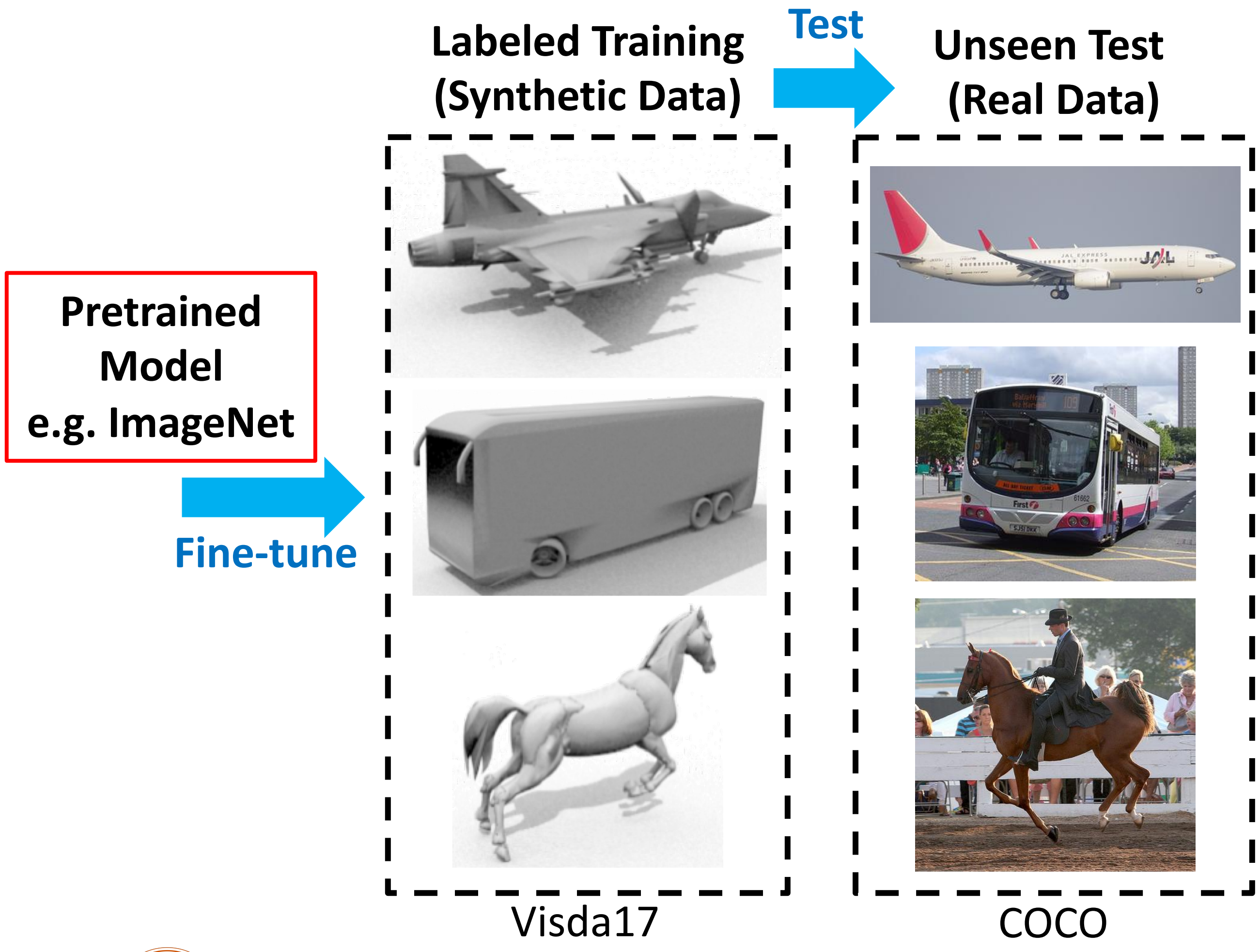
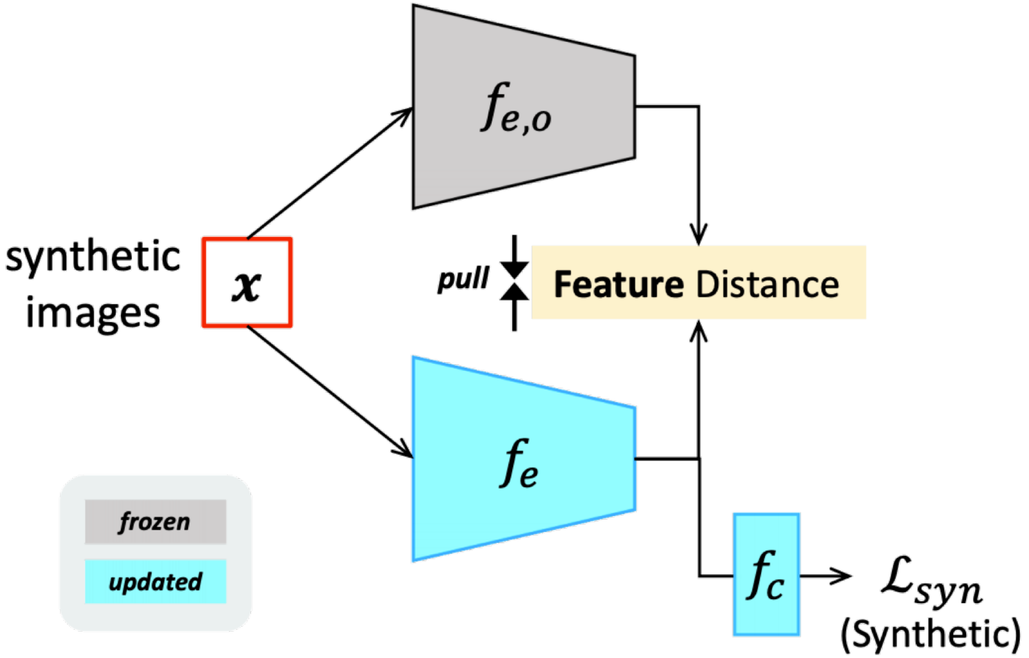
Contrastive Synthetic-to-Real Generalization

Wuyang Chen[1], Zhiding Yu[2], Shalini De Mello[2], Sifei Liu[2], Jose M. Alvarez[2],
Zhangyang Wang[1], Animashree Anandkumar[2,3]
[1]UT Austin [2]NVIDIA [3]Caltech
* Work done during the author's research internship in NVIDIA Inc.

Syn-to-Real Generalization: Problem & Previous Solution

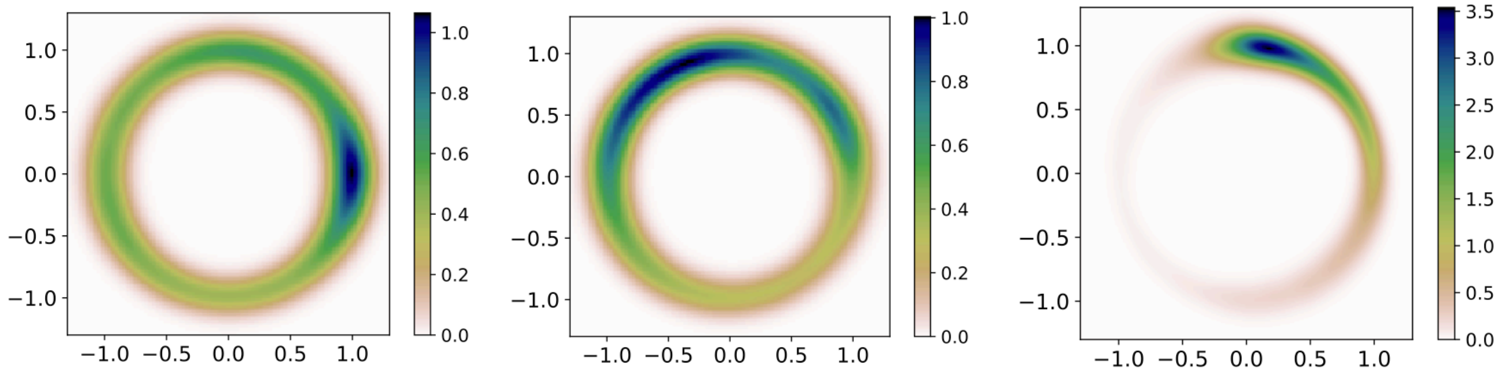




Chen, Wuyang, Zhiding Yu, Zhangyang Wang, and Animashree Anandkumar. "Automated synthetic-to-real generalization." ICML 2020.

But why Synthetic Training Fails? A Representation Learning Perspective

- Train model on natural images ➔ diverse representations.
- Train model on synthetic images ➔ collapsed representations!



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

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

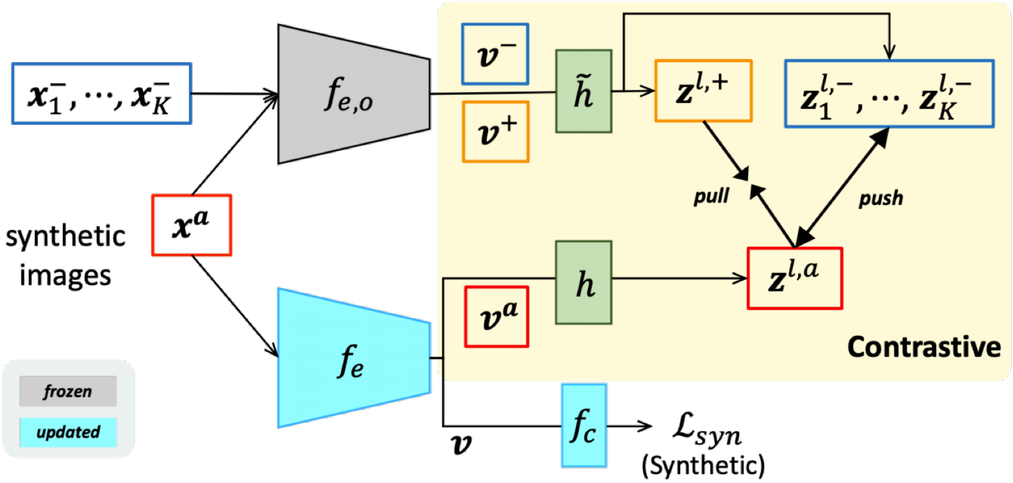
(c) Trained on VisDA-17 training set (synthetic). ($E_s = 0.4408$)

- E_s (Hyperspherical Energy): Lower the more diverse.

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

CSG: Contrastive Synthetic-to-Real Generalization

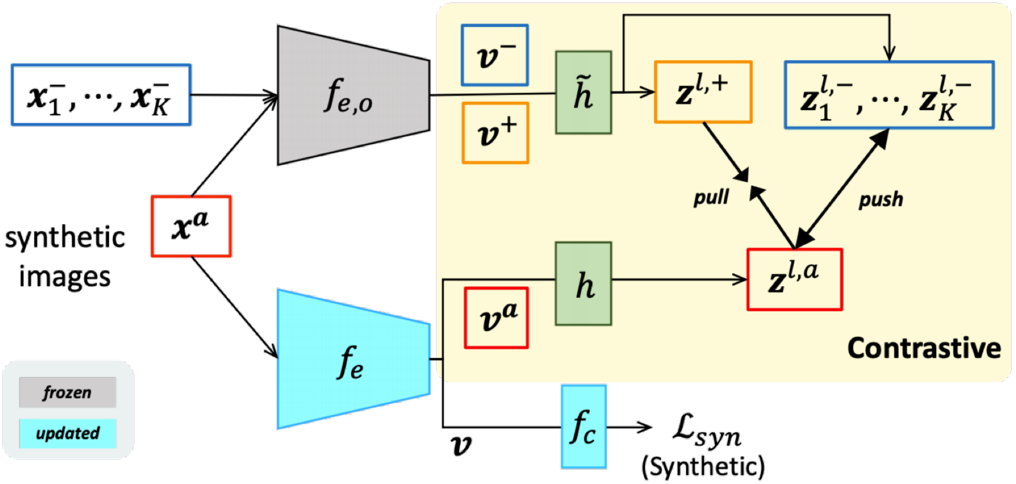
- How to *transfer real domain knowledge* + *promote feature diversity*?
- **Pull**: impose similarity b/w features from synthetic model v.s. ImageNet pre-trained model.
- **Push**: encourage feature diversity by pushing the feature embeddings away from each other across different images.



CSG: Contrastive Synthetic-to-Real Generalization

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

$$\mathcal{L} = \mathcal{L}_{\text{Task}} + \lambda \mathcal{L}_{\text{NCE}}$$



Multi-layer Contrastive Loss

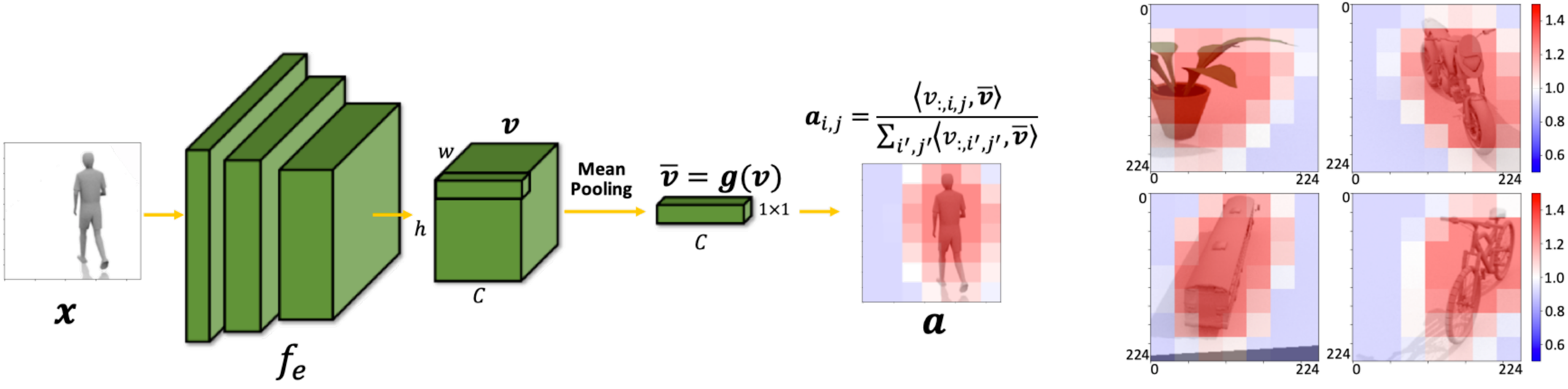
$$\mathcal{L}_{\text{NCE}} = \sum_{l \in \mathcal{G}} \mathcal{L}^l_{\text{NCE}} = \boxed{\sum_{l \in \mathcal{G}}} - \log \frac{\exp\left(\boldsymbol{z}^{l,a} \cdot \boldsymbol{z}^{l,+} / \tau\right)}{\exp\left(\boldsymbol{z}^{l,a} \cdot \boldsymbol{z}^{l,+} / \tau\right) + \sum_{\boldsymbol{z}^{l,-}}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

Dense Contrastive Loss for Segmentation

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

A-Pool: Attentional Pooling for Improved Representation

- A-Pool for non-linear projection head $h(\cdot)$ focus contrastive learning on more semantically meaningful regions.



(a) (b)

CSG: Best Performance Gain & Diverse Features

Classification: Visda17 ➔ COCO

Model	Power 0	Power 1	Power 2	Accuracy (%)
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., 2020c)	0.3251	1.1840	1.4229	61.1
CSG (Ours)	**0.3188**	**1.1806**	**1.4177**	**64.05**

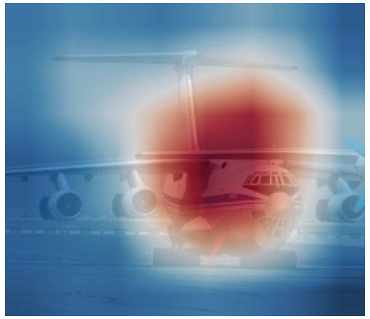
Segmentation: GTA5 ➔ Cityscapes

Methods	Backbone	mIoU %	mIoU ↑ %
No Adapt IBN-Net (Pan et al., 2018)	ResNet-50	22.17 29.64	7.47
No Adapt Yue et al. (Yue et al., 2019)	ResNet-50	32.45 37.42	4.97
No Adapt ASG (Chen et al., 2020c)	ResNet-50	25.88 29.65	3.77
No Adapt CSG (ours)	ResNet-50	25.88 35.27	**9.39**
No Adapt Yue et al. (Yue et al., 2019)	ResNet-101	33.56 42.53	8.97
No Adapt ASG (Chen et al., 2020c)	ResNet-101	29.63 32.79	3.16
No Adapt CSG (ours)	ResNet-101	29.63 38.88	**9.25**

CSG Improves Model Attention (GradCAM)

Input Image | Baseline | **CSG (Ours)**

Train ✗

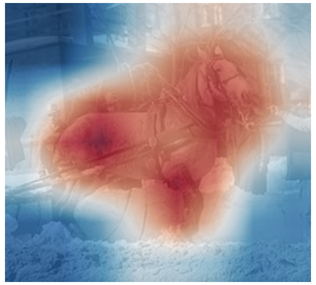


Airplane ✓

Motorcycle ✗

Bicycle ✓

Plant ✗



Horse ✓

Thank you!

Code

Paper