Application of Generative Models: (Selected) Advanced Topics

Hao Dong

Peking University

Application of Generative Models: (Selected) Advanced Topics

- Domain Adaptation
- Adversarial Attack
- Meta Learning
- Imitation Learning
- Reinforcement Learning

Application of Generative Models: (Selected) Advanced Topics

PEKING UNIVERSITY

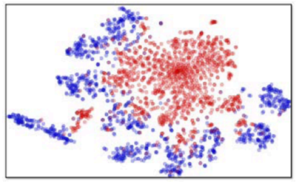
- Domain Adaptation
- Adversarial Attack
- Meta Learning
- Reinforcement Learning

Domain Adaptation

- **Single Source Domain Adaptation**

Source: Labelled

Target: Unlabelled



$$S(\mathbf{f}) = \{G_f(\mathbf{x};\theta_f)\,|\,\mathbf{x}\sim S(\mathbf{x})\}$$
$$T(\mathbf{f}) = \{G_f(\mathbf{x};\theta_f)\,|\,\mathbf{x}\sim T(\mathbf{x})\}$$

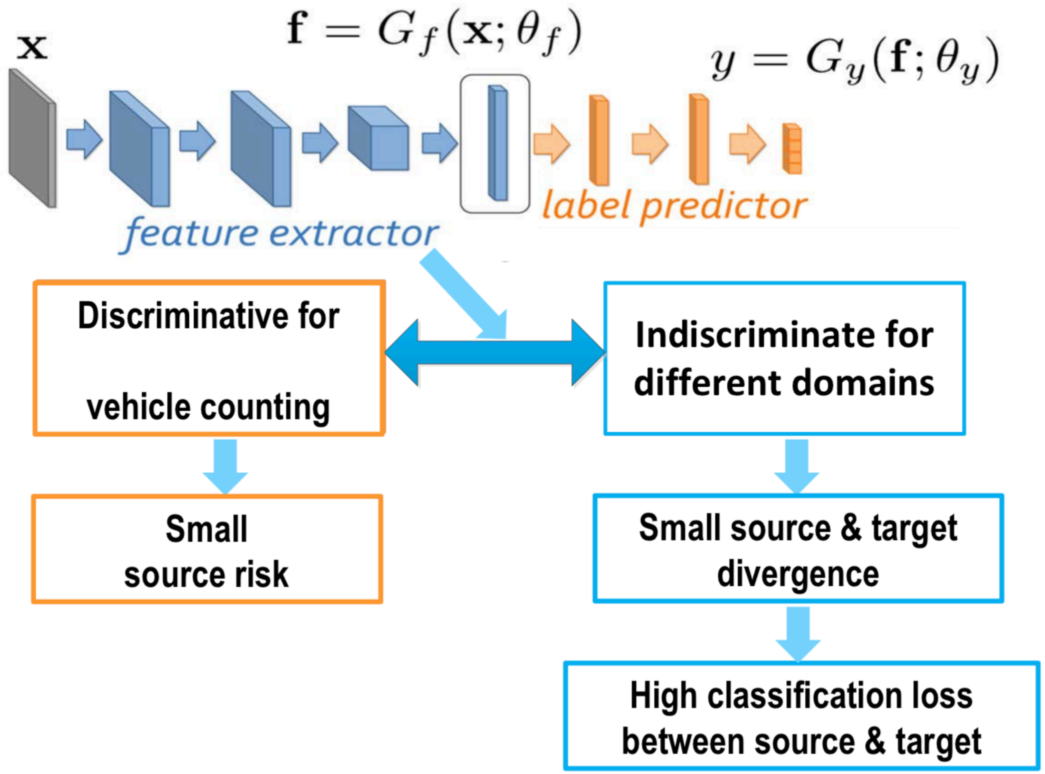
Domain shift among sources and target

Domain adaptation needed!

Domain-Adversarial Training of Neural Networks. *Y. Ganin, H. Ajakan et al. JMLR. 2016*

Domain Adaptation

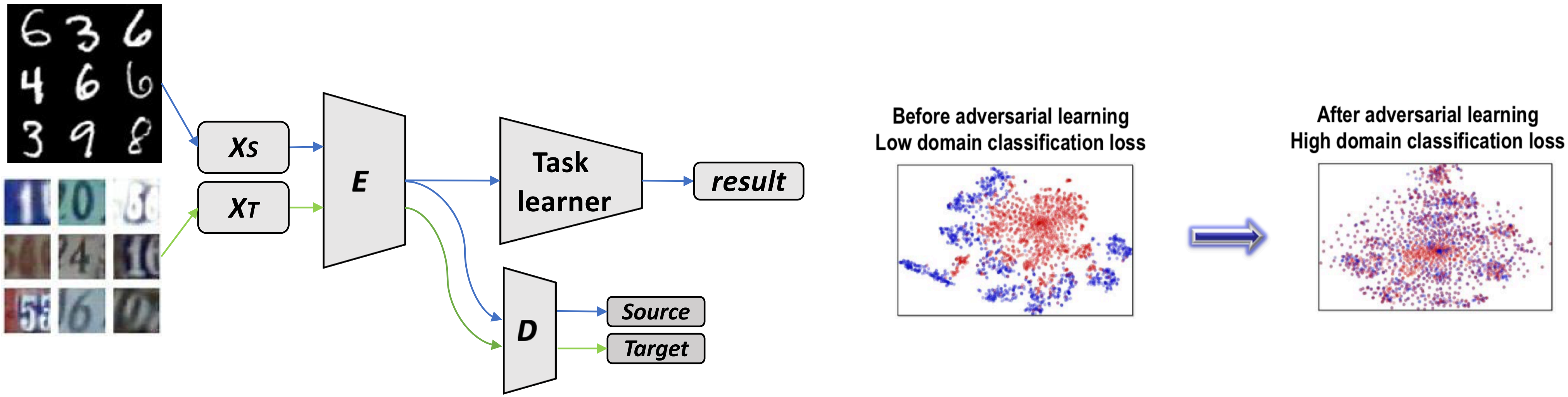
- **Learn domain-universal & task-discriminative features**



Domain-Adversarial Training of Neural Networks. *Y. Ganin, H. Ajakan et al. JMLR. 2016*

Domain Adaptation

- **Single Source Domain Adaptation**

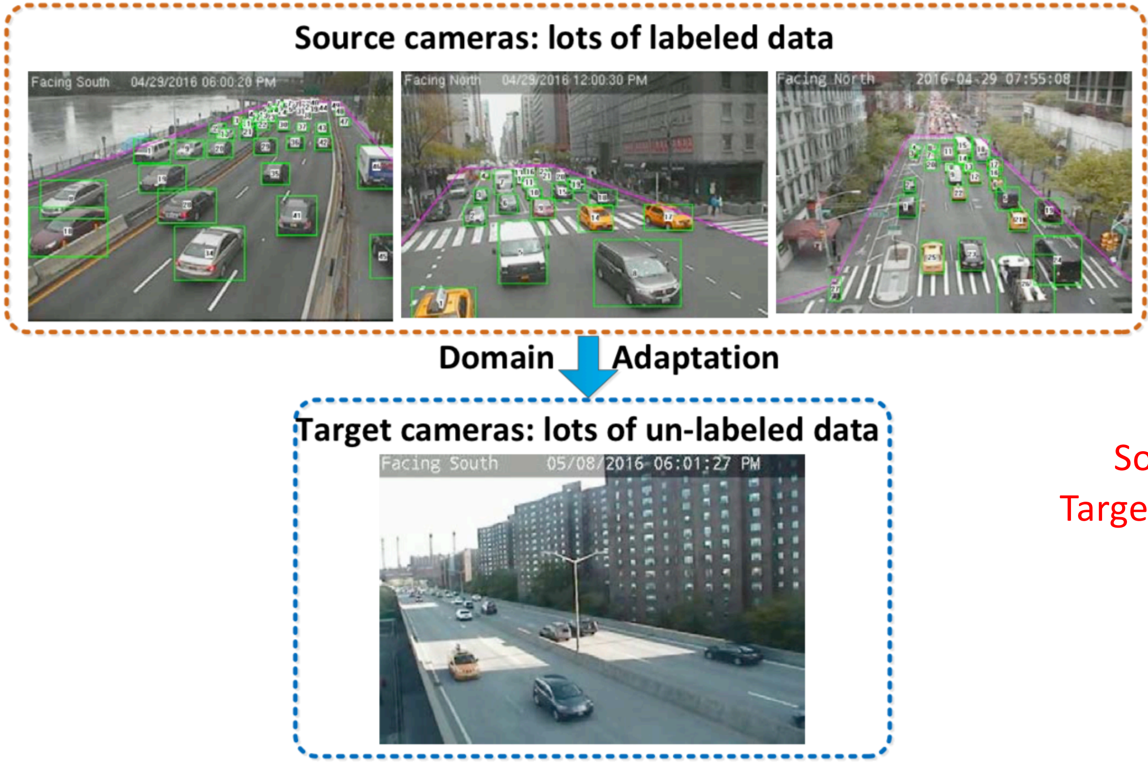


Source images have label
Target images do not have label

Domain-Adversarial Training of Neural Networks. *Y. Ganin, H. Ajakan et al. JMLR. 2016*

Domain Adaptation

- **Multiple Source Domain Adaptation**

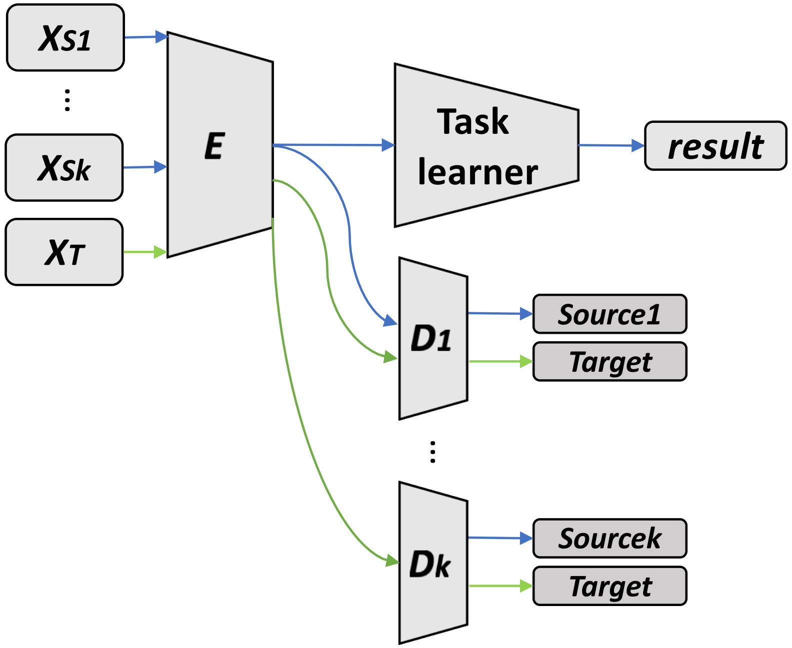


Source images have label
Target images do not have label

Multiple Source Domain Adaptation with Adversarial Learning. *S. Zhang, H. Zhao et al. NIPS. 2018.*

Domain Adaptation

- **Multiple Source Domain Adaptation**

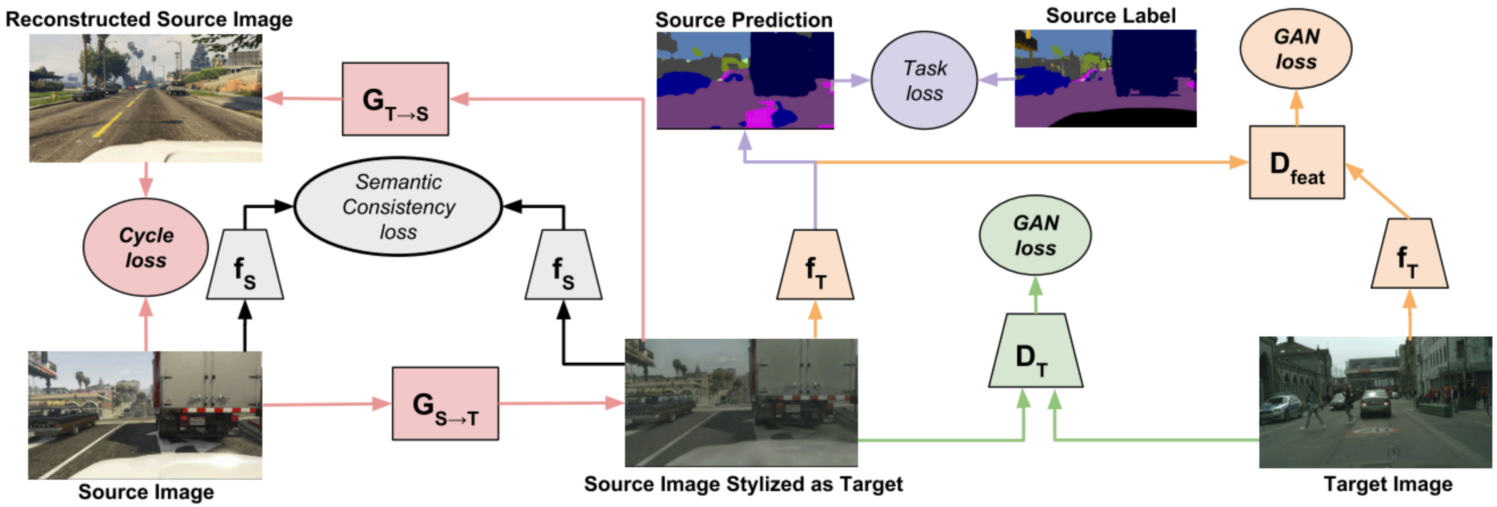


Source images have label
Target images do not have label

Multiple Source Domain Adaptation with Adversarial Learning. *S. Zhang, H. Zhao et al. NIPS. 2018.*

Domain Adaptation

- **Cross Domain Translation + Segmentation**



Source: GTA provides labeled maps
Target: real images

CyCADA: Cycle-Consistent Adversarial Domain Adaptation. Hoffman, Judy. Tzeng, Eric. Park, Taesung. Zhu, Jun-yan. Berkeley, U C. Isola, Phillip. Saenko, Kate. Efros, Alexei A. Darrell, Trevor. ICML 2018.

Application of Generative Models: (Selected) Advanced Topics

PEKING UNIVERSITY

- Domain Adaptation
- Adversarial Attack
- Meta Learning
- Reinforcement Learning

Adversarial Attack

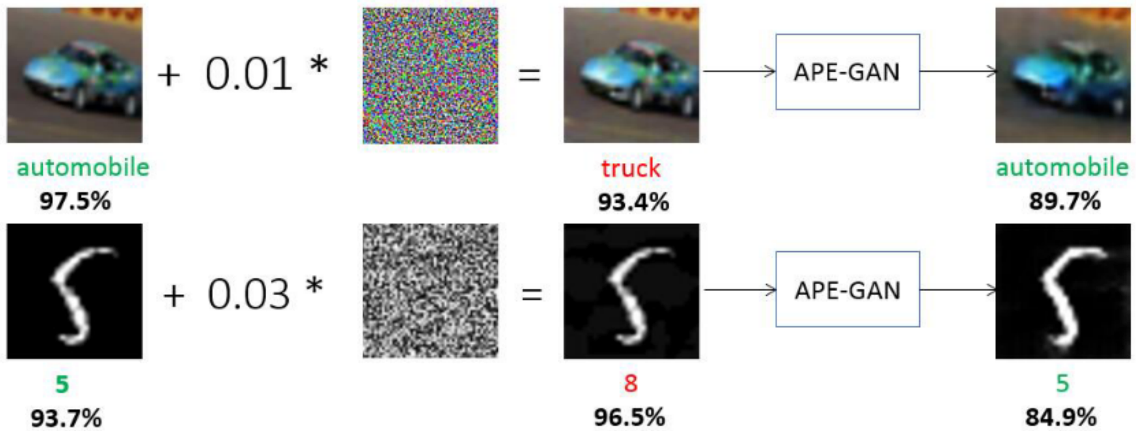
- WHITE-BOX ATTACK MODELS
  - White-box models assume that the attacker has complete knowledge of all the classifier parameters, i.e., network architecture and weights, as well as the details of any defense mechanism
  - targeted attack: they attempt to cause the perturbed image to be misclassified to a specific target class
  - untargeted attack: when no target class is specified

- BLACK-BOX ATTACK MODELS
  - black-box adversaries have no access to the classifier or defense parameters, It is further assumed that they do not have access to a large training dataset but can query the targeted DNN as a black-box.

Defense-Gan: Protecting classifiers against adversarial attacks using generative models

Adversarial Attack

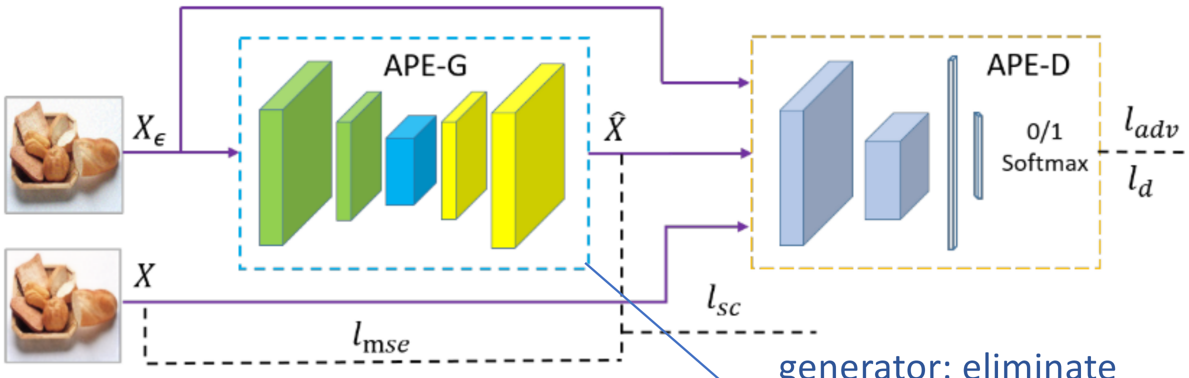
APE-GAN: adversarial perturbation elimination with GAN



The essence of the model is to eliminate the adversarial perturbations in the samples. The model use the adversarial samples themselves to generate corresponding real samples.

Adversarial Attack

APE-GAN: adversarial perturbation elimination with GAN



APE-GAN Training

generator: eliminate adversarial perturbation of the sample

$$l_{ape} = \xi_1 l_{mse} + \xi_2 l_{adv} + \xi_3 l_{sc}$$

$$\begin{cases} l_{mse} = \dfrac{1}{WH} \displaystyle\sum_{w=1}^{W} \sum_{h=1}^{H} (X_{w,h} - G_{\theta_G}(X_\epsilon)_{w,h}) \\ l_{adv} = \displaystyle\sum_{n=1}^{N} [1 - \log D_{\theta_D}(G_{\theta_G}(X_\epsilon))] \\ l_{sc} = \dfrac{1}{WH} \displaystyle\sum_{w=1}^{W} \sum_{h=1}^{H} \|\nabla G_{\theta_G}(X_\epsilon)_{w,h}\| \end{cases}$$

Total loss:

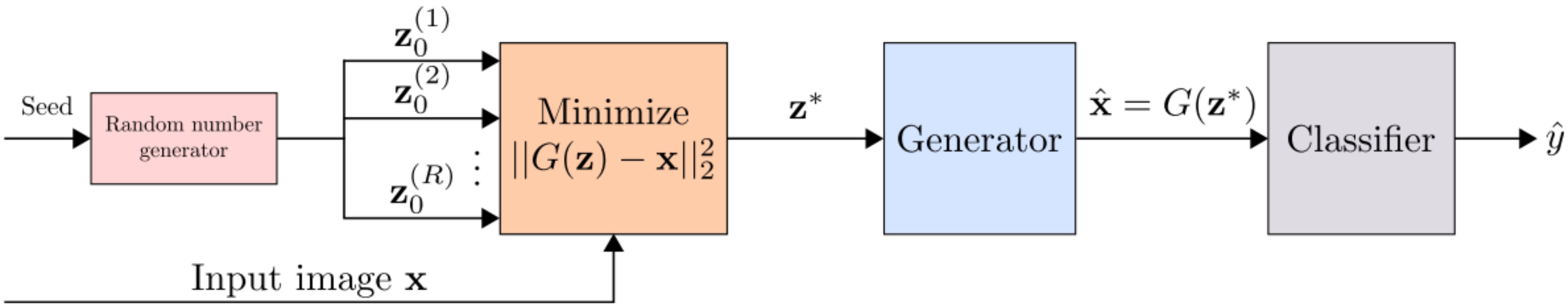
$$\hat{\theta}_G = \arg\min_{\theta_G} \frac{1}{N} \sum_{k=1}^{N} l_{ape}(G_{\theta_G}(X_\epsilon^k), X^k)$$

Adversarial training:

$$\min_{\theta_G} \max_{\theta_D} \quad \mathbb{E}_{X \sim p_{data}(X)} \log D_{\theta_D}(X) - \mathbb{E}_{X_\epsilon \sim p_G(X_\epsilon)} \log(D_{\theta_D}(G_{\theta_G}(X_\epsilon)))$$

Adversarial Attack

- Defense-GAN



use legitimate examples in training

- a new defense strategy which uses a WGAN trained on legitimate (un-perturbed) training samples to "denoise" adversarial examples.

Defense-Gan: Protecting classifiers against adversarial attacks using generative models

Adversarial Attack

PEKING UNIVERSITY

- APE-GAN:
  - Use **adversarial samples** as the input of the generator.
- Defense-GAN:
  - Use **multiple random noise** as the input of the generator.
  - Implement adversarial training without using adversarial samples as inputs.
- Both of the structures are based on WGAN.

Application of Generative Models: (Selected) Advanced Topics

- Domain Adaptation
- Adversarial Attack
- Meta Learning
- Reinforcement Learning

Meta Learning

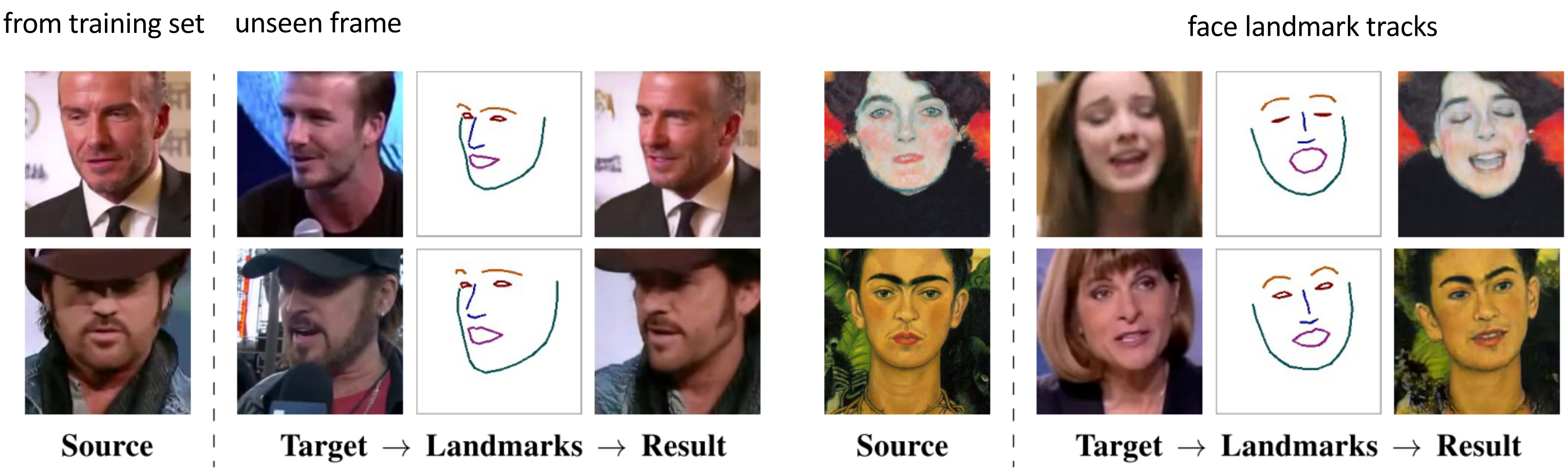
- **Definition**

  - In the context of machine learning, meta learning is the process of **learning to learn**.
  - Informally speaking, a meta learning algorithm uses experience to change certain aspects of a learning algorithm, or the learning method itself, such that the modified learner is better than the original learner at learning from additional experience.

http://www.scholarpedia.org/article/Metalearning

Meta Learning

- Meta Learning Architecture for few-shot learning with generative models

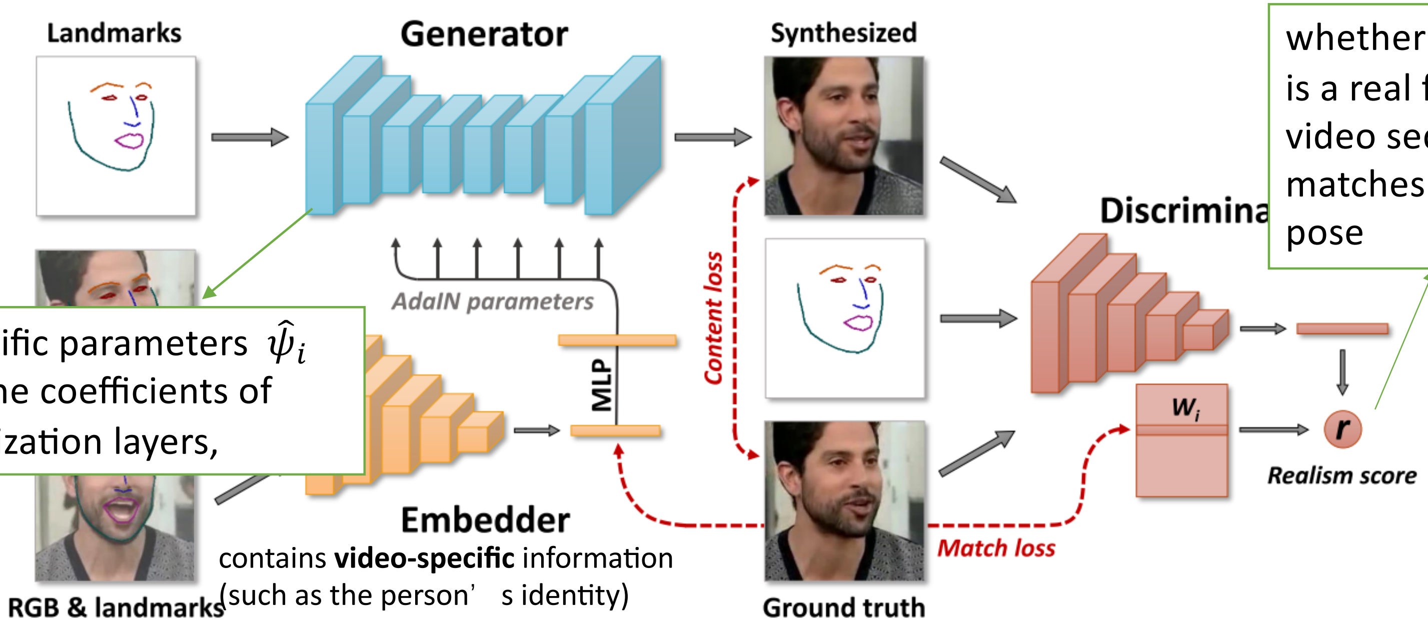


In fact, this system can generate a reasonable result based on a single photograph (one-shot learning), while adding a few more photographs increases the fidelity of personalization

Zakharov E, Shysheya A, Burkov E, et al. Few-shot adversarial learning of realistic neural talking head models[C]

Meta Learning

- Meta Learning Architecture for few-shot learning with generative models



whether the input frame is a real frame of the i-th video sequence & whether it matches the corresponding pose

The person-specific parameters $\hat{\psi}_i$ serve as the affine coefficients of instance normalization layers,

contains **video-specific** information (such as the person' s identity)

- All parameters of the generator are split into two sets: person-generic parameters ψ, and the person-specific parameters $\hat{\psi}_i$.
- During meta-learning, ψ are trained directly, while $\hat{\psi}_i$ are predicted from the embedding vector $\hat{e}_i$ using a trainable projection matrix $\mathbf{P}$: $\hat{\psi}_i = \mathbf{P}\hat{e}_i$

Meta Learning

- Meta Learning Architecture for few-shot learning with generative models

Meta Learning

- Meta Learning Architecture for few-shot learning with generative models

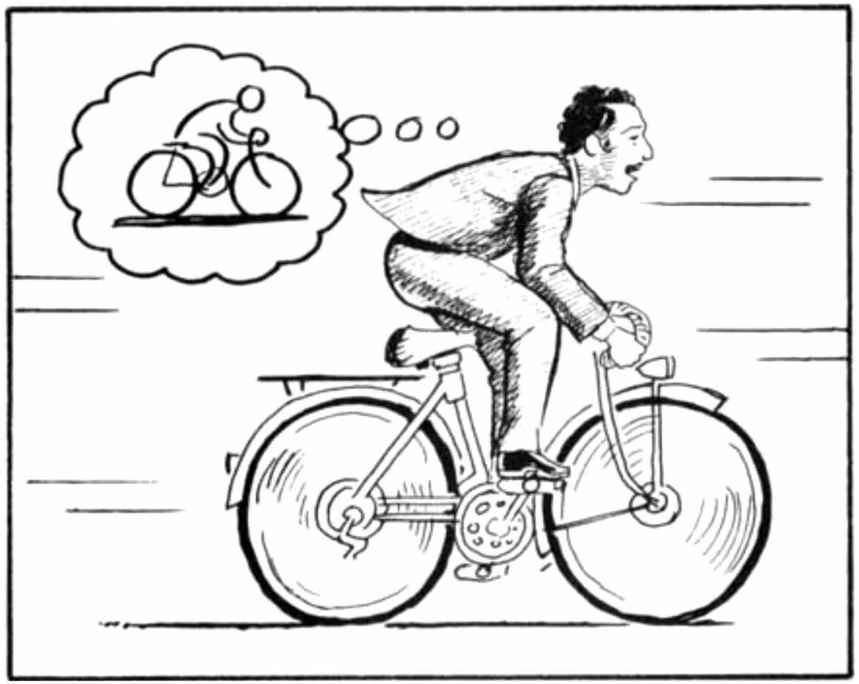
Application of Generative Models: (Selected) Advanced Topics

PEKING UNIVERSITY

- Domain Adaptation
- Adversarial Attack
- Meta Learning
- Reinforcement Learning

Reinforcement Learning

- **World Models**



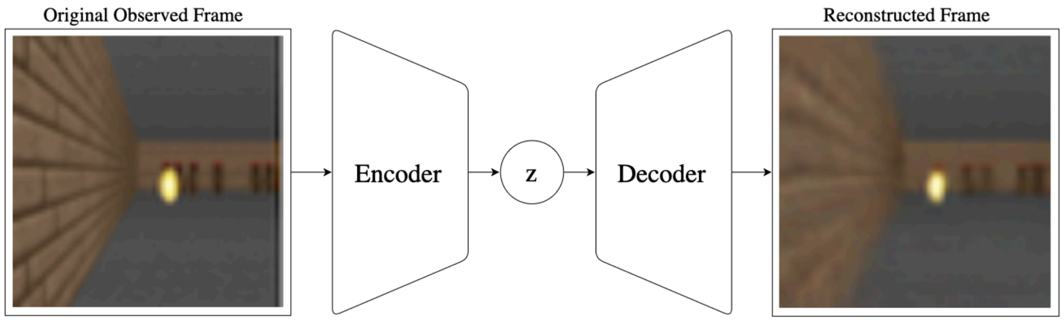
https://worldmodels.github.io

One way of understanding the predictive model inside of our brains is that it might not be about just predicting the future in general, but predicting future sensory data given our **current motor actions**

Learning in the imagination == Sampling efficiency

Reinforcement Learning

- **World Models**

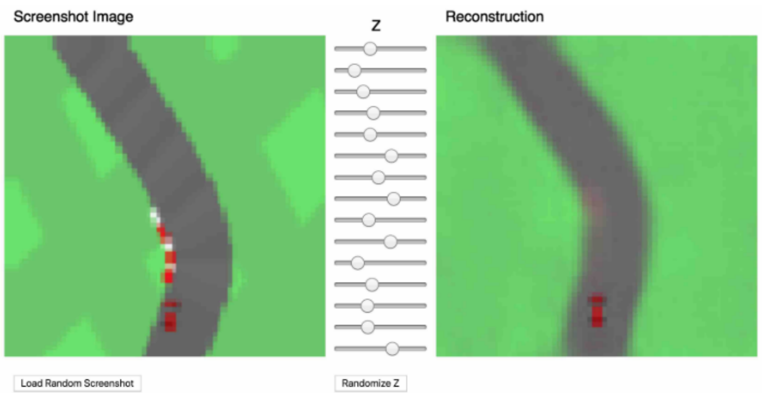


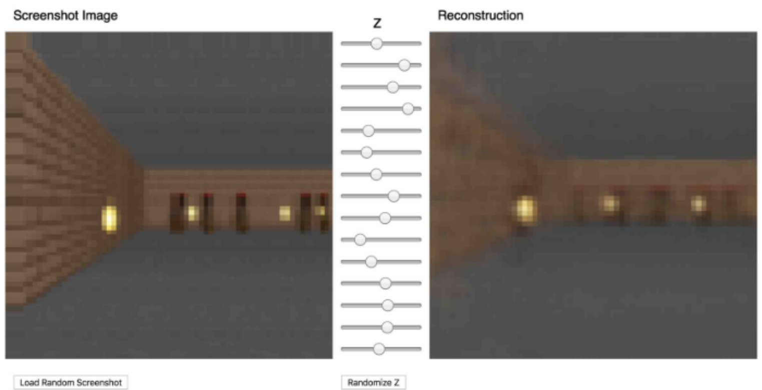
Learn the state representation

Here .. The encoder output is the state

Reinforcement Learning

- **World Models**





In this model, the agent has a visual sensory component that compresses what it sees into a small representative code.

Ha D, Schmidhuber J. World models[J]. arXiv preprint arXiv:1803.10122, 2018.

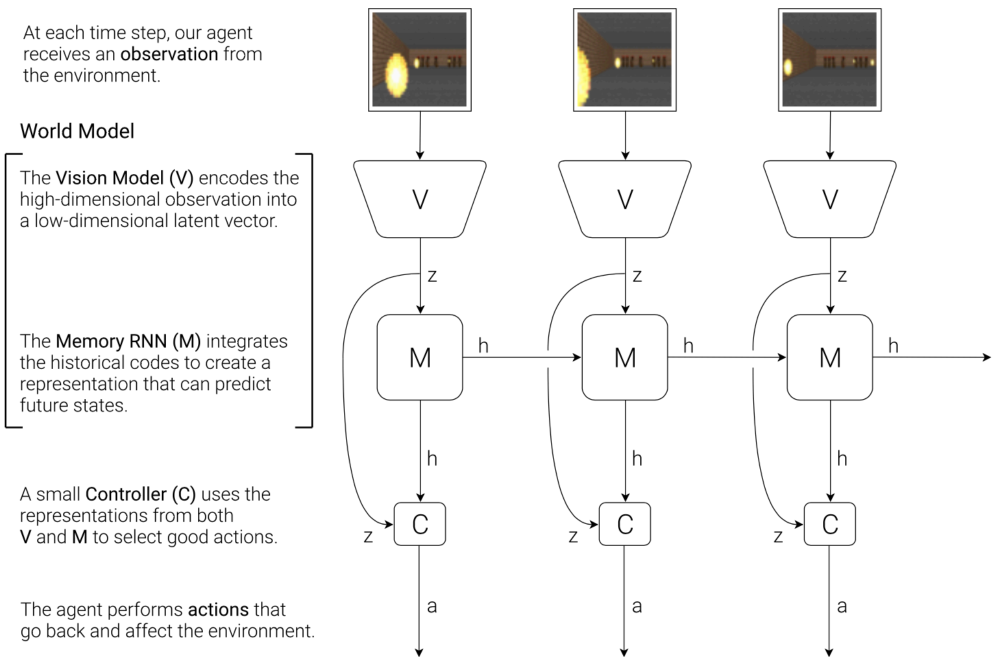
Reinforcement Learning

- **World Models**

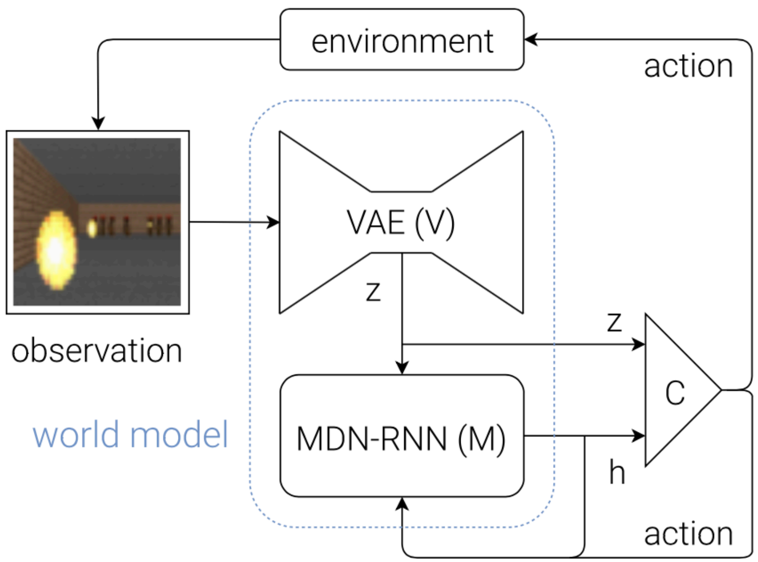
Learn the state representation

Here .. The encoder output is the state

RNN predicts the action



Our agent consists of three components that work closely together: Vision (V), Memory (M), and Controller (C).



Summary

- Domain Adaptation
- Adversarial Attack
- Meta Learning
- Imitation Learning
- Reinforcement Learning

北京大學
PEKING UNIVERSITY

Thanks