The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering

Spring 2021

ADVANCED TOPICS IN COMPUTER VISION

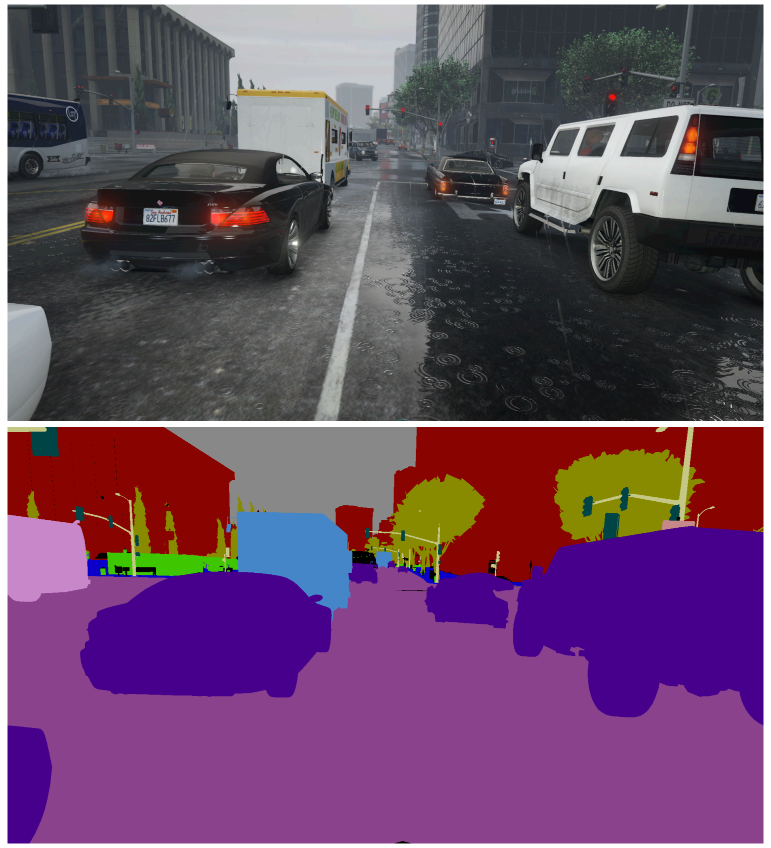
Atlas Wang
Assistant Professor, The University of Texas at Austin

Visual Informatics Group@UT Austin
https://vita-group.github.io/

Why Synthetic Training

- Collections of real data are costly
  - Massive real image
  - Classification / Segmentation / Detection
- Synthetic data are relatively cheap to generate

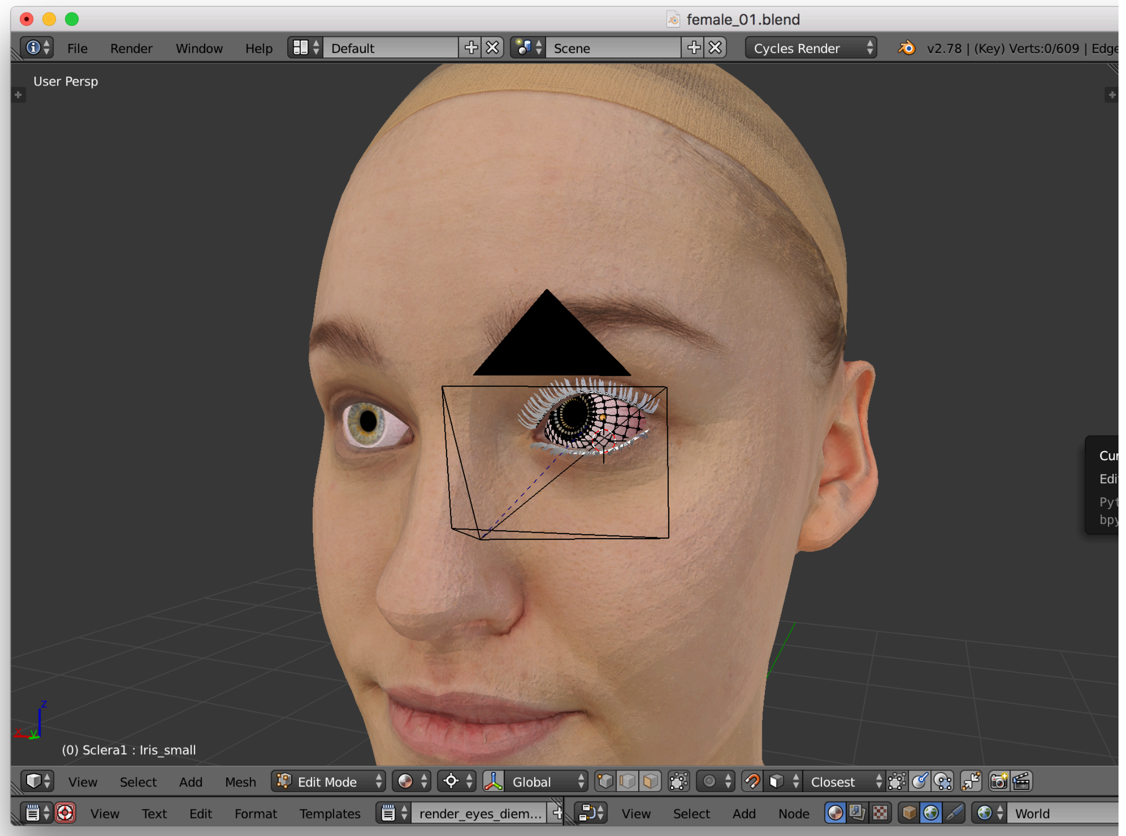
Cityscapes (3K annotations)



GTA5 (24,966 annotations)

Why Synthetic Training

- In some cases, synthetic data is all you have...
- EyeGaze / Depth / Flow / 3D Mesh reconstruction / Robotics

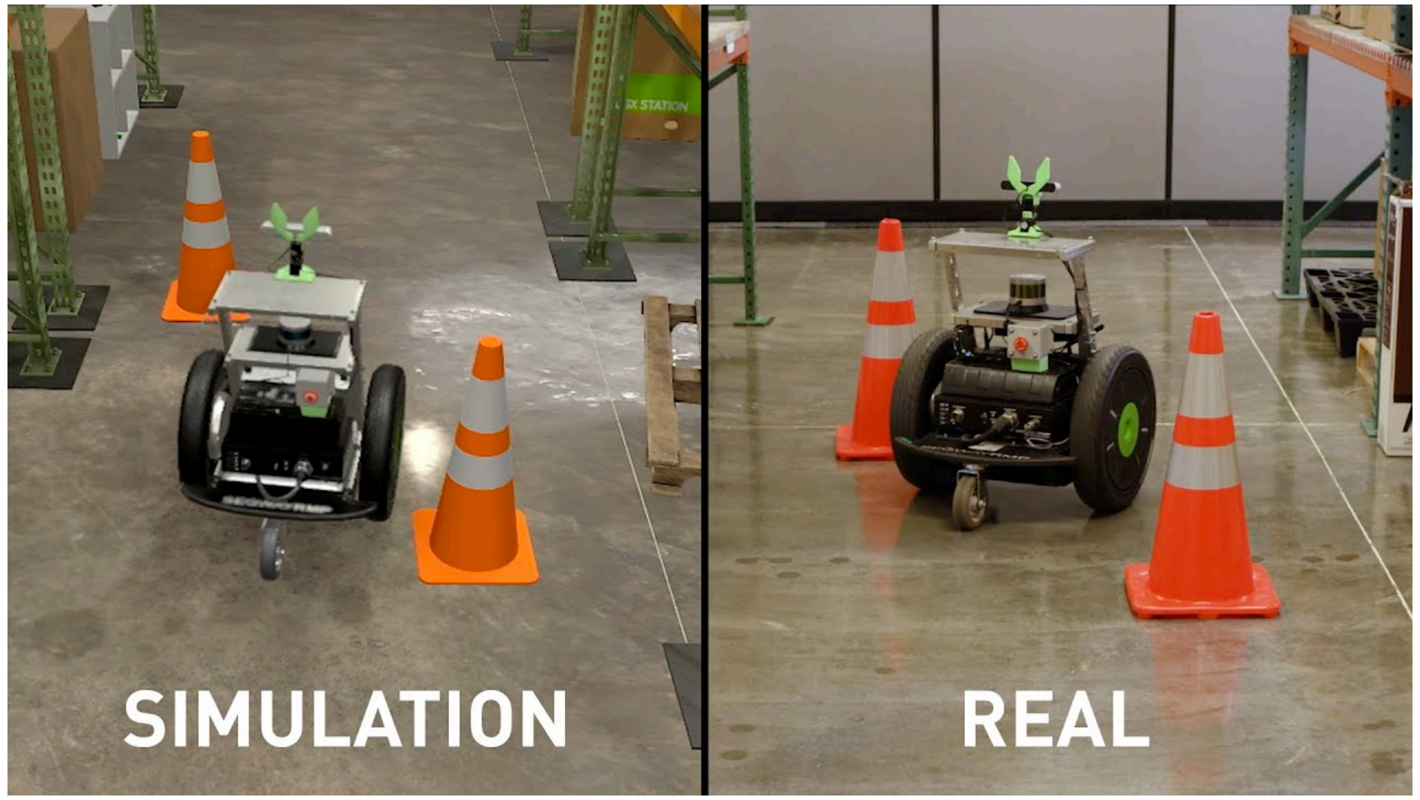


Wood et al. ICCV 2015

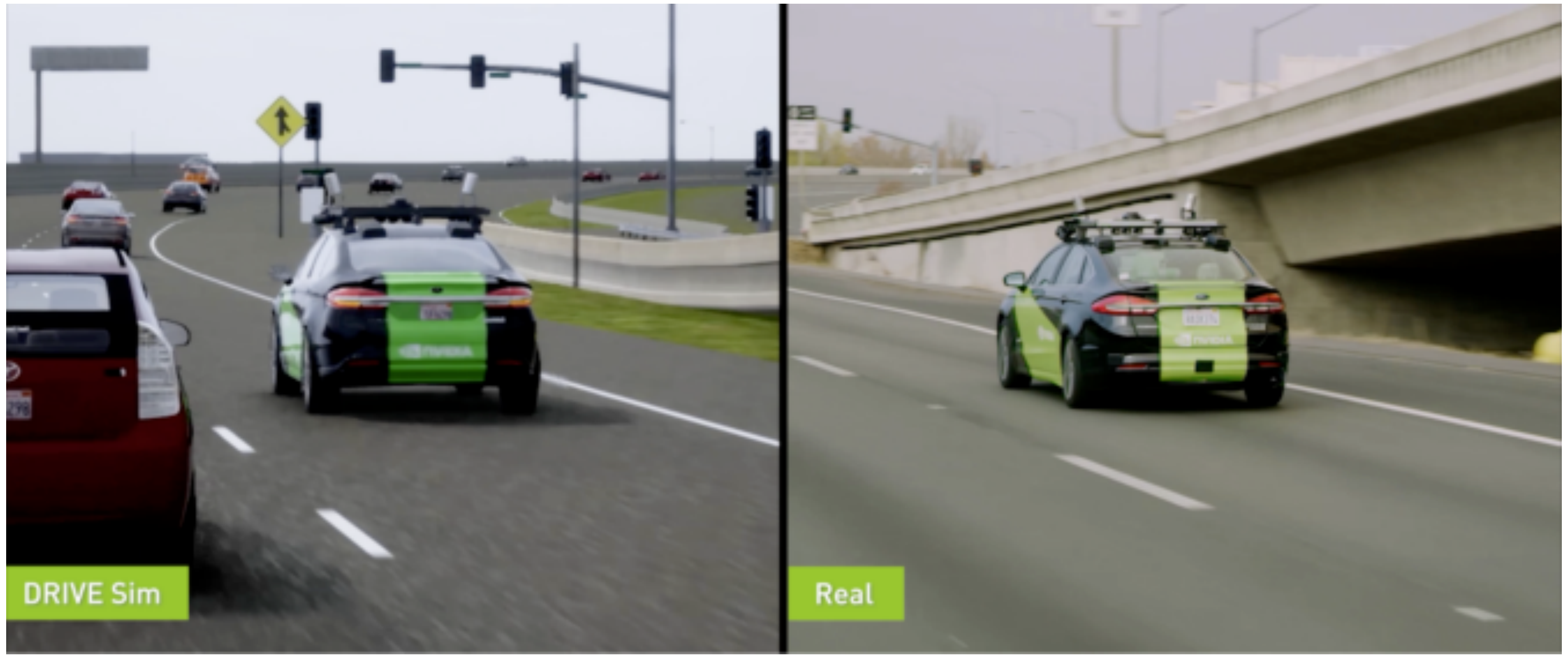
Habitat (Facebook)

Synthetic Simulation Empowers Some Most Important Applications

- Autonomous Driving: Omniverse, ISAAC, DRIVE Sim, etc.



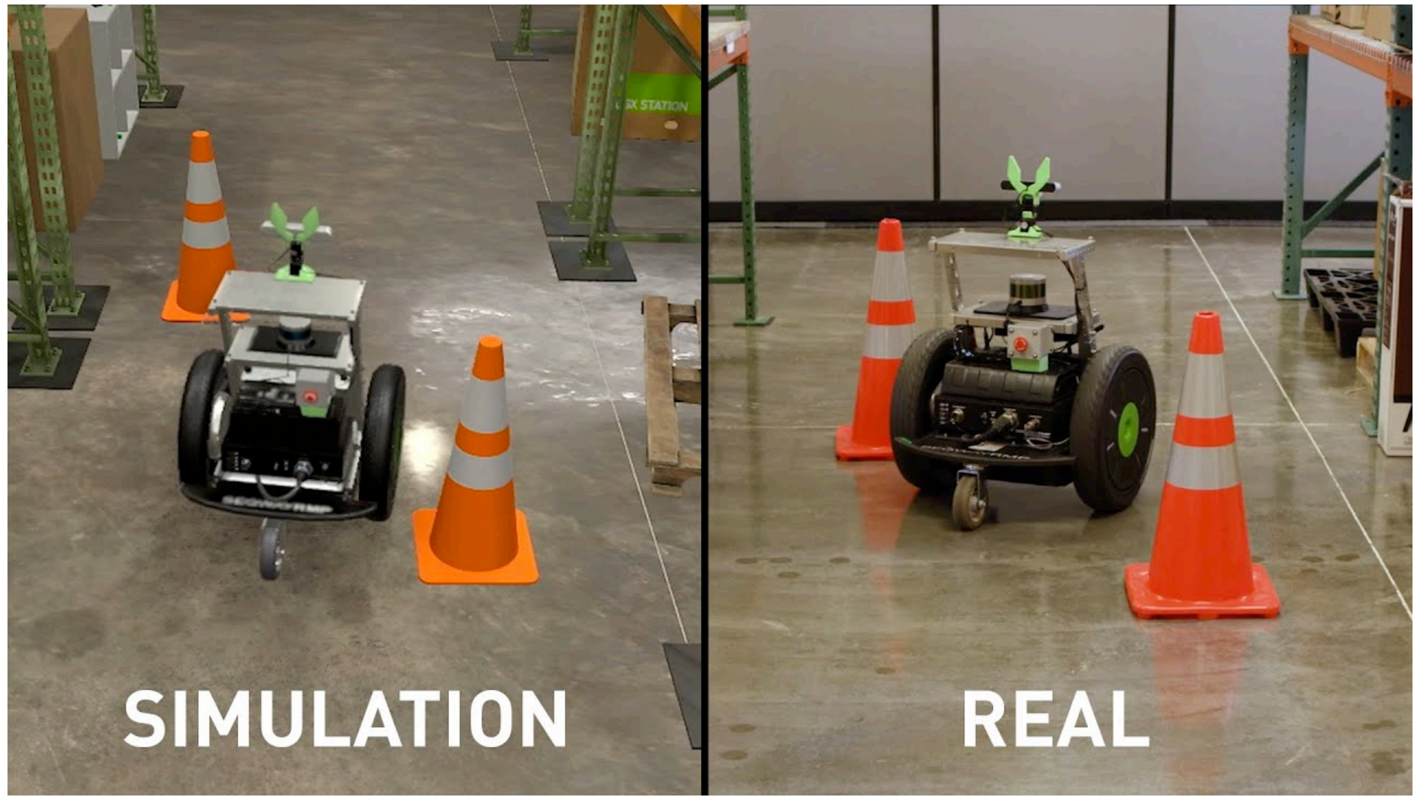
ISAAC platform



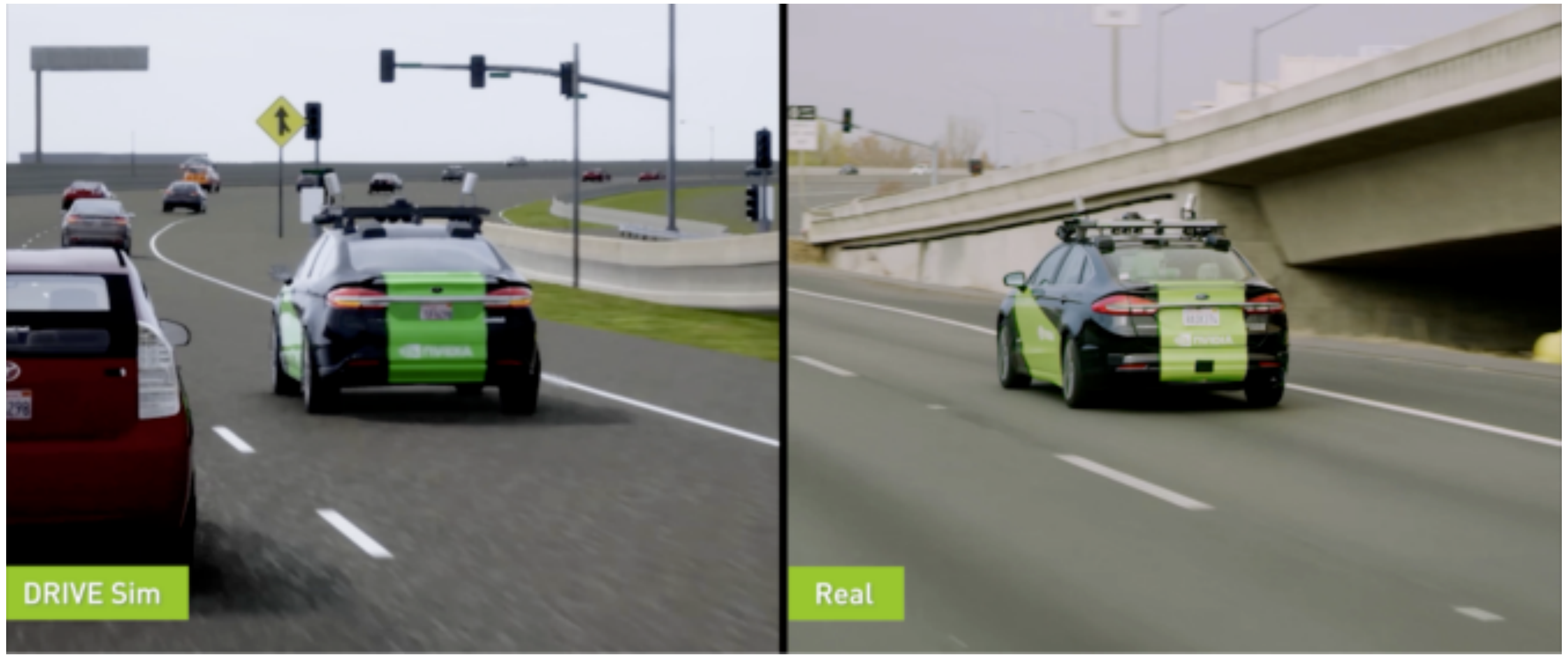
DRIVE Sim

Synthetic Simulation Empowers Some Most Important Applications

- Autonomous Driving: Omniverse, ISAAC, DRIVE Sim, etc.



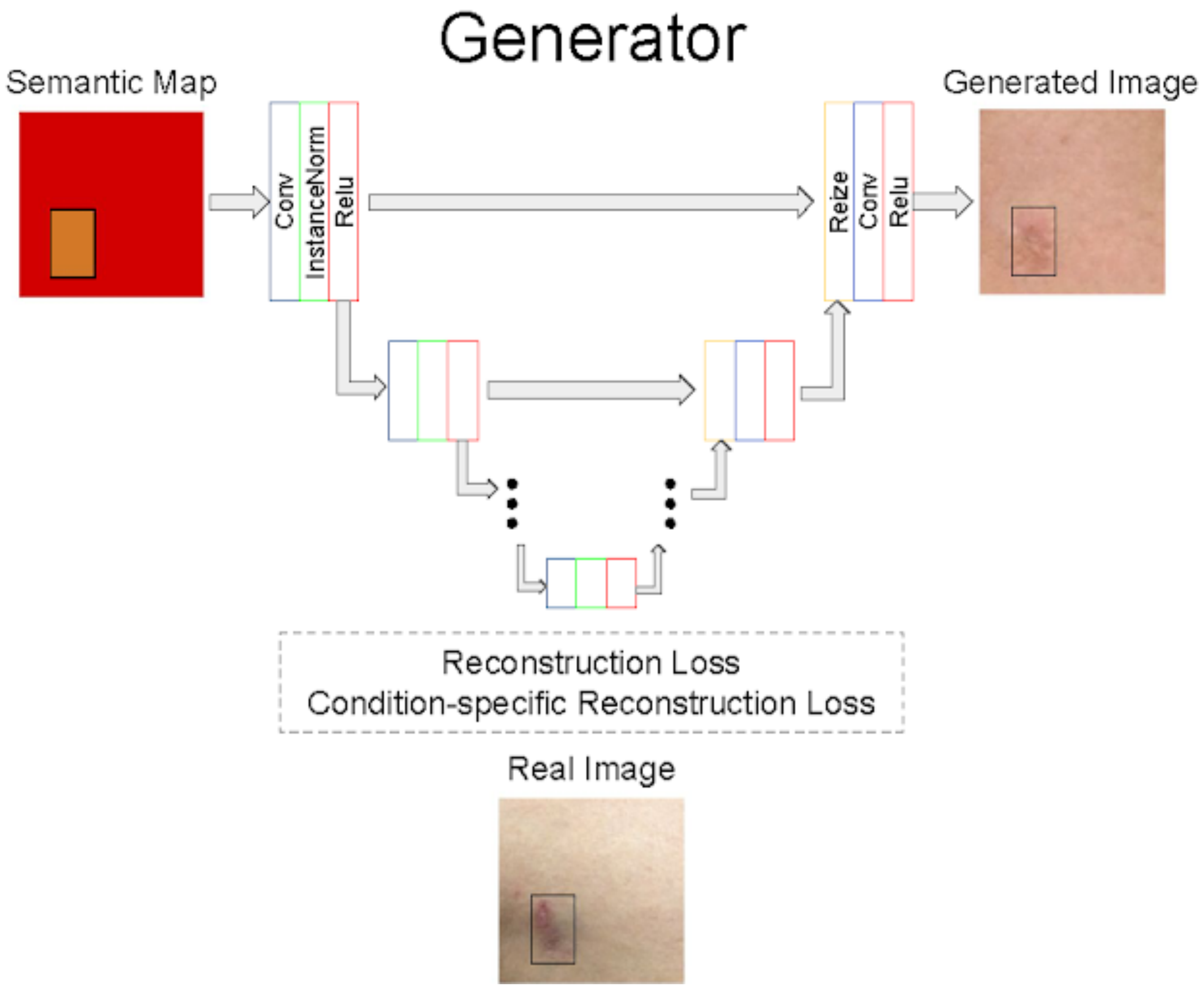
ISAAC platform

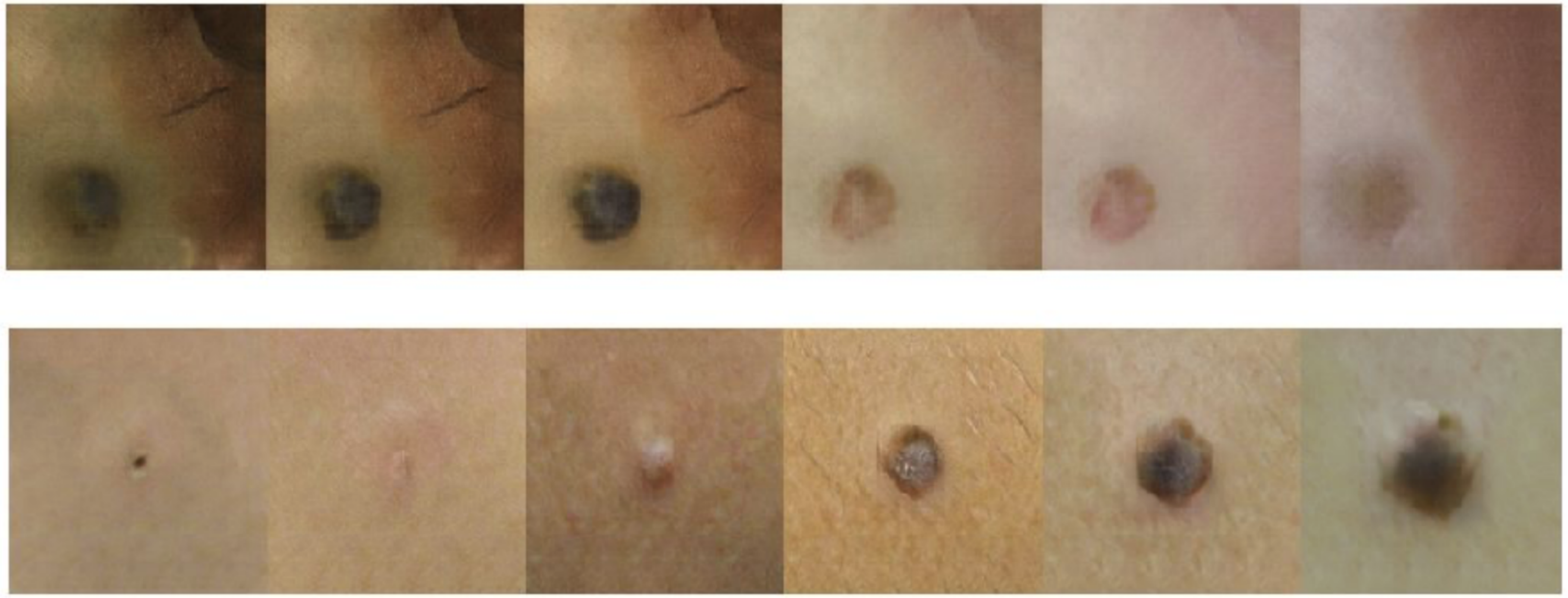


DRIVE Sim

Synthetic Simulation Empowers Some Most Important Applications

- Medical Image Analysis: cover more corner cases, resolve privacy concerns…



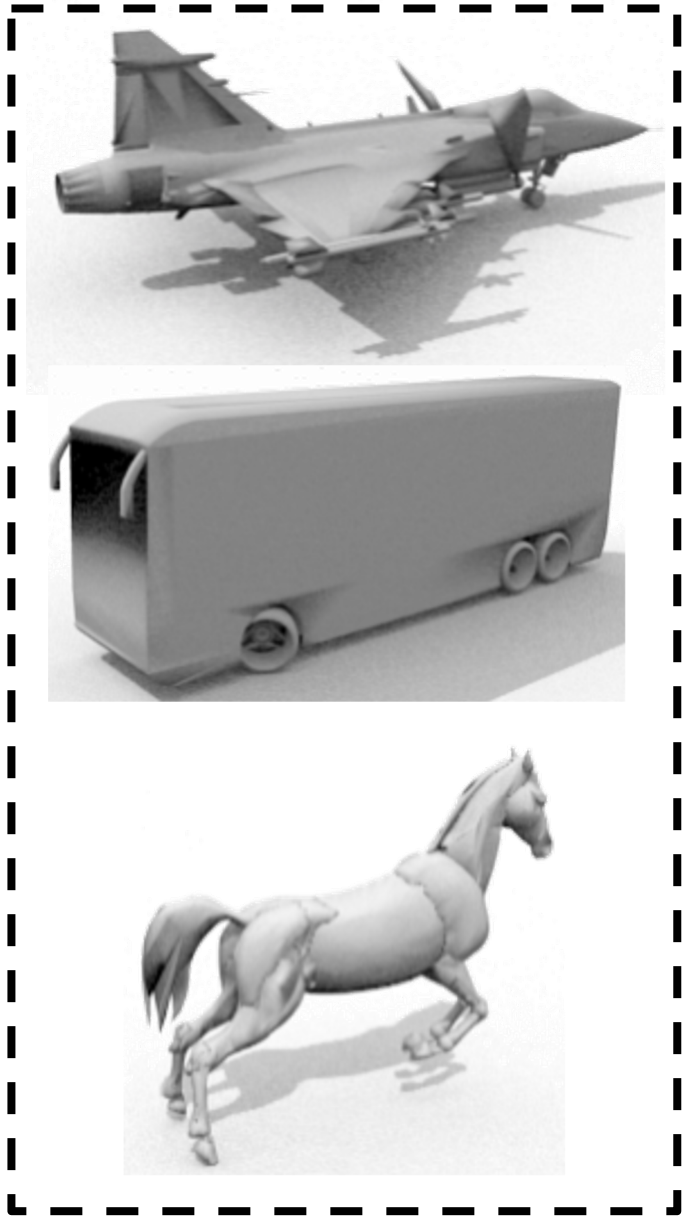


Challenging Domain Gap: Synthetic vs Real

ImageNet Pretrained Model

Fine-tune

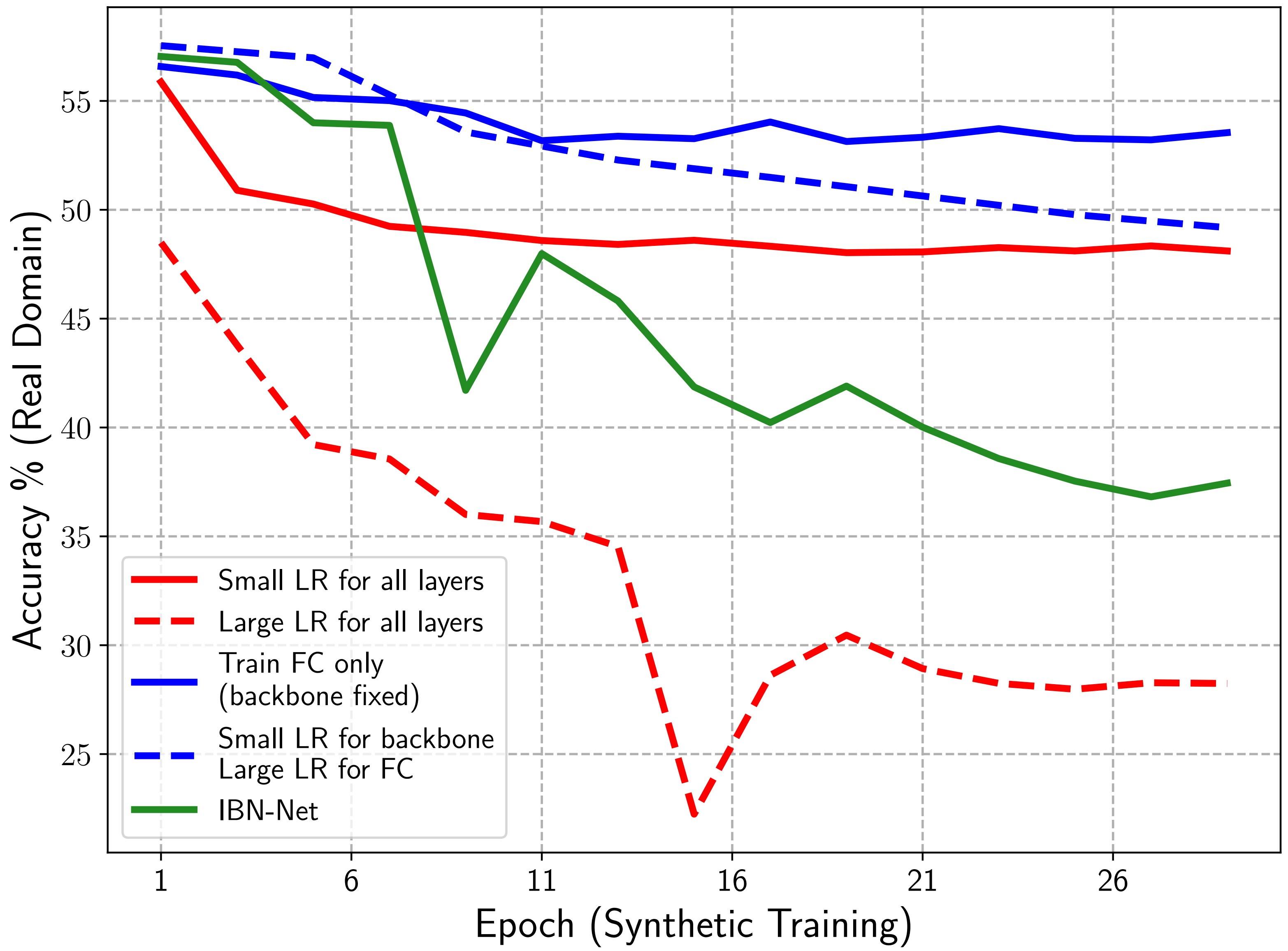
Supervised Training (Synthetic Data)



Visda17

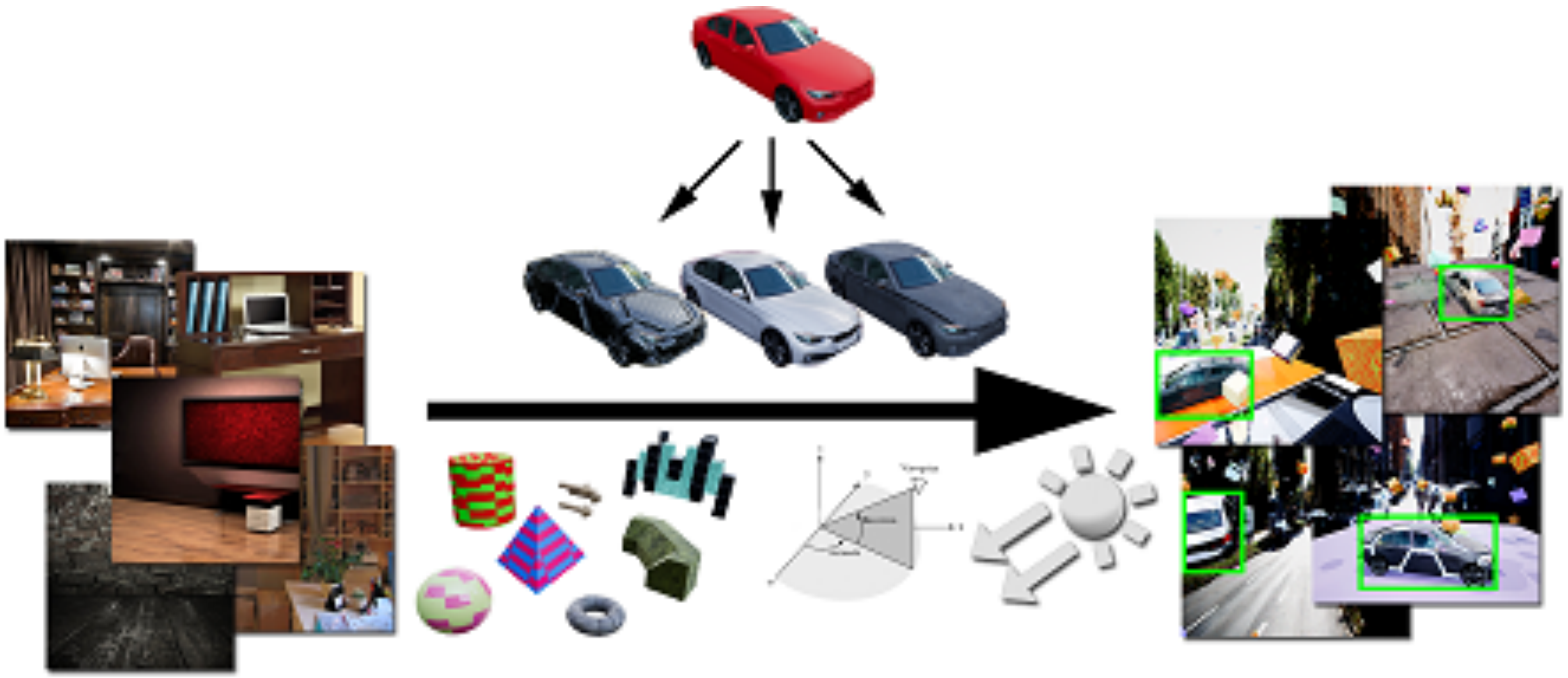
Unseen Test (Real Data)

COCO



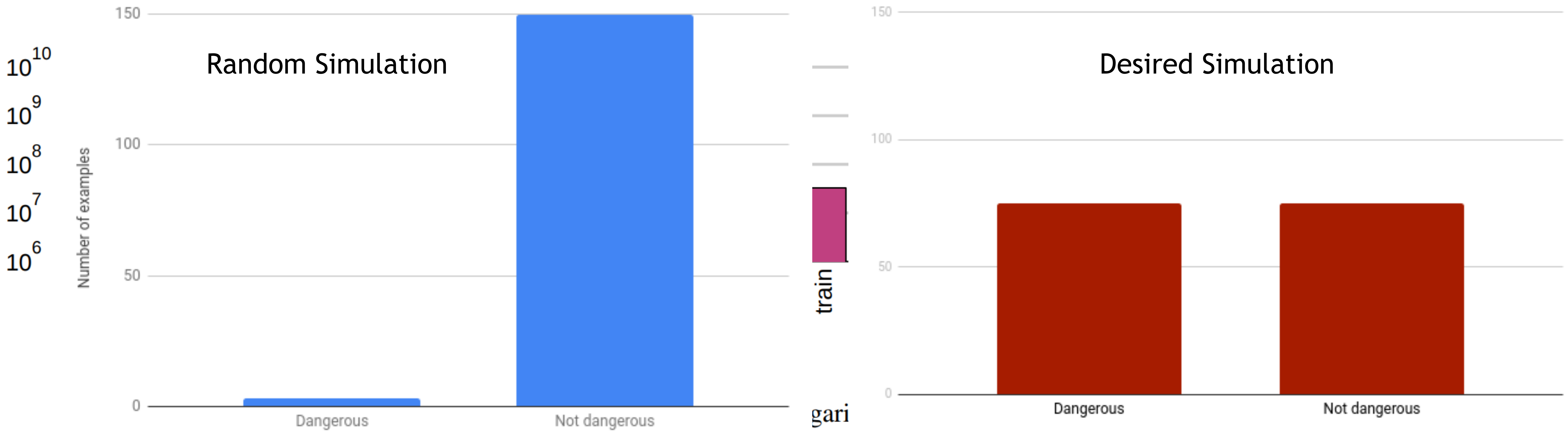
Domain Randomization (IROS'17)

- To handle the variability in real-world data, the simulator parameters (lighting, pose, object textures, etc) are randomized in non-realistic ways to force the learning of essential diverse features.



Can We Do Better than Random?

- Learn to simulate better data for a particular downstream task?
- Learn to simulate edge cases?



Learning to Simulate (ICLR'19)

➢ **We want to solve the following bi-level optimization problem.**

Simulation parameters

Loss of main task model trained in simulation and evaluated on real data

$$\boxed{\psi^*} = \arg\min_{\psi} \boxed{\sum_{(\boldsymbol{x},\boldsymbol{y})\in D_{\text{val}}} \mathcal{L}\left(y, h_{\boldsymbol{\theta}}(\boldsymbol{x}; \boldsymbol{\theta}^*(\psi))\right)} \longrightarrow \text{meta-learner}$$

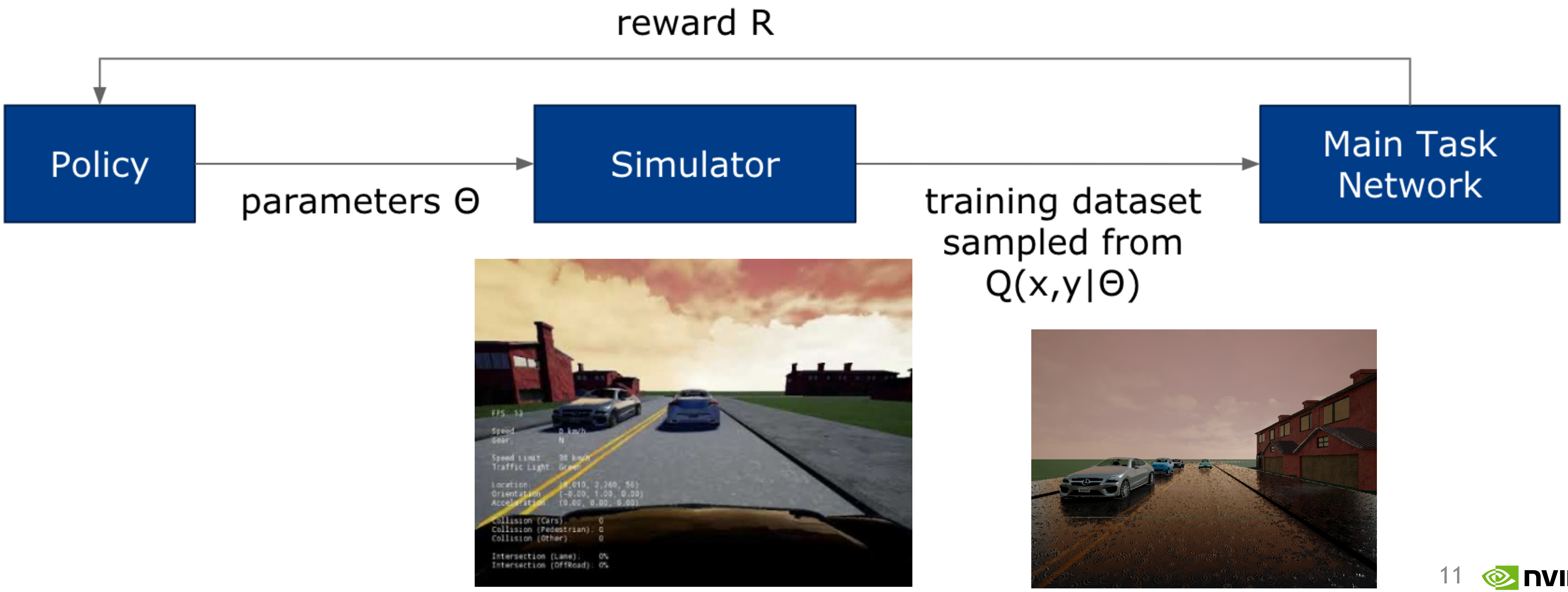
$$\text{s.t.} \quad \boxed{\boldsymbol{\theta}^*(\psi)} = \arg\min_{\boldsymbol{\theta}} \sum_{(\boldsymbol{x},\boldsymbol{y})\in \boxed{D_{q(\mathbf{x},\mathbf{y}|\,\psi)}}} \mathcal{L}\left(\boldsymbol{y}, h_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta})\right), \longrightarrow \text{main task model } h_{\boldsymbol{\theta}} \text{ trained on simulated data}$$

Main task model parameters

Dataset generated by simulator

Learning to Simulate (ICLR'19)

- Train the policy of selecting simulator parameters, using policy gradient, since the simulator is often non-differentiable



Are better simulators enough?

Models overfit to any difference

High quality is expensive

Virtual KITTI Dataset

Multi-object tracking accuracy:

Sim: 63.7%

Real: 78.1%

Virtual Worlds as Proxy for Multi-Object Tracking Analysis
[Gaidon*, Wang*, Cabon, Vig, 2016]

Jungle Book:

30M render hours

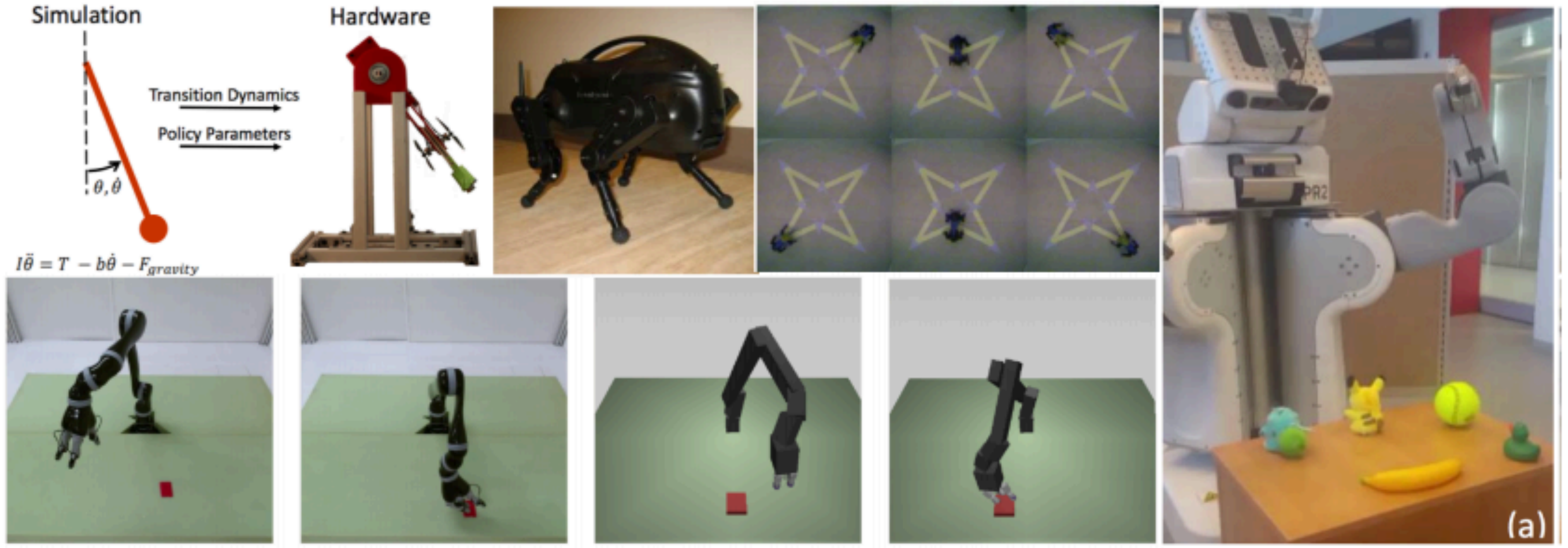
19 hours per frame

800 artist-years of effort

Jungle Book, 2016

Supervised domain adaptation

Fine-tuning



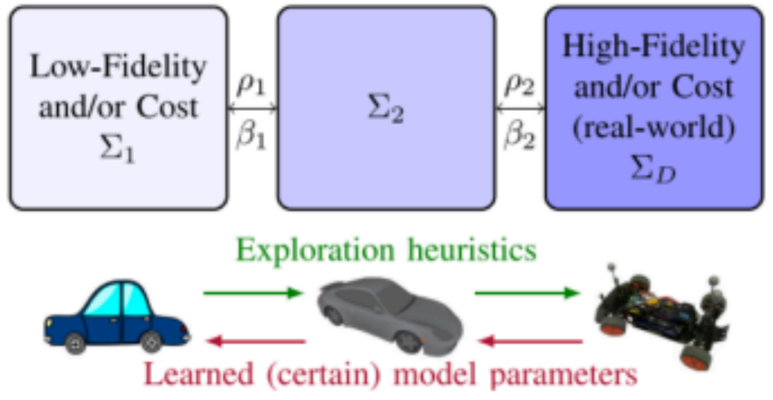
Learning Omnidirectional Path Following Using Dimensionality Reduction [Kolter, Ng, 2003]

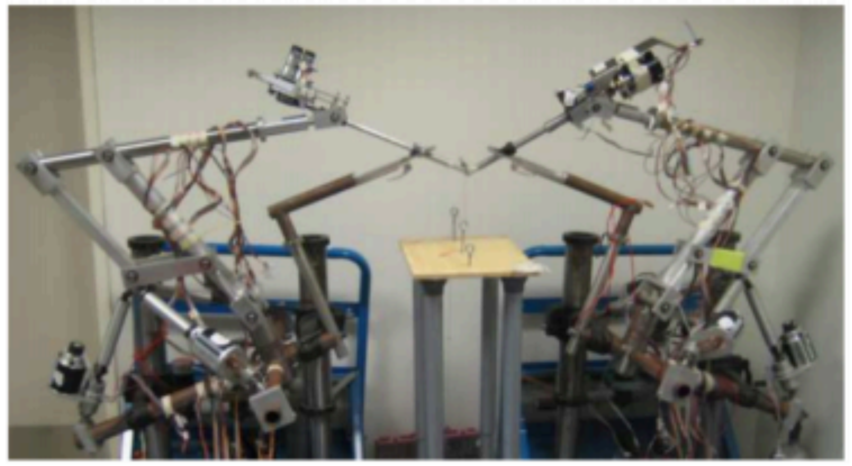
Efficient Reinforcement Learning for Robotics using Informative Simulated Priors [Cutler, How, 2015]

Sim-to-Real Robot Learning from Pixels with Progressive Nets [Rusu et al. 2016]

Deep Predictive Policy Training using Reinforcement Learning [Ghadirzadeh, Maki, Kragic, Bjorkman, 2017]

Iterative learning control





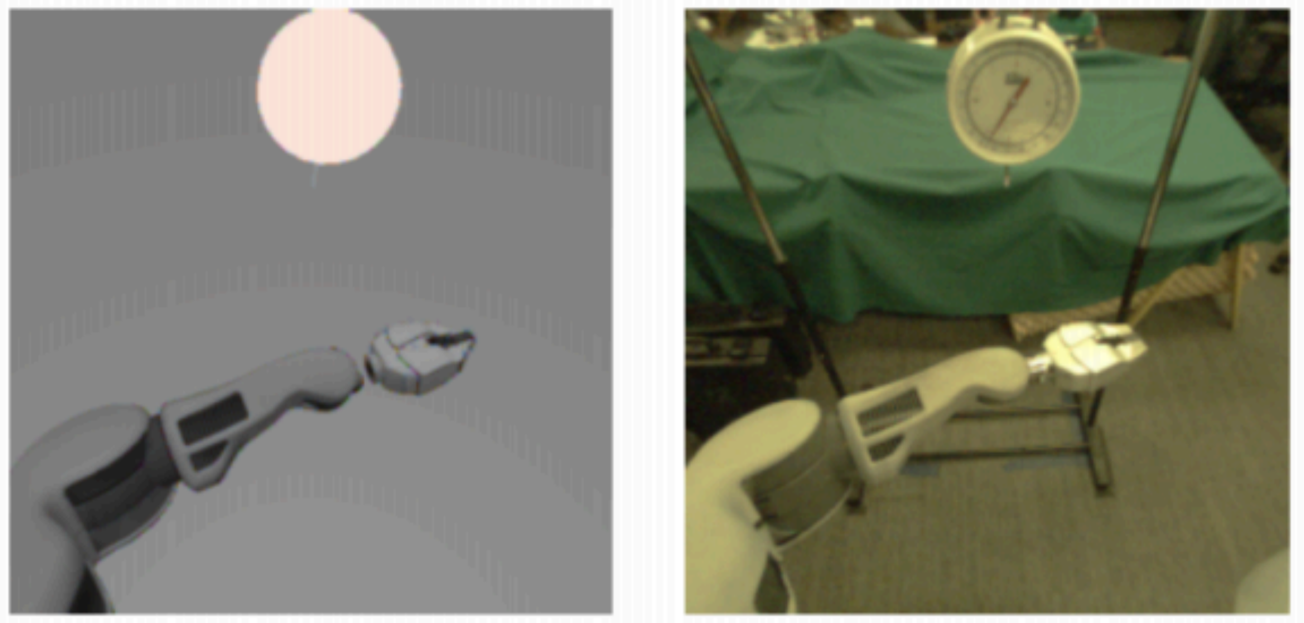
Using inaccurate models in reinforcement learning [Abbeel, Quigley, Ng, 2006]

Reinforcement learning with multi-fidelity simulators [Cutler, Walsh, How 2014]

Superhuman performance of surgical tasks by robots using iterative learning from human-guided demonstrations [Van Den Berg, Miller, Duckworth, Hu, Wan, Fu, Goldberg, Abbeel, 2010]

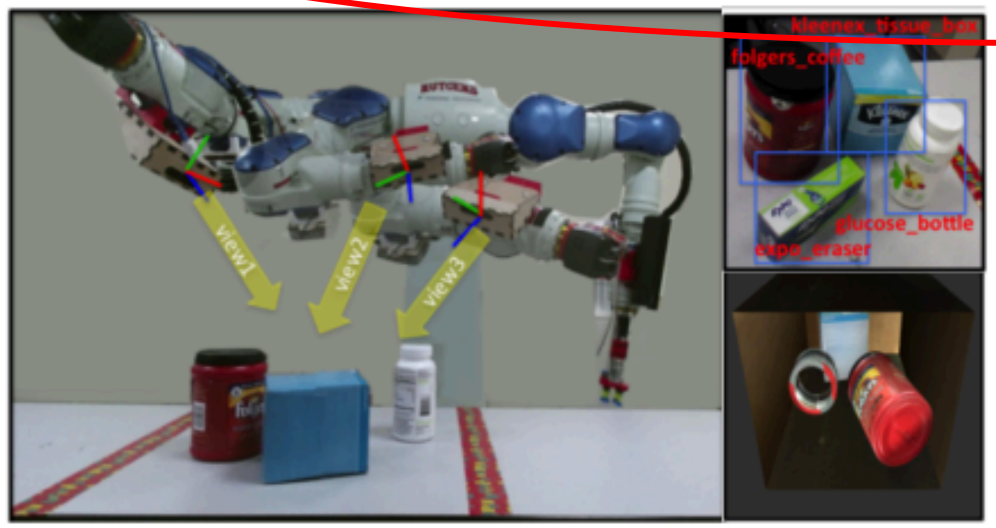
(Less) supervised domain adaptation

Weakly Supervised



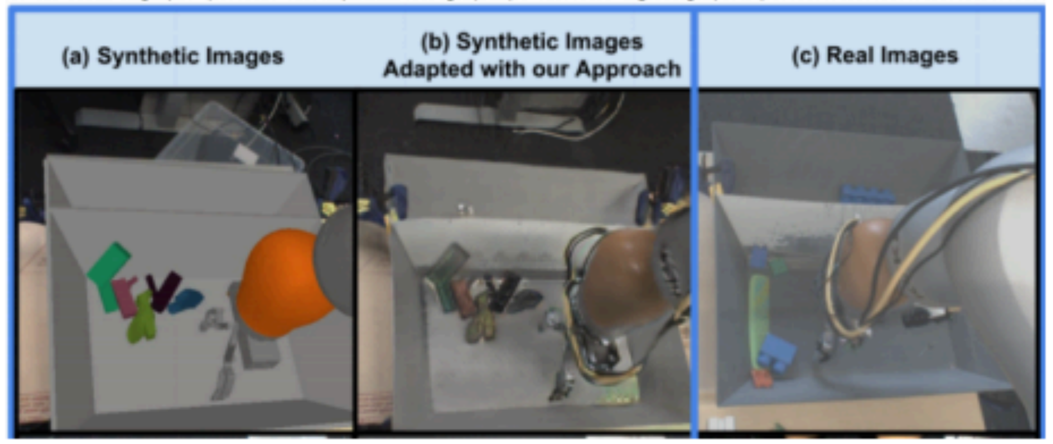
Adapting Deep Visuomotor Representations with Weak Pairwise Constraints [Tzeng, Devin, Hoffman, Finn, Abbeel, Levine, Saenko, Darrell, 2016]

Self-Supervised



A Self-supervised Learning System for Object Detection using Physics Simulation and Multi-view Pose Estimation [Mitash, Bekris, Boularias, 2017]

Unsupervised



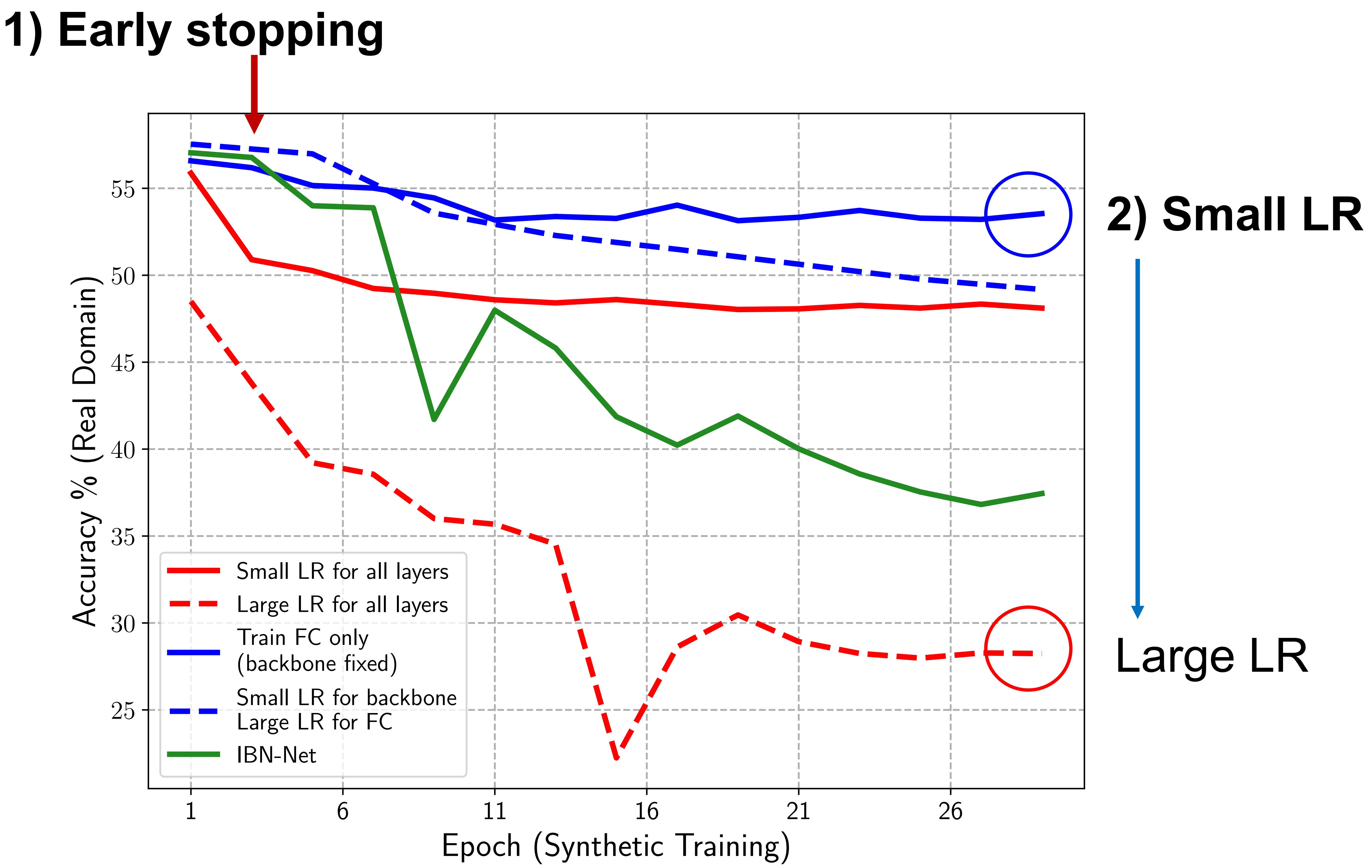
CyCADA [Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrel, 2017]
Using Simulation and Domain Adaptation to Improve Efficiency of Deep Robotic Grasping [Bousmalis et al., 2017]

Automated Synthetic-to-Real Generalization

ICML 2020

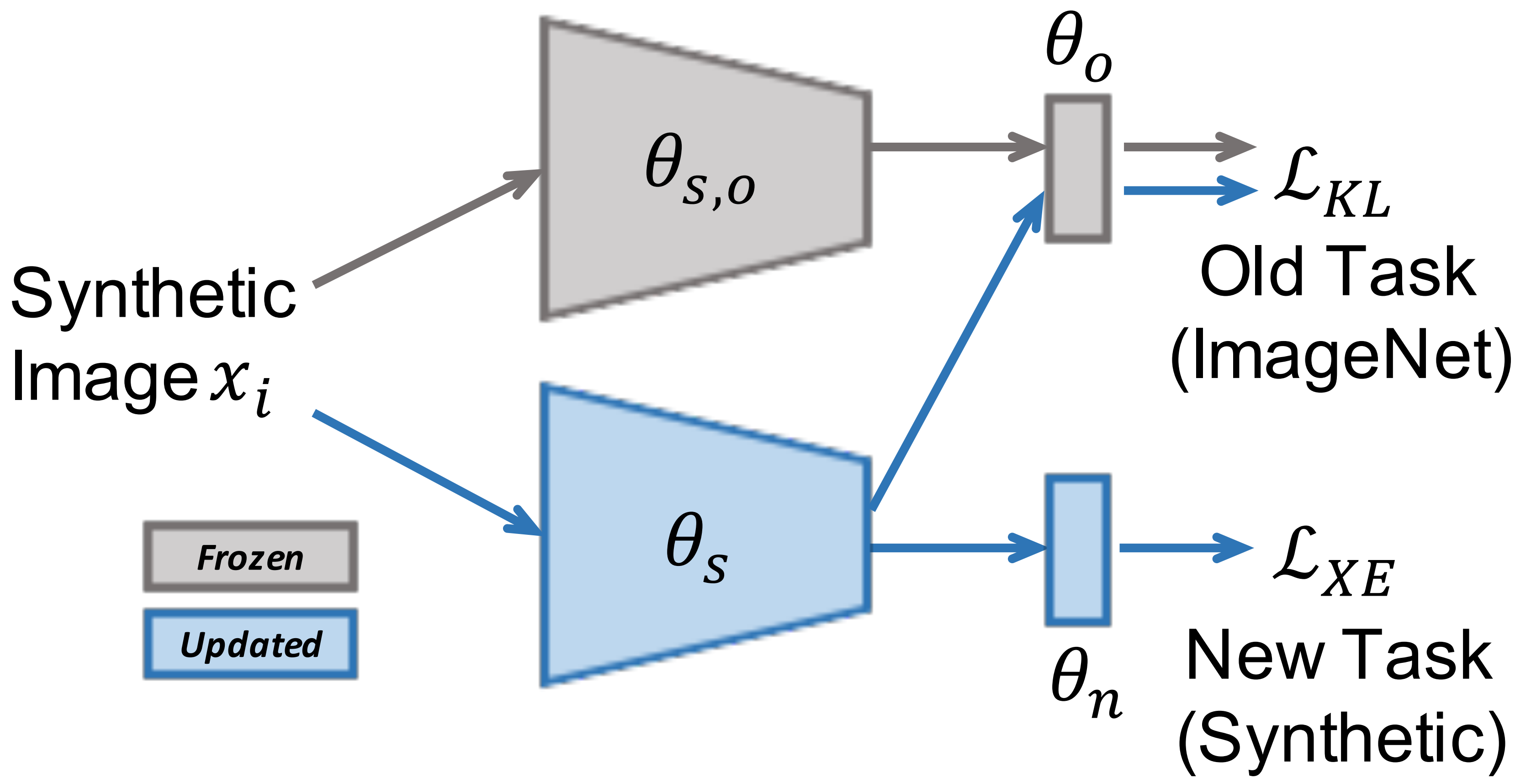
Wuyang Chen, Zhiding Yu, Zhangyang “Atlas” Wang, Anima Anandkumar

Previous solutions: Heuristic Hand-tuning



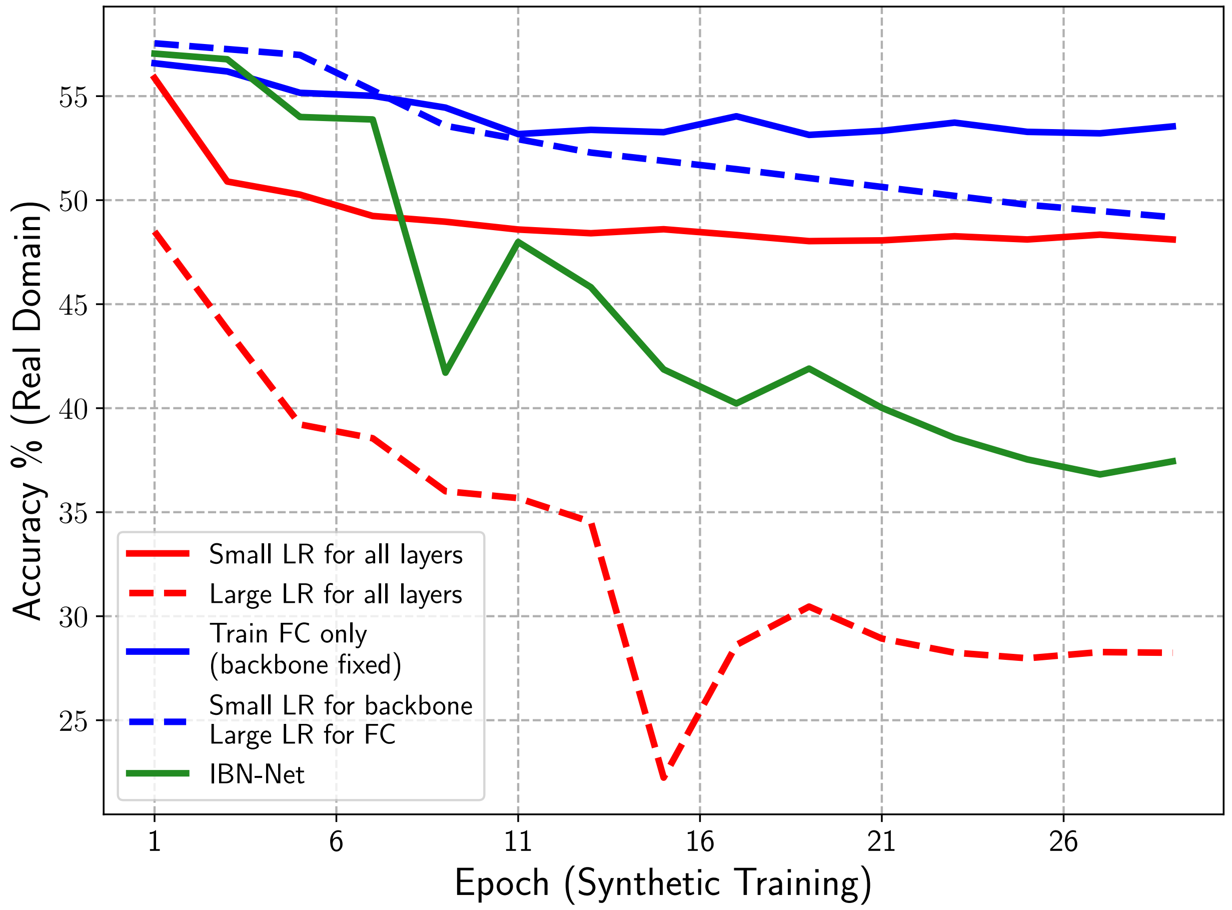
ImageNet as Proxy Guidance

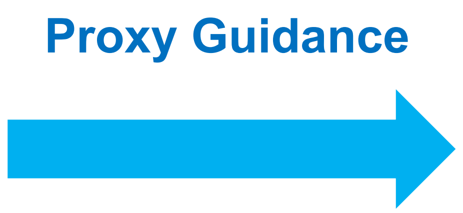
- Why early stopping?
  - ➔ Keep weights close to ImageNet initialization.
- We minimize $\mathcal{L}_{KL}$: new model vs ImageNet initialization
  - ➔ ImageNet as proxy guidance in syn2real training.

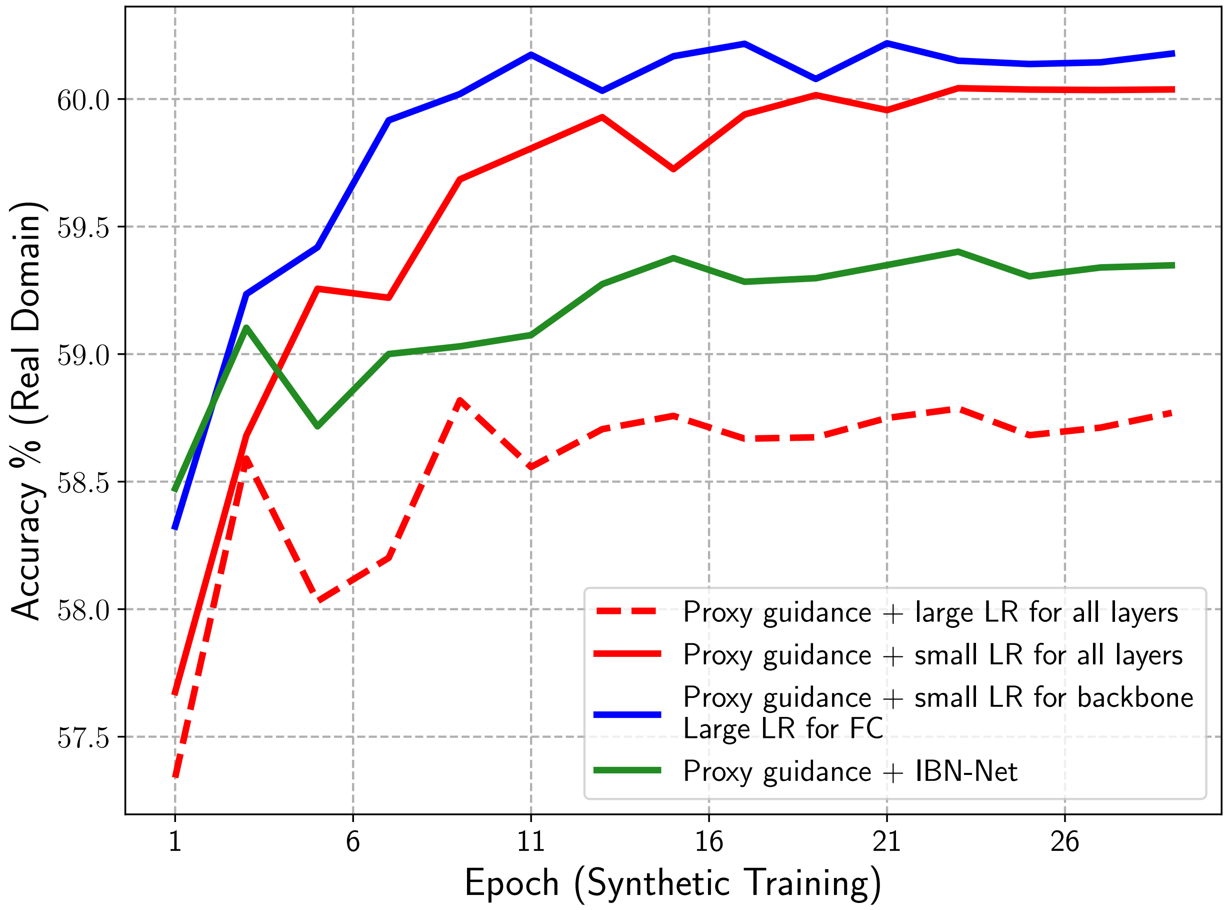


ImageNet as Proxy Guidance

- Why early stopping?
  - ➔ Keep weights close to ImageNet initialization.
- We minimize $\mathcal{L}_{KL}$: new model vs ImageNet initialization
  - ➔ ImageNet as proxy guidance in syn2real training.

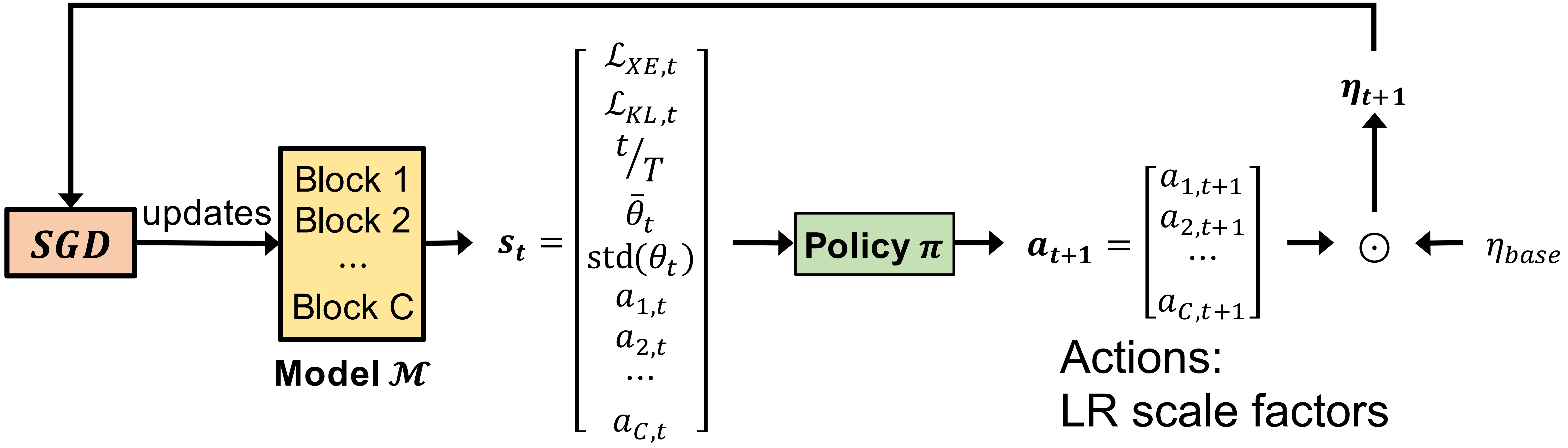






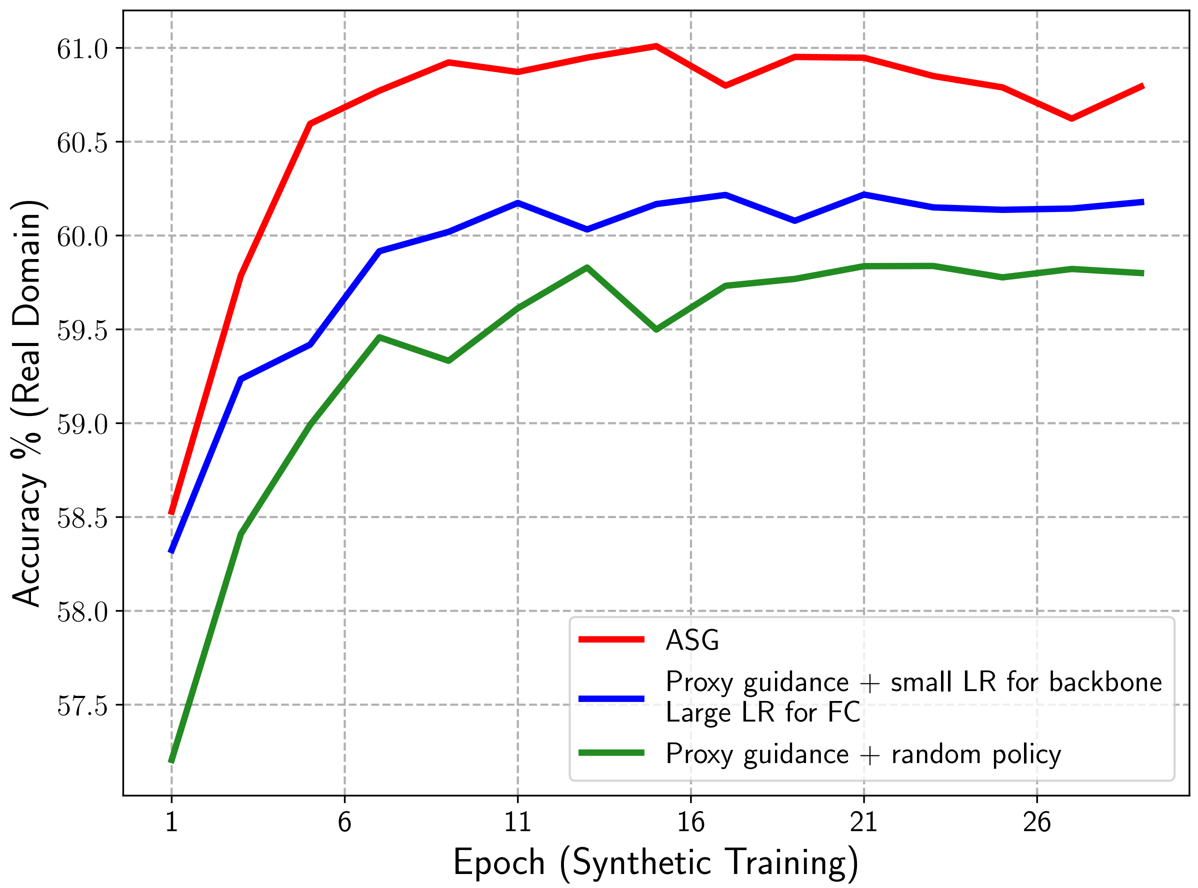
L2O: automatic control of layer-wise learning rate

- Why small learning rate?
  - ➔ Keep weights close to ImageNet initialization.
- But how small for which layer?
  - ➔ L2O (learning-to-optimize): automatic control of layer-wise learning rate

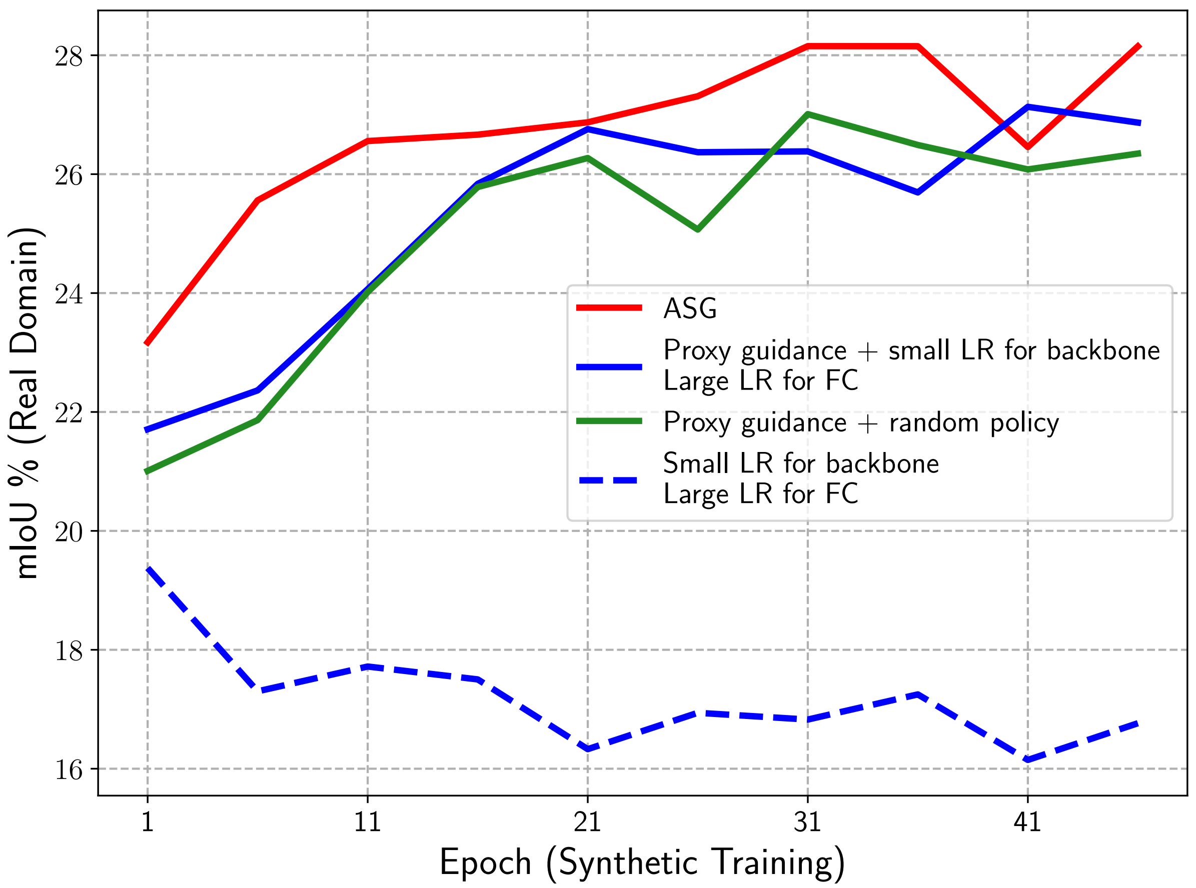


Automated Synthetic-to-Real Generalization (ASG)

- Why small learning rate?
  - ➔ Keep weights close to ImageNet initialization
- But how small for which layer?
  - ➔ L2O (learning-to-optimize): automatic control of layer-wise learning rate



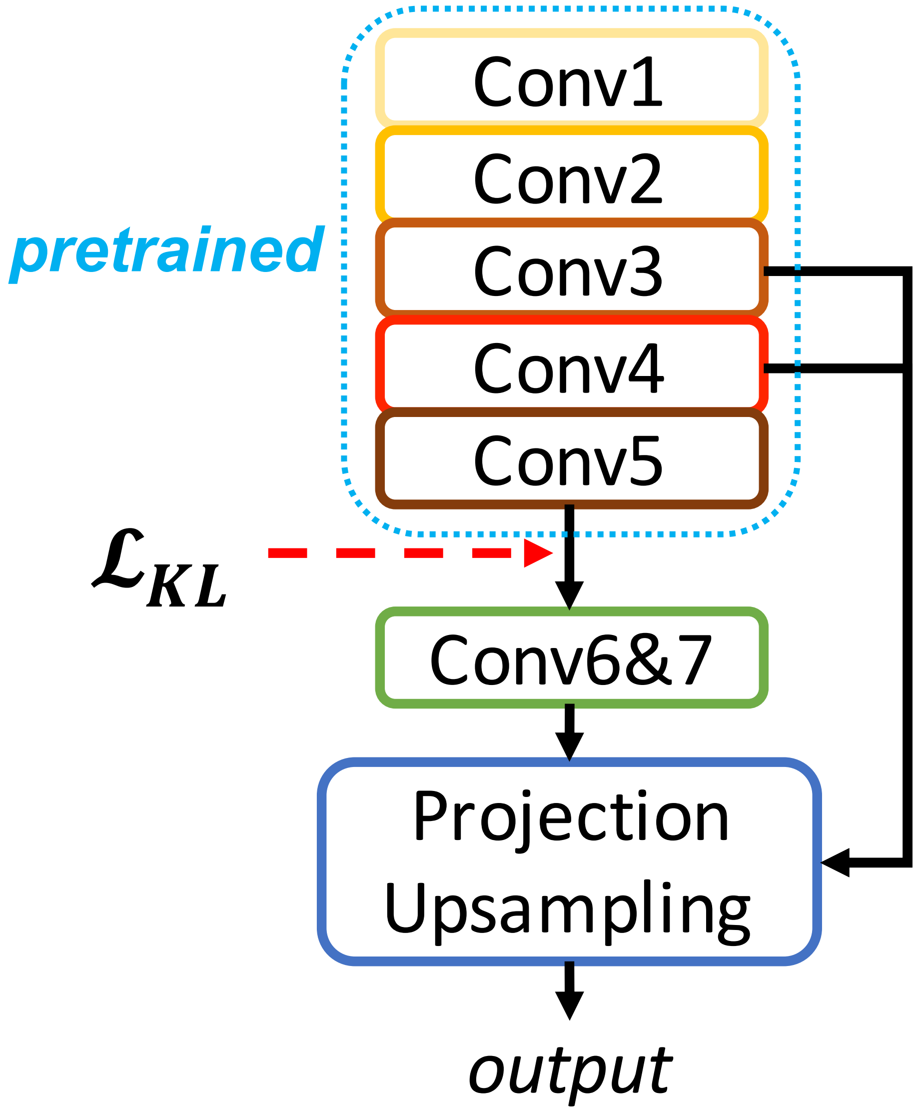
Visda17 ➔ COCO



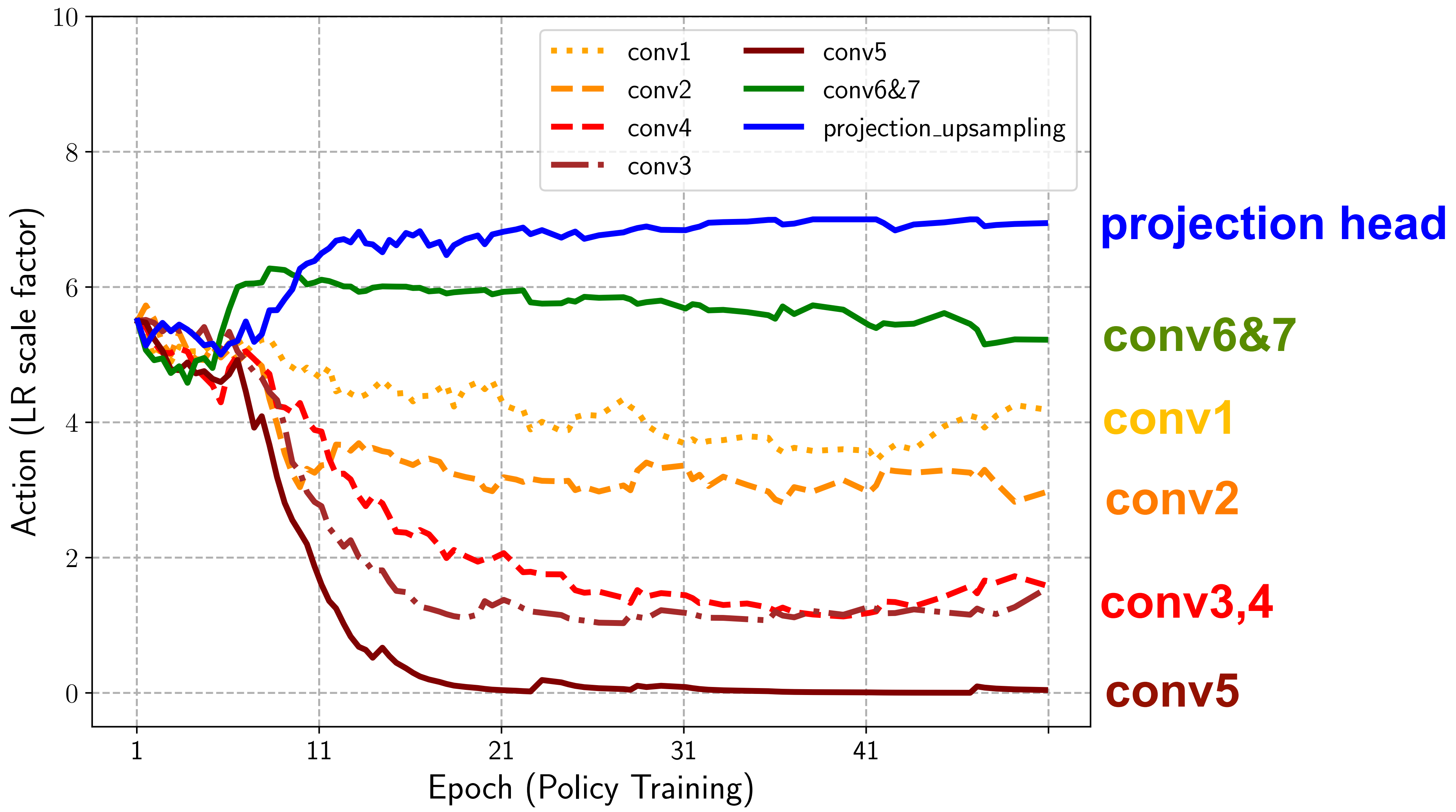
GTA5 ➔ Cityscapes

Action Behavior of RL-L2O Policy

- Backbone (ImageNet pretrained): closer to $\mathcal{L}_{KL}$ → smaller LR
- Projection head: large LR



FCN-Vgg16



Why ASG Works? Retaining ImageNet Information

#	Model	Visda-17	ImageNet
1.	Large LR for all layers	28.2	0.8
2.	+ our Proxy Guidance	58.7 (+30.5)	76.2 (+75.4)
3.	Small LR for backbone and large LR for FC	49.3	33.1
4.	+ our Proxy Guidance	60.2 (+10.9)	76.5 (+43.4)
5.	Oracle on ImageNet[2]	53.3 (+4.0)	**77.4**
6.	ROAD (Chen et al., 2018)	57.1 (+7.8)	**77.4**
7.	Vanilla L2 distance	56.4 (+7.1)	49.1
8.	SI (Zenke et al., 2017)	57.6 (+8.3)	53.9
9.	ASG (ours)	**61.5**	76.7

ASG Benefits Domain Adaptation & Self-Training

- ASG serves as better initialization
  1. ImageNet ➔ Self-training for DA
  2. ImageNet ➔ ASG ➔ Self-training for DA

Method	Tgt Img	Accuracy
Source-Res101 (Zou et al., 2019)	✗	51.6
CBST (Zou et al., 2018)	✓	76.4 (0.9)
MRKLD (Zou et al., 2019)	✓	77.9 (0.5)
MRKLD + LRENT (Zou et al., 2019)	✓	78.1 (0.2)
ASG (ours)	✗	61.5
ASG + CBST	✓	82.5 (0.7)
ASG + MRKLD	✓	**84.6** (0.4)
ASG + MRKLD + LRENT	✓	84.5 (0.4)

DISTRIBUTIONALLY ROBUST LEARNING FOR UNSUPERVISED DOMAIN ADAPTATION

Haoxuan Wang *
Shanghai Jiao Tong University
hatch25@sjtu.edu.cn

Anqi Liu *
Caltech
anqiliu@caltech.edu

Zhiding Yu
NVIDIA
zhidingy@nvidia.com

Yisong Yue
Caltech
yyue@caltech.edu

Anima Anandkumar
Caltech & NVIDIA
anima@caltech.edu
aanandkumar@nvidia.com

Method	Mean
Source (Saito et al., 2018a)	52.4
MMD (Long et al., 2015)	61.1
MCD (Saito et al., 2018b)	71.9
ADR (Saito et al., 2018a)	74.8
CBST (Zou et al., 2020)	76.4
CRST (Zou et al., 2020)	78.1
AVH (Chen et al., 2020a)	81.5
DRST (proposed)	83.75
ASG (Chen et al., 2020b)	61.17
CBST-ASG (Chen et al., 2020b)	82.23
CRST-ASG (Chen et al., 2020b)	84.21
DRST-ASG (proposed)	**85.25**

Contrastive Syn-to-Real Generalization

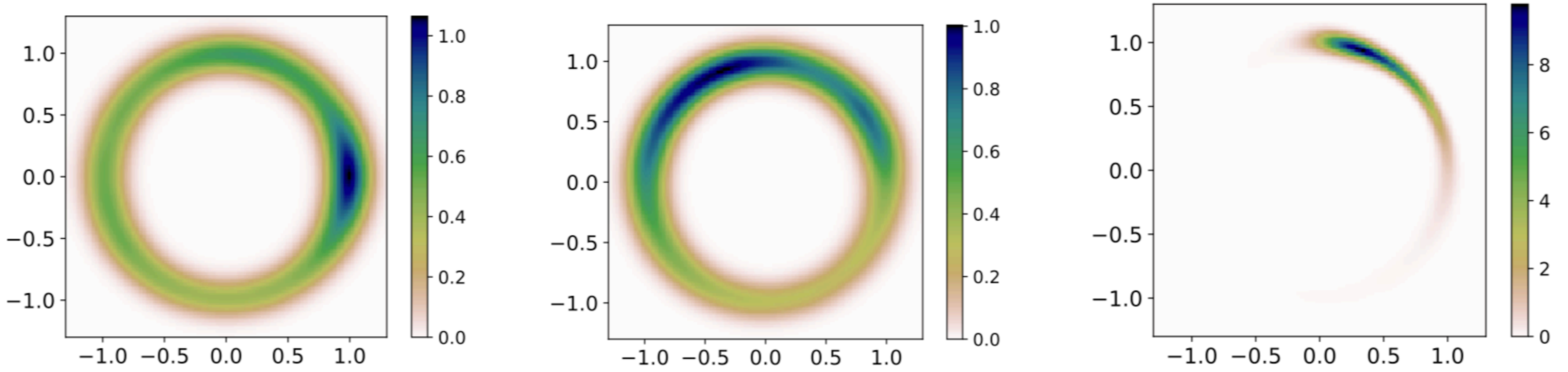
ICLR 2021

Wuyang Chen, Zhiding Yu, Shalini De Mello, Sifei Liu, Jose M. Alvarez,

Zhangyang “Atlas” Wang, Anima Anandkumar

Deeper Look Into Domain Gap

- Synthetic images leads to collapsed feature space!



(a) ImageNet pre-trained (real). ($E_s = 0.2541$)

(b) Trained on VisDA-17 validation set (real). ($E_s = 0.3355$)

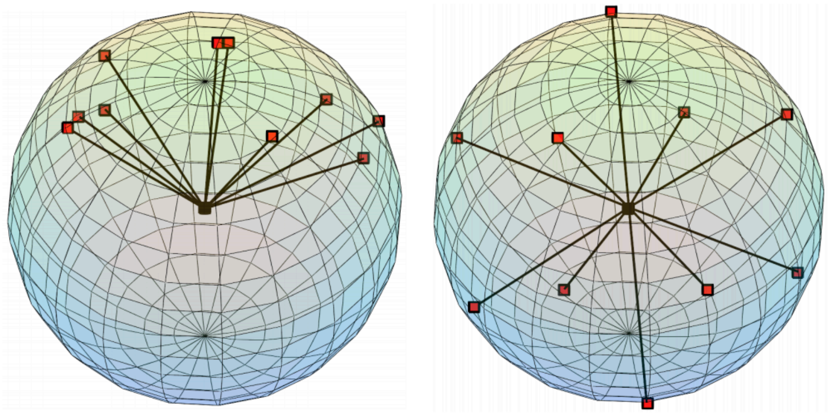
(c) Trained on VisDA-17 training set (synthetic). ($E_s = 0.4542$)

Figure 2: Feature diversity in $\mathbb{R}^2$ with Gaussian kernel density estimation (KDE). Darker areas have more concentrated features. E_s: hyperspherical energy of features, lower the more diverse.

Hyperspherical Energy (HSE, E_s)

Low E_s ➔ diverse features

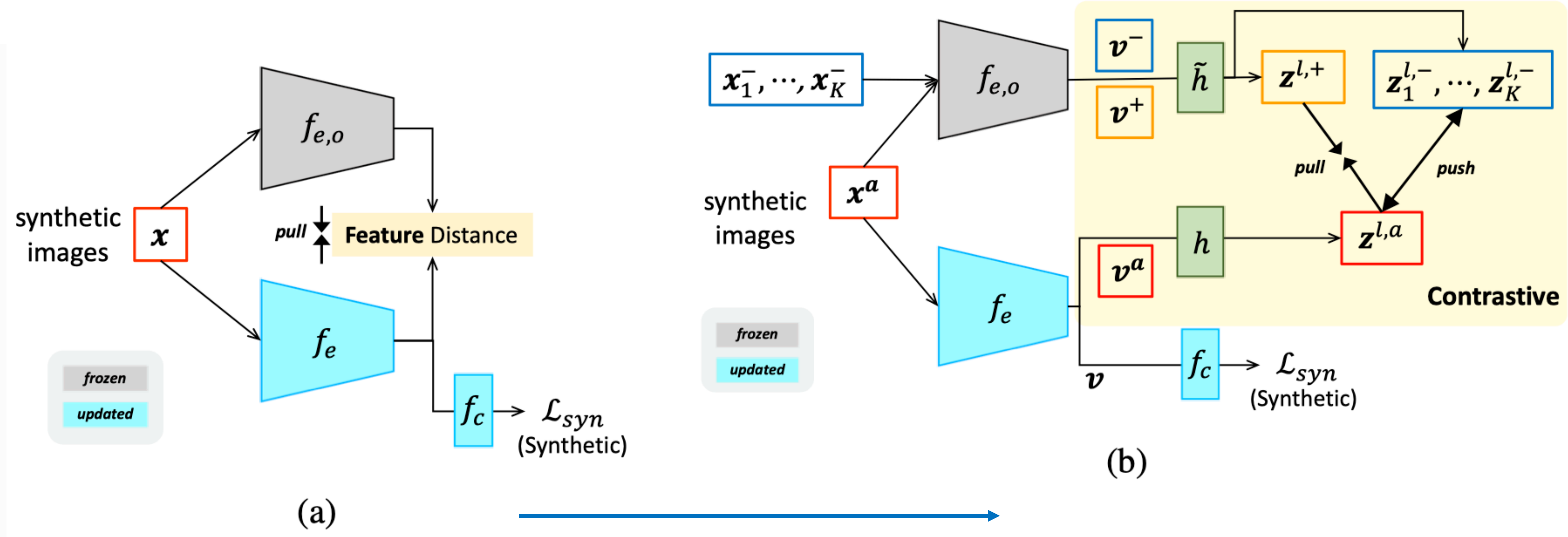
$$E_s\left(\bar{\boldsymbol{v}}_i|_{i=1}^N\right) = \sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} e_s\left(\|\bar{\boldsymbol{v}}_i - \bar{\boldsymbol{v}}_j\|\right) = \begin{cases} \sum_{i\neq j} \|\bar{\boldsymbol{v}}_i - \bar{\boldsymbol{v}}_j\|^{-s}, & s > 0 \\ \sum_{i\neq j} \log\left(\|\bar{\boldsymbol{v}}_i - \bar{\boldsymbol{v}}_j\|^{-1}\right), & s = 0 \end{cases}$$



Liu et al. Neurips 2018

ImageNet Distillation + Feature Diversity

- Synthetic-to-real with a "push and pull" strategy



Contrastive Loss

- InfoNCE

$$\mathcal{L}_{\text{NCE}} = -\log \frac{\exp\left(\boldsymbol{z}^a \cdot \boldsymbol{z}^+ / \tau\right)}{\exp\left(\boldsymbol{z}^a \cdot \boldsymbol{z}^+ / \tau\right) + \sum_{\boldsymbol{z}^-} \exp\left(\boldsymbol{z}^a \cdot \boldsymbol{z}^- / \tau\right)}, \tag{3}$$

$$\mathcal{L} = \mathcal{L}_{\text{syn}} + \lambda \mathcal{L}_{\text{NCE}} \tag{4}$$

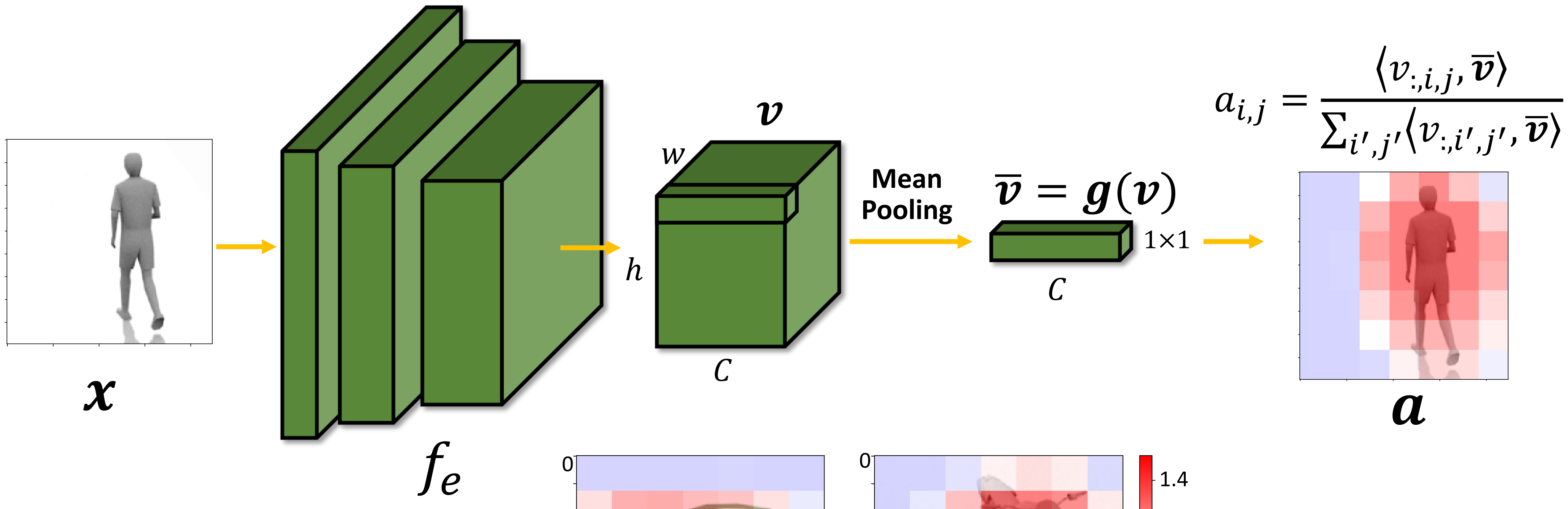
- Multi-layer InfoNCE

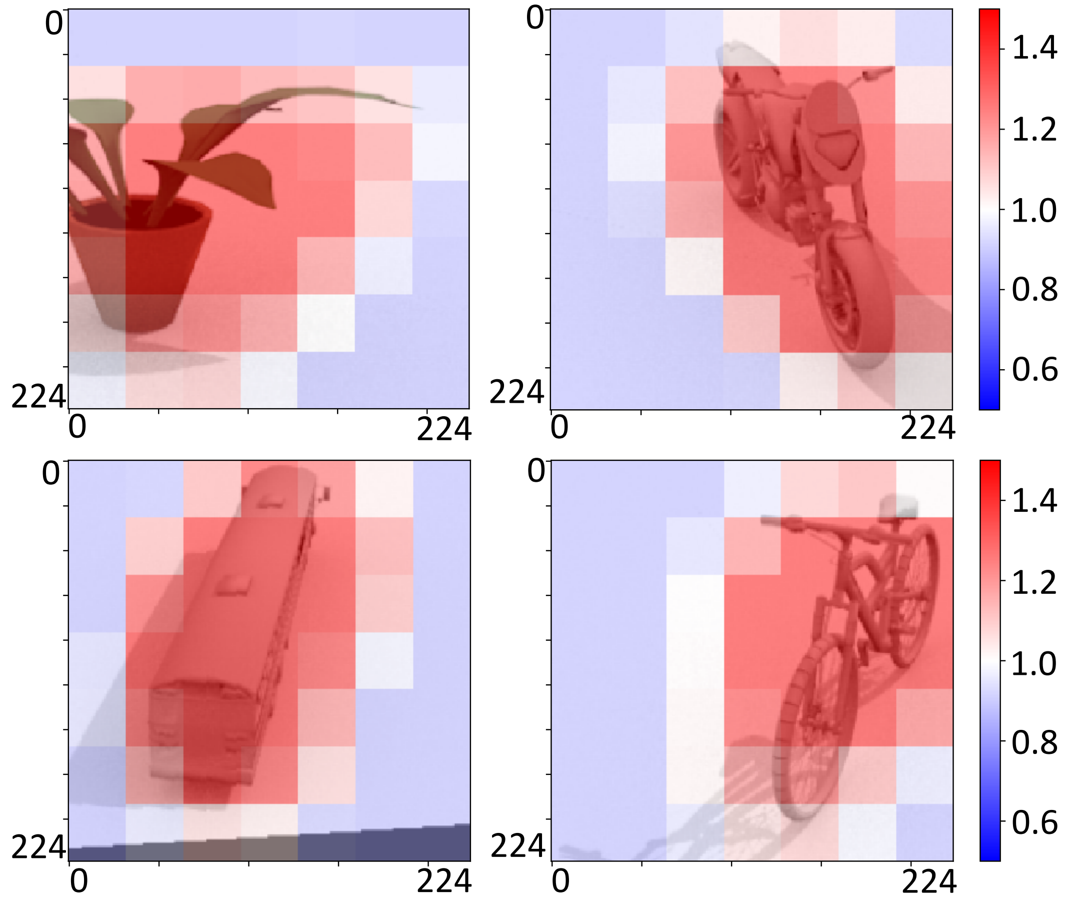
$$\mathcal{L}_{\text{NCE}} = \sum_{l \in \mathcal{G}} \mathcal{L}^{l}_{\text{NCE}} = \sum_{l \in \mathcal{G}} -\log \frac{\exp\left(\boldsymbol{z}^{l,a} \cdot \boldsymbol{z}^{l,+} / \tau\right)}{\exp\left(\boldsymbol{z}^{l,a} \cdot \boldsymbol{z}^{l,+} / \tau\right) + \sum_{\boldsymbol{z}^{l,-}} \exp\left(\boldsymbol{z}^{l,a} \cdot \boldsymbol{z}^{l,-} / \tau\right)} \tag{5}$$

- Dense InfoNCE (segmentation)

$$\mathcal{L}_{\text{NCE}} = \sum_{l \in \mathcal{G}} \sum_{i} \mathcal{L}^{l,i}_{\text{NCE}} = \sum_{l \in \mathcal{G}} \sum_{i} -\frac{1}{N_l} \log \frac{\exp\left(\boldsymbol{z}_i^{l,a} \cdot \boldsymbol{z}_i^{l,+} / \tau\right)}{\exp\left(\boldsymbol{z}_i^{l,a} \cdot \boldsymbol{z}_i^{l,+} / \tau\right) + \sum_{\boldsymbol{z}_i^{l,-}} \exp\left(\boldsymbol{z}_i^{l,a} \cdot \boldsymbol{z}_i^{l,-} / \tau\right)} \tag{6}$$

Attention-guided Global Pooling





Results: Feature Diversity vs Generalization

- Model preserves diverse features ➔ generalize better on real domain

Table 1: Generalization performance and hyperspherical energy of the features extracted by different models (lower is better). Dataset: VisDA-17 (Peng et al., 2017) validation set. Model: ResNet-101.

Model	Power			Accuracy (%)
	0	1	2	
Oracle on ImageNet[3]	-	-	-	53.3
Baseline (vanilla synthetic training)	0.4245	1.2500	1.6028	49.3
Weight $l2$ distance (Kirkpatrick et al., 2017)	0.4014	1.2296	1.5302	56.4
Synaptic Intelligence (Zenke et al., 2017)	0.3958	1.2261	1.5216	57.6
Feature $l2$ distance (Chen et al., 2018)	0.3337	1.1910	1.4449	57.1
ASG (Chen et al., 2020b)	0.3251	1.1840	1.4229	61.1
CSG (Ours)	**0.3188**	**1.1806**	**1.4177**	**64.05**

Results: Segmentation

- GTA5→Cityscapes

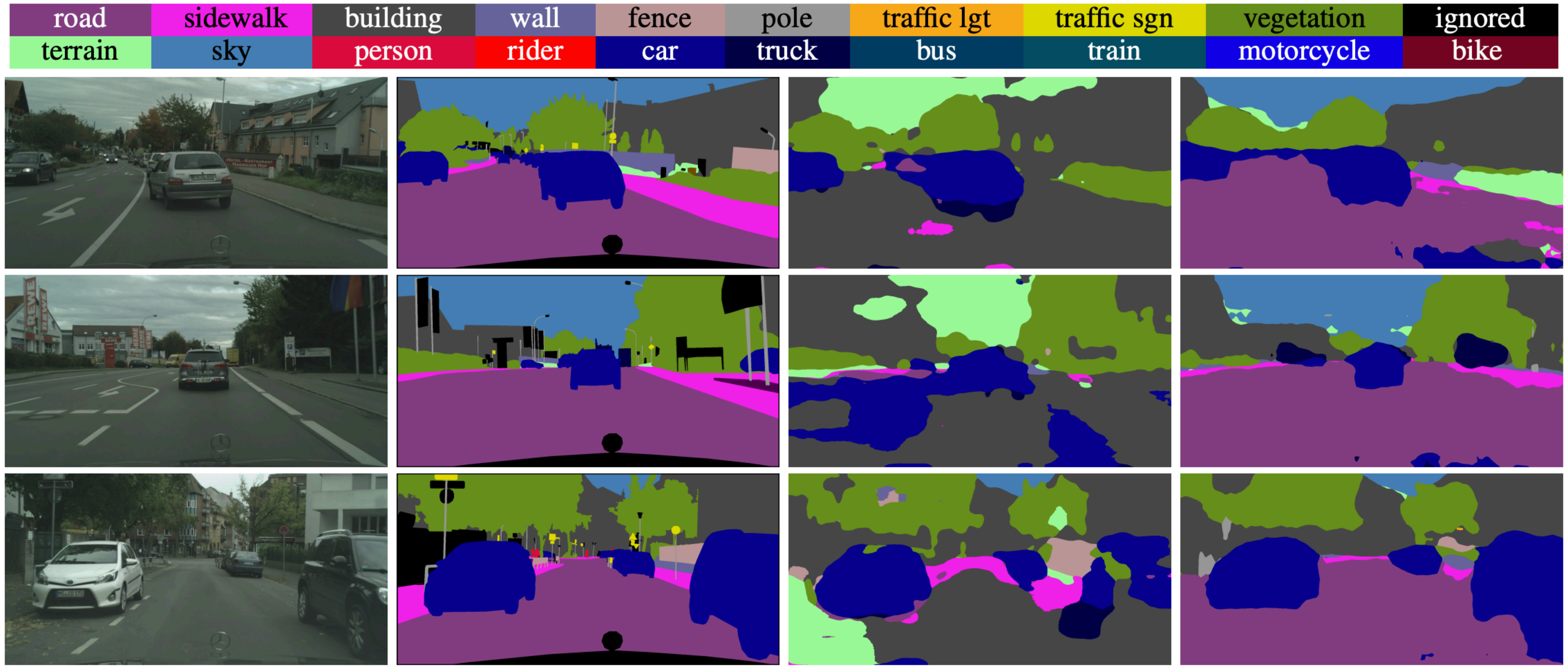


Table 5: Comparison to prior domain generaliz

Methods			
No Adapt IBN-Net (Pan et al., 2018			
No Adapt Yue et al. (Yue et al., 201			
No Adapt ASG (Chen et al., 2020b			
No Adapt CSG (ours)			
No Adapt Yue et al. (Yue et al., 201			
No Adapt ASG (Chen et al., 2020b)	ResNet-101	27.94 32.79	4.85
No Adapt CSG (ours)		28.94 38.88	**9.94**

Future Works

- More applications: Gaze, Detection, Robotics, etc.
- Joint training with domain adaptation.
- Better leveraging multiple sources
  - labeled real domain (ImageNet)
  - labeled synthetic domain
  - Unlabeled target real domain

The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering