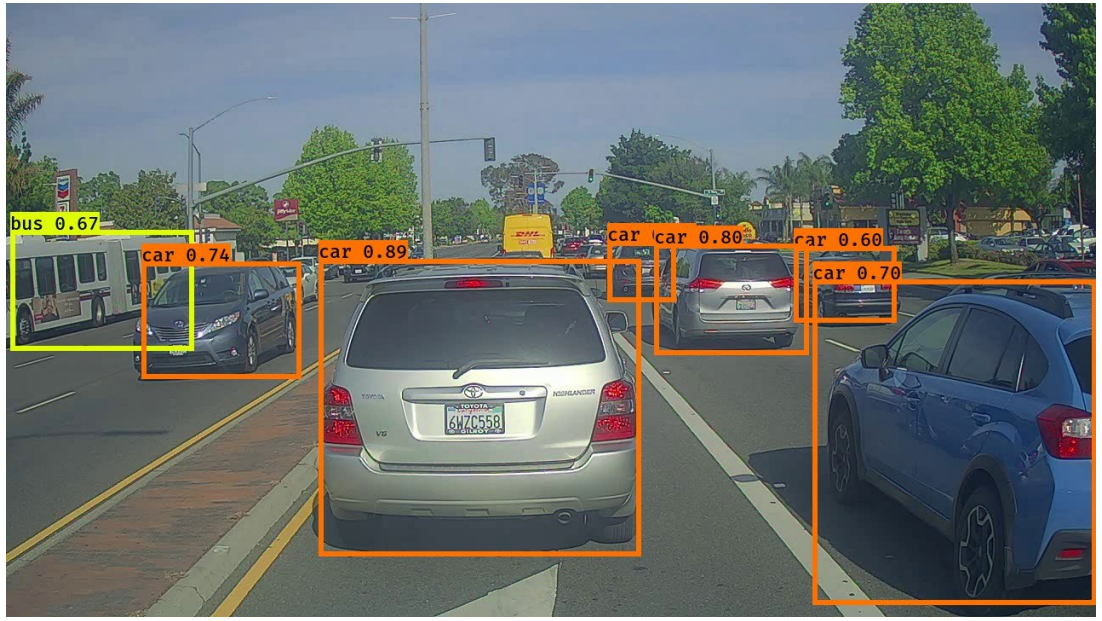
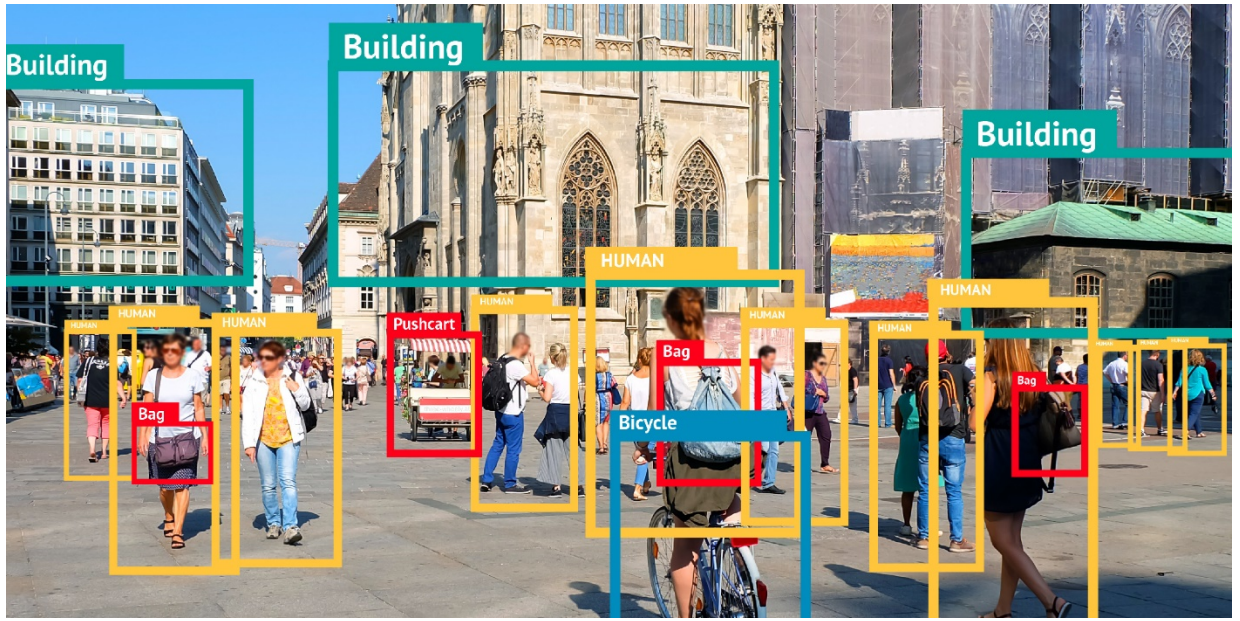
Overcoming Dataset Bias in Machine Learning

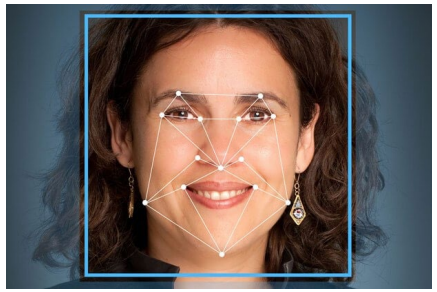
Kate Saenko

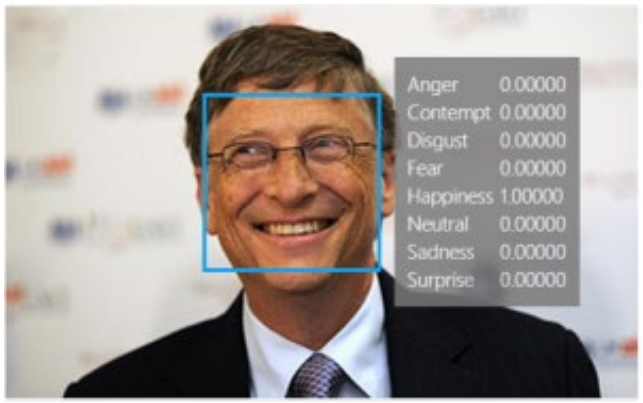
Boston University & MIT-IBM Watson AI Lab

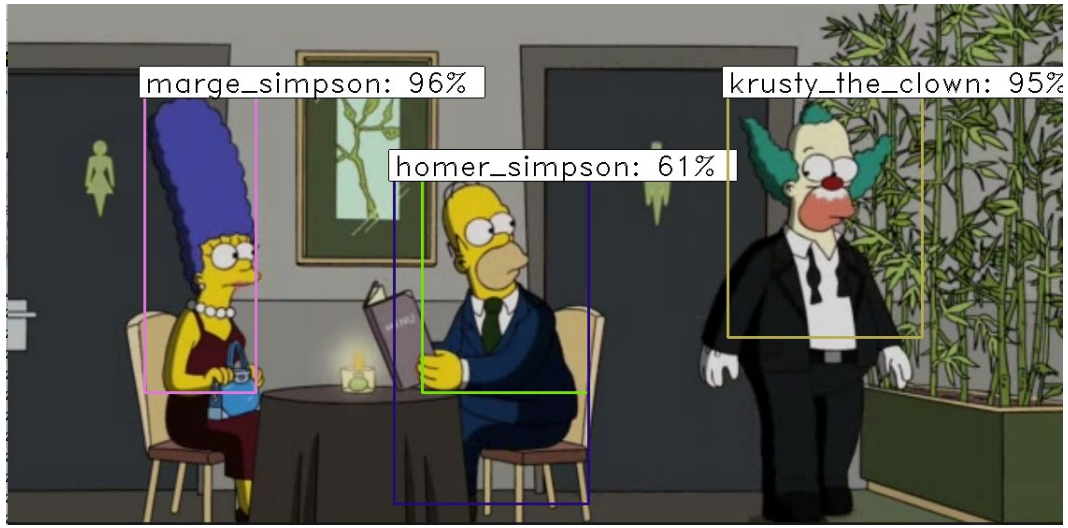
Successes of AI and computer vision







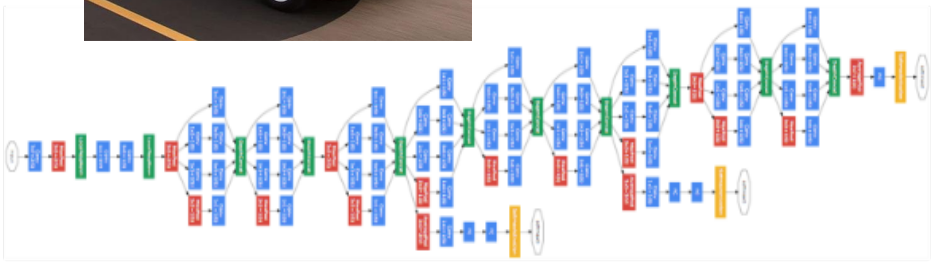




Problem: dataset bias

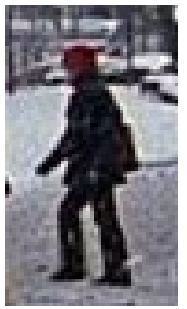
What your net is trained on

What it's asked to label



"Dataset Bias"
"Domain Shift"

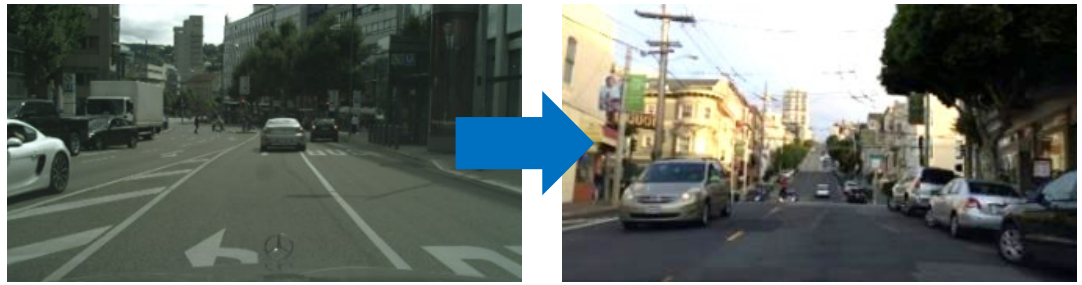
vs



Adapting visual category models to new domains K Saenko, B Kulis, M Fritz, T Darrell, ECCV, 2010

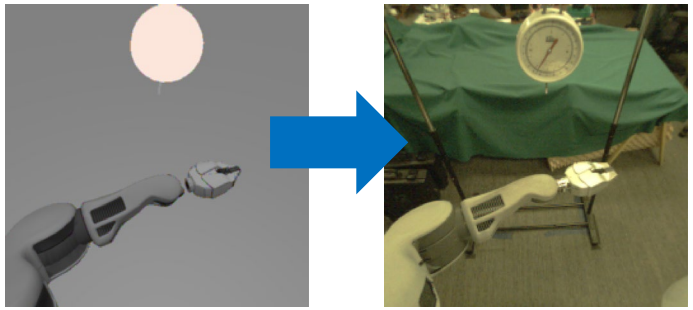
When does dataset bias happen?

From one city to another



From web to robot

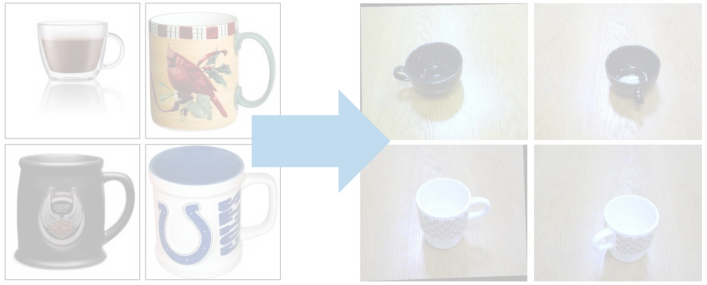
From simulated to real control



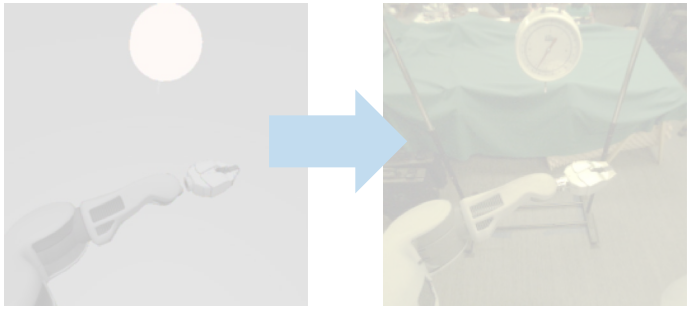
When does dataset bias happen?

From one city to another

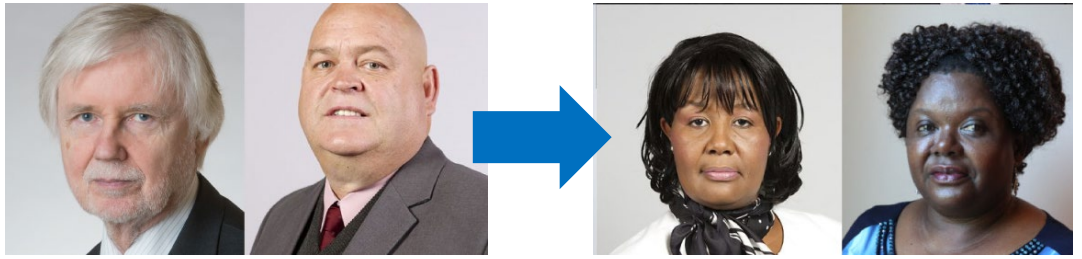
From web to robot



From simulated to real control

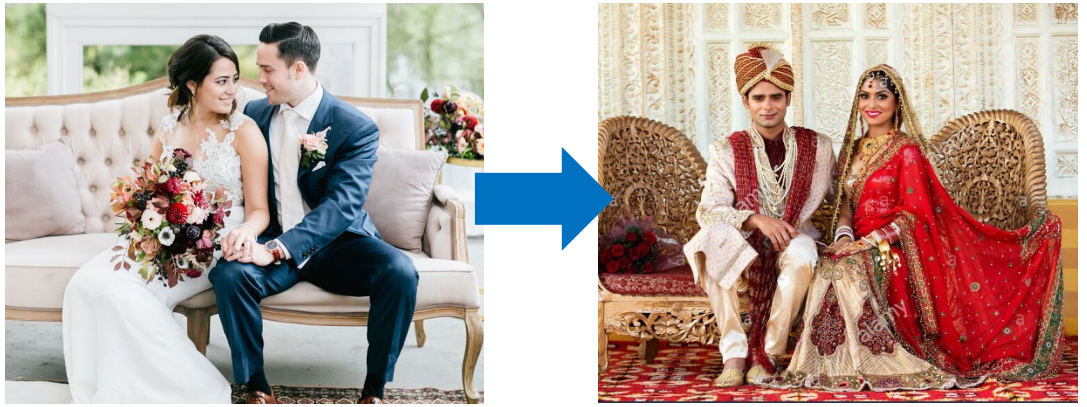


From one demographic to another

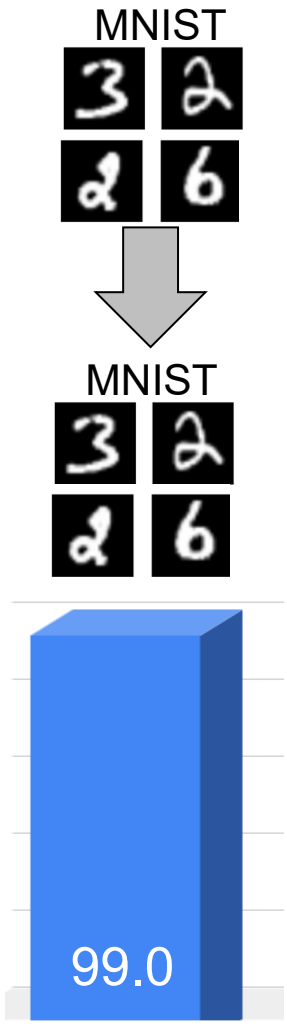


https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html

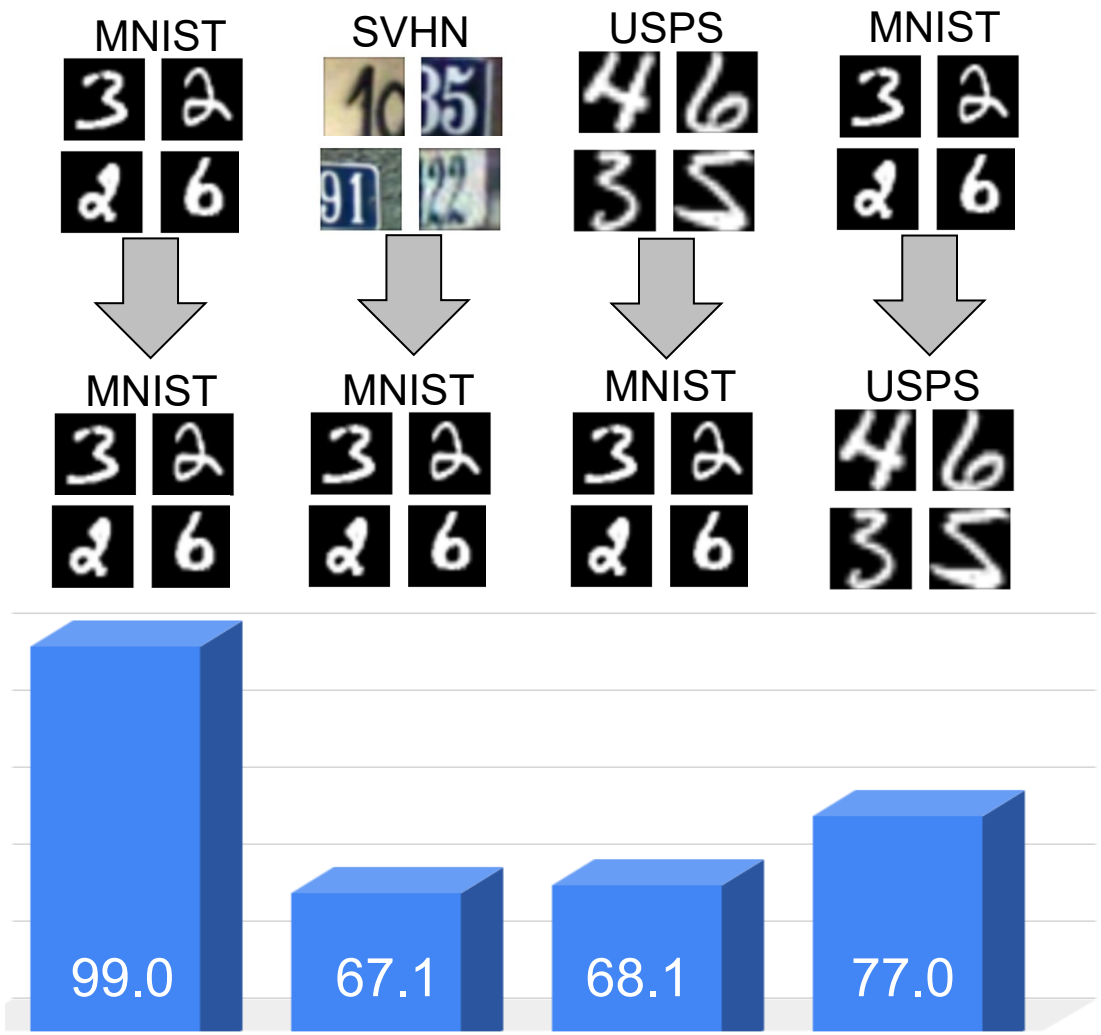
From one culture to another



Domain bias reduces accuracy



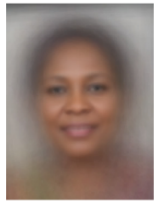
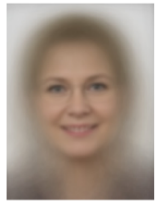
Domain bias reduces accuracy



Real-world implications of dataset bias

August 2018 Accuracy on Facial Analysis Pilot Parliaments Benchmark

amazon

98.7%	68.6%	100%	92.9%
			
DARKER MALES	DARKER FEMALES	LIGHTER MALES	LIGHTER FEMALES

Amazon Rekognition Performance on Gender Classification

BBC NEWS

A US government study suggests facial recognition algorithms are far less accurate at identifying African-American and Asian faces compared to Caucasian faces.

… The National Institute of Standards and Technology (Nist) tested 189 algorithms from 99 developers, including Intel, Microsoft, Toshiba, and Chinese firms Tencent and DiDi Chuxing.

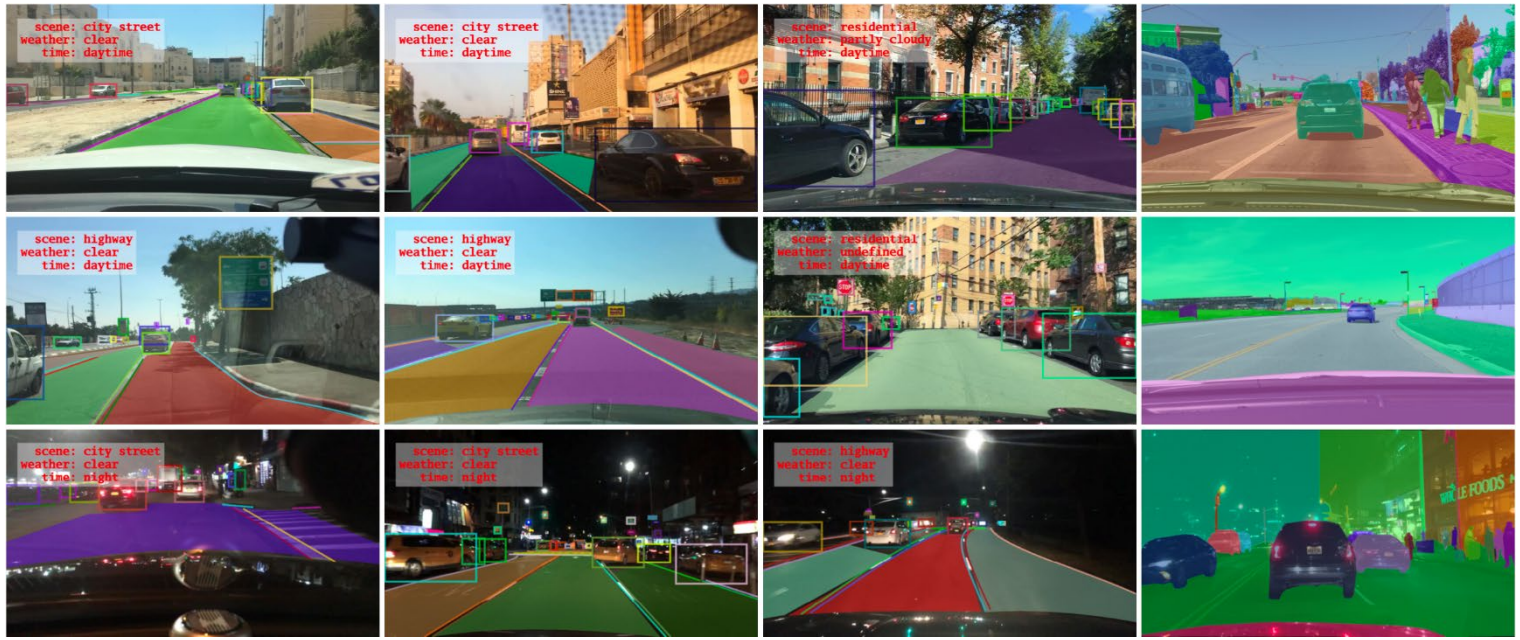
WIRED **Uber's Self-Driving Car Didn't Know Pedestrians Could Jaywalk**

THE SOFTWARE INSIDE the Uber self-driving SUV that killed an Arizona woman last year was not designed to detect pedestrians outside of a crosswalk, according to new documents released as part of a federal investigation into the incident.

AAA NewsRoom

New research from AAA reveals that automatic emergency braking systems with pedestrian detection perform inconsistently, and proved to be completely ineffective at night.

Can’t we fix it by collecting more data?



Labeling 1,000 pedestrians costs ~$1,000

Pose x Gender x Age x Race x Clothing style x Weather x City x Time of day x … x Riding bicycle x …

What causes poor performance?

- Train and test data distributions are different
- Model lacks discriminative features

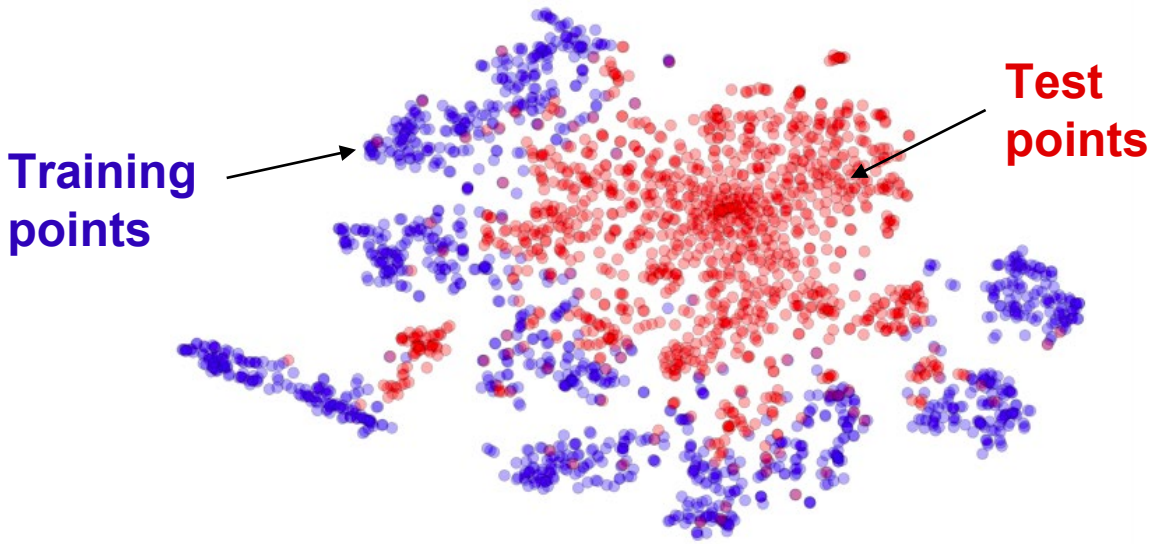
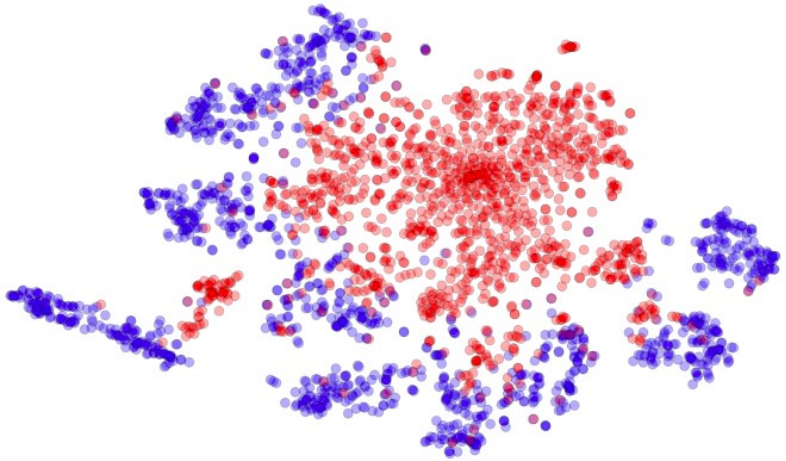


Figure from Ganin and Lempitsky. "Unsupervised domain adaptation by backpropagation." ICML 2015

Techniques that help deal with data bias

- Collect some labelled data from target domain
- Better backbone CNNs
- Batch Normalization ([Li'17], [Chang'19])
- Instance Normalization + Batch Normalization [Nam'19]
- Data Augmentation, Mix Match [Berthelot'19]
- Semi-supervised methods, such as Pseudo labeling [Zou'19]
- Domain Adaptation (this talk)



Solution: adapt knowledge to new domains

SVHN

Source Domain

lots of labeled data

$$D_S = \{(\mathbf{x}_i, y_i), \forall i \in \{1, \ldots, N\}\}$$

$\neq$

MNIST

Target Domain

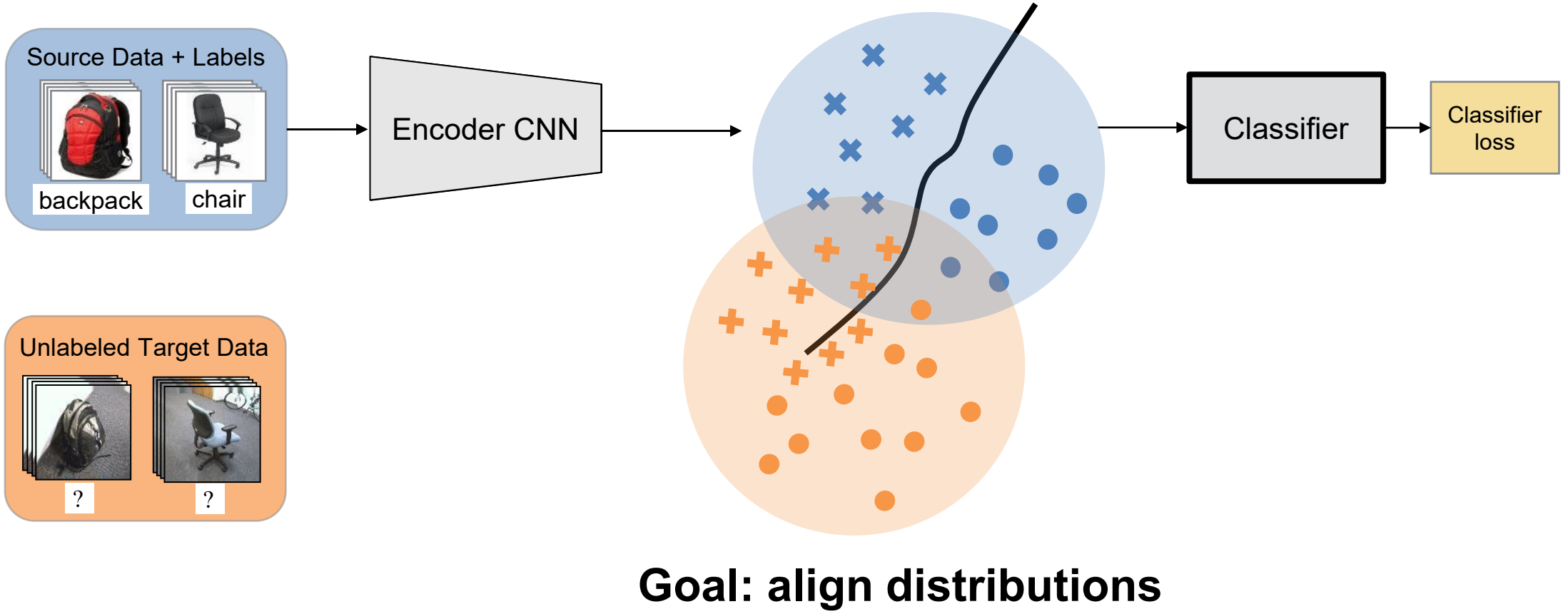
unlabeled data

$$D_T = \{(\mathbf{z}_j, ?), \forall j \in \{1, \ldots, M\}\}$$

Goal: learn a classifier f that achieves low expected loss under distribution D_T

Assume: we get to see the unlabeled target data, but not its labels

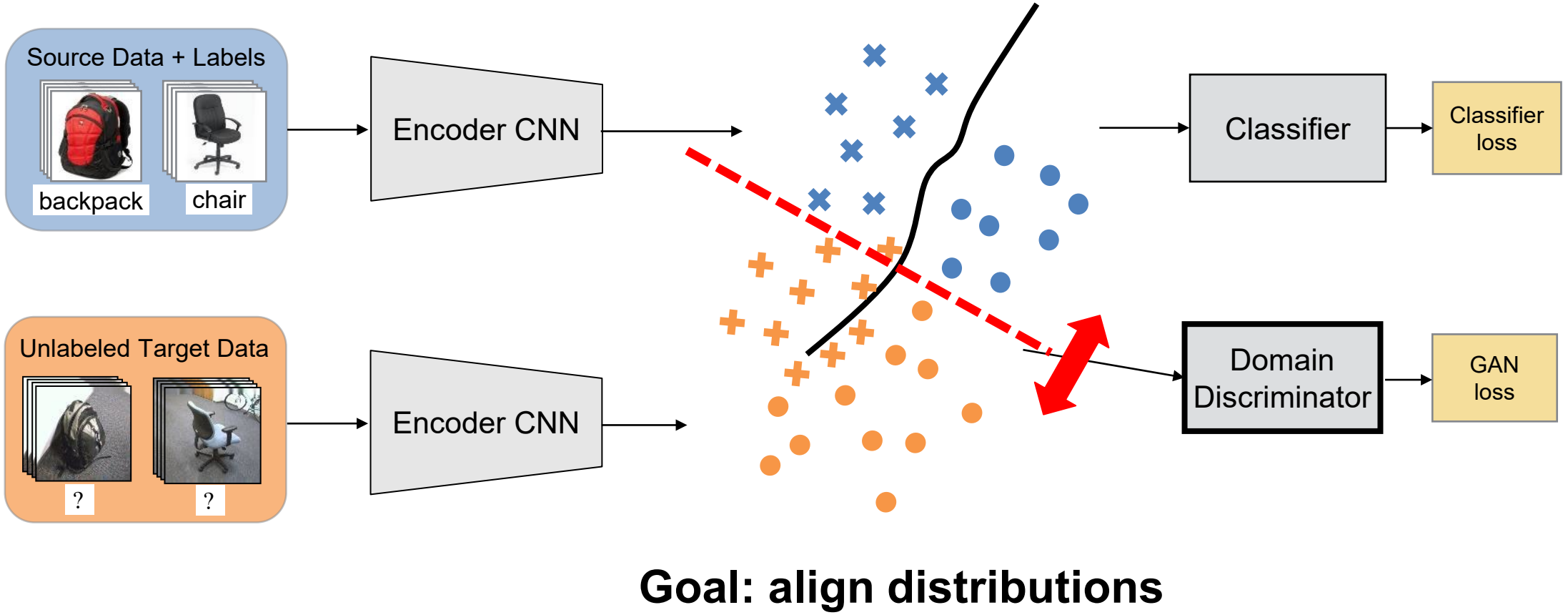
Adversarial domain alignment



Goal: align distributions

ADDA: Tzeng, Eric, et al. "Adversarial discriminative domain adaptation." CVPR 2017.

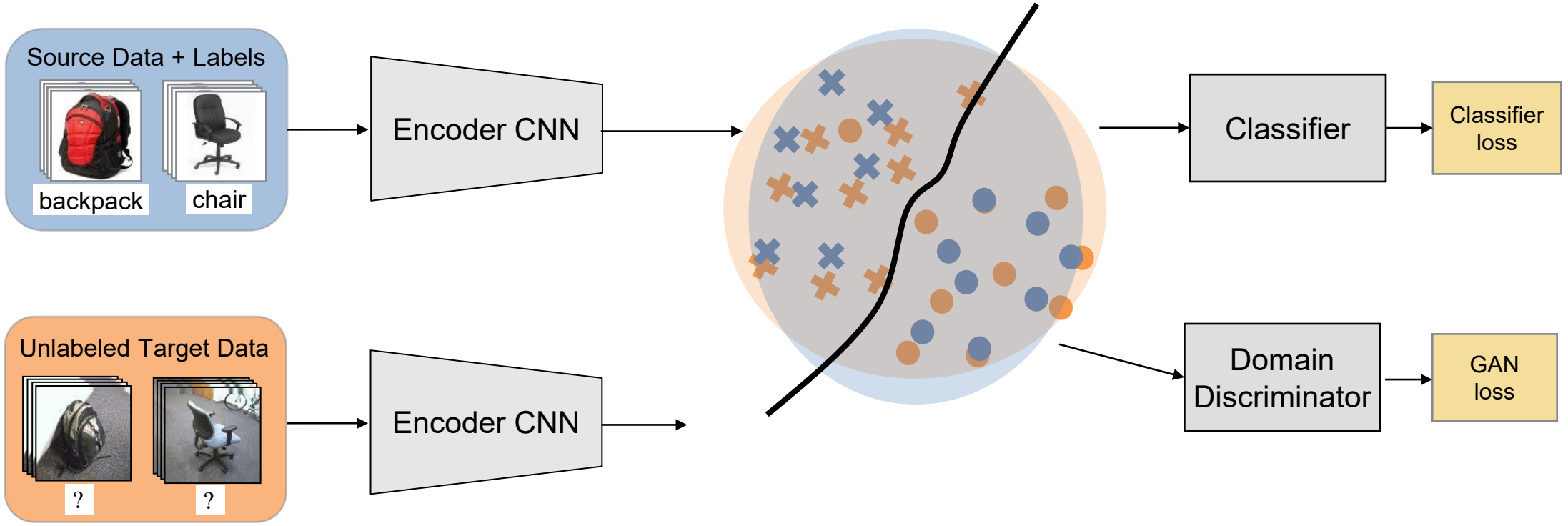
Adversarial domain alignment



Goal: align distributions

ADDA: Tzeng, Eric, et al. "Adversarial discriminative domain adaptation." CVPR 2017.

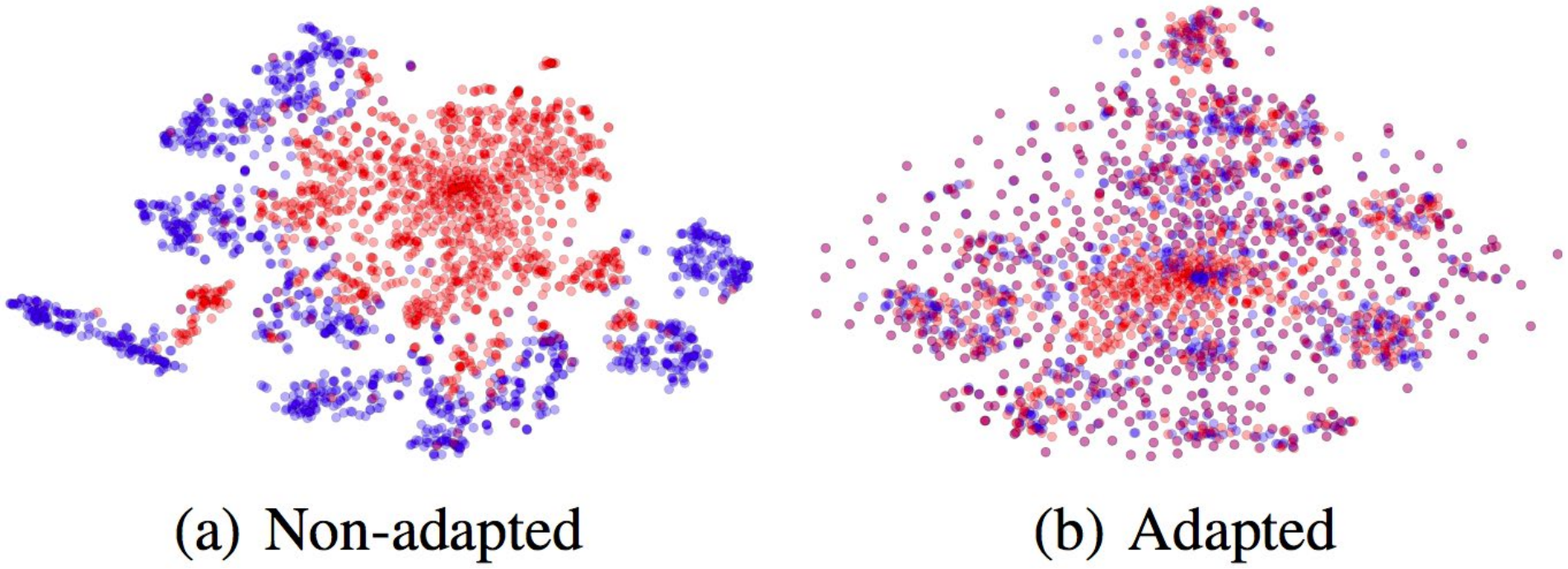
Adversarial domain alignment



Goal: align distributions

ADDA: Tzeng, Eric, et al. "Adversarial discriminative domain adaptation." CVPR 2017.

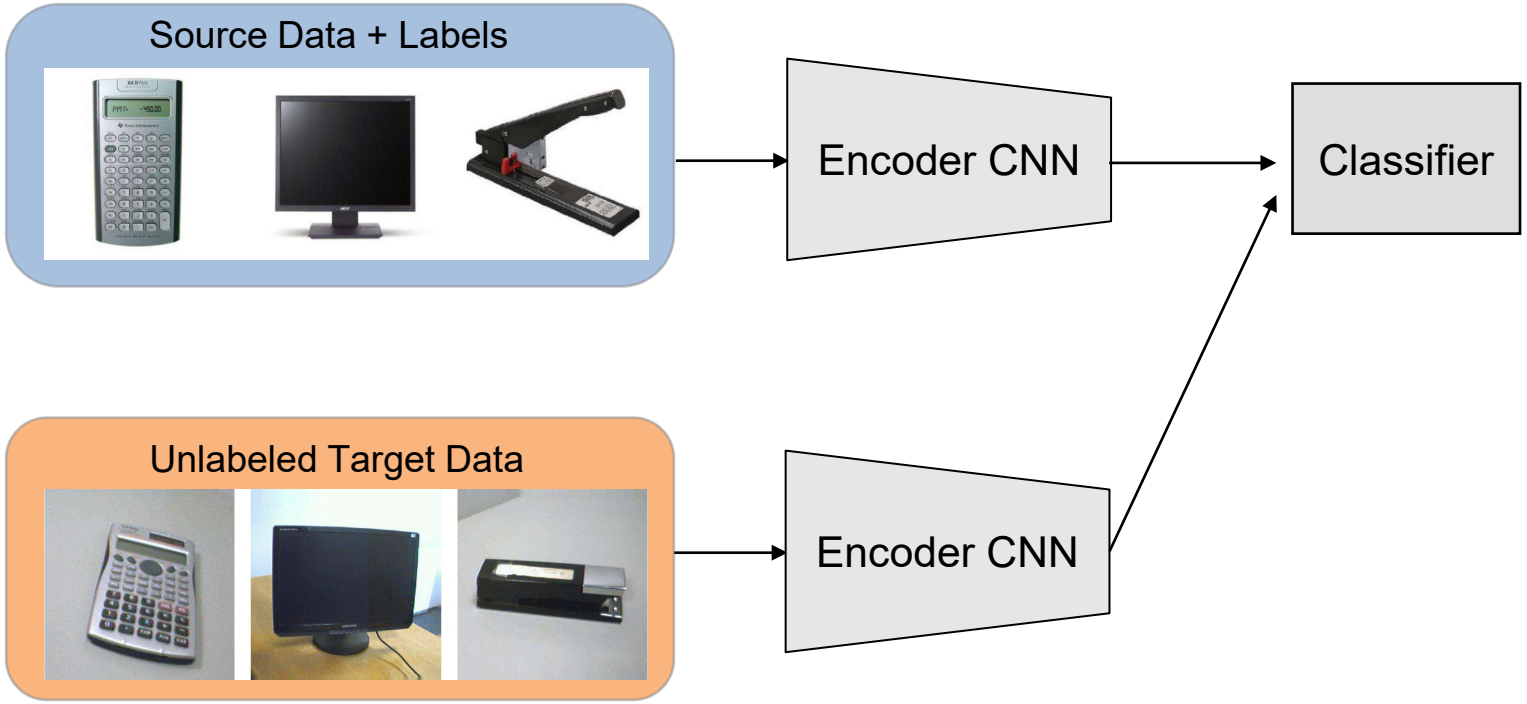
Domain alignment: feature visualization on digits



Effect of adaptation on features in MNIST → MNIST-M shift
(top feature extractor layer)

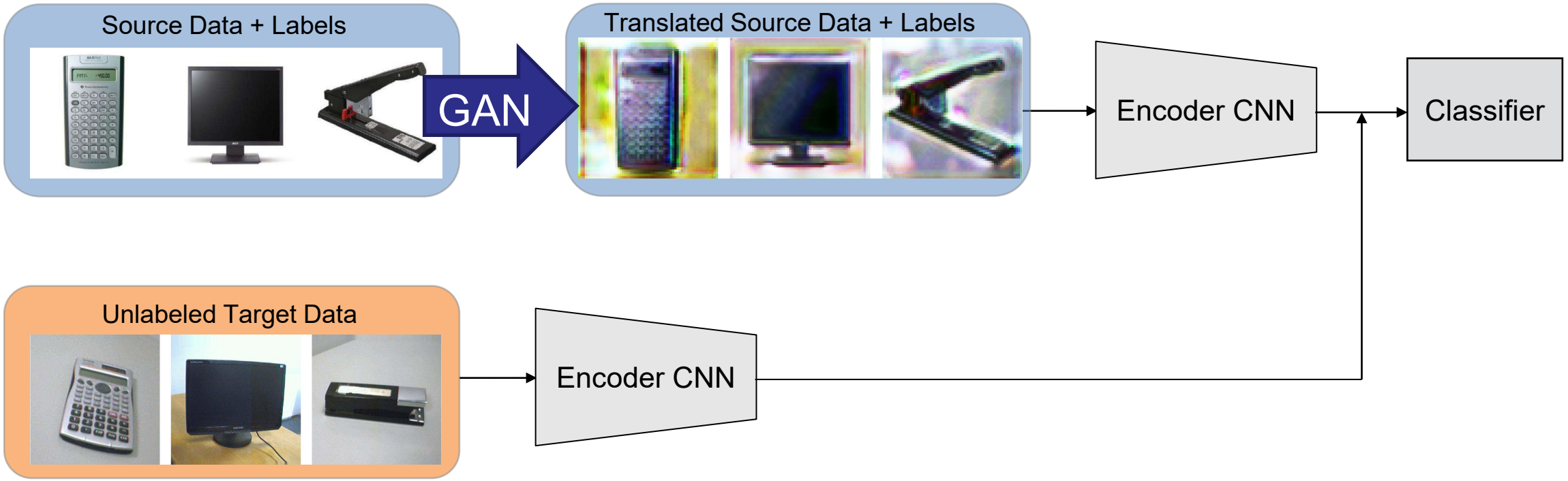
Figure from Ganin, Yaroslav, and Victor Lempitsky. "Unsupervised domain adaptation by backpropagation." ICML 2015

Pixel-space domain alignment



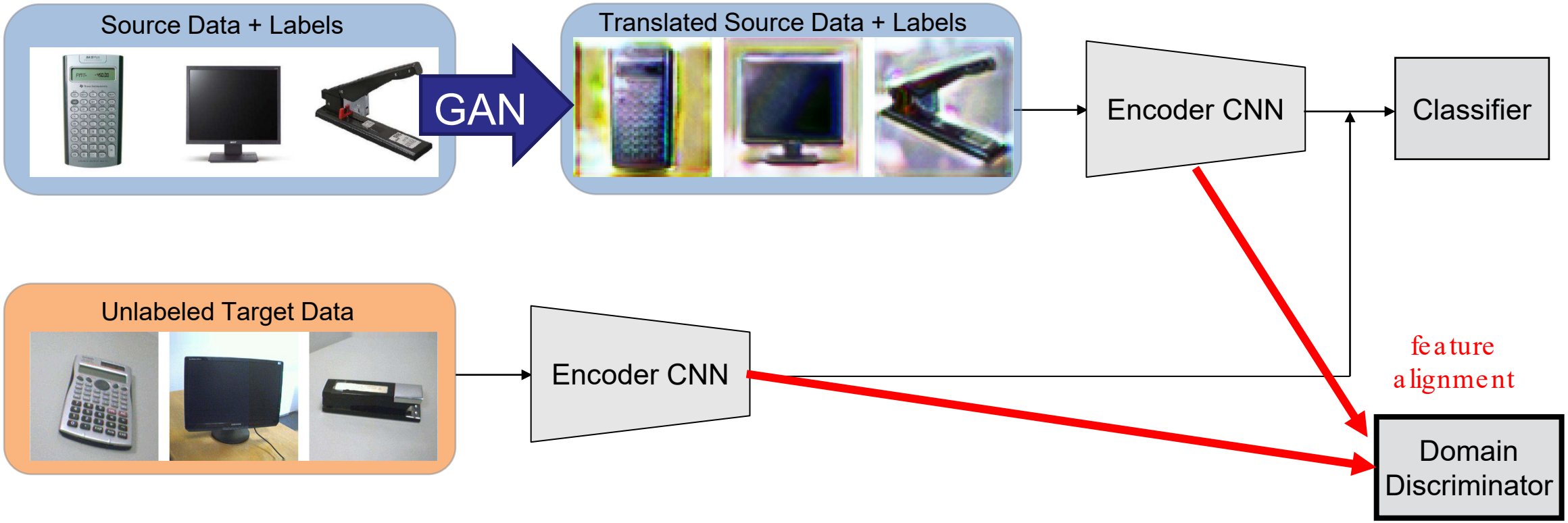
CyCADA, Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, ICML 2018
From source to target and back: Symmetric Bi-Directional Adaptive GAN, Russo et al. CVPR 2018

Pixel-space domain alignment



CyCADA, **Hoffman**, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, ICML 2018
From source to target and back: Symmetric Bi-Directional Adaptive GAN, Russo et al. CVPR 2018

Pixel-space domain alignment



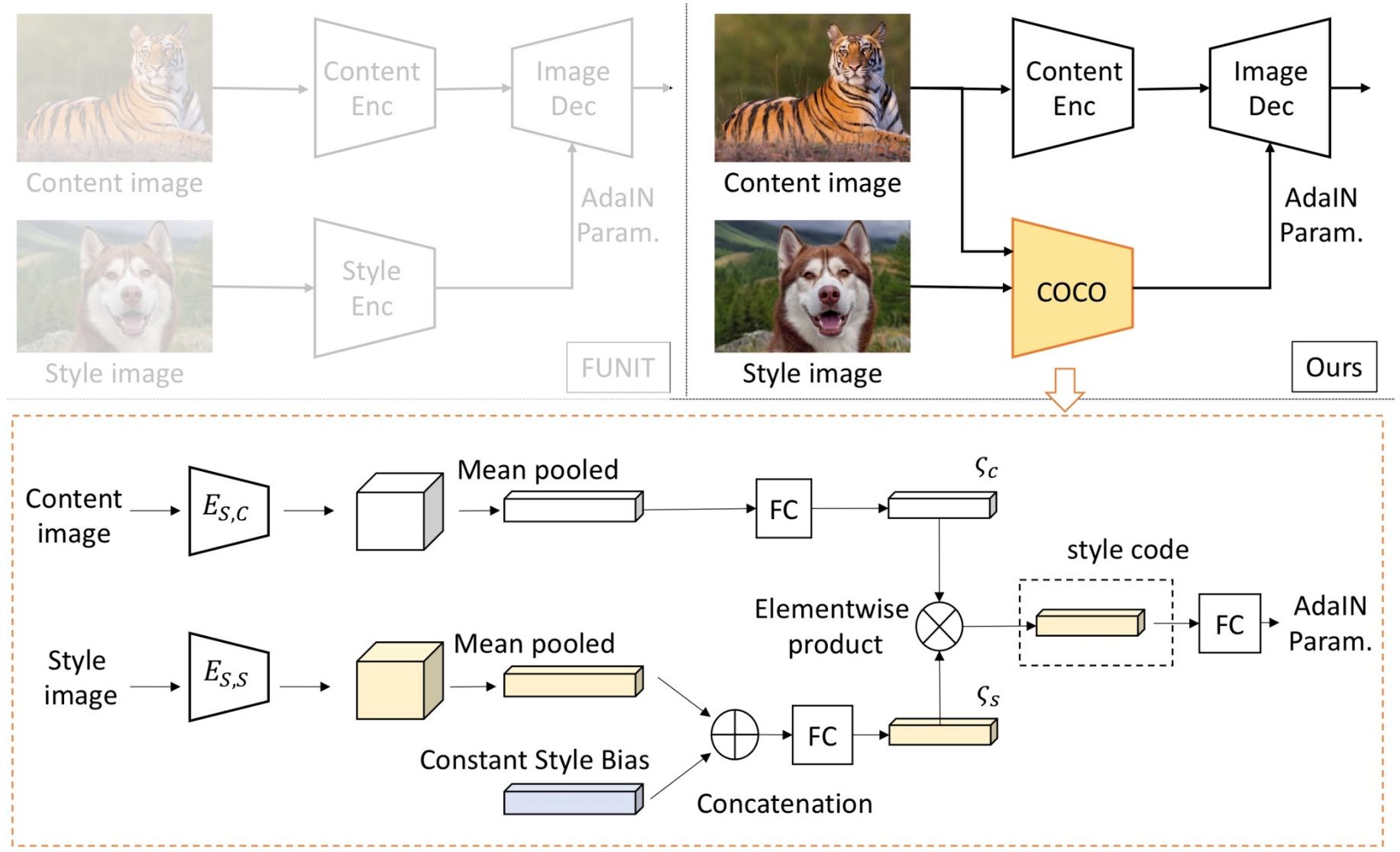
CyCADA, **Hoffman**, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, ICML 2018
From source to target and back: Symmetric Bi-Directional Adaptive GAN, Russo et al. CVPR 2018

Few-shot domain translation

- So far we have assumed lots of unlabeled target data
- What if we only have 1-5 images of the target domain?

COCO-FUNIT: Few-Shot Unsupervised Image Translation with a Content Conditioned Style Encoder, Kuniaki Saito, Kate Saenko, Ming-Yu Liu. ECCV2020 Spotlight https://nvlabs.github.io/COCO-FUNIT/

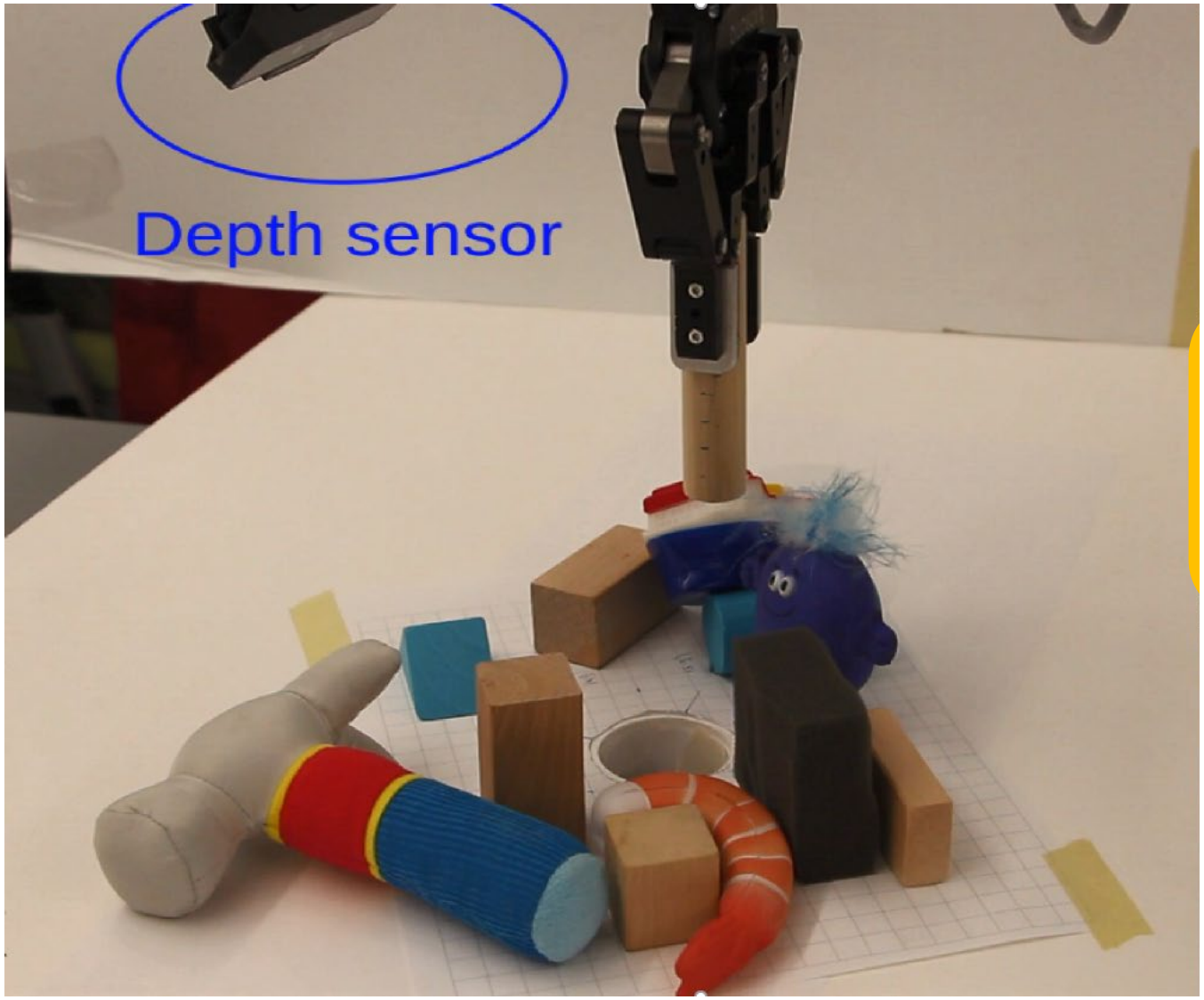
COCO-FUNIT



COCO-FUNIT: Few-Shot Unsupervised Image Translation with a Content Conditioned Style Encoder, Saito, Saenko, Liu. ECCV2020 Spotlight


Style
Content
COCO-
FUNIT
Ours

Sim2Real adaptation for robotics



Domain alignment
between simulated and real depth images

Learning visual servo policies via planner cloning, Ulrich Viereck, Kate Saenko, Robert Platt, RSS 2020 VLRRM Workshop

Sim2Real adaptation for robotics

Real RGB view

Real Depth view

Simulated

Real -> Sim

Learning visual servo policies via planner cloning, Ulrich Viereck, Kate Saenko, Robert Platt, RSS 2020 VLRRM Workshop

Summary

- Dataset bias is a major problem for machine learning
- Domain adaptation: attempt to transfer knowledge using unlabeled data
  - Feature-space
  - Pixel-space
- More general ethics problems with data
  - e.g. very large data that might contain offensive material (Bender et al.)

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific

alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent

MIT-IBM Watson AI Lab

Honda Research Institute JP


Donghyun Kim


Xingchao Peng


Kuniaki Saito

References

- ADDA: Adversarial discriminative domain adaptation. Tzeng, Hoffman, Saenko, Darrell. CVPR 2017
- CyCADA: Cycle-Consistent Adversarial Domain Adaptation. Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, ICML 2018
- COCO-FUNIT: Few-Shot Unsupervised Image Translation with a Content Conditioned Style Encoder, Saito, Saenko, Liu. ECCV'20
- Strong-Weak DA: Strong-Weak Distribution Alignment for Adaptive Object Detection. Saito, Yoshitaka Ushiku, Harada, Saenko, CVPR'19
- ADR: Adversarial Dropout Regularization, Saito, Ushiku, Harada, Saenko, ICLR 2018
- MME: Semi-Supervised Domain Adaptation via Minimax Entropy, Saito, Kim, Sclaroff, Darrell and Saenko, ICCV 2019
- Universal Domain Adaptation through Self Supervision, Saito, Kim, Sclaroff, Saenko, arXiv:2002.07953m, 2020
- DomainNet: Moment Matching for Multi-Source Domain Adaptation, Peng et al. ICCV 2019
- Domain2Vec: Domain2Vec: Domain Embedding for Unsupervised Domain Adaptation, Peng et al. ECCV 2020