Today: Outline

- **Explainable AI and Domain Adaptation**
- **Exam Details**

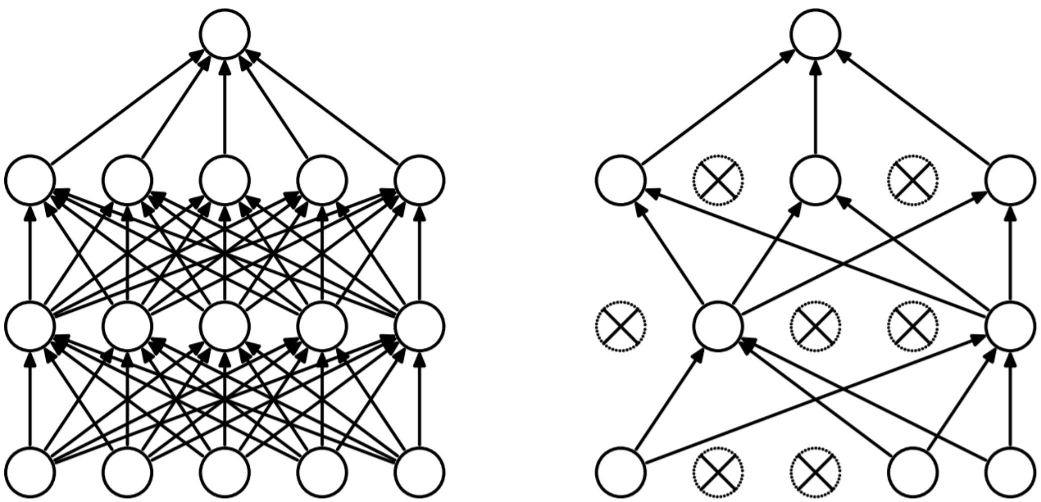
- Reminders:
  - Problem Set 1, due: Oct 12 by midnight
  - No class on Oct 13 per BU Calendar
    (Substitute Mon Schedule of Classes)
  - Midterm Exam, in class, Oct 20
  - Practice problems will be posted tomorrow
  - Thu Oct 15 will be a revision session + study group

Neural Networks VI

Explainability to Improve Network Performance

Dropout: A Classical Regularization Technique

- Many Deep Models employ dropout ***at training time*** to avoid overfitting, allowing for better generalization.



(a) Standard Neural Net (b) After applying dropout.

- In this work, we propose a scheme for biasing this neuron selection.

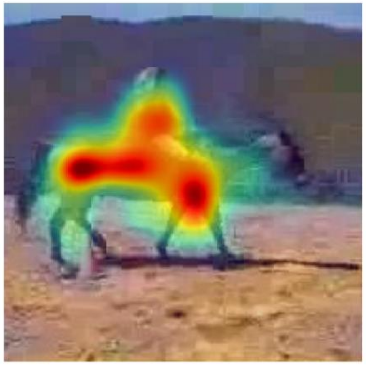
[Srivastava *et al.*]

Excitation Dropout

- We target answering the question: *Which neurons to drop out?*
  - *Neurons that have a higher contribution to the ground-truth prediction.*
  - *Example for ground-truth class HorseRiding:*

$image$

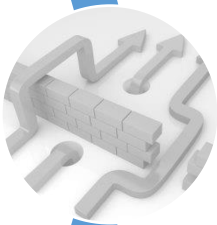
$evidence$: p_{EB}



Our Approach

Excitation Dropout

Dropping neurons having stronger evidence,



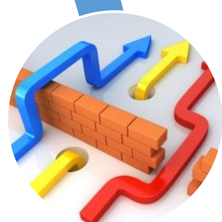
forces a network to learn alternative paths in order to maintain loss minimization,

leading to plasticity-like behavior, a characteristic of human brains too.

Our Approach

Excitation Dropout

Dropping neurons having stronger evidence,



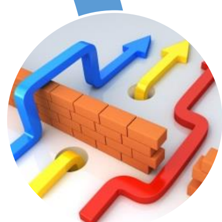
forces a network to learn alternative paths in order to maintain loss minimization,

leading to plasticity-like behavior, a characteristic of human brains too.

Our Approach

Excitation Dropout

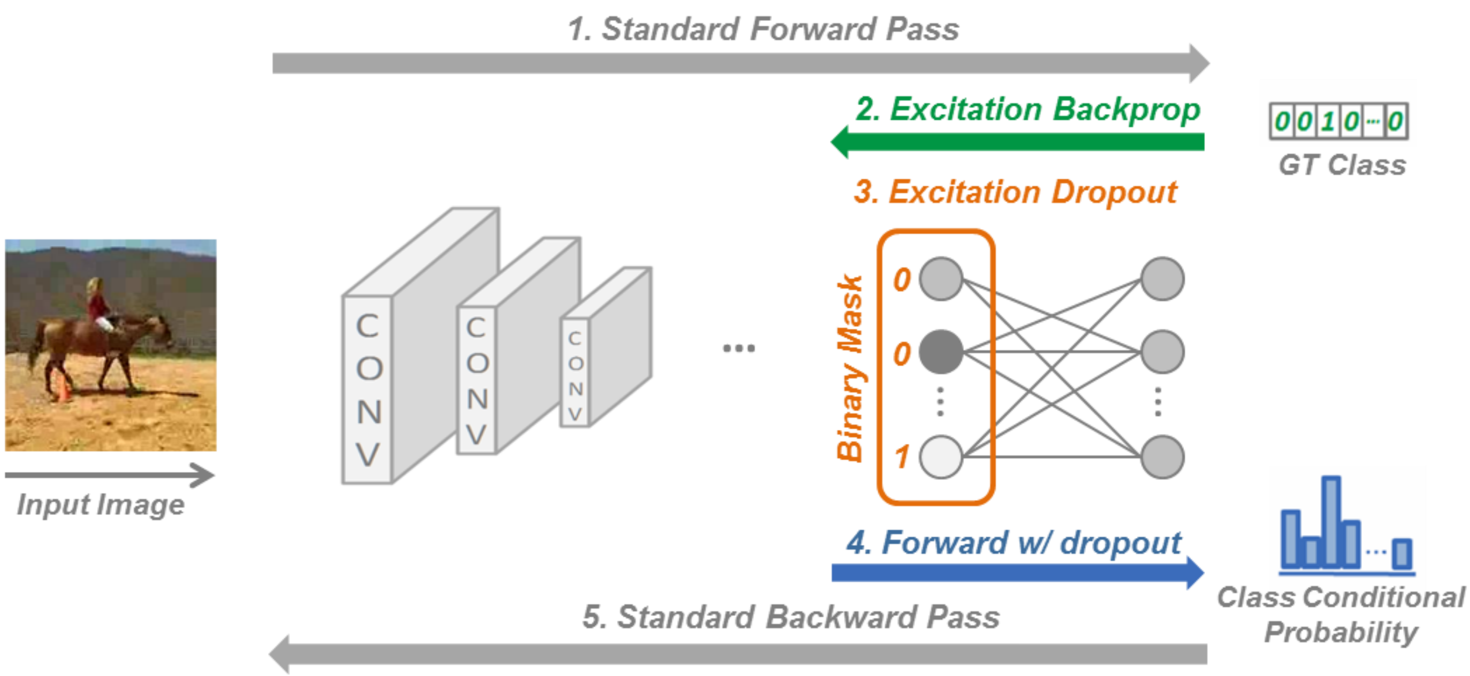
Dropping neurons having stronger evidence,



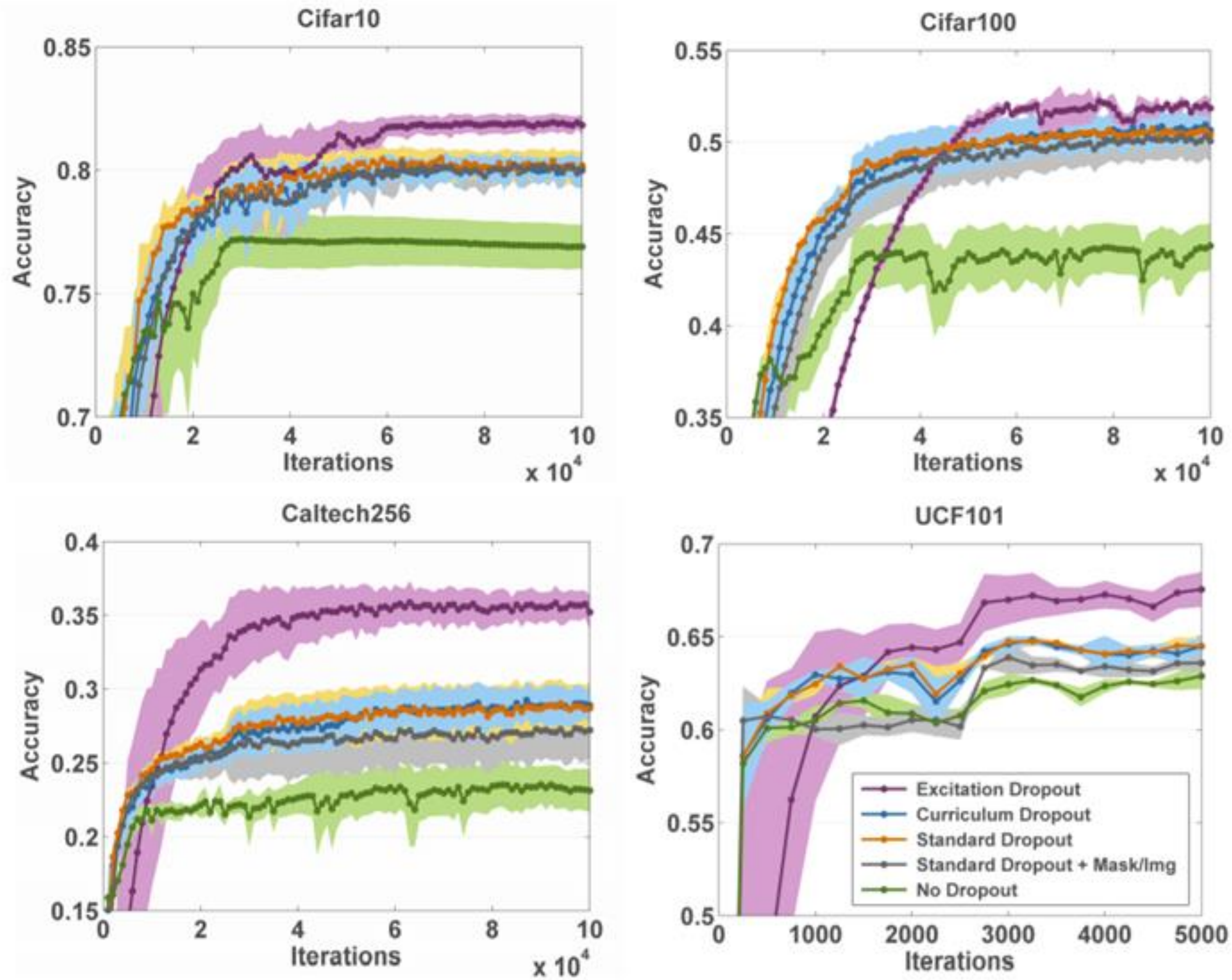
forces a network to learn alternative paths in order to maintain loss minimization,

leading to plasticity-like behavior, a characteristic of human brains too.

Excitation Dropout Pipeline



Improved Generalization



Neural Networks VI

Domain Adaptation

Has deep learning solved vision?

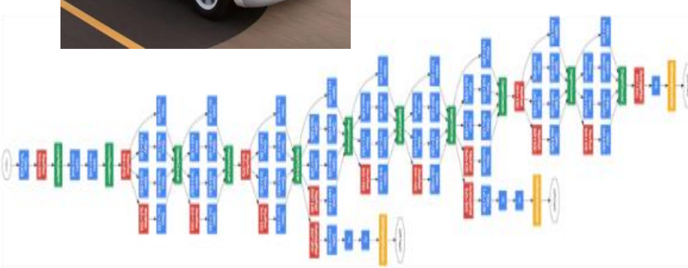
pedestrian detection FAIL

https://www.youtube.com/watch?v=w2pwxv8rFkU

“What you saw is not what you get”

What your net is trained on

What it’s asked to label



“Dataset Bias”
“Domain Shift”

Problem: Domain Shift

Input Image

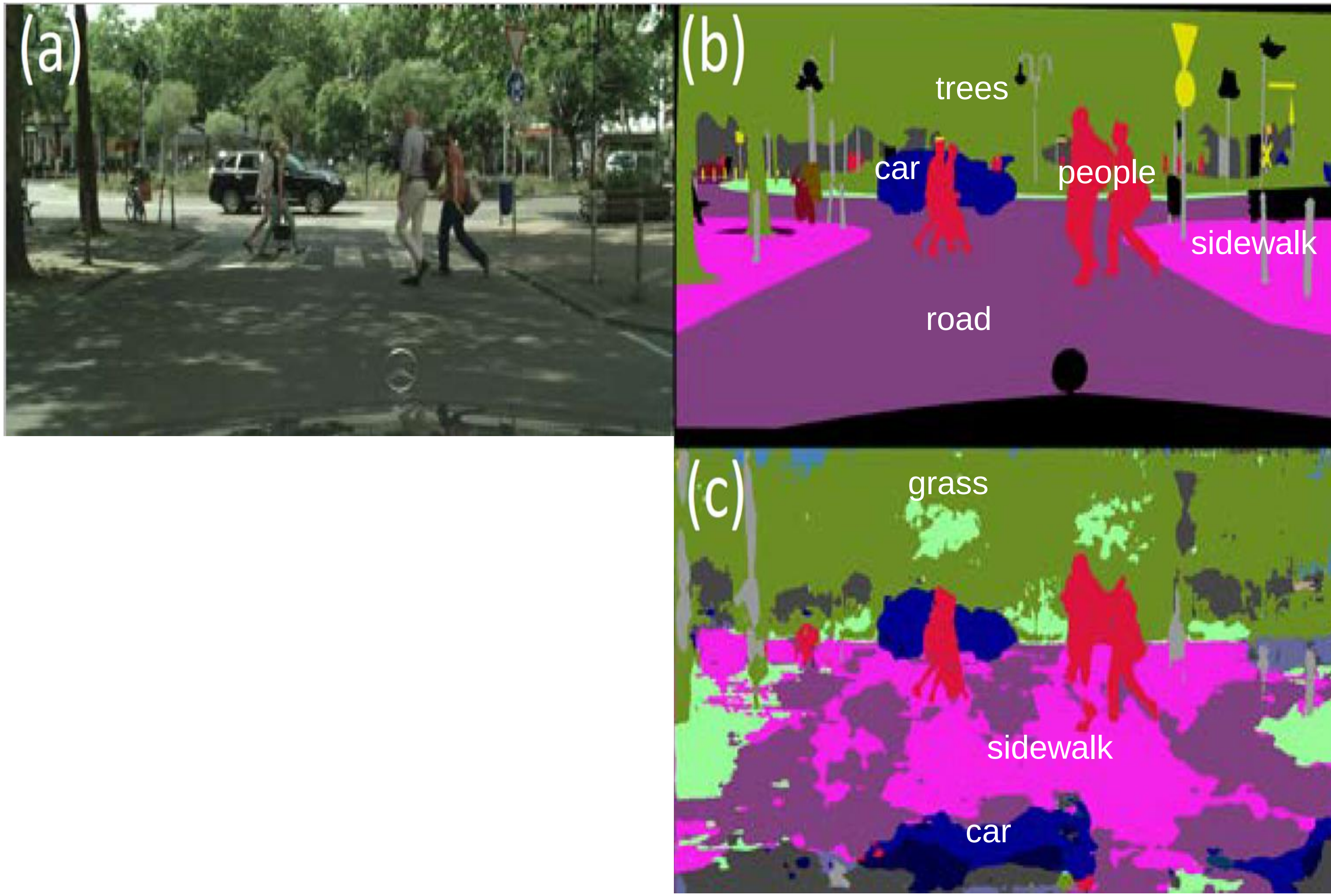
True Segmentation

Model Output

Solution: Domain Adaptation

Input Image

True Segmentation



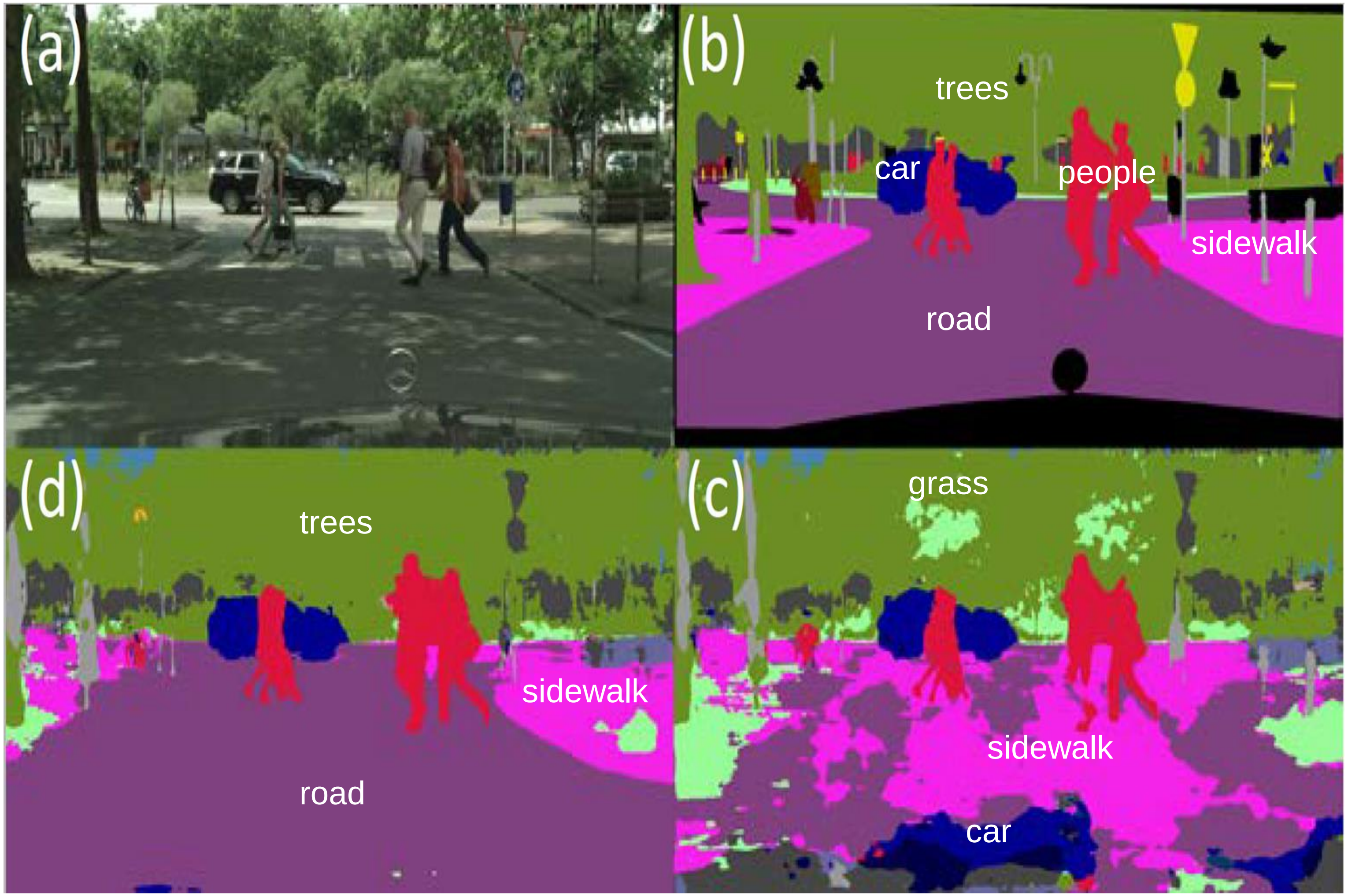
Adapted Model Output

Model Output

Solution: Domain Adaptation

Input Image

True Segmentation

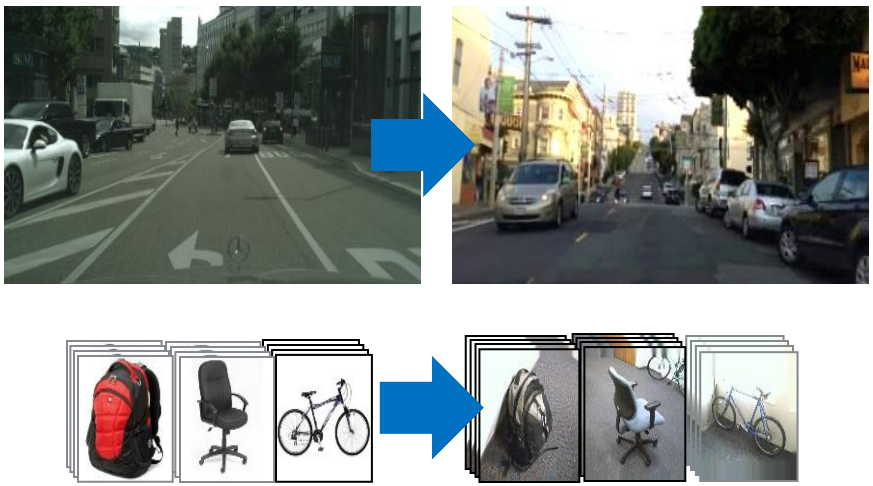


Adapted Model Output

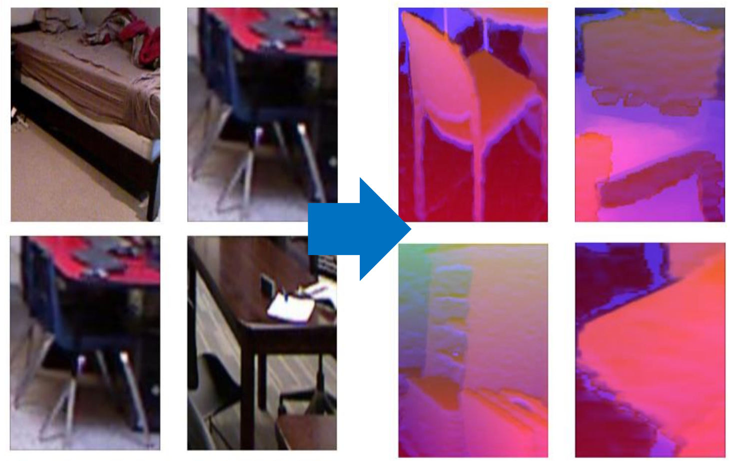
Model Output

Applications of Domain Adaptation

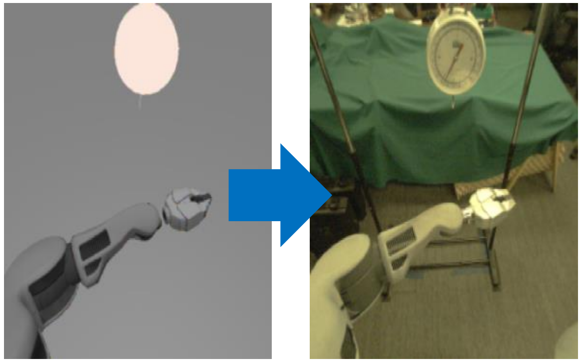
From dataset to dataset



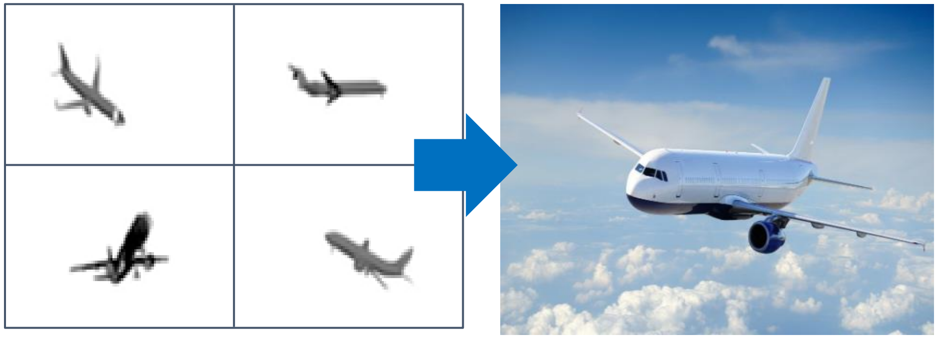
From RGB to depth



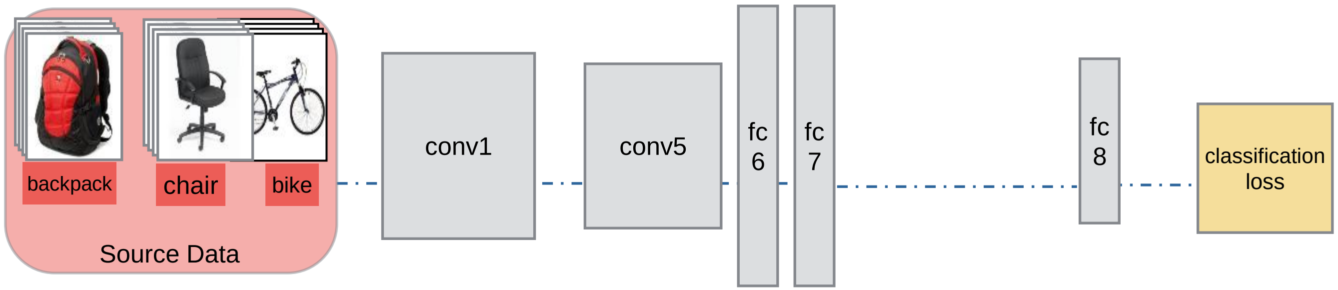
From simulated to real control



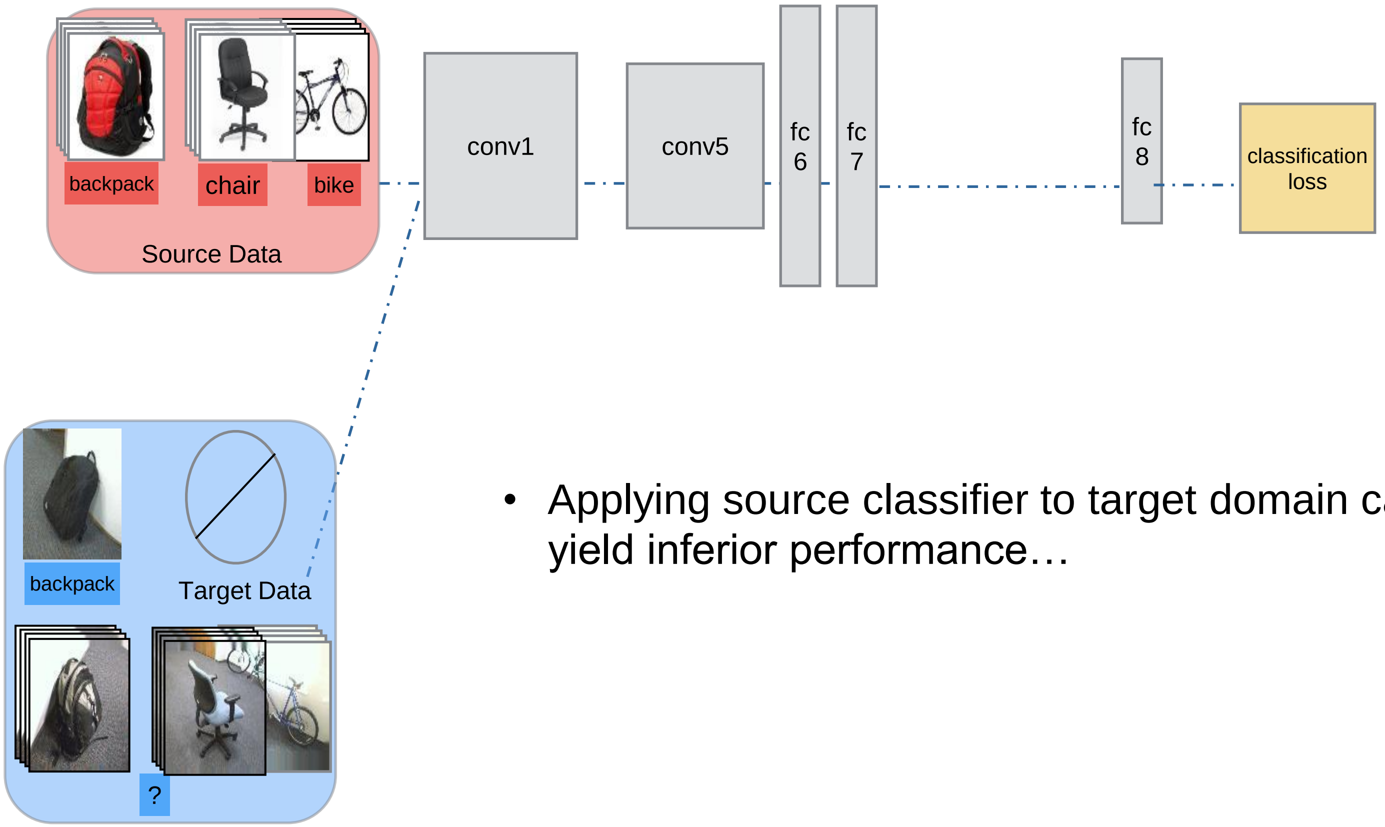
From CAD models to real images



How to adapt a deep network?

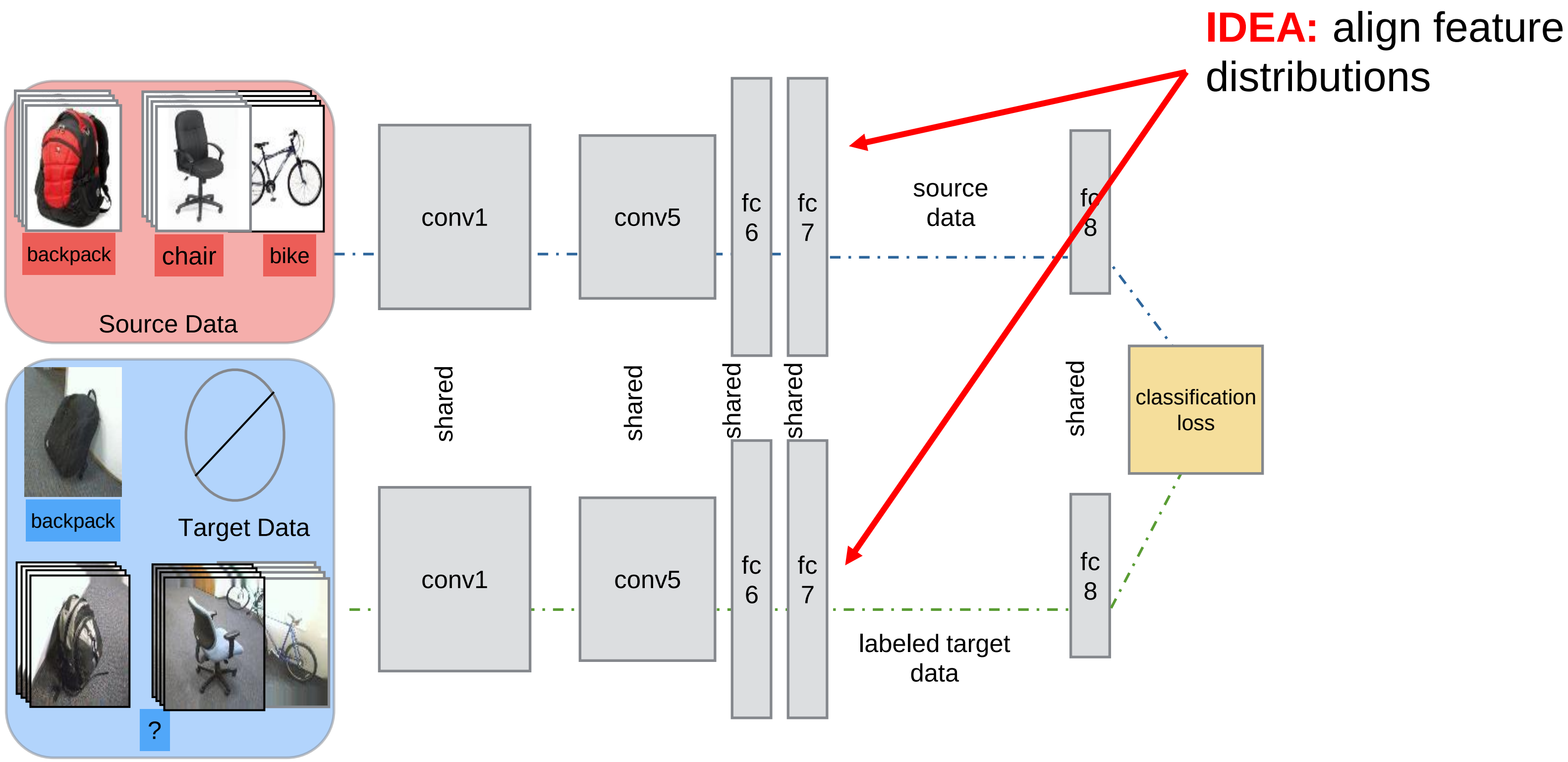


How to adapt a deep network?



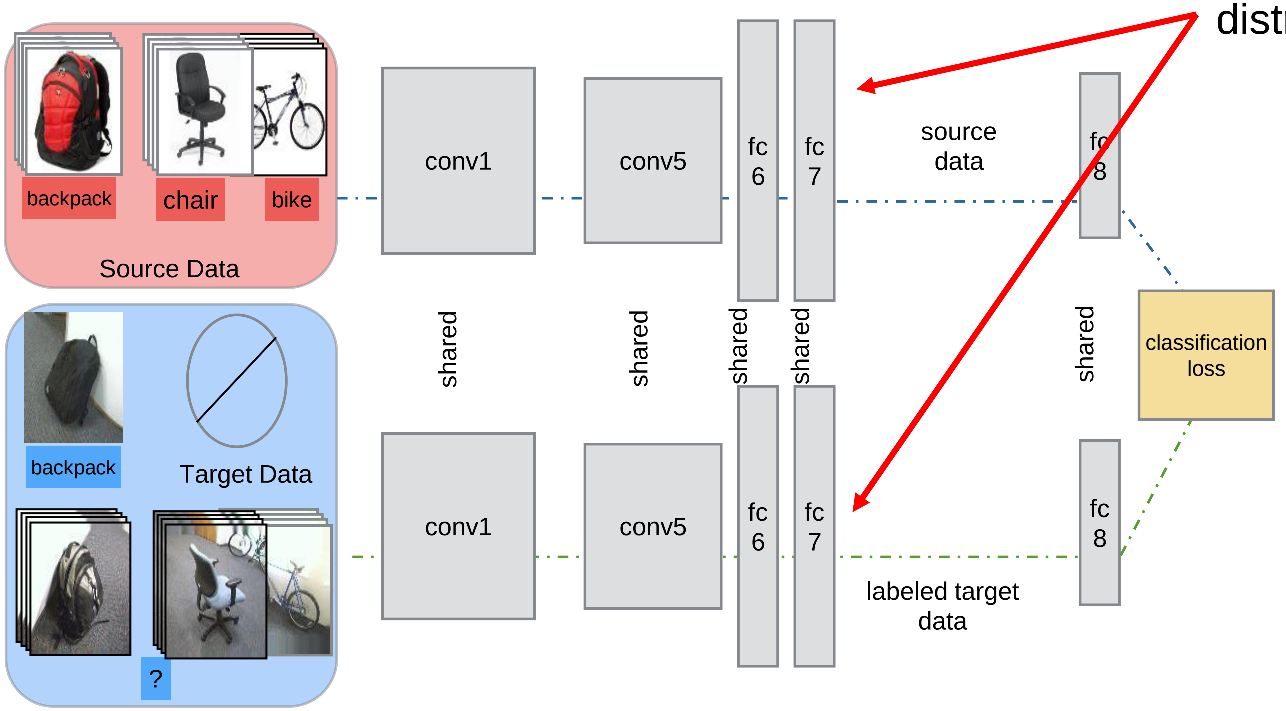
- Applying source classifier to target domain can yield inferior performance…

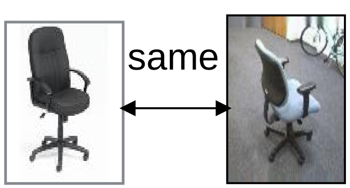
How to adapt a deep network?



How to adapt a deep network?

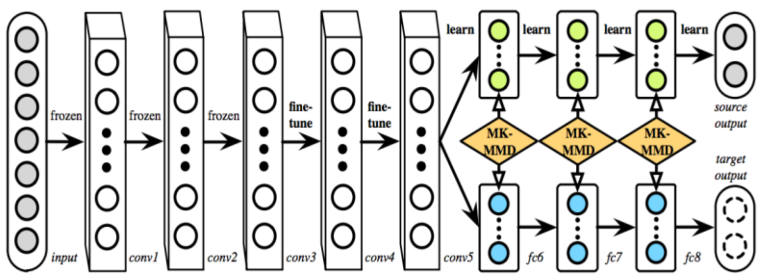
IDEA: align feature distributions



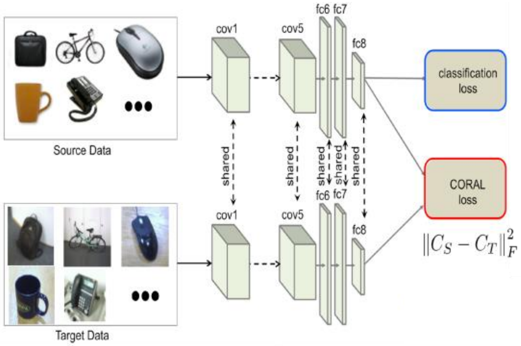


Solution: align deep feature distributions

- by minimizing **distance** between distributions, e.g.

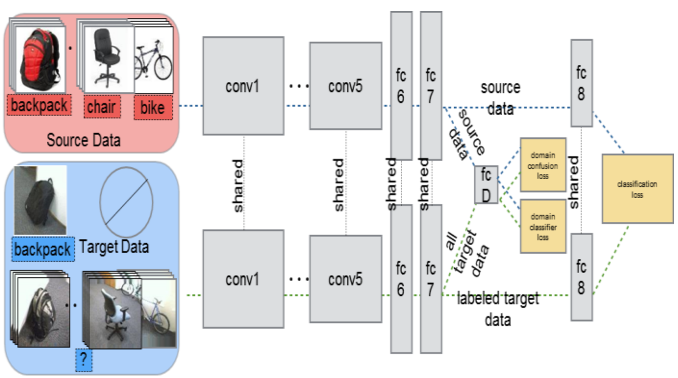


Maximum Mean Discrepancy M. Long, et al. ICML 2015

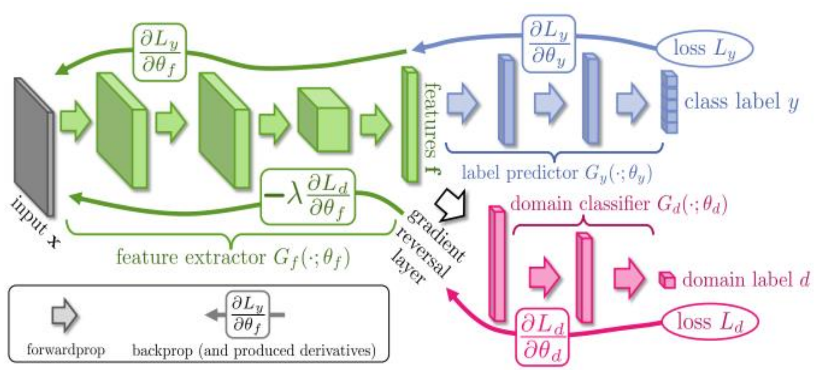


CORrelation ALignment Sun and Saenko, AAAI 2016

- …or by **adversarial** domain alignment, e.g.



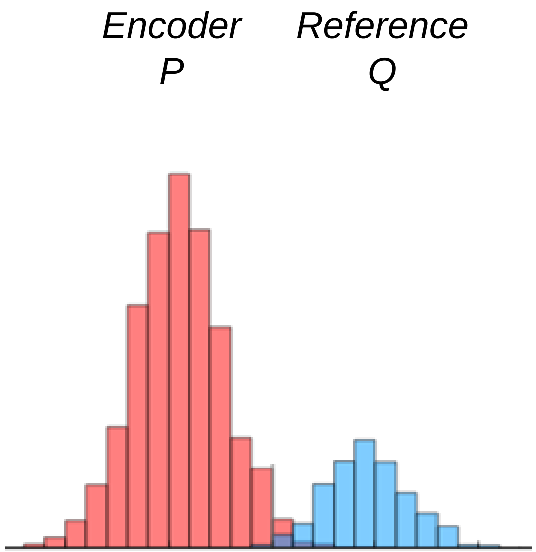
Domain Confusion E. Tzeng et al. ICCV 2015

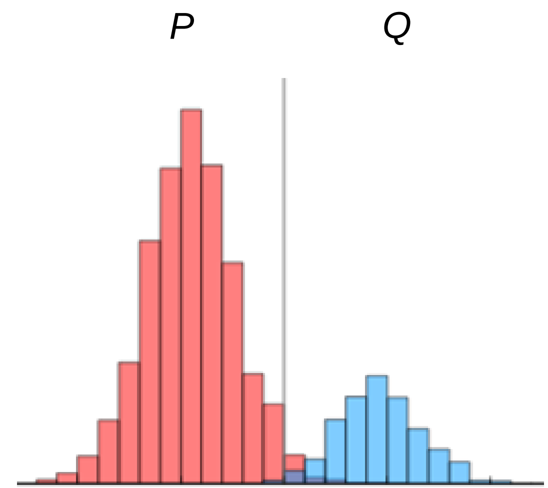


Reverse Gradient Y. Ganin and V. Lempitsky ICML 2015

Adversarial Feature Alignment

Adversarial networks

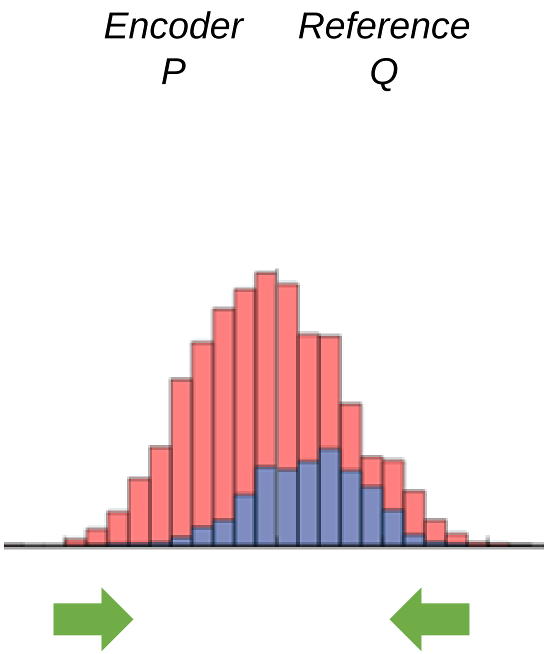


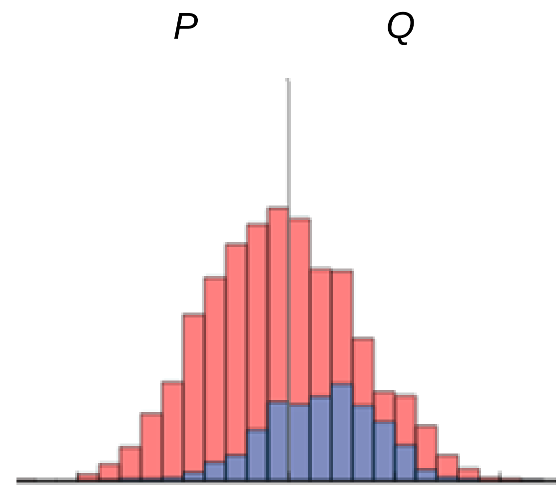


Encoder
Generates features such that their distribution P matches reference distribution Q

Adversary
Tries to discriminate between samples from P and samples from Q

Adversarial networks





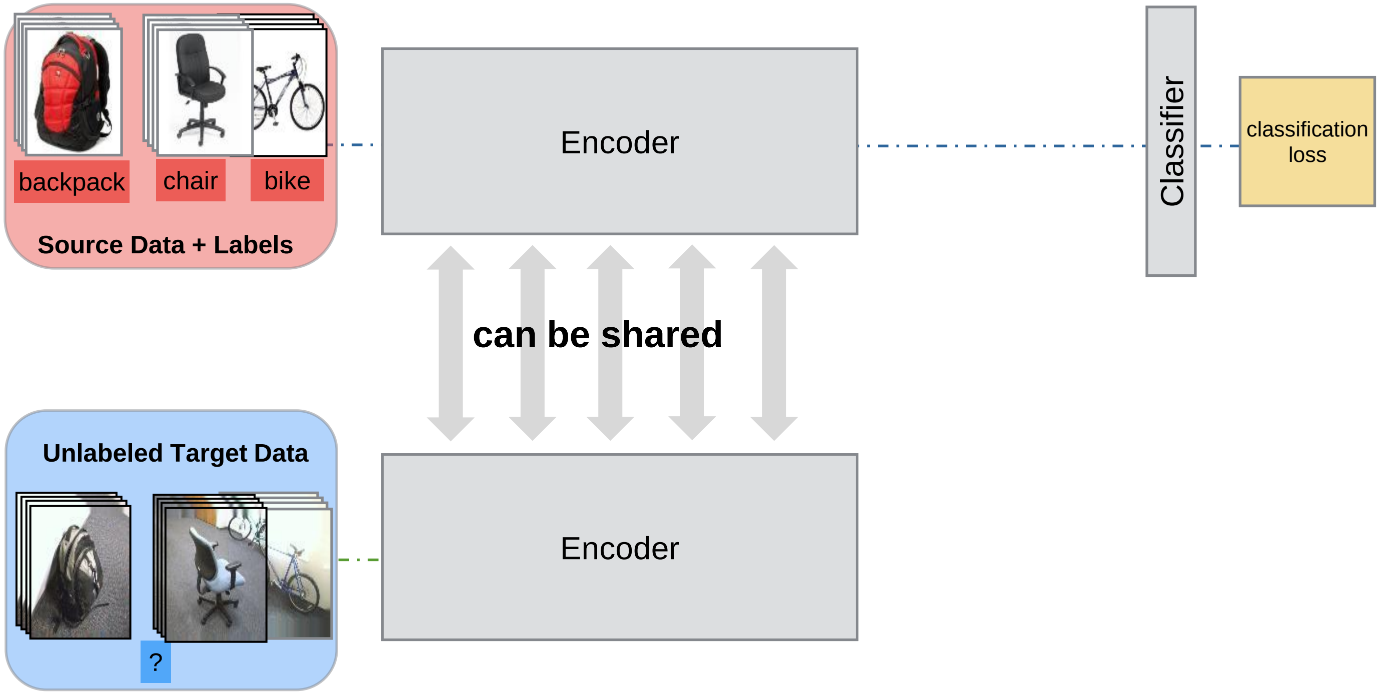
Encoder
Generates features such that their distribution P matches reference distribution Q

fools adversary

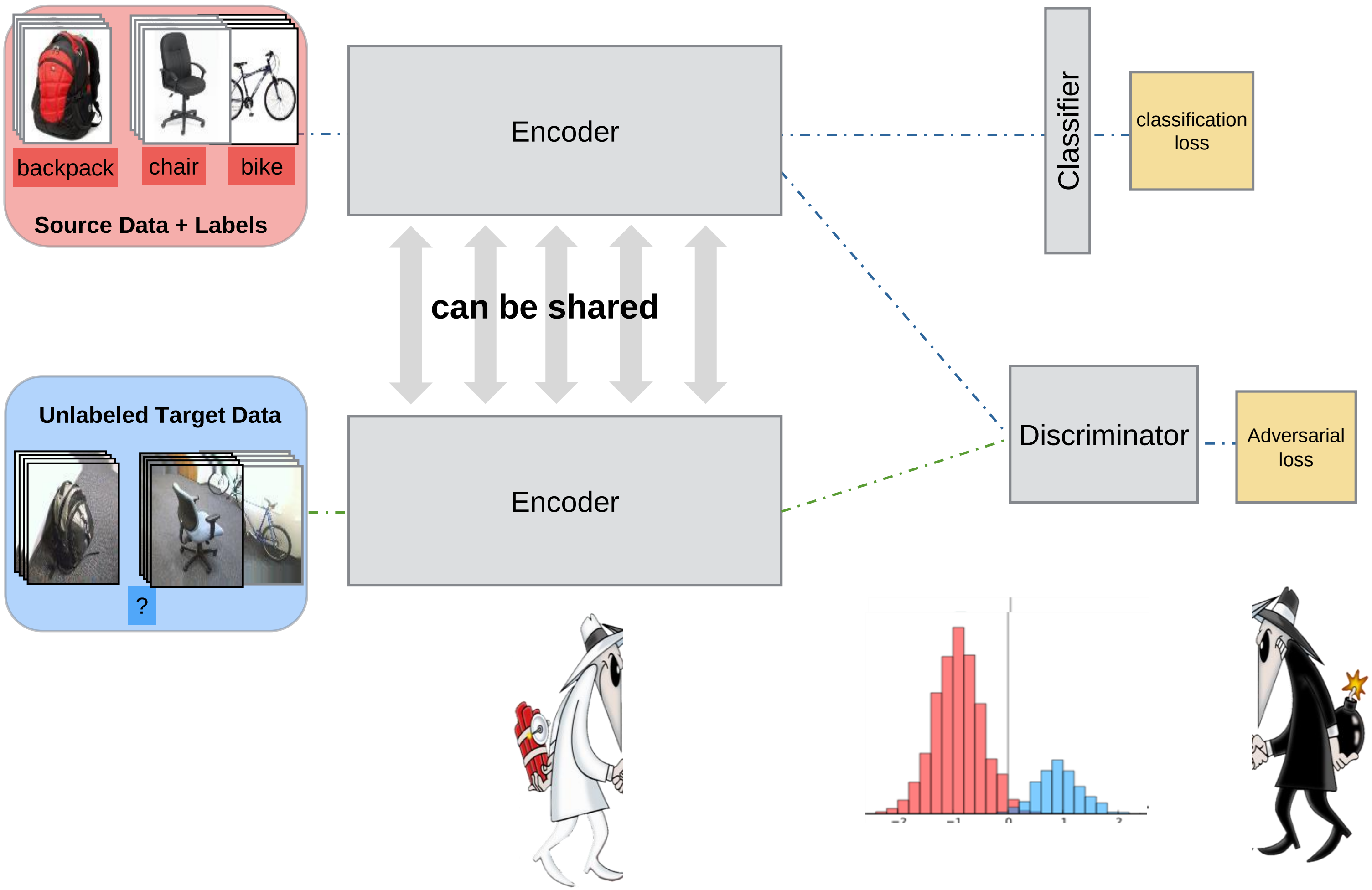
Adversary
Tries to discriminate between samples from P and samples from Q

tries harder

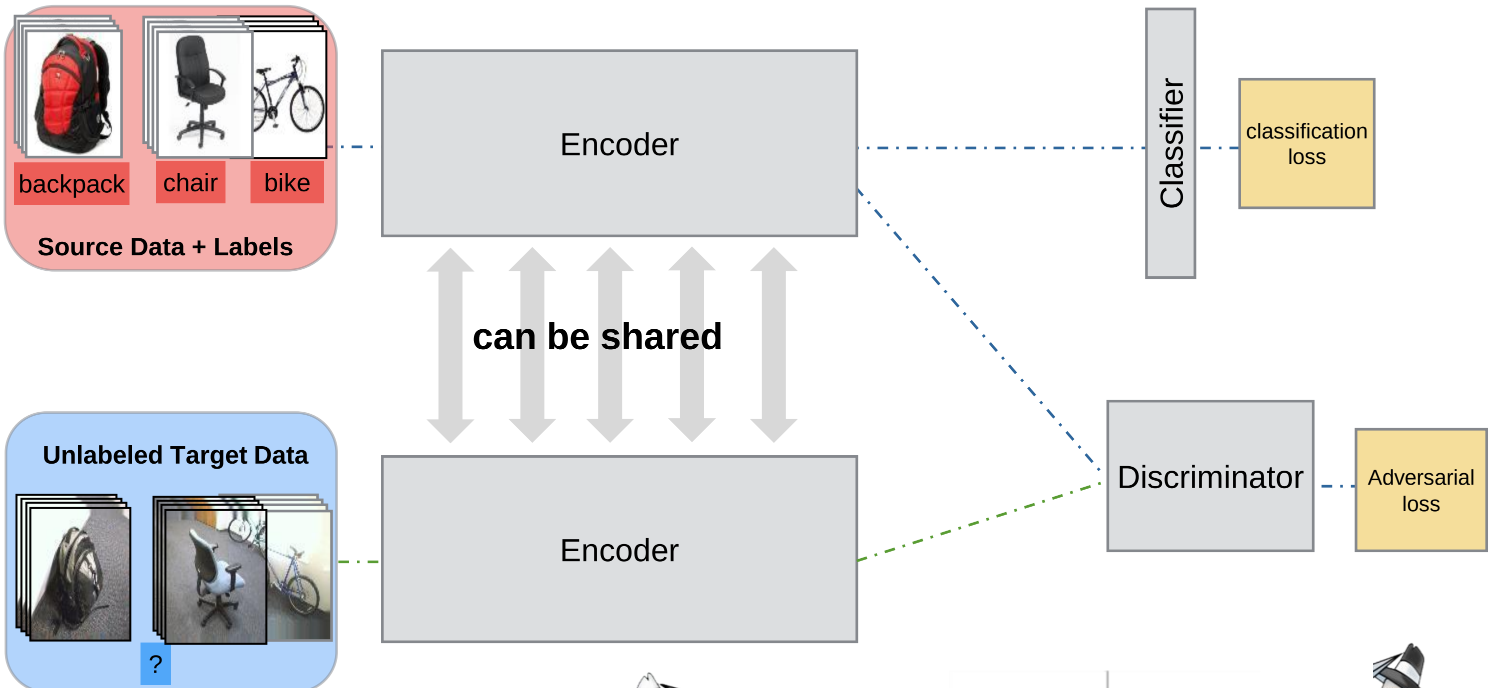
Adversarial domain adaptation

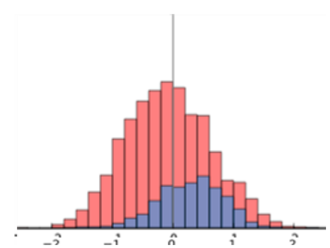


Adversarial domain adaptation

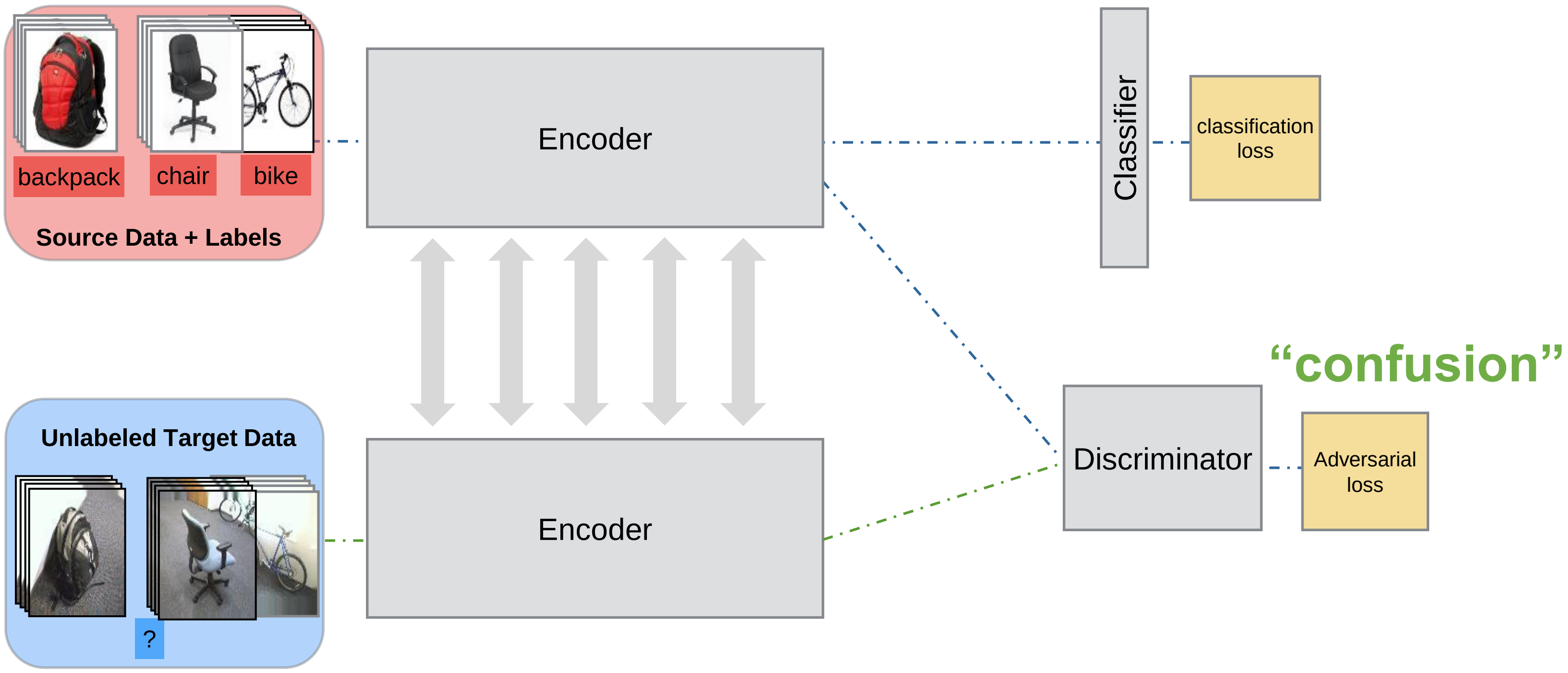


Adversarial domain adaptation





Design choices in adversarial adaptation



[13] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks, NIPS 2016

Domain Adaptation and Generalization

- In domain adaptation one needs to know a priori the target distribution, which may not be available in practice.

- In standard domain generalization techniques, one needs several source domains for training, both of which may not be available in practice.

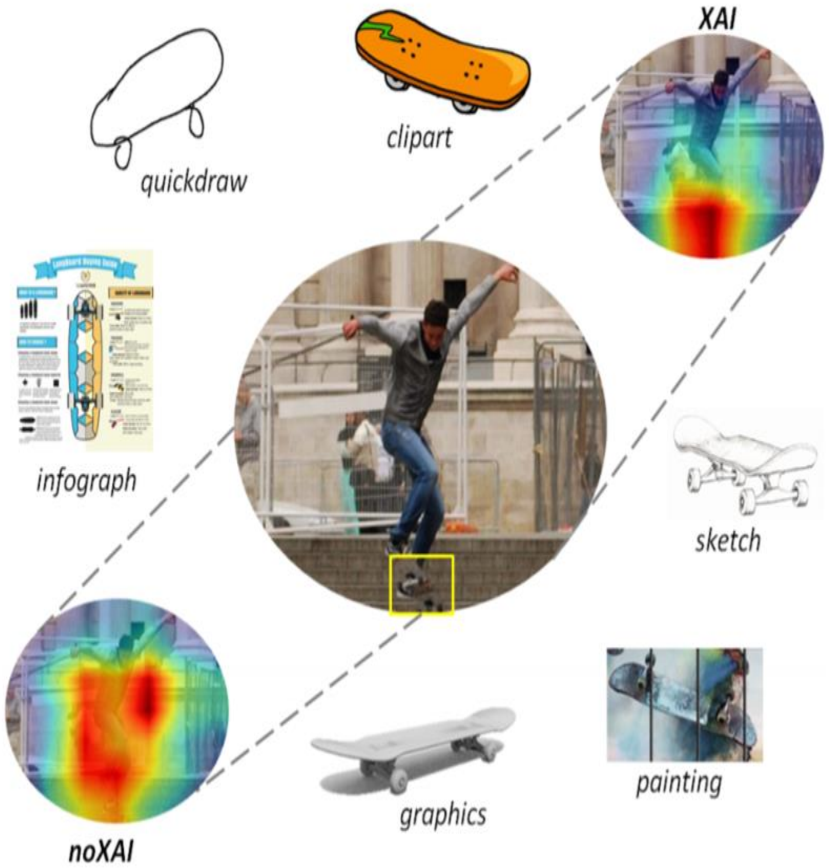
- A more generic formulation is single-source domain generalization, where one would like to avoid learning dataset bias for better generalization, but only has access to a single source distribution.

Neural Networks VI

Explainability and Domain Generalization

Explainable AI (XAI) for Domain Generalization

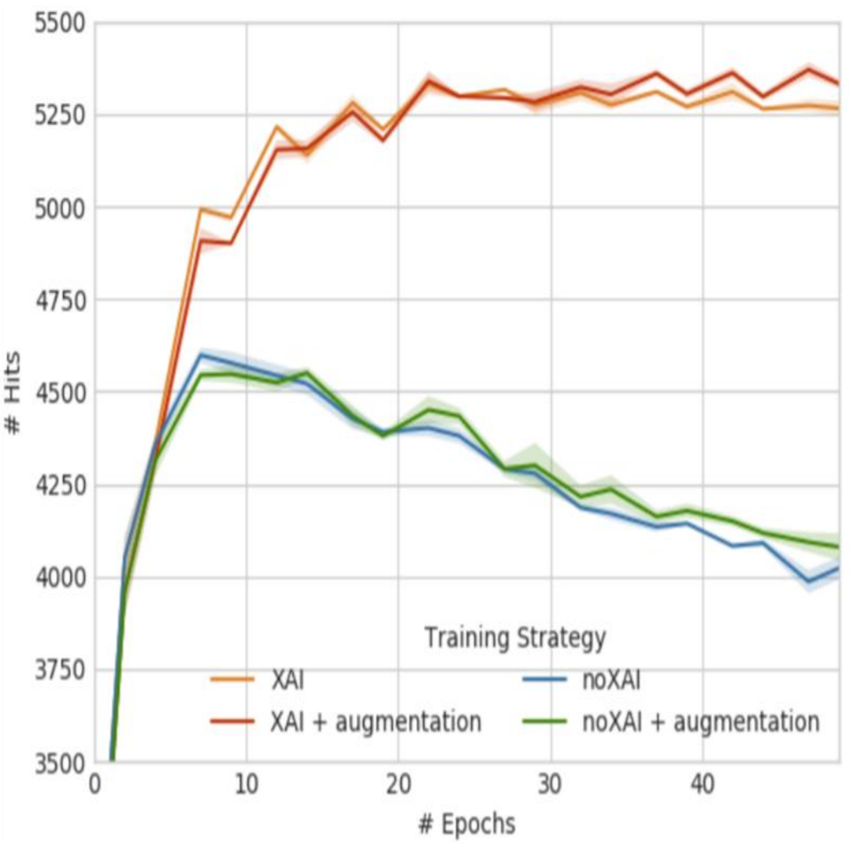
Training a deep neural network model to enforce explainability, *e.g.* focusing on the skateboard region (red is most salient, and blue is least salient) for the ground-truth class skateboard in the central training image, enables improved generalization to other domains where the background is not necessarily class-informative.



Explainability Results: Quantitative *[Automated]*

- The number of unseen MSCOCO images, among the 16K validation set, where the model is able to provide an accurate explanation for, among the correctly classified ones during training.

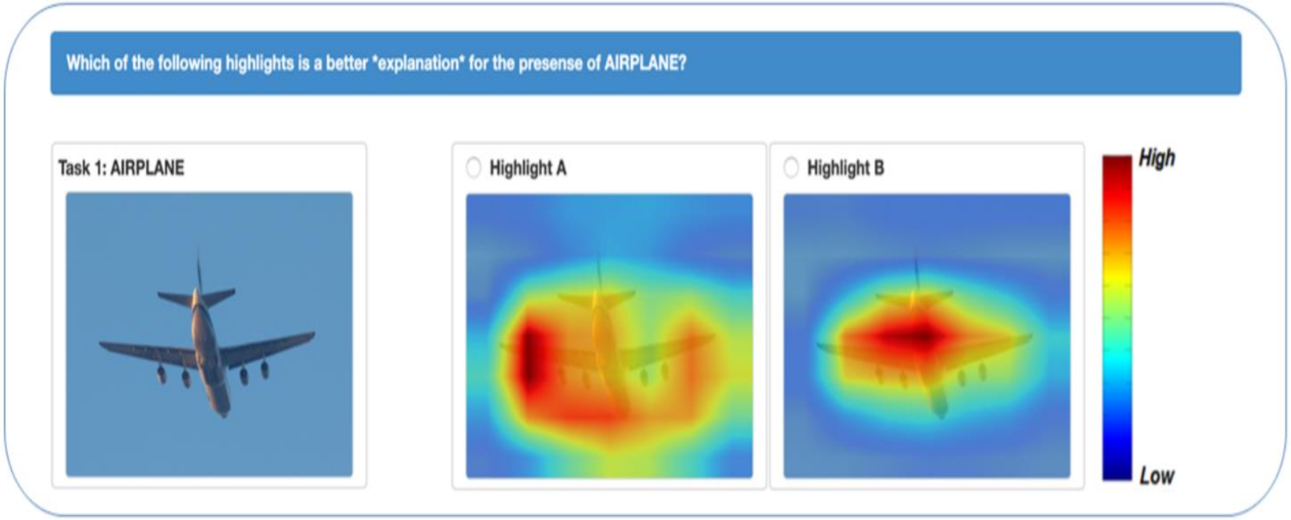
- We can see that the noXAI model fits the dataset bias at training time, while the XAI model improves its explainability over time for validation data.



Explainability Results: Quantitative *[Human Judgment]*

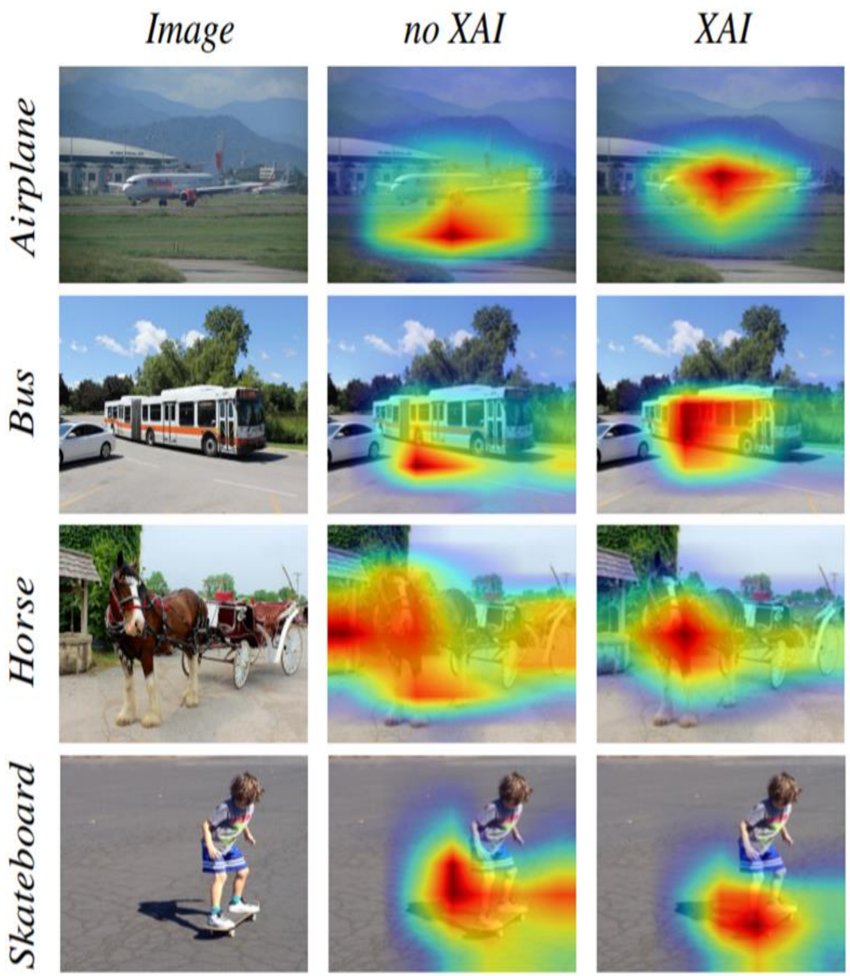
- The interface asks the users to select the evidence (“highlight”) they think is a better explanation for the presence of an object.

- 80% of the images with a winner choice favored the XAI explanation over the noXAI explanation.



Explainability Results: Qualitative

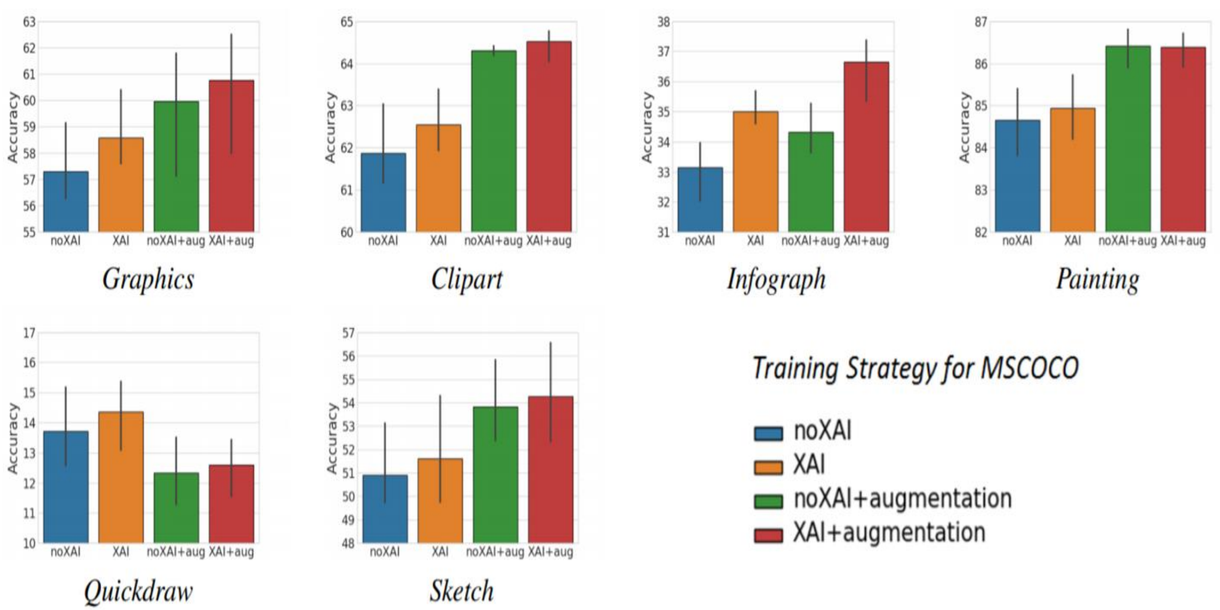
- The XAI model, based on human spatial annotations, provides feedback that enables saliency to be better localized over the objects corresponding to the ground-truth class compared to the noXAI vanilla training of a deep model, for unseen validation data.



Single-Source Domain Generalization Results

- Domain generalization on six *unseen* target domains from the Syn2Real and DomainNet datasets.

- Training has been conducted on a single source: the MSCOCO dataset, and no data from any of the target domains is used for training.



Style Transfer

Figure 3. Images that combine the content of a photograph with the style of several well-known artworks. The images were created by finding an image that simultaneously matches the content representation of the photograph and the style representation of the artwork. The original photograph depicting the Neckarfront in Tübingen, Germany, is shown in **A** (Photo: Andreas Praefcke). The painting that provided the style for the respective generated image is shown in the bottom left corner of each panel. **B** *The Shipwreck of the Minotaur* by J.M.W. Turner, 1805. **C** *The Starry Night* by Vincent van Gogh, 1889. **D** *Der Schrei* by Edvard Munch, 1893. **E** *Femme nue assise* by Pablo Picasso, 1910. **F** *Composition VII* by Wassily Kandinsky, 1913.

https://openaccess.thecvf.com/content_cvpr_2016/papers/Gatys_Image_Style_Transfer_CVPR_2016_paper.pdf

Exam 1

- Oct 20 ***in class***
- Covering everything up to and including today's lecture
- Practice problems will be posted tomorrow
- Thu Oct 15 will be a revision session + study group
- Will be completely remote for everyone
  (please do not appear in person)

Exam Details

- **Administering the exam:**
  During lecture time
  Open: Video camera + Microphone
    Your hands must be in the camera's field of view
  Open exam pdf
  Take photos of your solutions on paper
  Submit a pdf of the photos on gradescope, just like submitting PS1.
  Confirm we received your submission before you leave (through private chat)

- **What do you need?**
  Internet + Pen/pencil + Empty sheets of paper (~10)
  New question, new page
  Computer to see exam questions
  Cell phone to take photos of your solutions at the end for submission