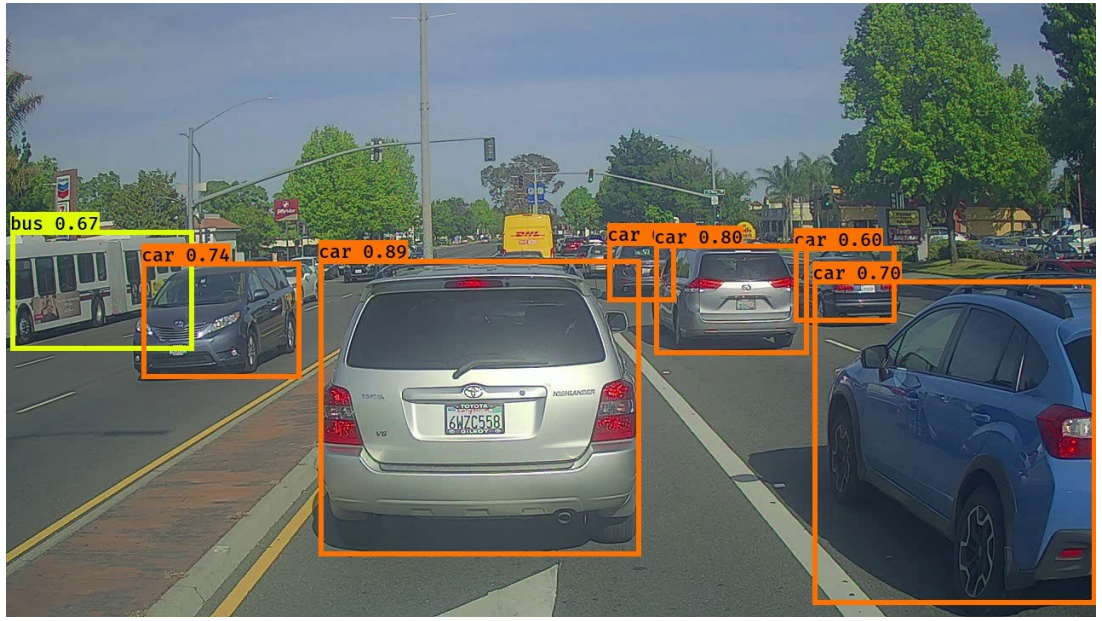
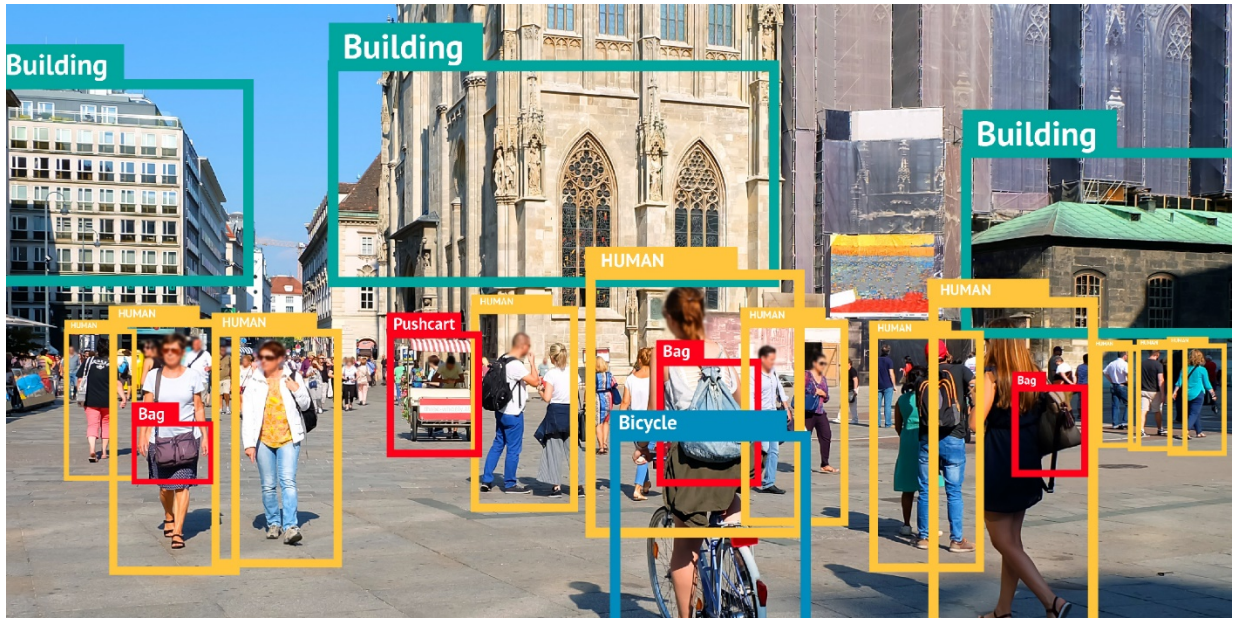
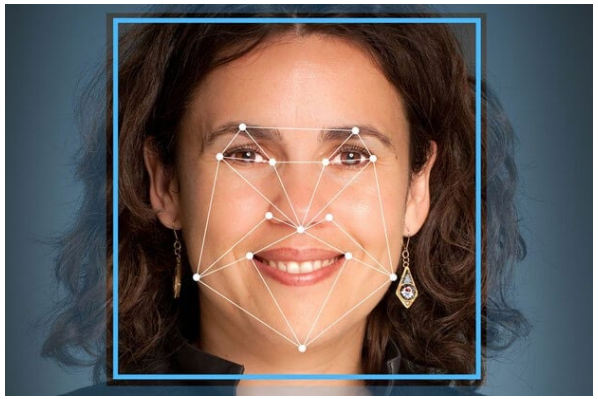
Fighting Dataset Bias in Computer Vision

Kate Saenko
Boston University & MIT-IBM Watson AI Lab

Successes of AI and computer vision

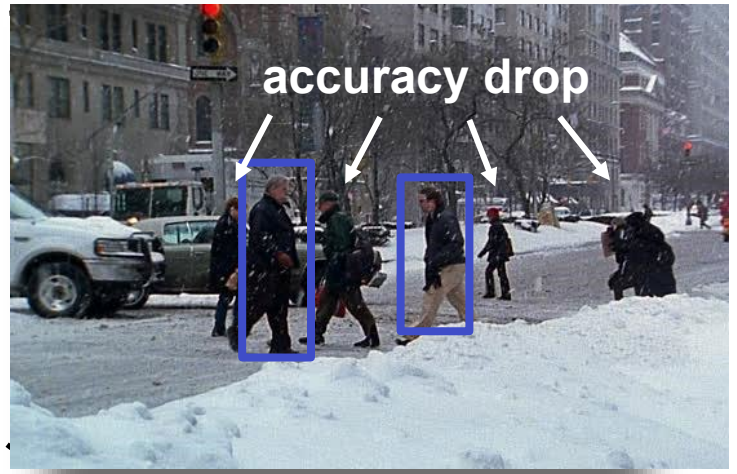




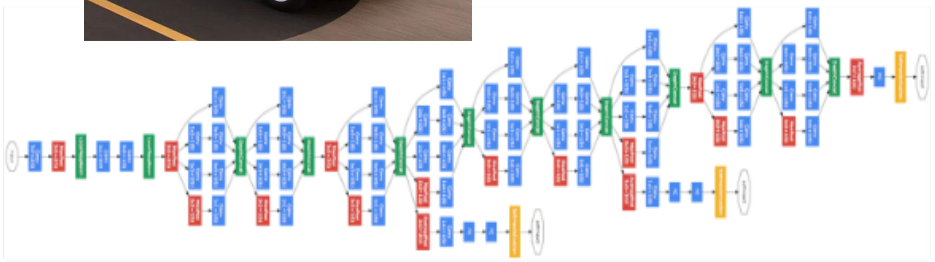


Problem: dataset bias

What your net is trained on

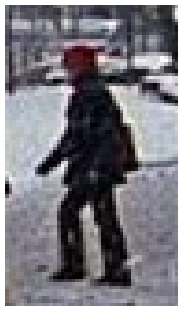


What it's asked to label



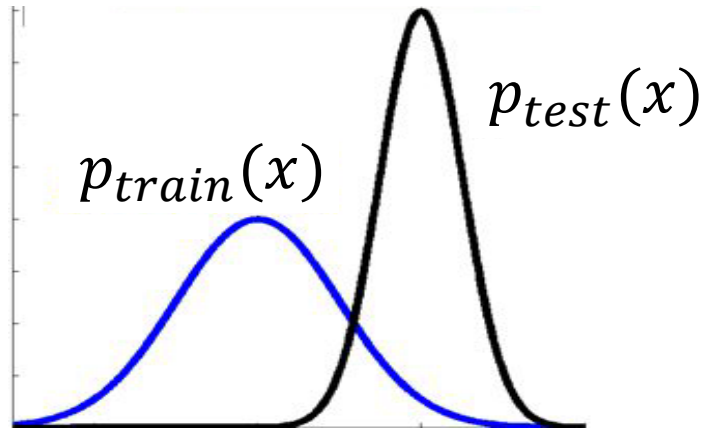
"Dataset Bias"
"Domain Shift"

vs

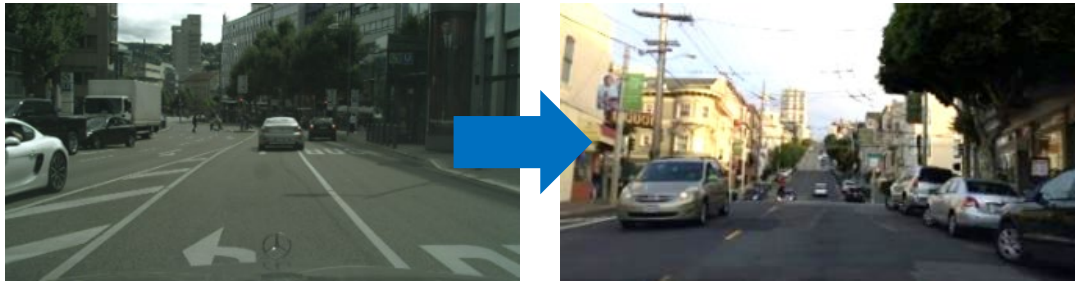


Adapting visual category models to new domains K Saenko, B Kulis, M Fritz, T Darrell, ECCV, 2010

When does dataset bias happen?

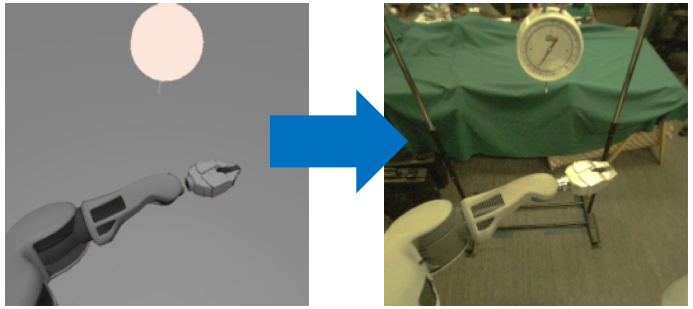


From one city to another



From web to robot

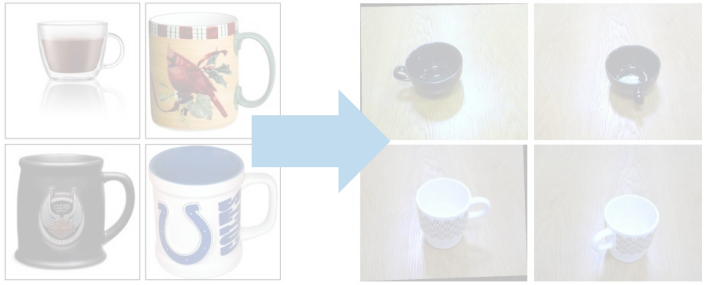
From simulated to real control



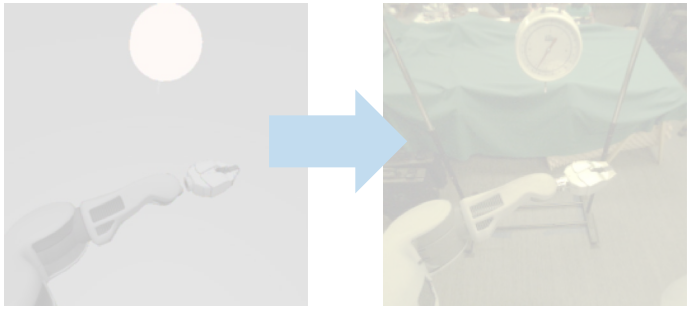
When does dataset bias happen?

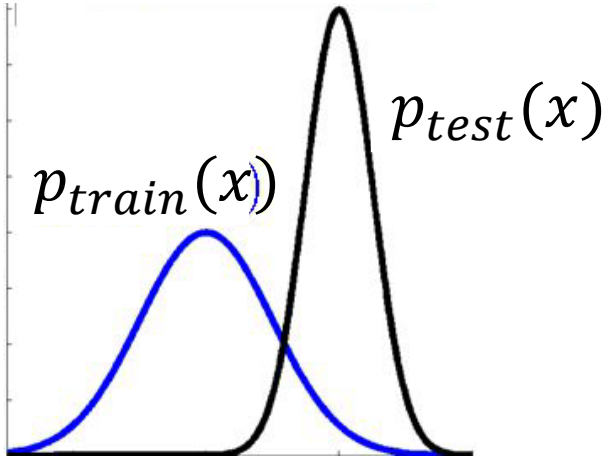
From one city to another

From web to robot

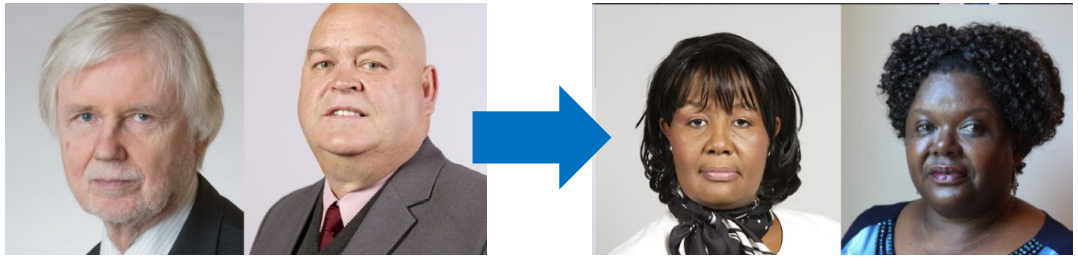


From simulated to real control



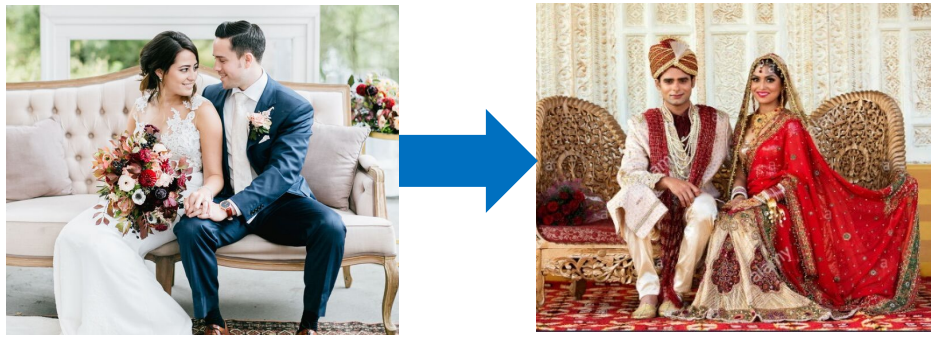


From one demographic to another

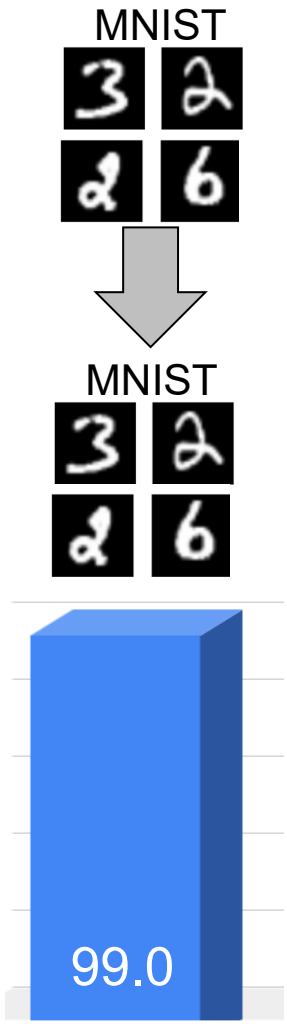


https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html

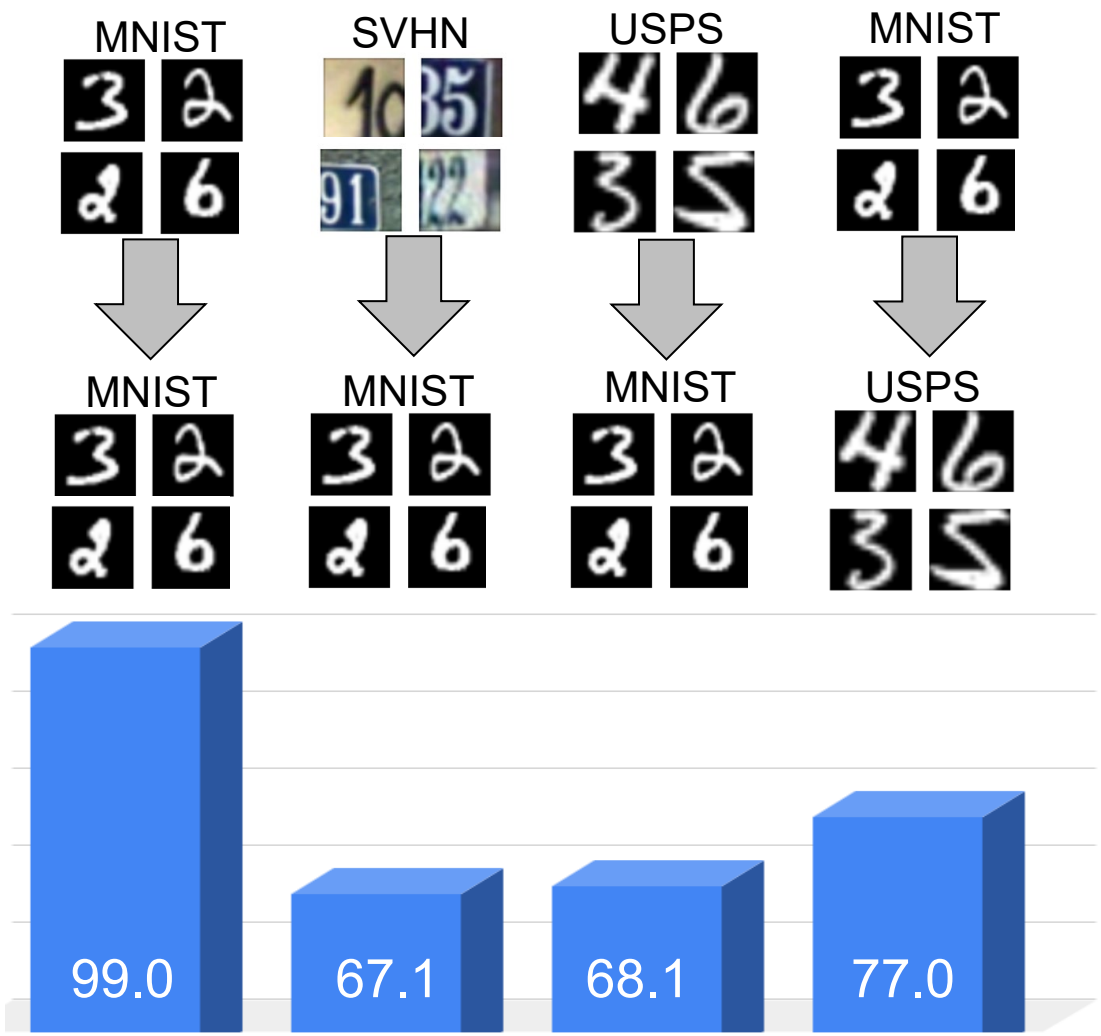
From one culture to another



Dataset bias reduces accuracy



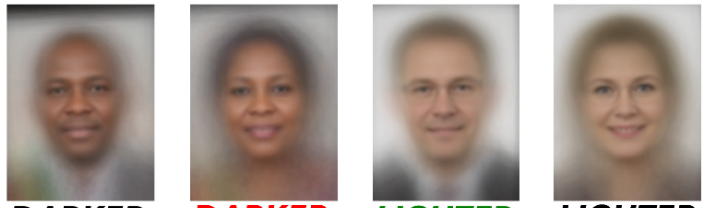
Domain bias reduces accuracy



Real-world implications of dataset bias

August 2018 Accuracy on Facial Analysis Pilot Parliaments Benchmark

amazon



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.

Real-world implications of dataset bias

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 pedestrian polygons costs ~$1,000

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

…

What causes poor performance?

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

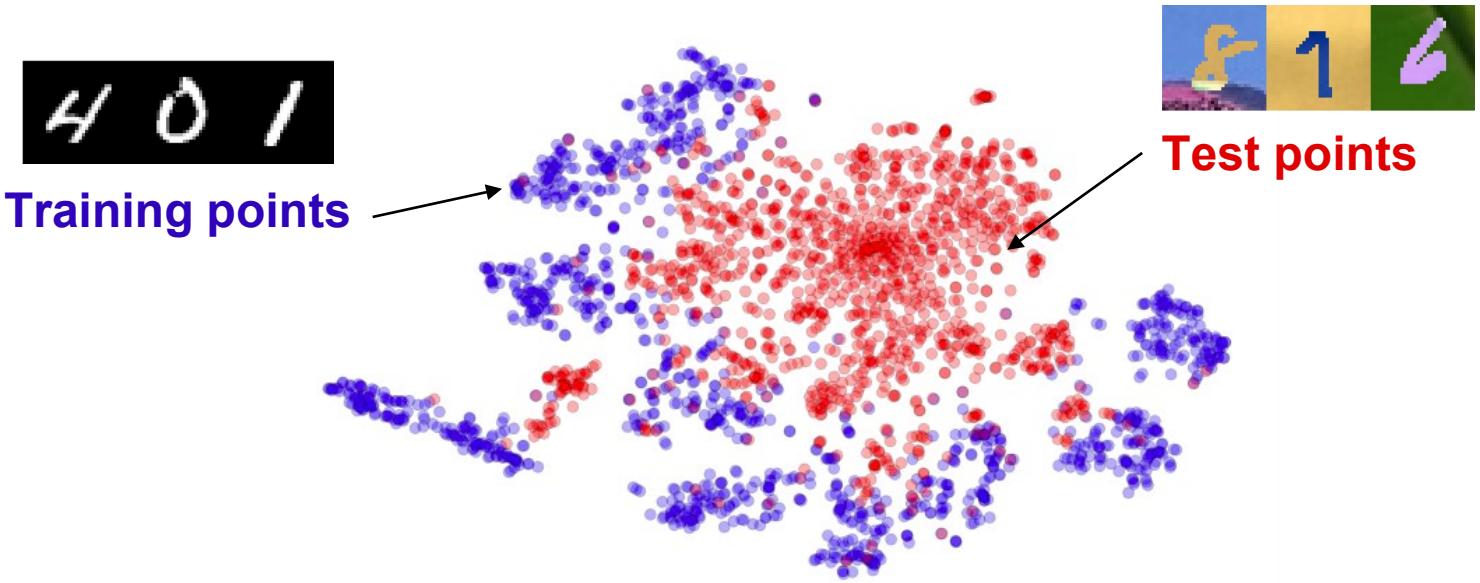
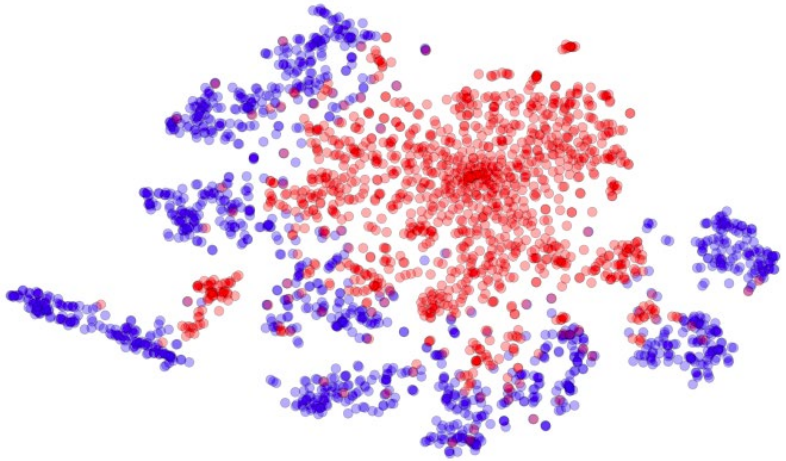


Figure from Ganin and Lempitsky. "Unsupervised domain adaptation by backpropagation." ICML 2015

Techniques that help deal with data bias



- Collect some labelled data from target domain
- Better source model
- 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)

Domain adaptation: adapt knowledge to new domain

Source Domain D_S

lots of labeled data

$\neq$

Target Domain D_T

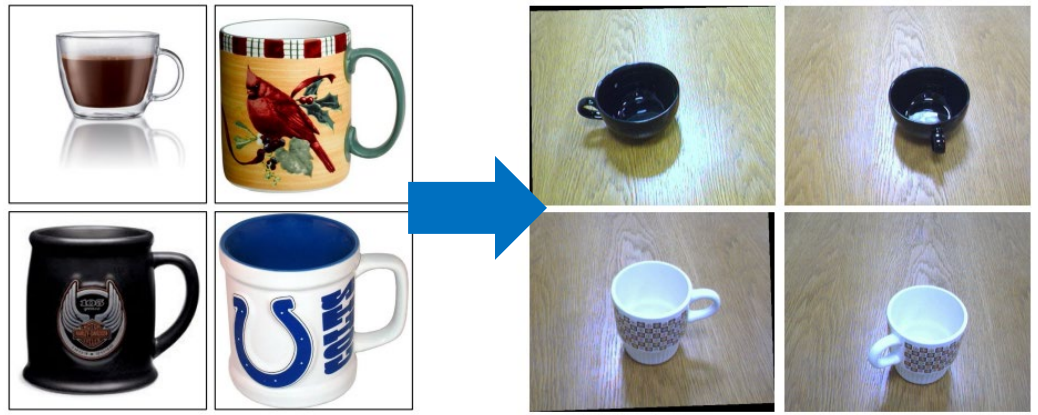
unlabeled data

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

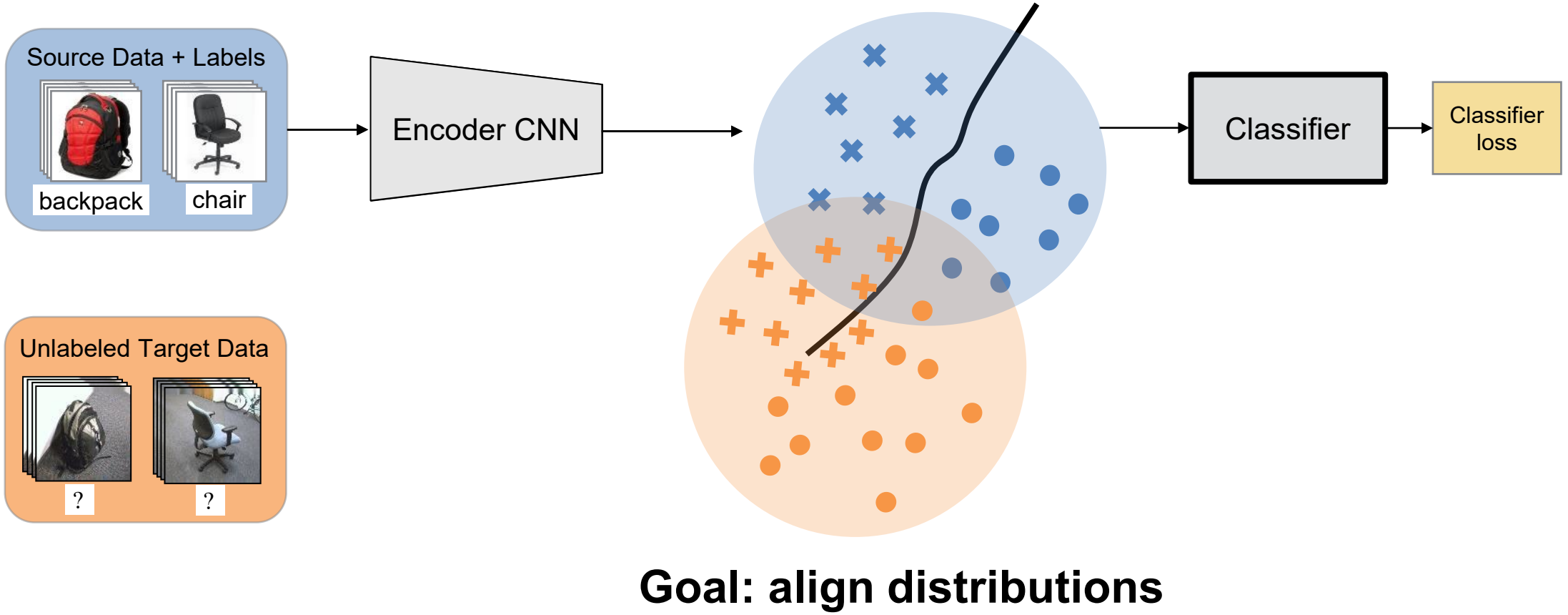
Assume: we get to see the unlabeled target data, but not its labels

Outline

- Adversarial domain alignment
  - Feature-space
  - Pixel-space
  - Few-shot pixel alignment
- Beyond alignment
  - Self-supervised learning
  - Consistency



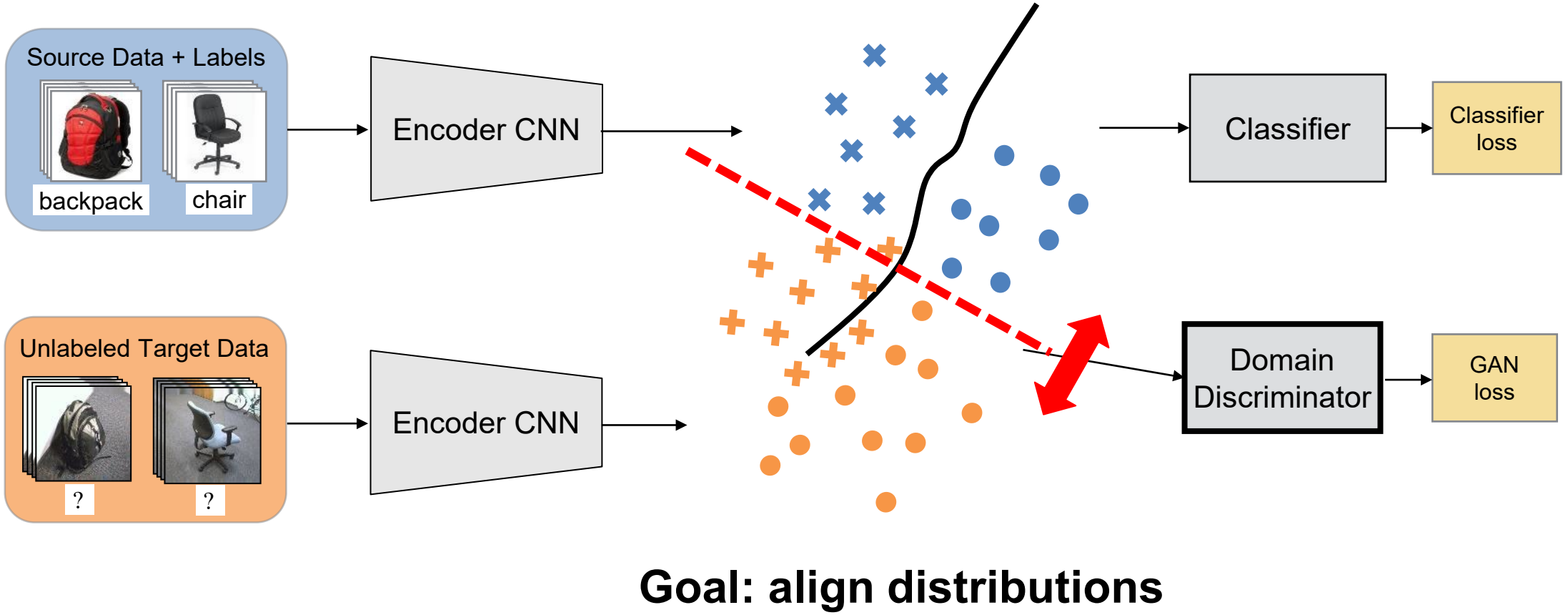
Adversarial domain alignment



Goal: align distributions

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

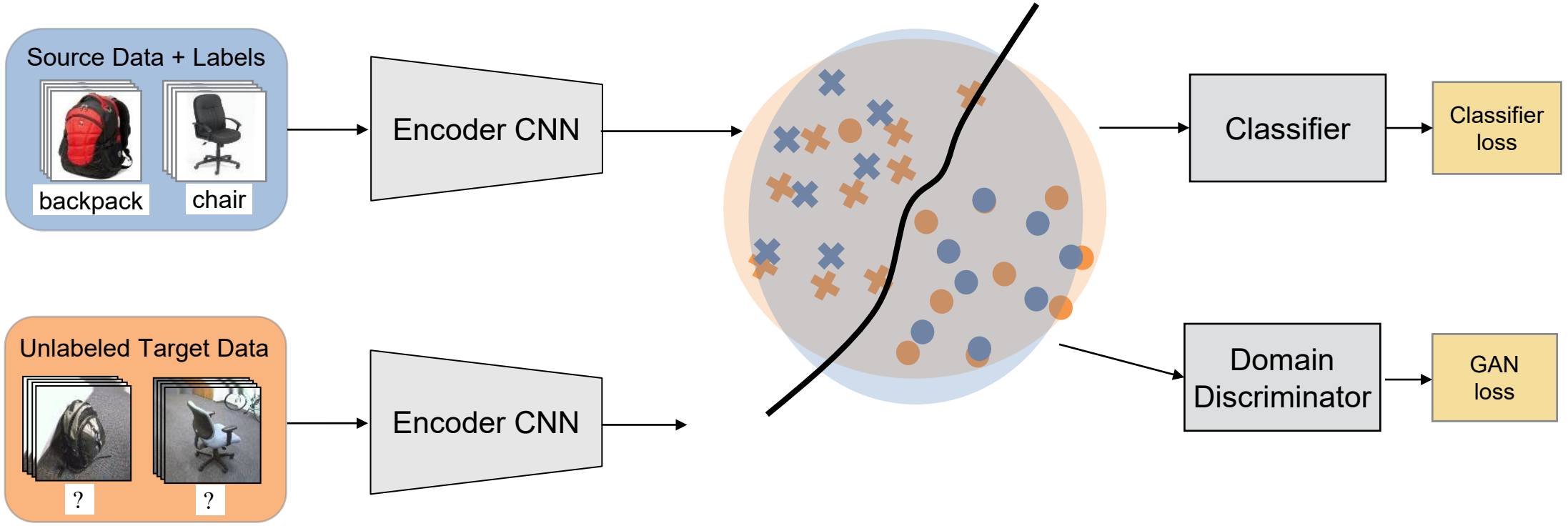
Adversarial domain alignment



Goal: align distributions

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

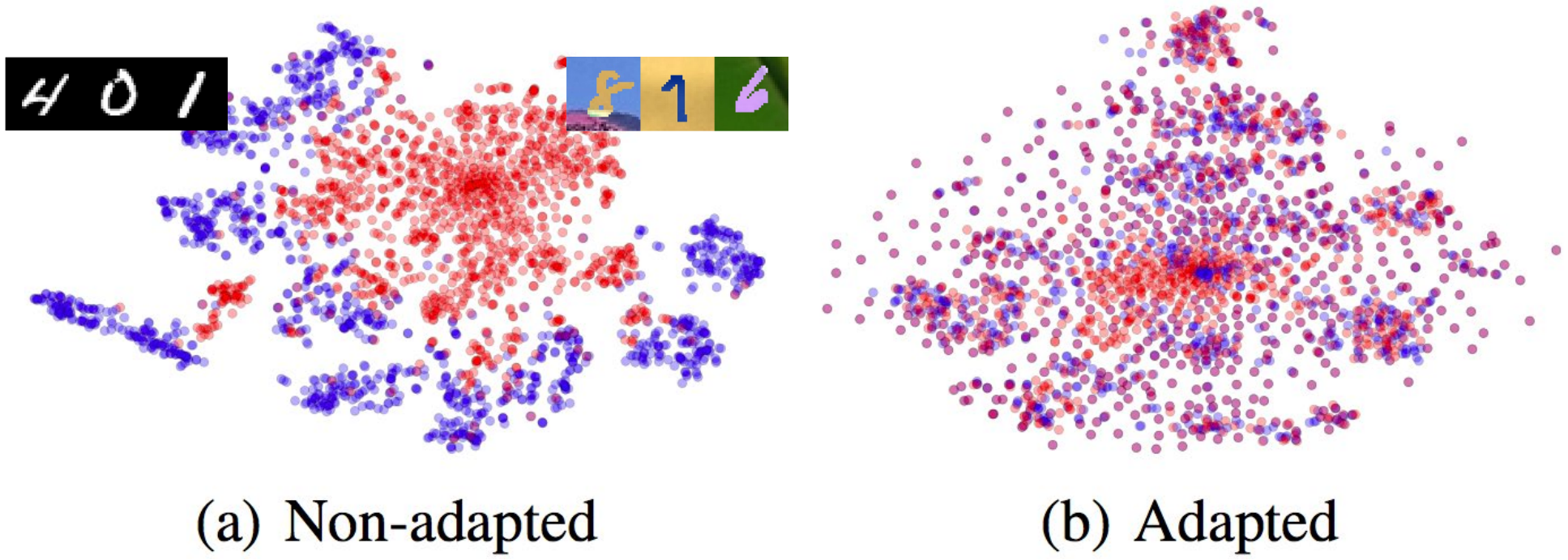
Adversarial domain alignment



Goal: align distributions

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

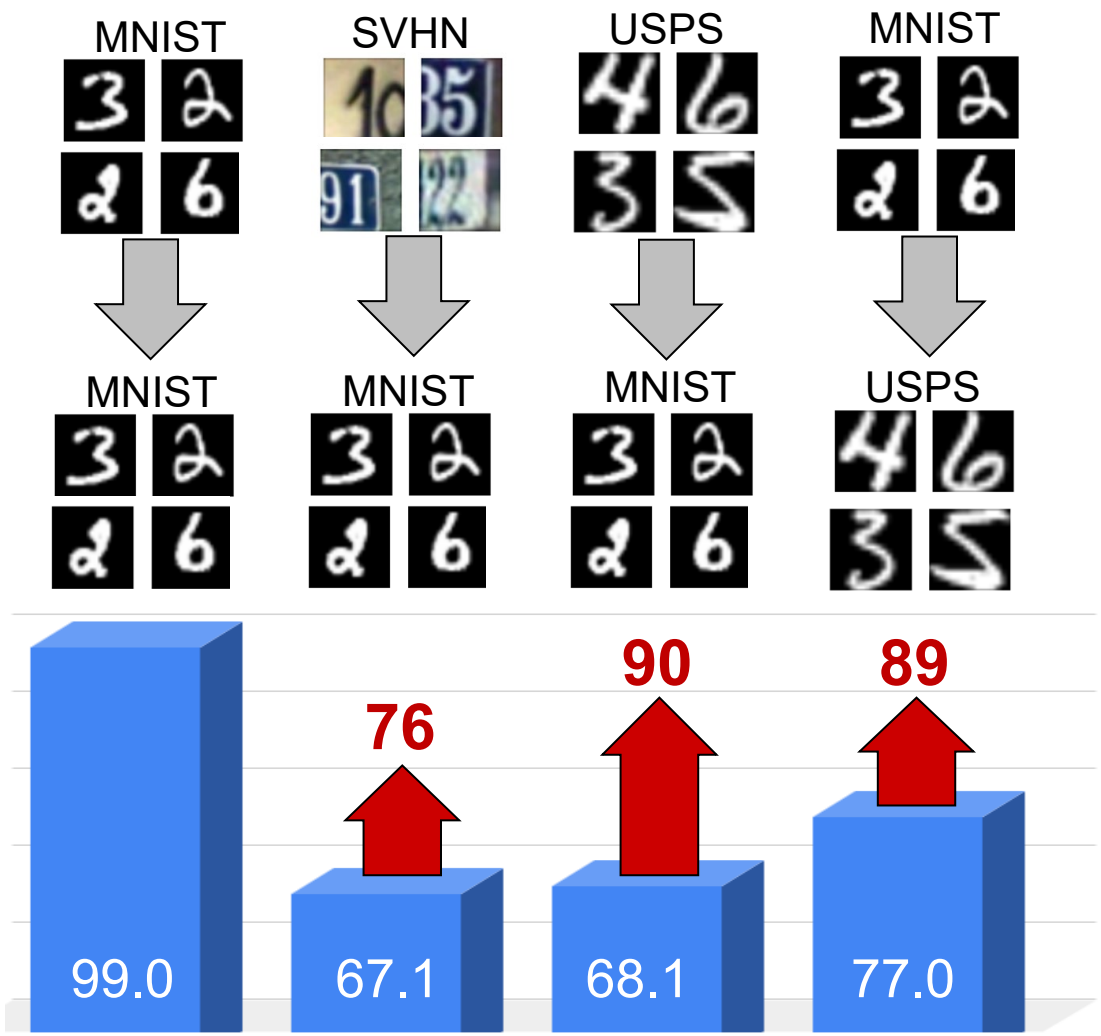
Domain alignment: feature visualization on digits



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

Domain alignment results on digits



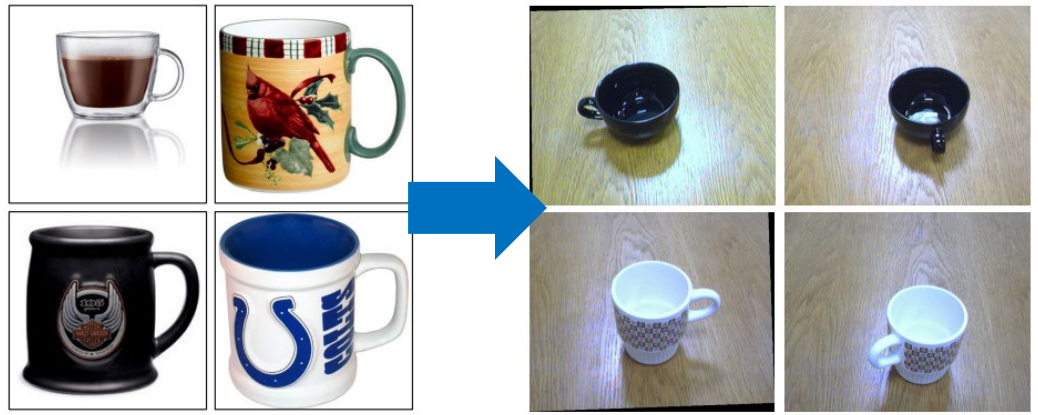
Takeaway:

Domain adaptation can improve accuracy on target data without any labels; "unsupervised fine-tuning"

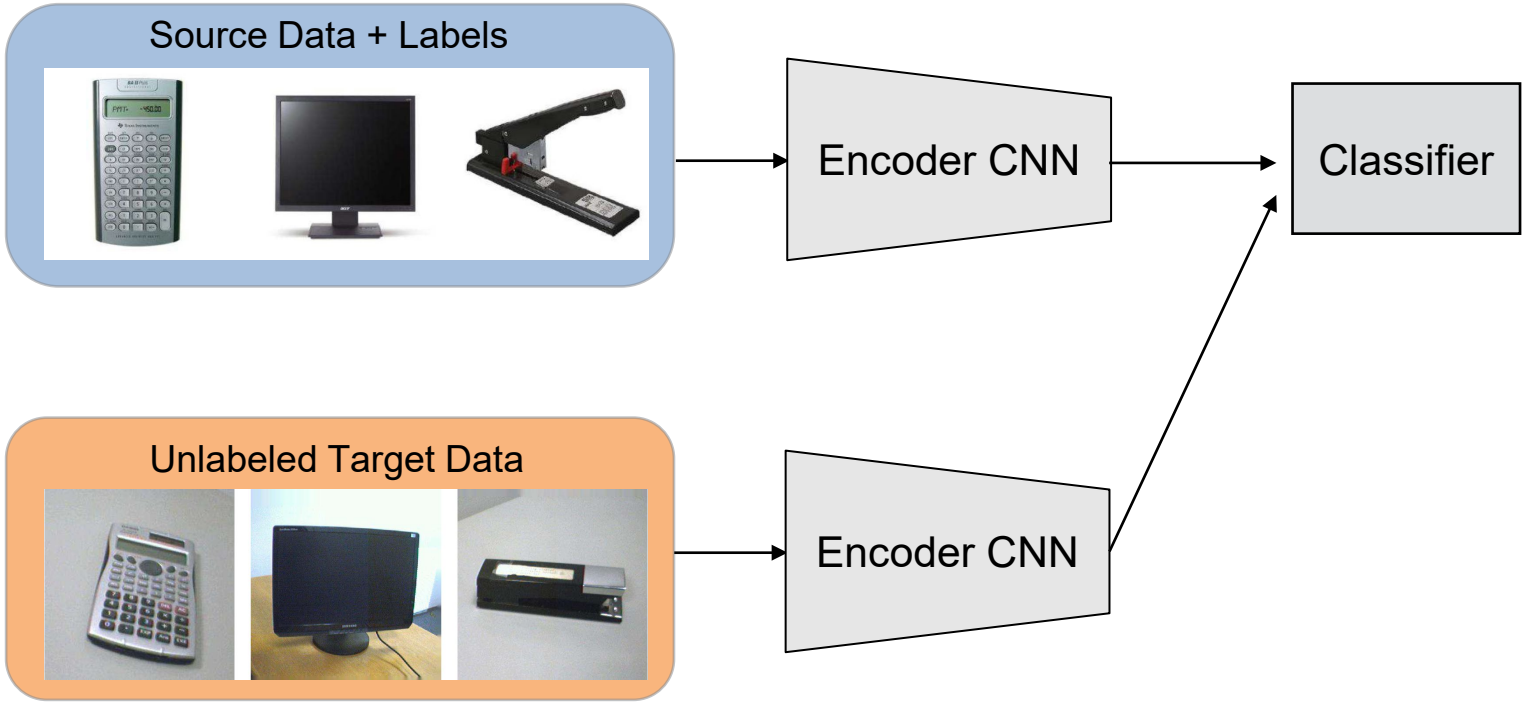
ADDA: Tzeng, Eric, et al. "Adversarial discriminative domain adaptation." CVPR 2017.

Outline

- Adversarial domain alignment
  - Feature-space
  - Pixel-space
  - Few-shot pixel alignment
- Beyond alignment
  - Self-supervised learning
  - Consistency

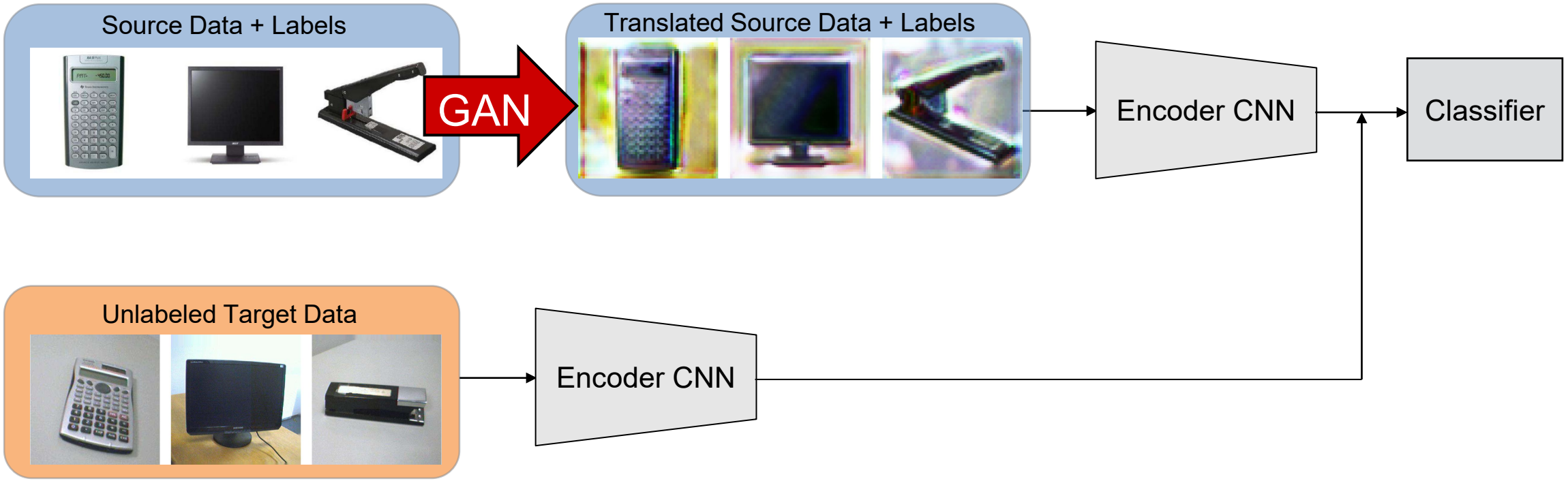


Pixel-space domain alignment



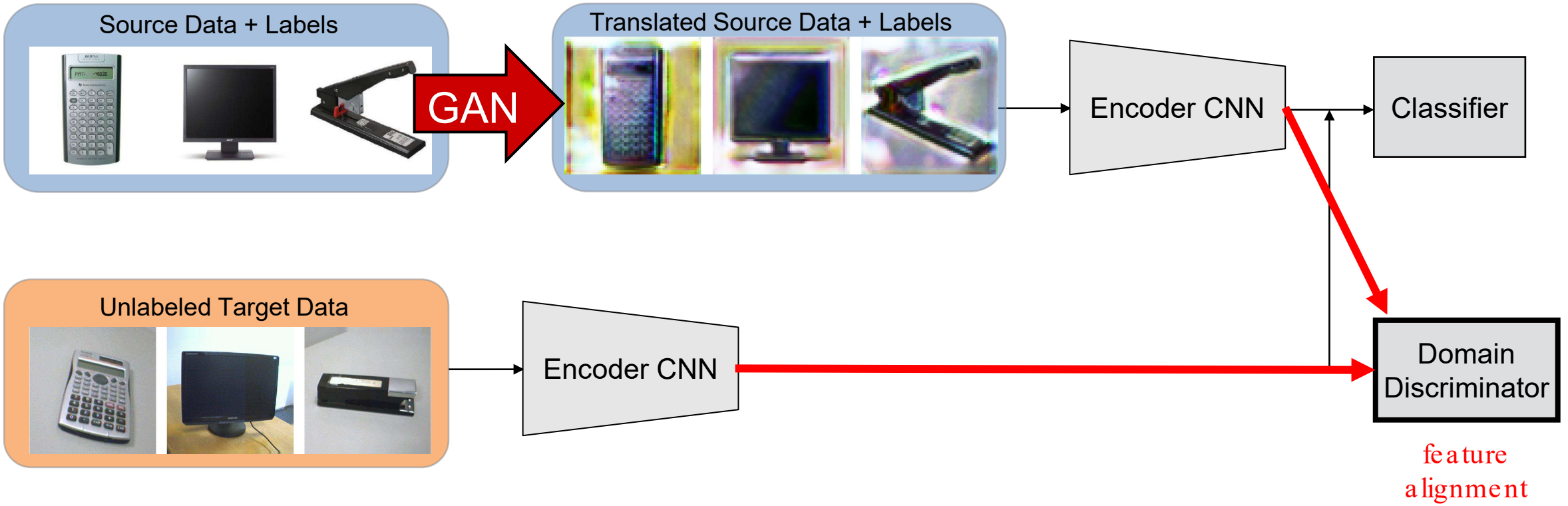
CyCADA, Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, ICML 2018
From source to target and back: Symmetric Bi-Directional Adaptive GAN, Russo et al. CVPR 2018

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

Synthetic to real pixel adaptation

Train

GTA
(synthetic)

Test

CityScapes (Germany)

CyCADA, Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, ICML 2018

Real to synthetic translation

https://youtu.be/lCR9sT9mbis

CycleGAN. Zhu*, Park*, Isola, Efros. ICCV 2017.

Adaptation with CyCADA

Source image (GTA5)

Adapted source image (**Ours**)

Target image (CityScapes)

Pixel accuracy on target

Source-only: 54.0%

Adapted (**ours**): **83.6%**

CyCADA, Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, ICML 2018

Adaptation with CyCADA

Source image (GTA5)

Adapted source image (**Ours**)

Target image (CityScapes)

Pixel accuracy on target
Source-only: 54.0%
Adapted (**ours**): **83.6%**

Source images (SVHN)

Adapted source images (**Ours**)

Target images (MNIST)

Accuracy on target
Source-only: 67.1%
Adapted (**ours**): **90.4%**

(cf. 76% without pixel alignment)

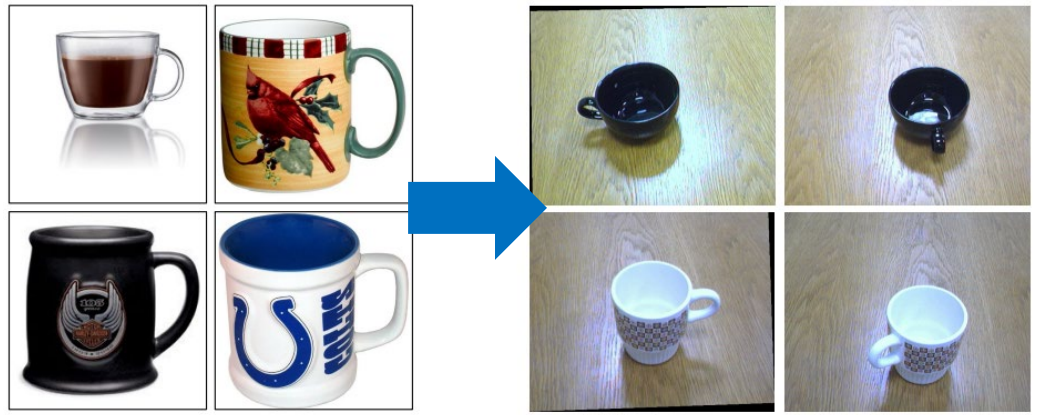
Takeaway:

Unsupervised image-to-image translation can discover and align corresponding structures in the domains

CyCADA, Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, ICML 2018

Outline

- Adversarial domain alignment
  - Feature-space
  - Pixel-space
  - Few-shot pixel alignment
- Beyond alignment
  - Self-supervised learning
  - Consistency



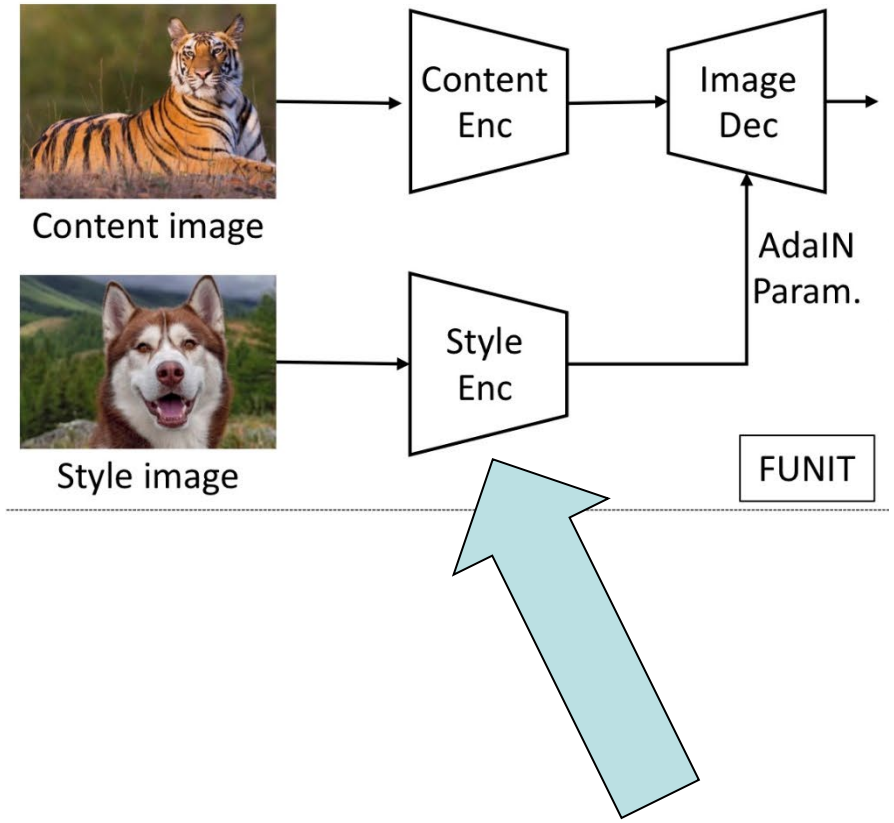
Few-shot domain translation

- 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/

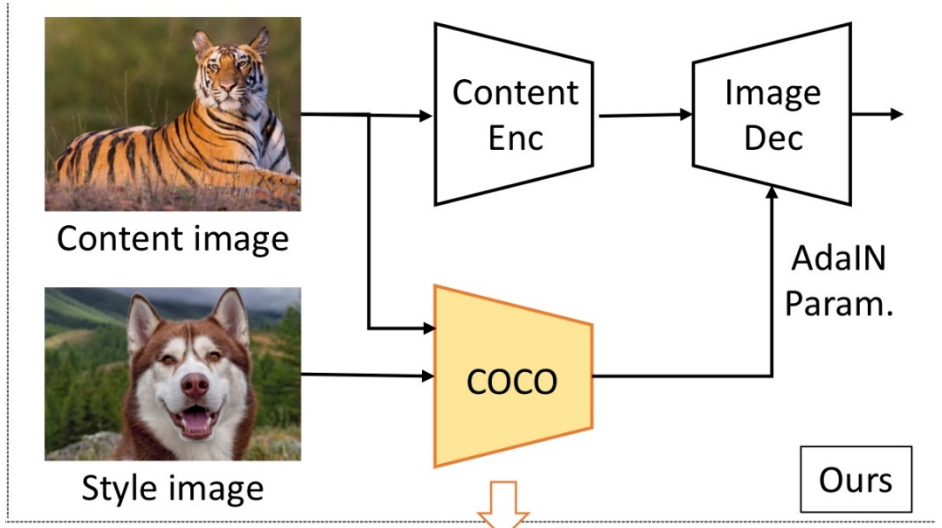
COCO-FUNIT

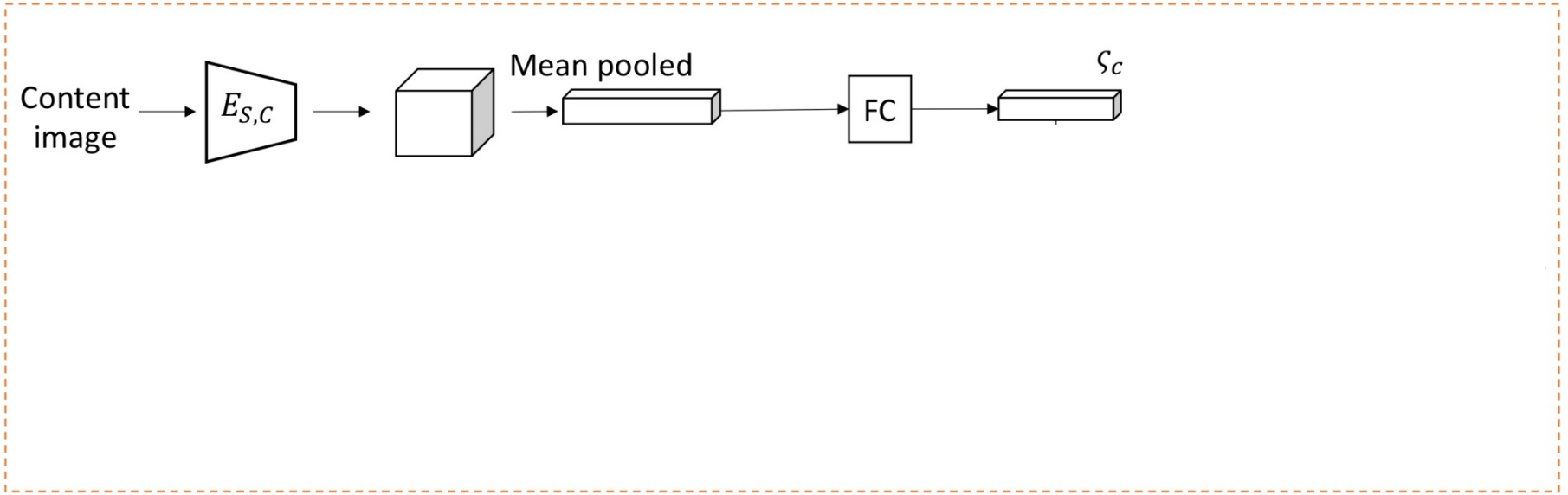
Previous work: FUNIT



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

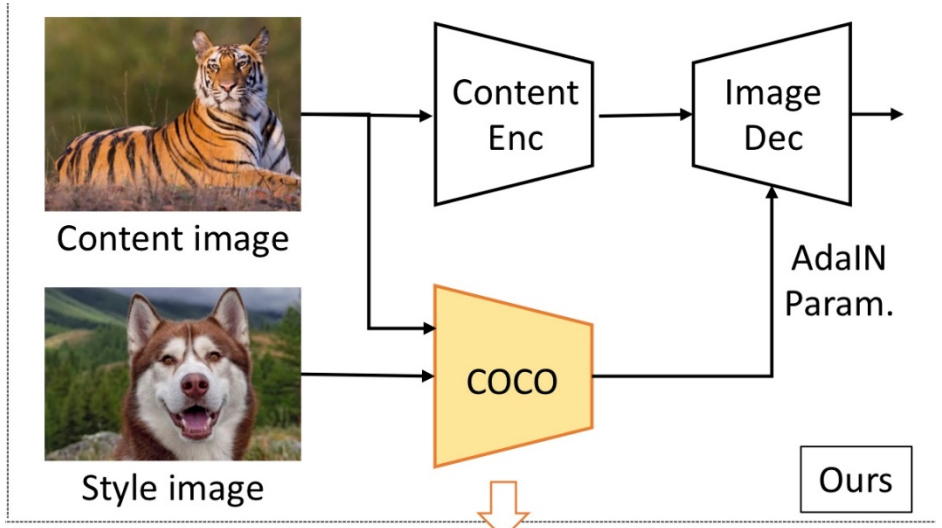
COCO-FUNIT

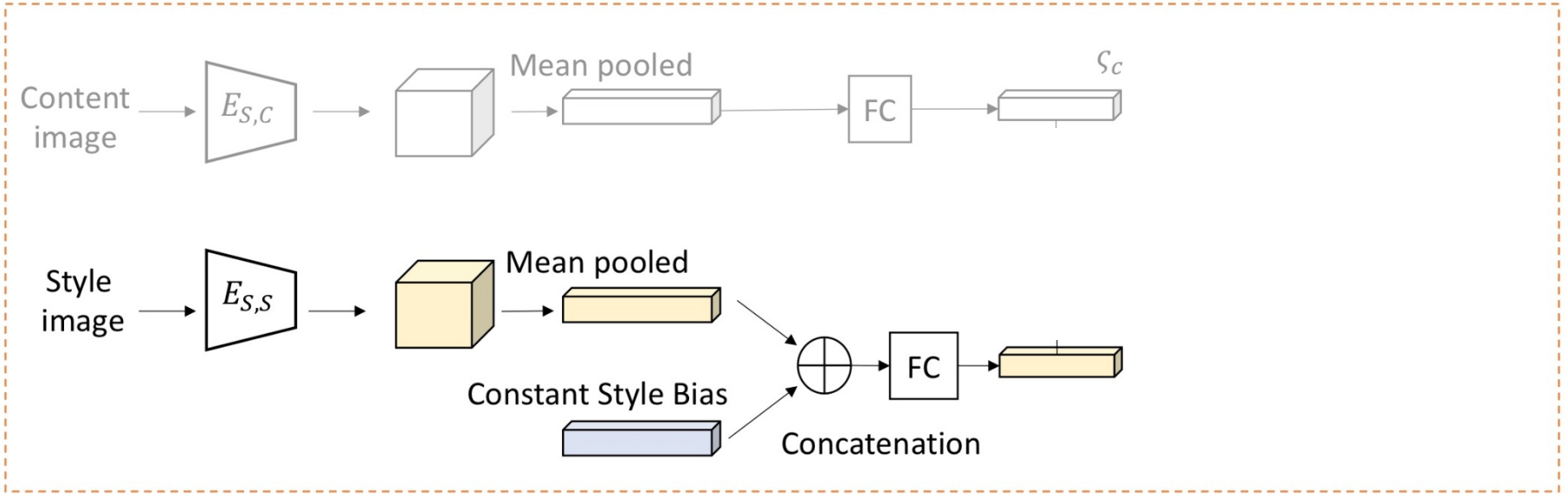




COCO-FUNIT: Few-Shot Unsupervised Image Translation with a Content Conditioned Style Encoder, Saito, Saenko, Liu. ECCV2020 Spotlight

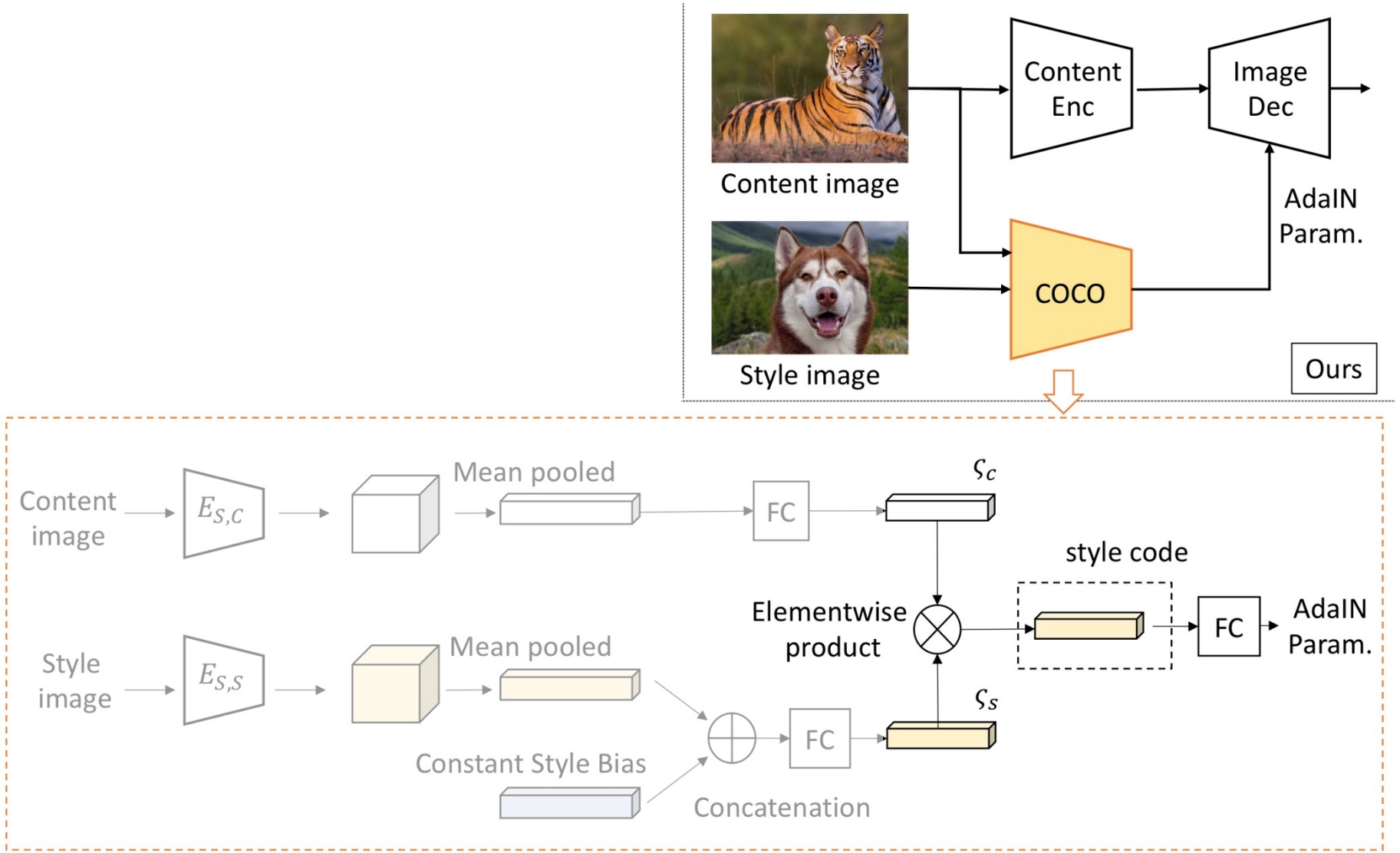
COCO-FUNIT





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

COCO-FUNIT



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

COCO-FUNIT

Few-shot domain translation with COCO-FUNIT

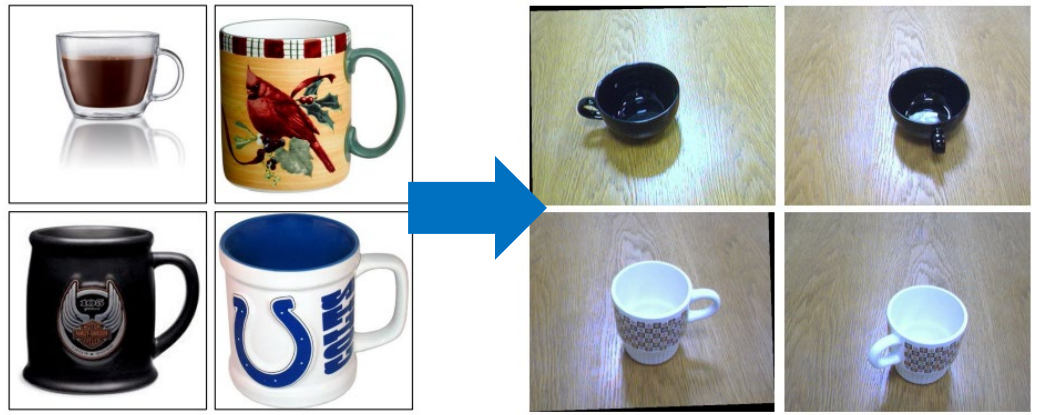
Takeaway:

Conditioning on content and style image results in better encoding of style

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

Outline

- Adversarial domain alignment
  - Feature-space
  - Pixel-space
  - Few-shot pixel alignment
- Beyond alignment
  - Self-supervised learning
  - Consistency

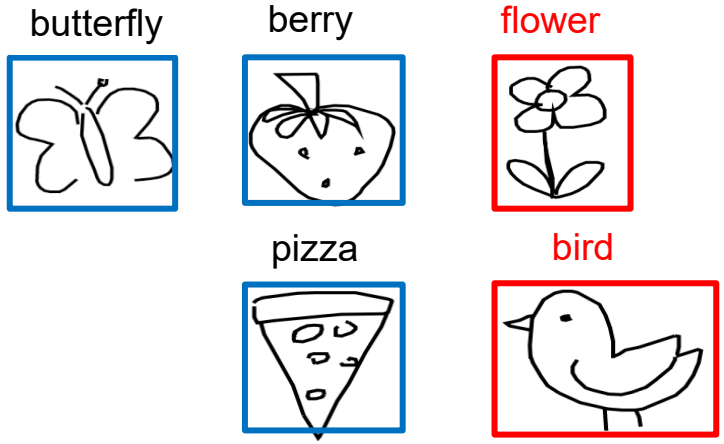


Problem: Category shift

When categories are not the same in source and target

Source domain

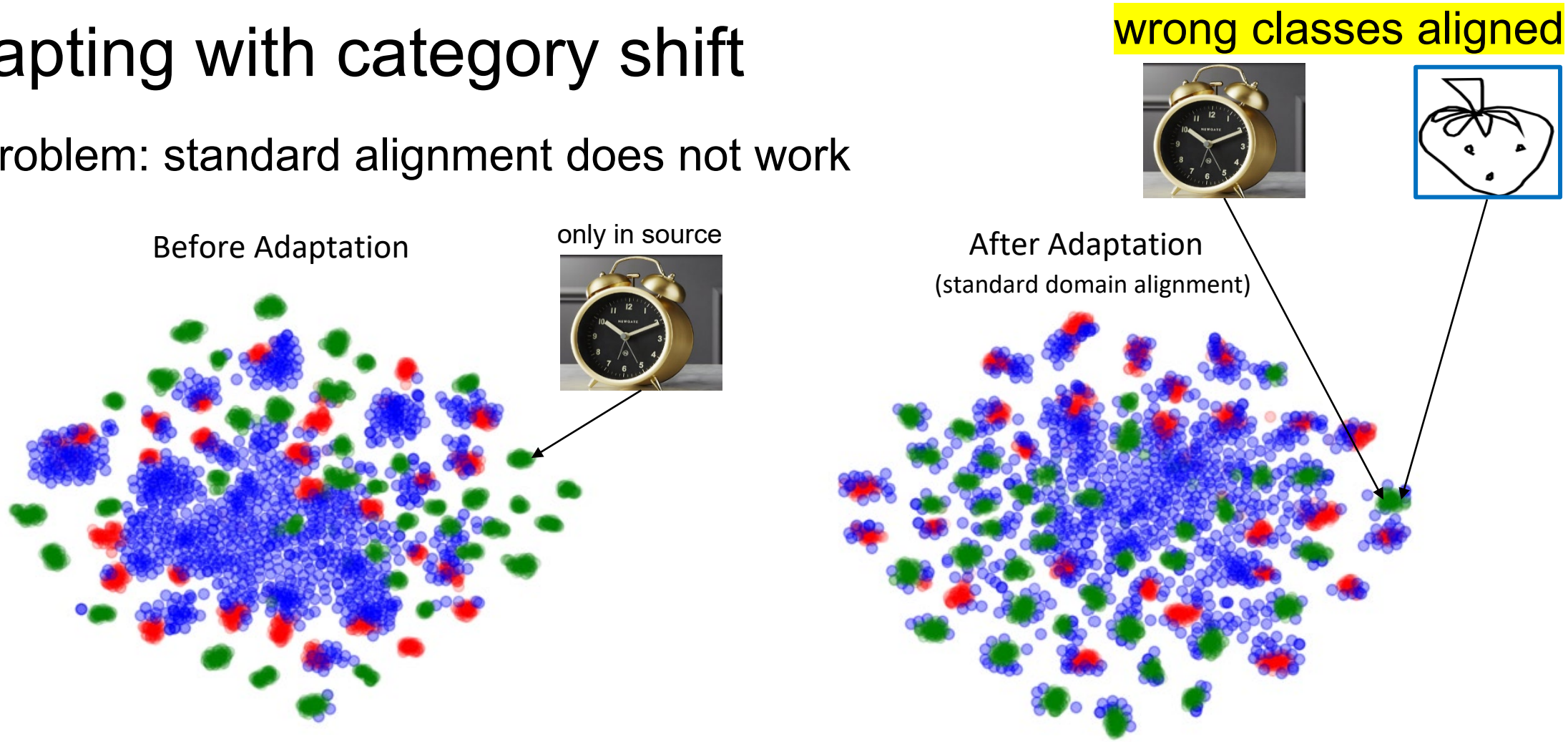
Transfer



Target domain

Adapting with category shift

- Problem: standard alignment does not work



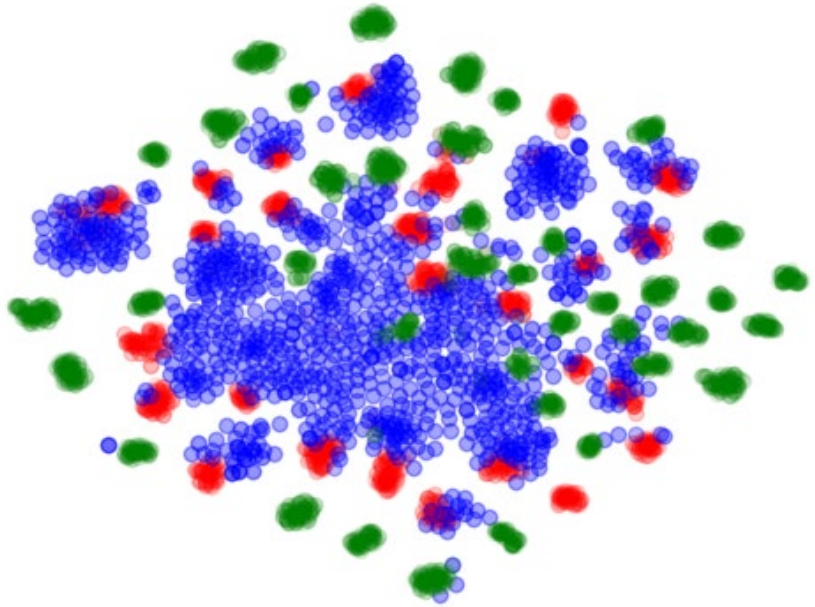
Feature visualization with t-SNE on OfficeHome partial domain adaptation

Red: Features of the source classes shared with target. Blue: Features of the target domain. Green: Features of the source-only classes.

Adapting with category shift

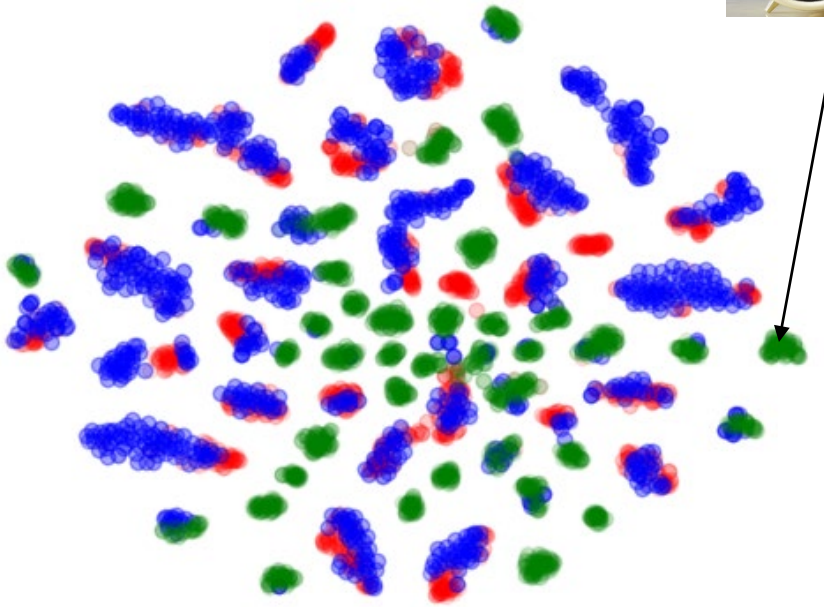
- Our method avoids alignment with missing source classes

Before Adaptation



After Adaptation

(our method)

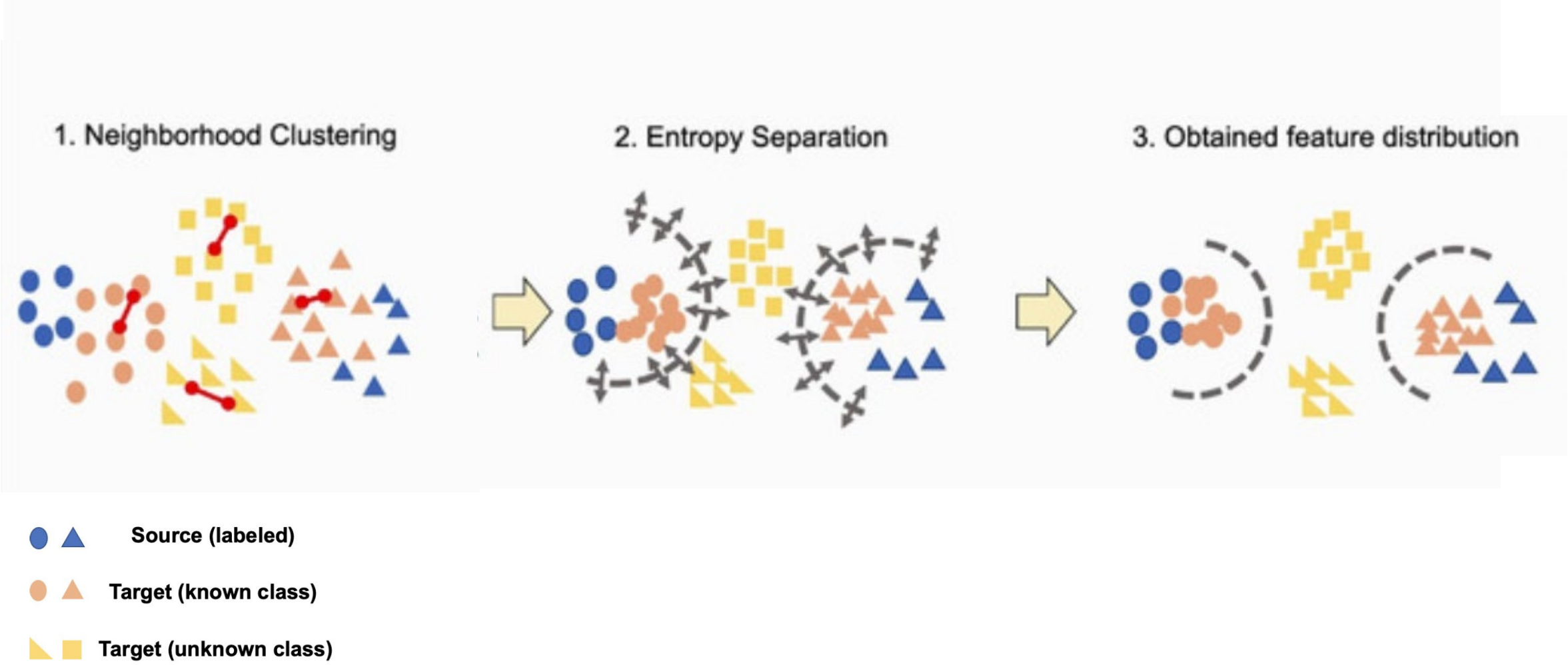


not aligned

Feature visualization with t-SNE on OfficeHome partial domain adaptation

Red: Features of the source classes shared with target. Blue: Features of the target domain. Green: Features of the source-only classes.

DANCE: domain adaptation with neighborhood clustering



https://github.com/VisionLearningGroup/DANCE

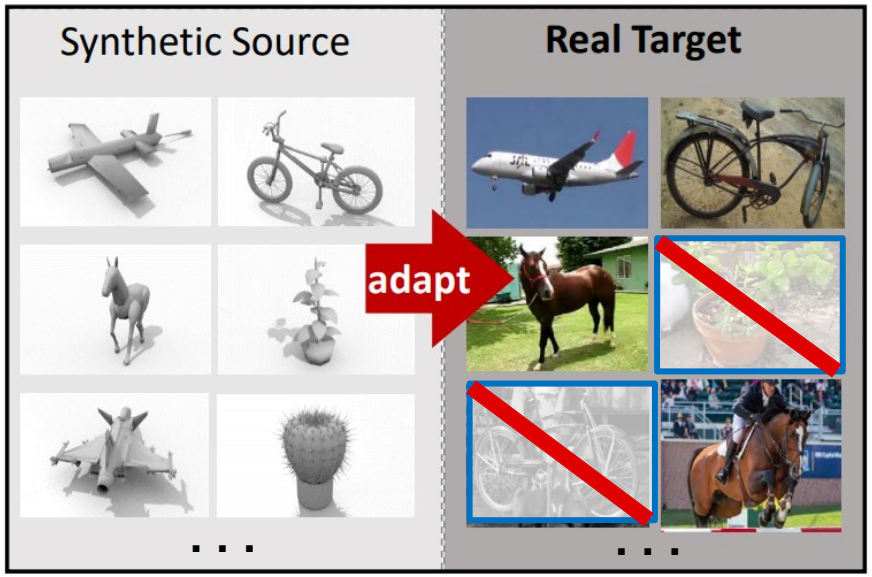
Universal Domain Adaptation through Self Supervision, Saito, Kim, Sclaroff, Saenko, NeurIPS, 2020

Adapting with category shift

- Example: syn2real object recognition
- 6 categories missing in target (“partial” shift)

- **Our method “DANCE” improves accuracy compared to SOTA on this and other shifts**

VisDA Challenge



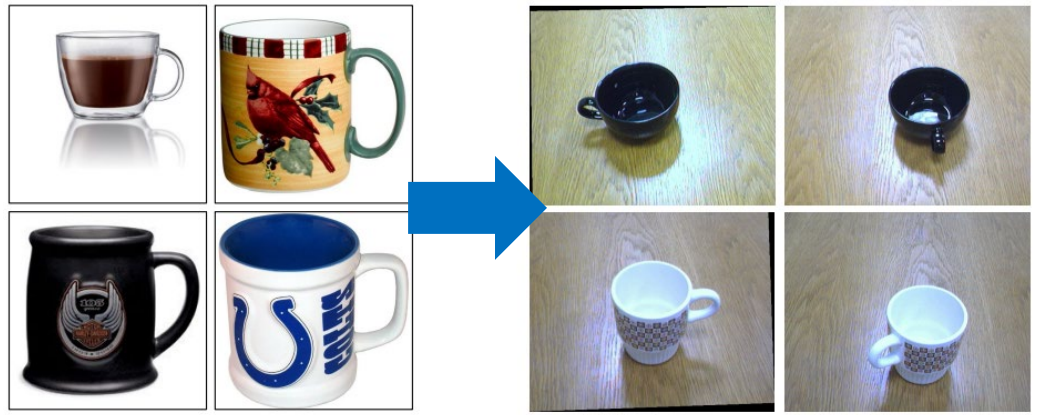
Source Only	49.9
DANN (Ganin et al., 2016)	38.7
ETN (Zhangjie Cao, 2019)	59.8
STA (Liu et al., 2019)	48.2
UAN (You et al., 2019)	39.7
DANCE (ours)	**73.7**

+14%

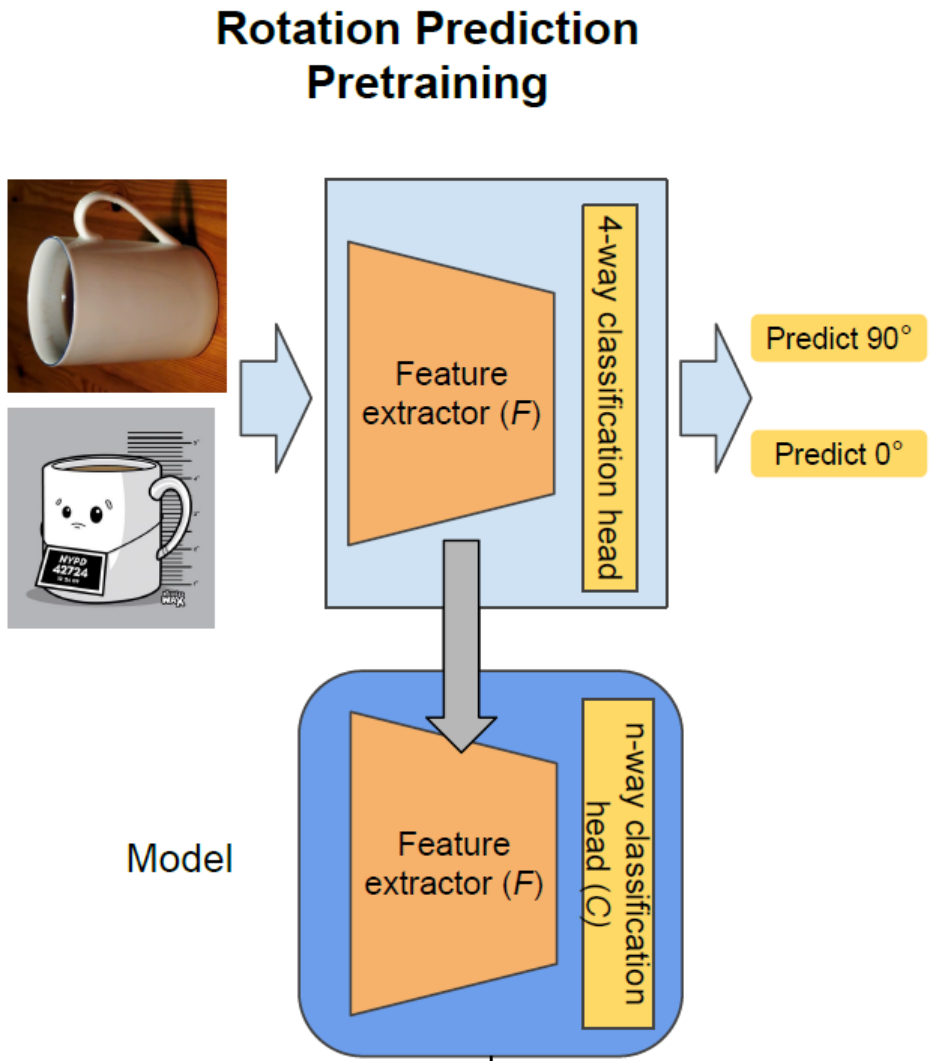
Universal Domain Adaptation through Self Supervision, Saito, Kim, Sclaroff, Saenko, NeurIPS, 2020

Outline

- Adversarial domain alignment
  - Feature-space
  - Pixel-space
  - Few-shot pixel alignment
- Beyond alignment
  - Self-supervised learning
  - Consistency

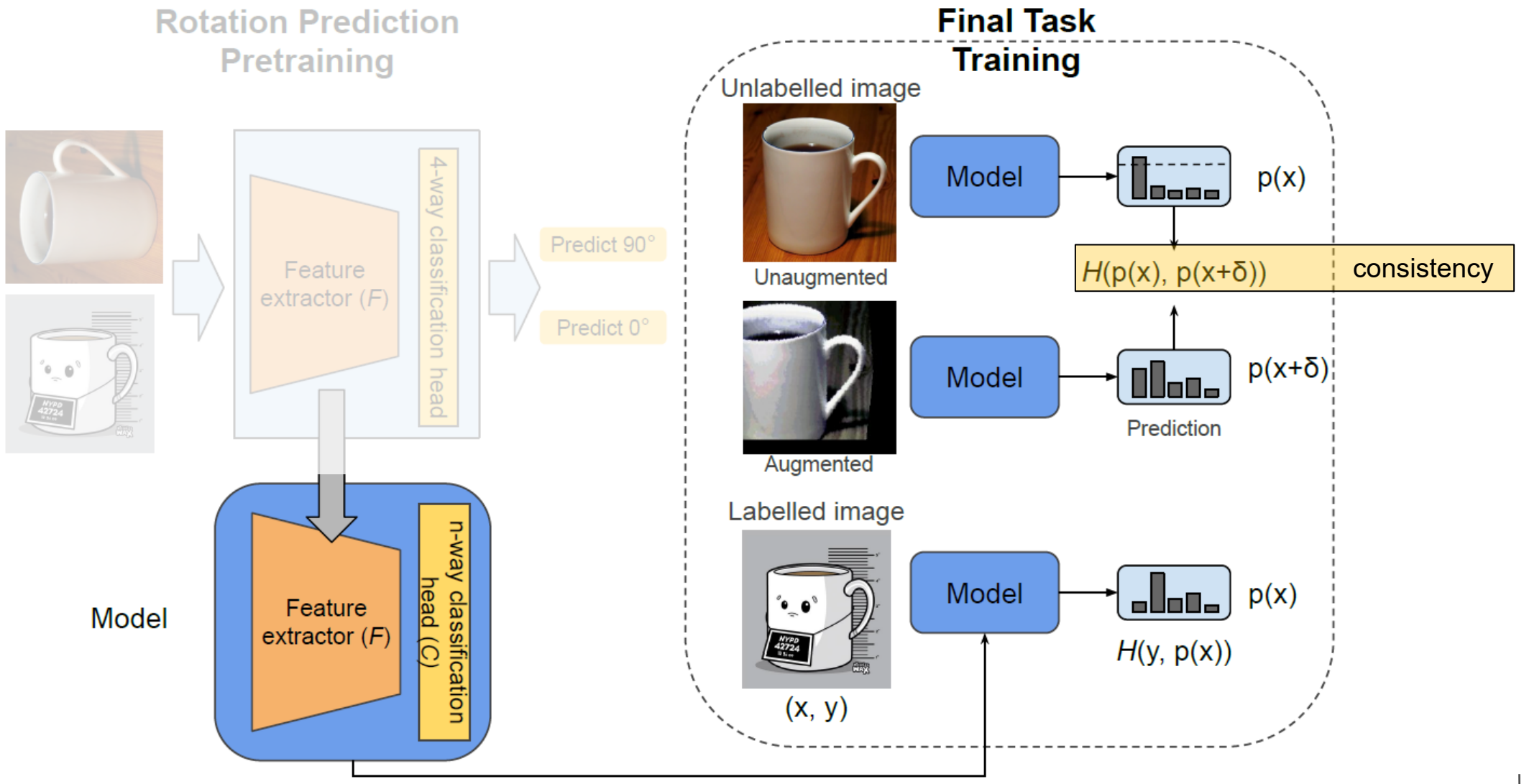


Self-supervised pretraining



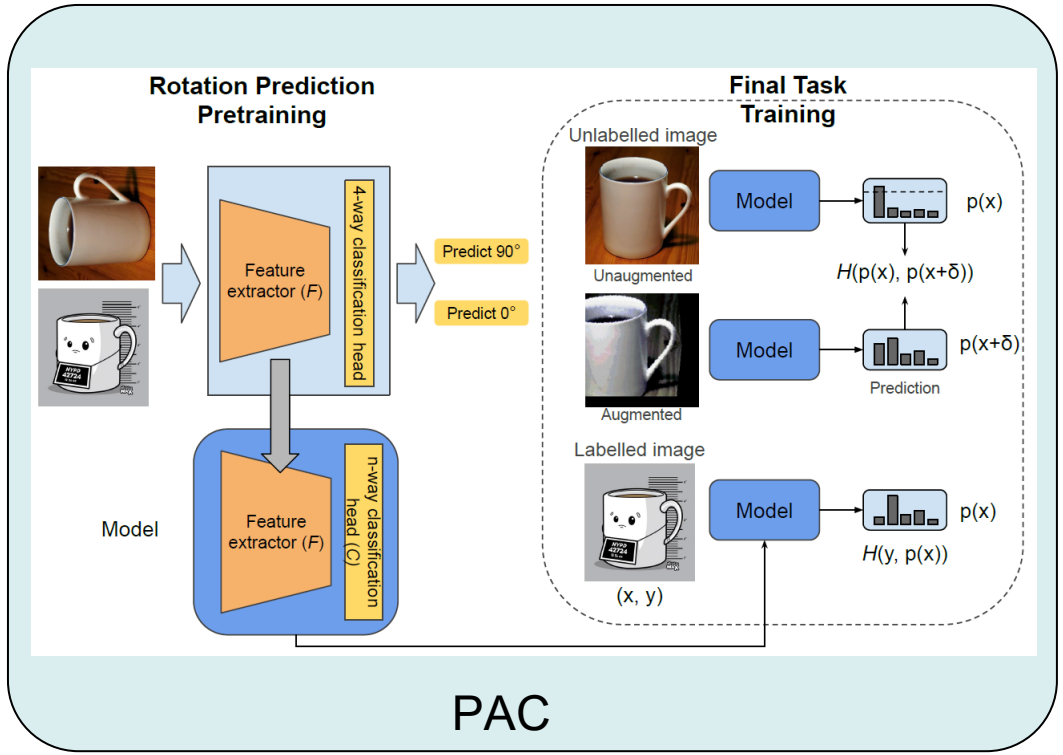
In submission, 2020

Self-supervised pretraining + consistency loss

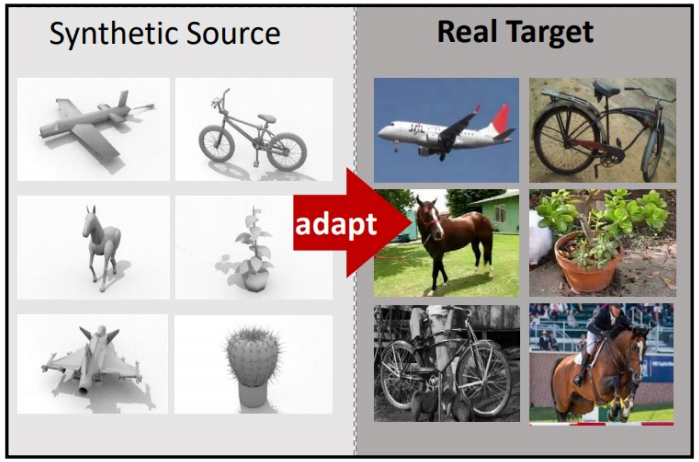


In submission, 2020

What if we have a few labels for real data?



VisDA Challenge



Method	Overall Accuracy 1-shot	3-shot
S+T	57.7	59.9
MME	69.7	70.7
PAC	**75.2**	**80.4**

In submission, 2020

Summary

- Dataset bias is a major problem!
- Domain adaptation: transfer knowledge using unlabeled data, i.e. “unsupervised fine-tuning”
- Adversarial domain alignment
  - Feature-space, Pixel-space, Few-shot pixel alignment
- Beyond alignment
  - Self-supervised learning, Consistency

MIT-IBM Watson AI Lab

HRI Honda Research Institute JP

NVIDIA


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