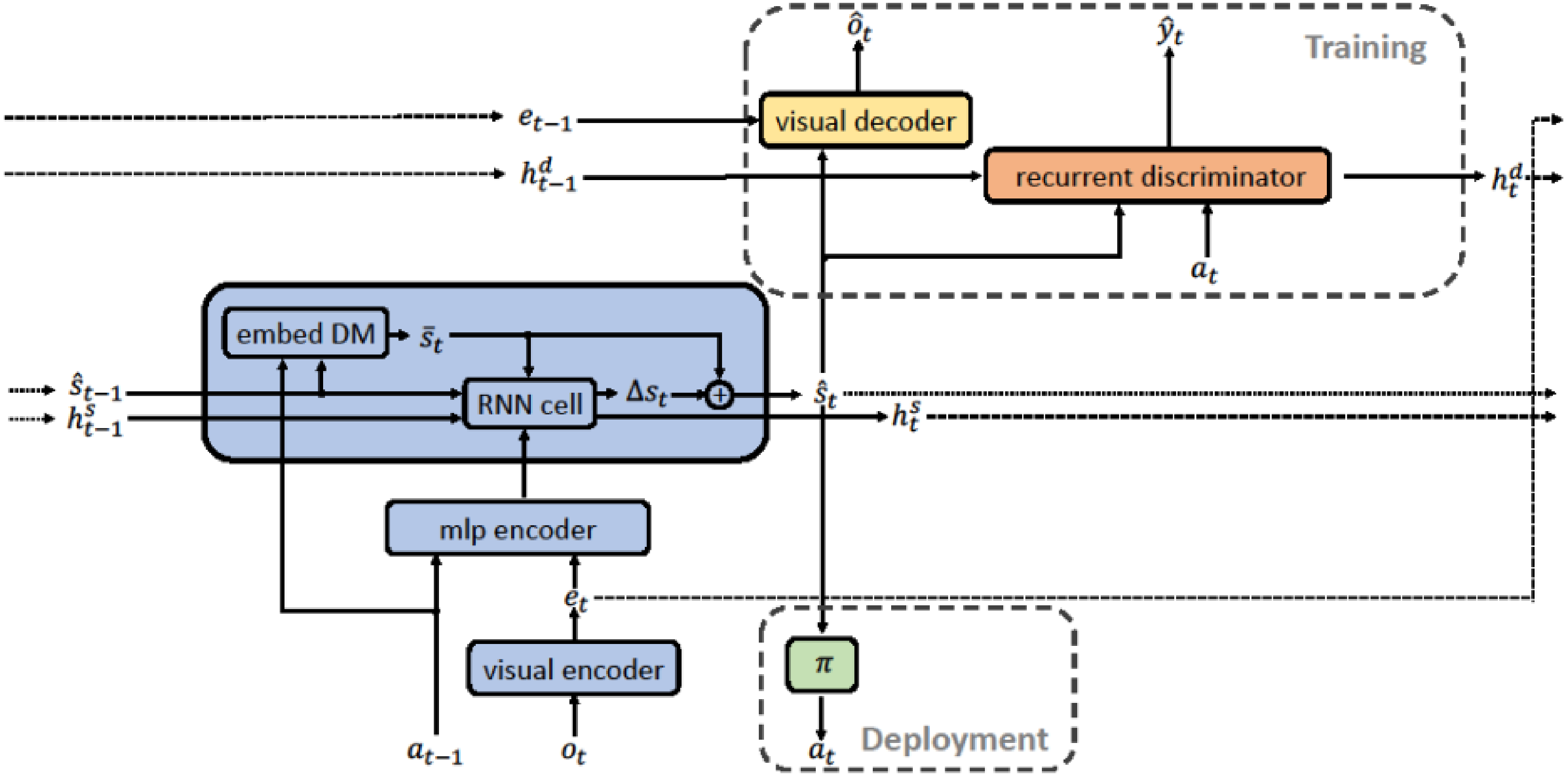
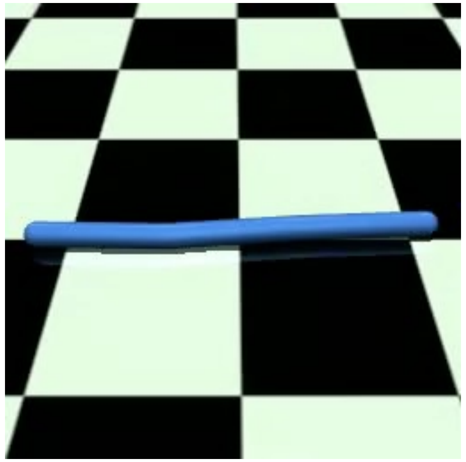


Table of Contents

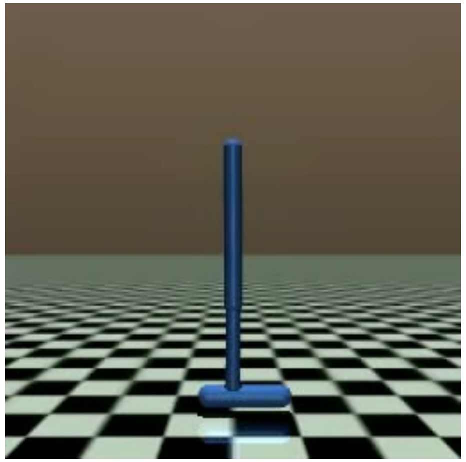
NANJING UNIVERSITY

Learning And Mining from DatA

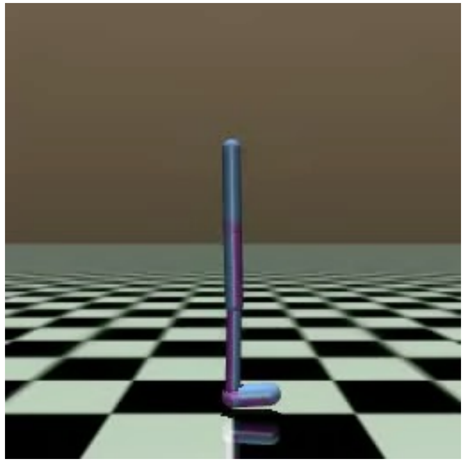
Performance on MuJoCo Tasks



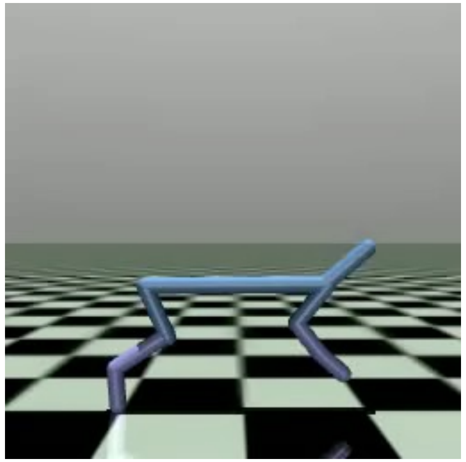
Swimmer



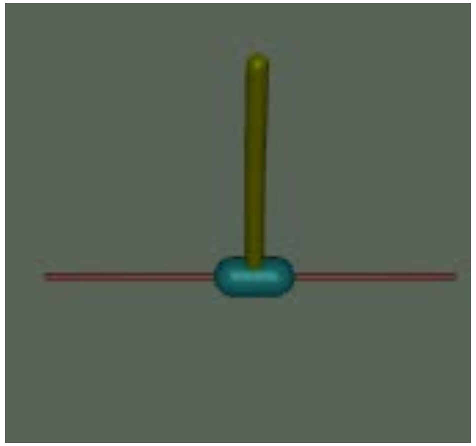
Hopper



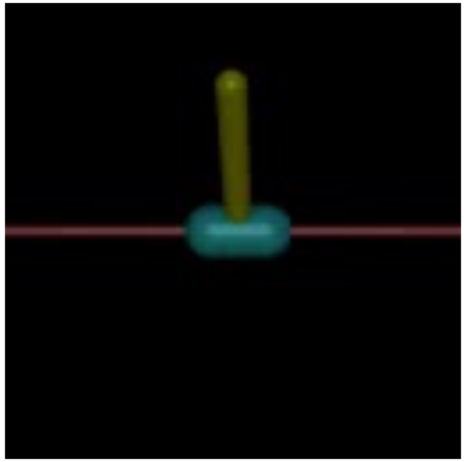
Walker2d



Half Cheetah



Inverted Double Pendulum



Inverted Pendulum

Comparative Evaluation in MuJoCo Tasks

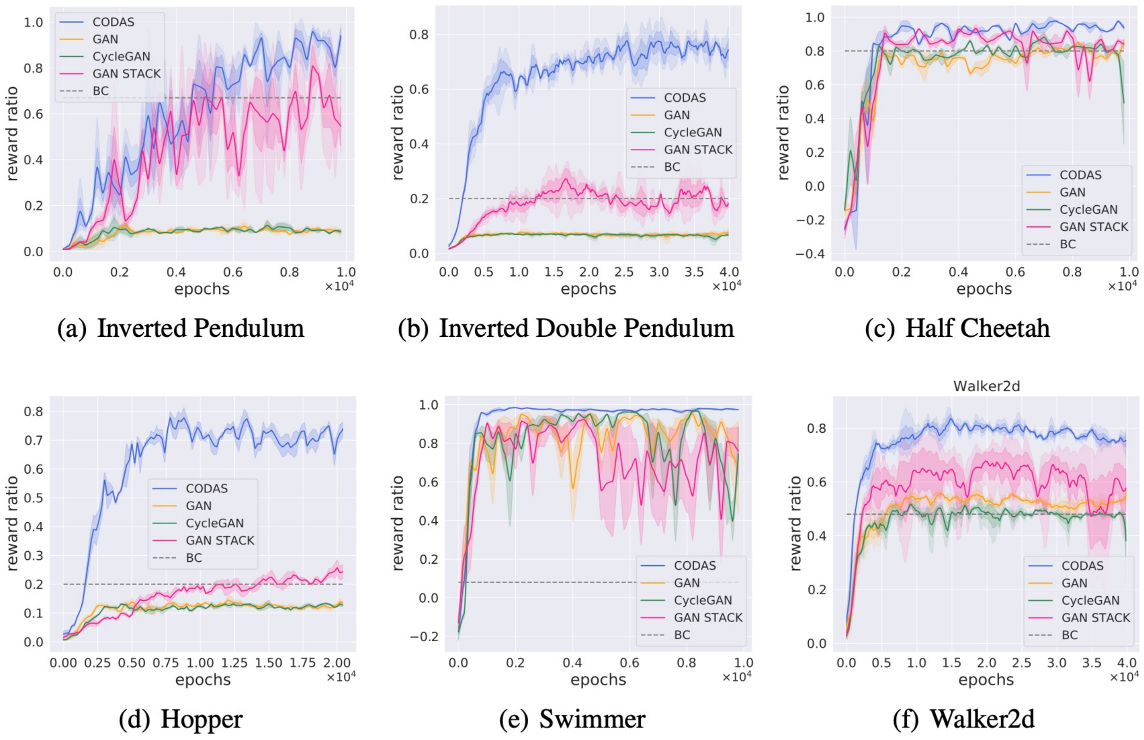


Figure 4: Training curves of different methods on MuJoCo.

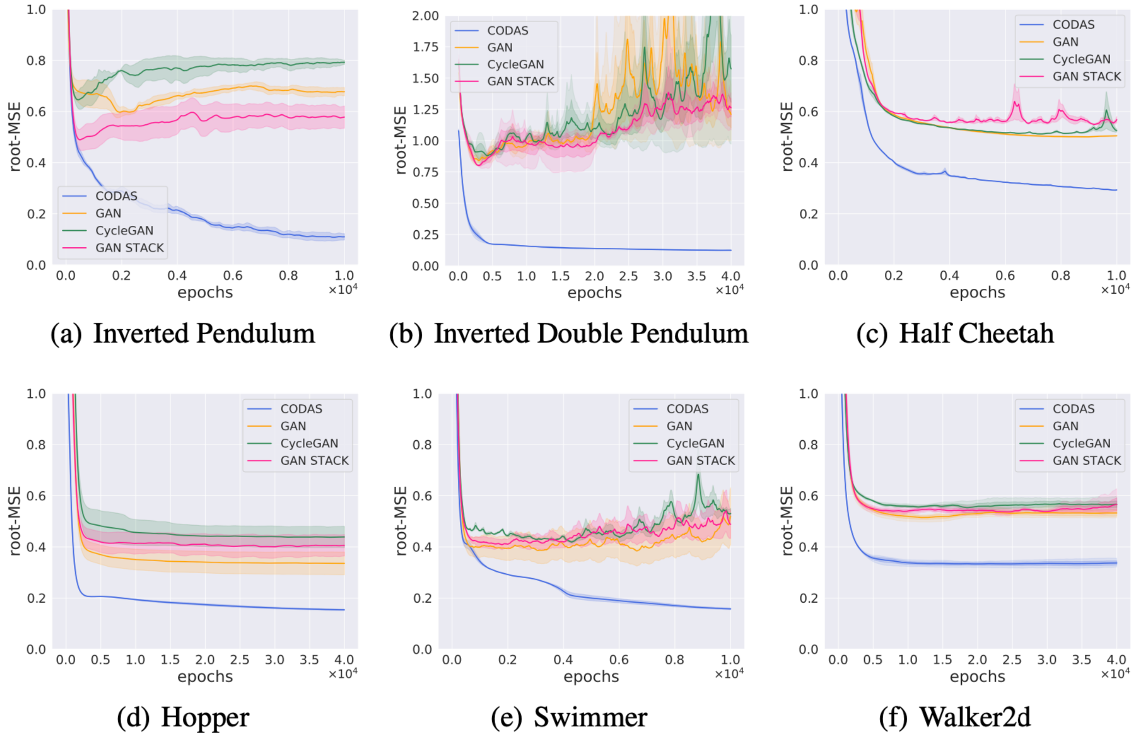


Figure 6: Root mean squared error between mapped states and ground-truth states. The solid lines denote the mean value. The shadows denote the standard deviation.

For all of the tasks, CODAS can map the correct states (i.e., with the smallest MSE-loss to the oracle states) and the performance of the deployment policies reach reasonable performance (75%~100%)

NANJING UNIVERSITY

Learning And Mining from DatA

Visualization of the Learned Mapping on MuJoCo Tasks

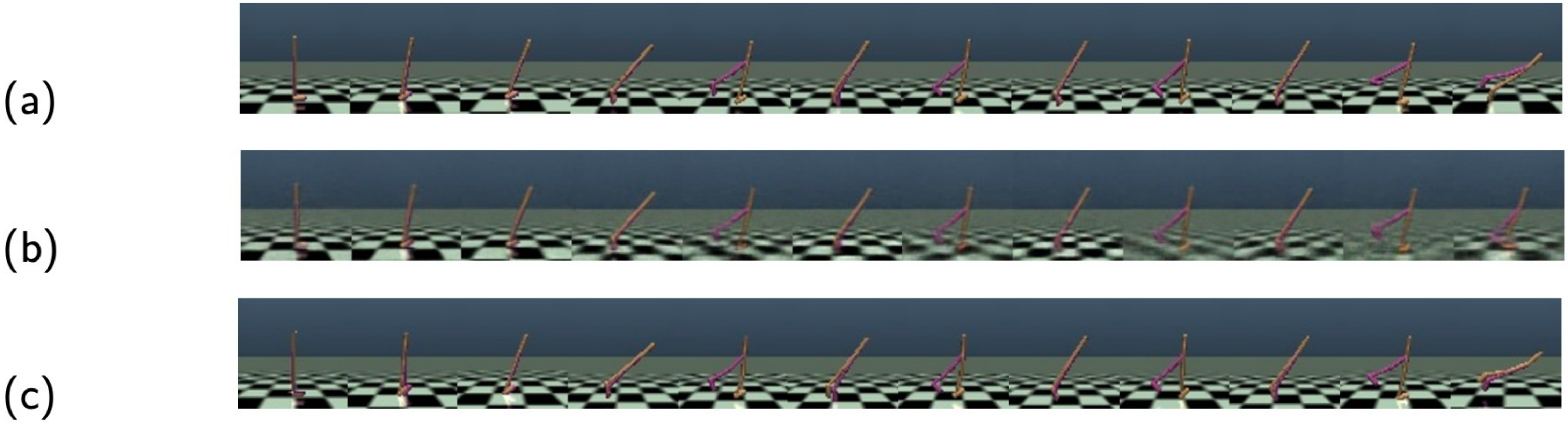
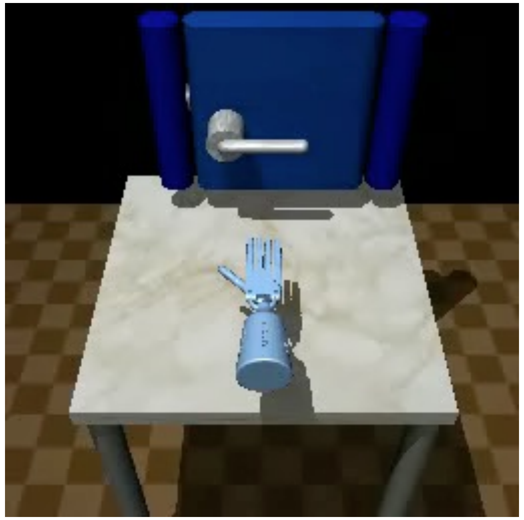


Fig. A visual illustration of (a) original images, (b) reconstructed images, and (c) re-rendered images of the mapped states in Hopper.

Both reconstructed images and re-rendered images match the original ones well. Re-rendered images can even match the original ones well in the last falling frames which are sparse in the dataset.

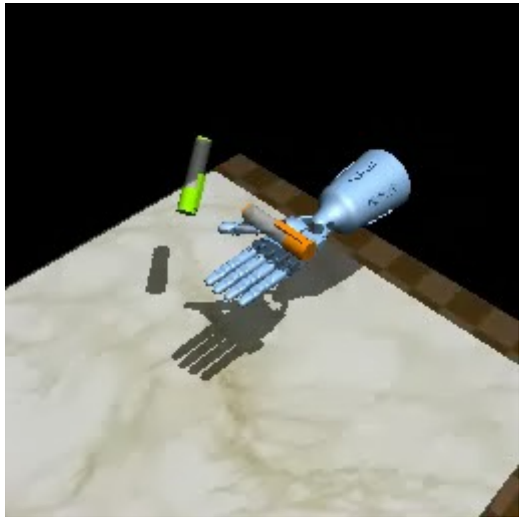
Performance on Robot Hand Manipulation Tasks

Relocate



Door

Hammer



Pen

NANJING UNIVERSITY

Learning And Mining from DatA

Performance on Robot Hand Manipulation Tasks

Data collecting policy

Re-rendered video

Reconstructed video

Tasks	hammer	pen	door	relocate
Reward Ratio	0.820	0.701	0.886	0.090

In three out of four tasks, CODAS yields reasonable mapping functions for policy deployment.

南京大學 NANJING UNIVERSITY

LAMDA Learning And Mining from DatA

Take-home Message

KEY point:

The Formulation of Variational Inference which considered the sequential structure in MDP can handle the ill-posedness of the objective and solve the UDA problem without relying on the knowledge of modality consistency.

Future work:

1. CODAS solve the UDA problem in a general way, it can be adopted to image-to-image UDA in theory and the practical adoption can be tried.
2. In the current formulation, we assume the policies/dynamics in the source and target domains are the same, which might not hold in real-world applications. By modeling the mismatching of dynamics models and data-collected policies into the CODAS framework, we can build a more practical UDA algorithm.

>> Thanks