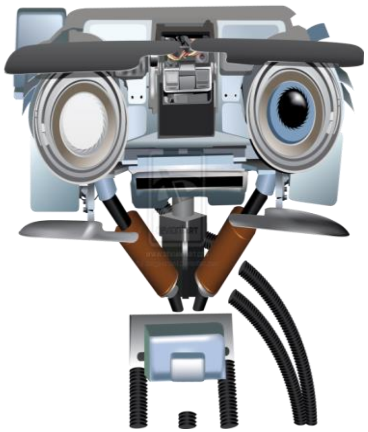
Using Zoom for Lectures

- **Sign in using:**
  - your name
- **Please mute both:**
  - your video cameras for the entire lecture
  - your audio/mics unless asking or answering a question
- **Asking/answering a question, option 1:**
  - click on Participants
  - use the hand icon to raise your hand
  - I will call on you and ask you to unmute yourself
- **Asking/answering a question, option 2:**
  - click on Chat
  - type your question, and I will answer it

Today: Outline

- **Explainability and Domain Adaptation/Generalization**

- **Reminders:** Class Challenge, due Apr 24
  One more pre-lecture material
  Midterm Scores Next week
  Wed Apr 22 is a Mon Schedule (No class)



Explainability

Sarah Adel Bargal

Importance of *Explainability*

- An important action to be detected in the vision systems of autonomous vehicles is: *Pedestrian Crossing*

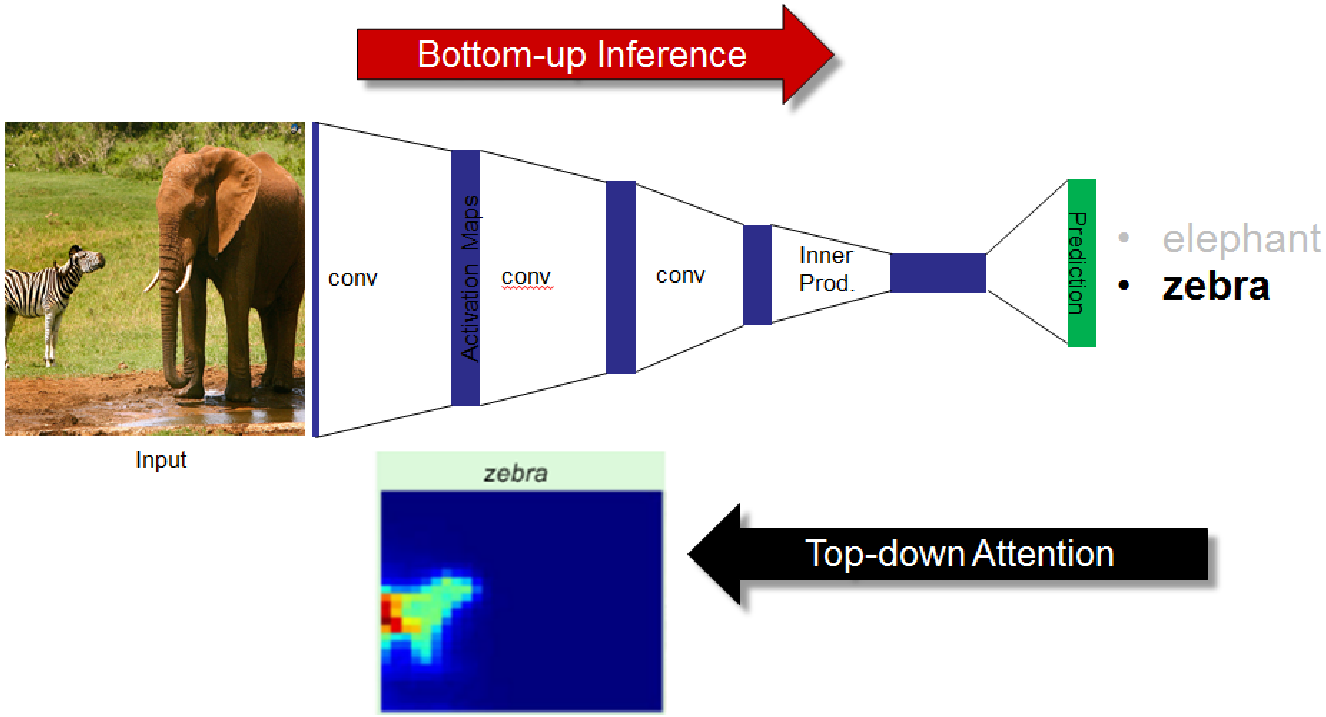
Importance of *Explainability*

Sample Misclassification

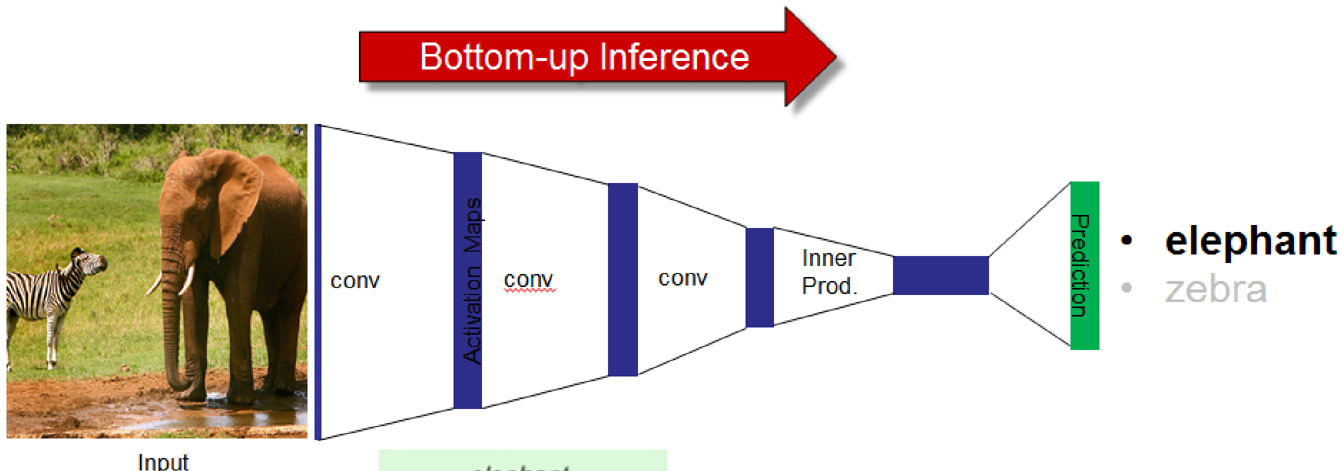
Ground Truth:
BabyCrawling

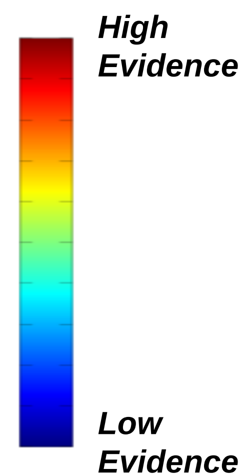
Classified as:
Pushups

Spatial Grounding

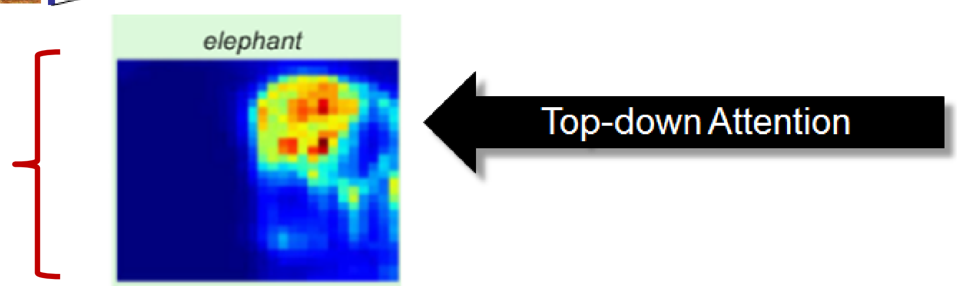


Spatial Grounding

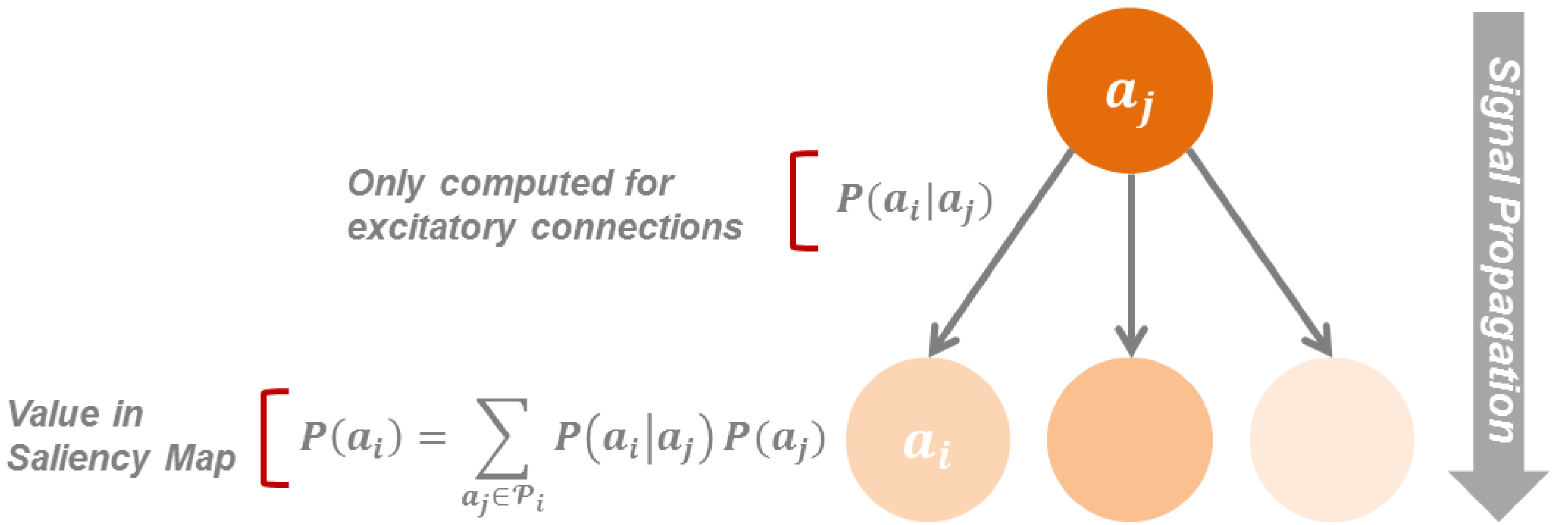




- *Evidence*
- *Grounding*
- *Saliency*

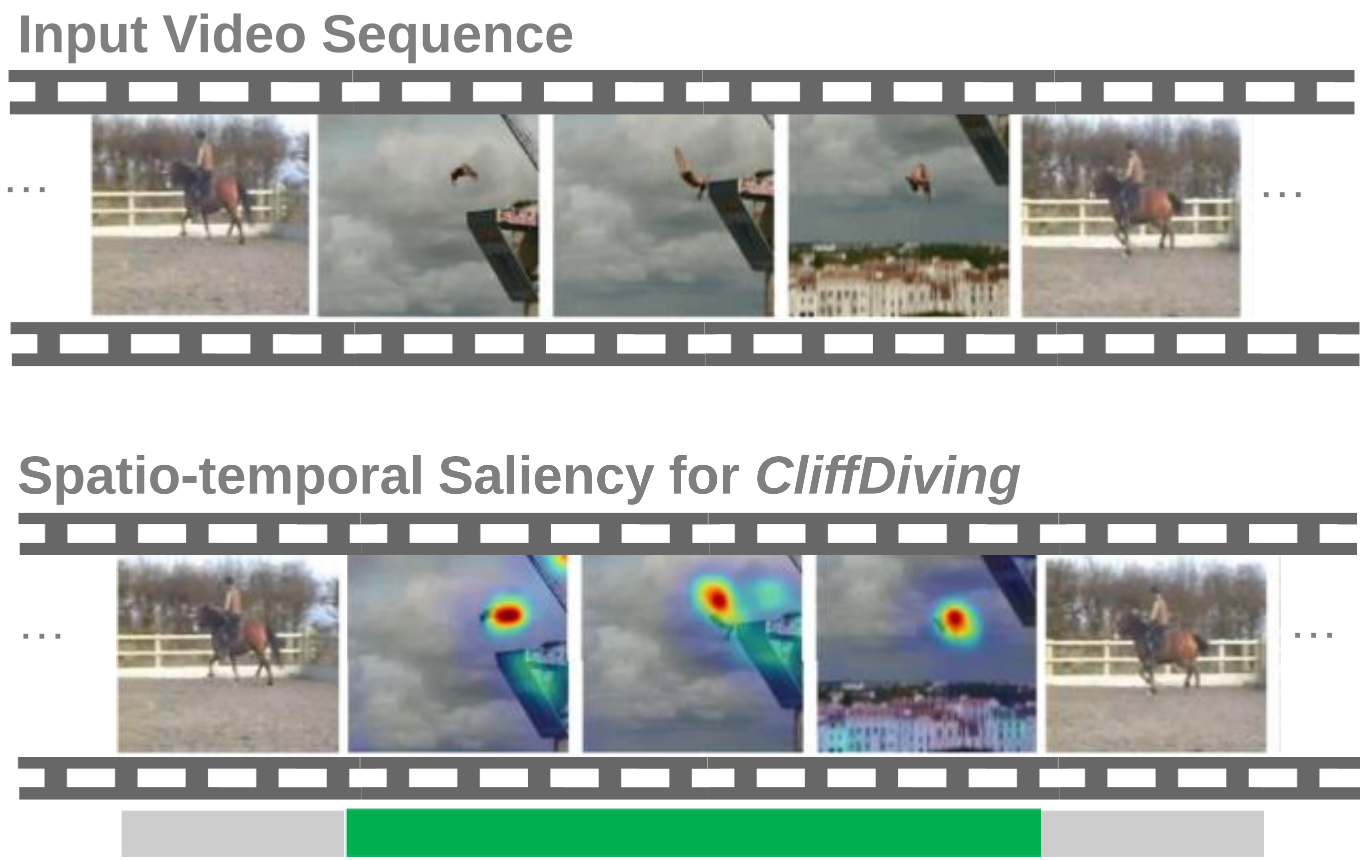


Excitation Backprop (EB)



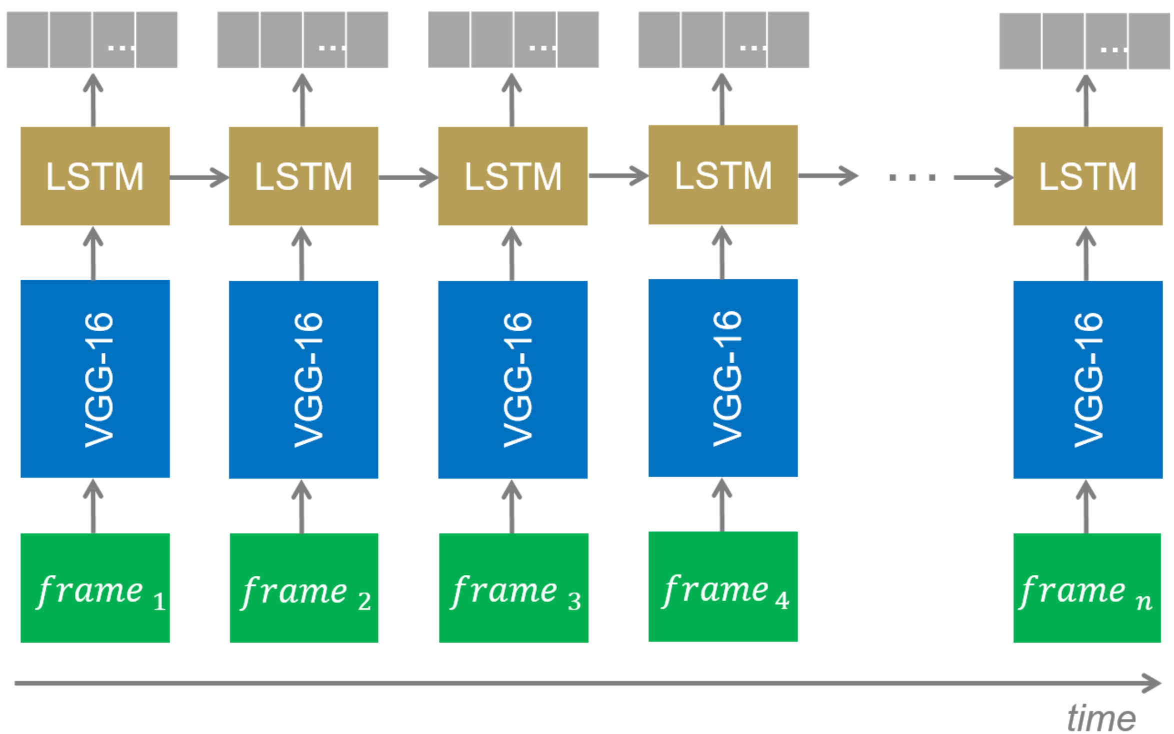
[Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, Stan Sclaroff. "*Top-down Neural Attention by Excitation Backprop., ECCV'16*]

Spatiotemporal Grounding

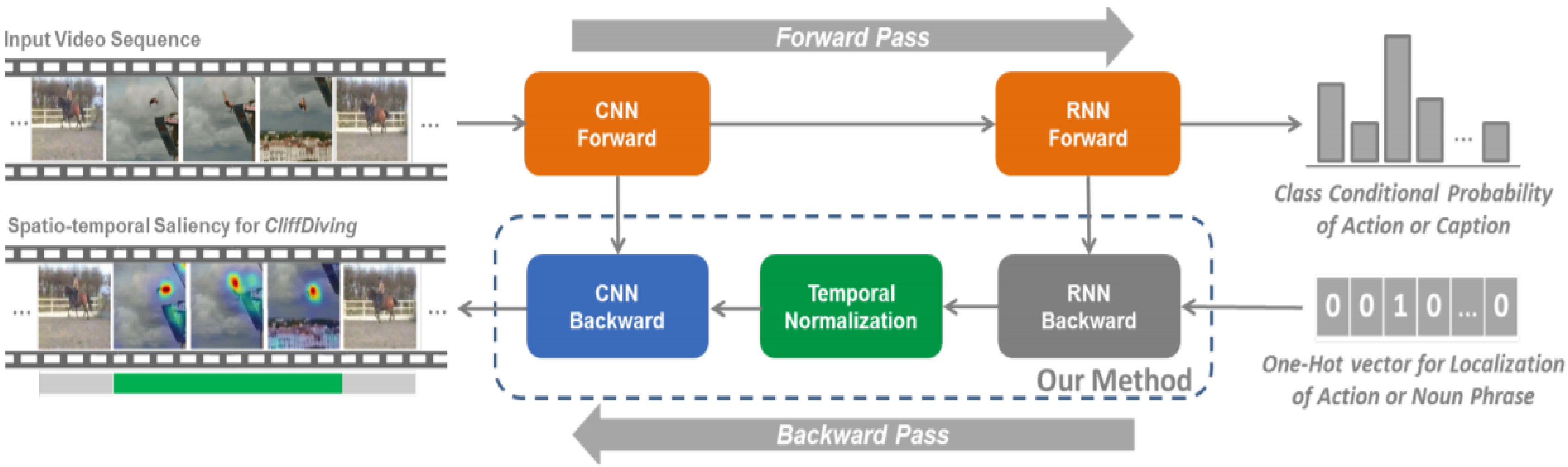


Architecture: Forward Pass

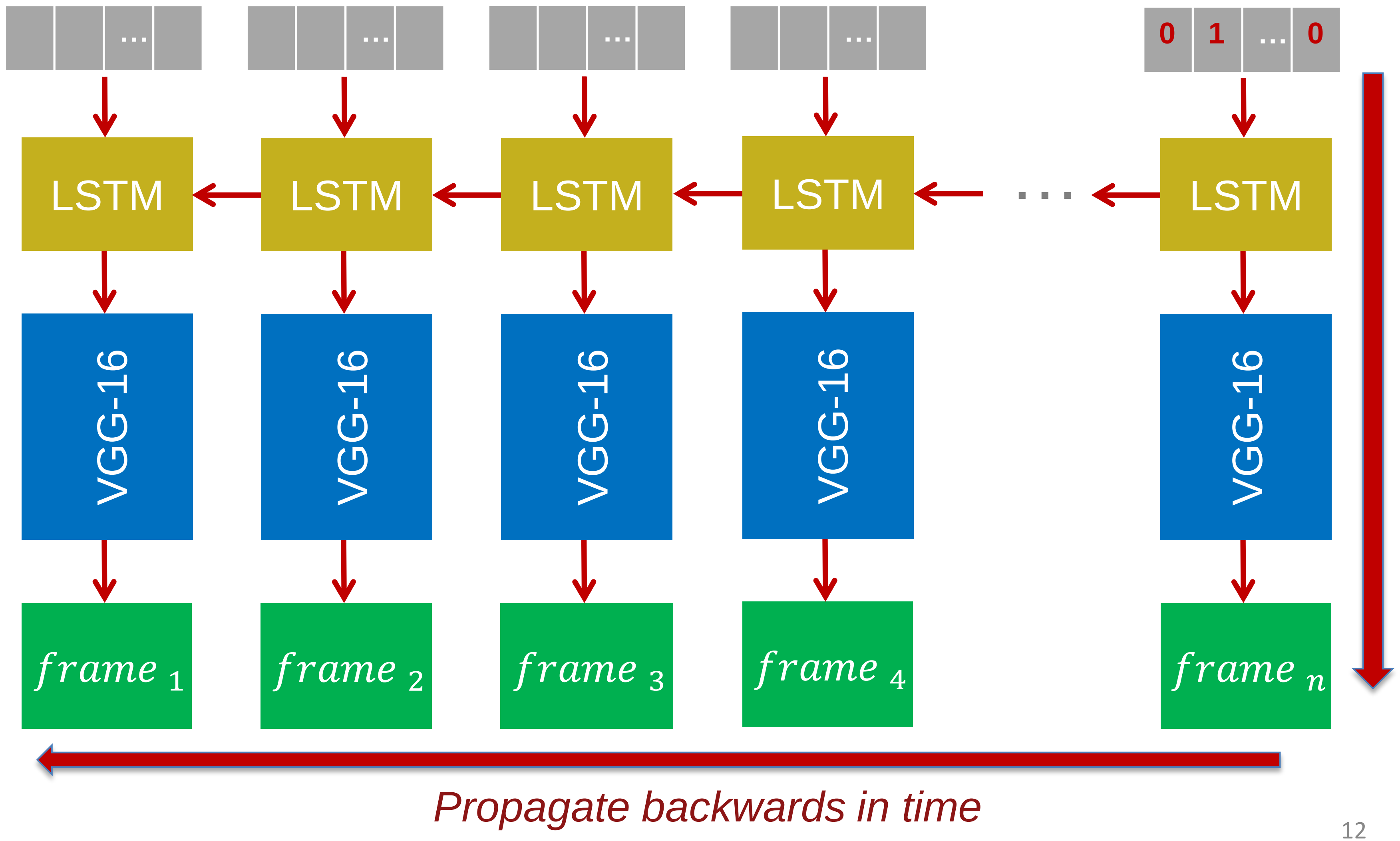
- CNN-LSTM is trained for the action recognition task.
- Resulting grounding is weakly-supervised.



Excitation Backprop in RNNs



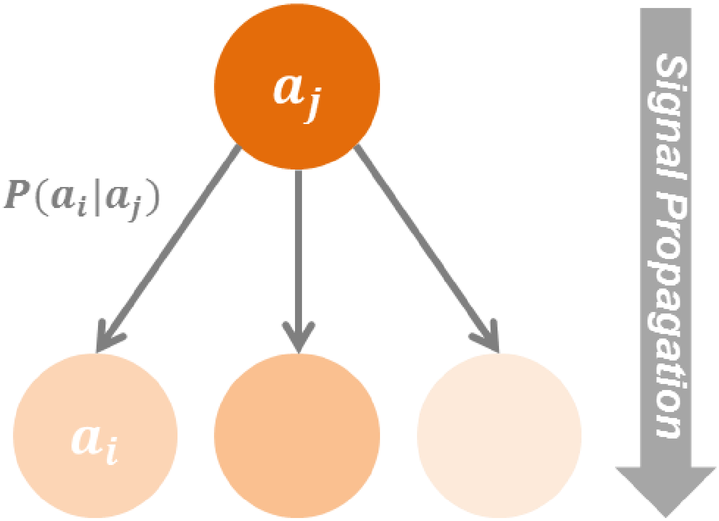
Architecture: Backward Grounding Pass



Excitation Backprop (EB)

Only computed for excitatory connections $\left[\, P(a_i|a_j) \right.$

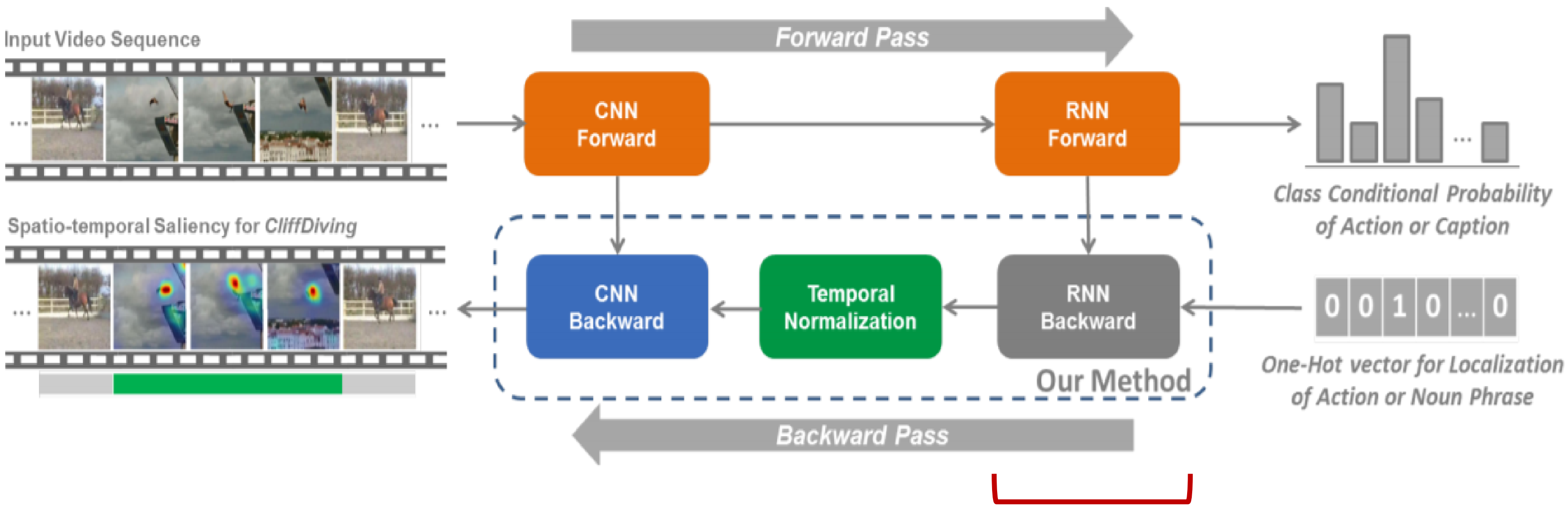
Value in Saliency Map $\left[\, P(a_i) = \sum_{a_j \in \mathcal{P}_i} P(a_i|a_j)\, P(a_j) \right.$



[Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, Stan Sclaroff. "*Top-down Neural Attention by Excitation Backprop., ECCV'16*]

RNN Backward

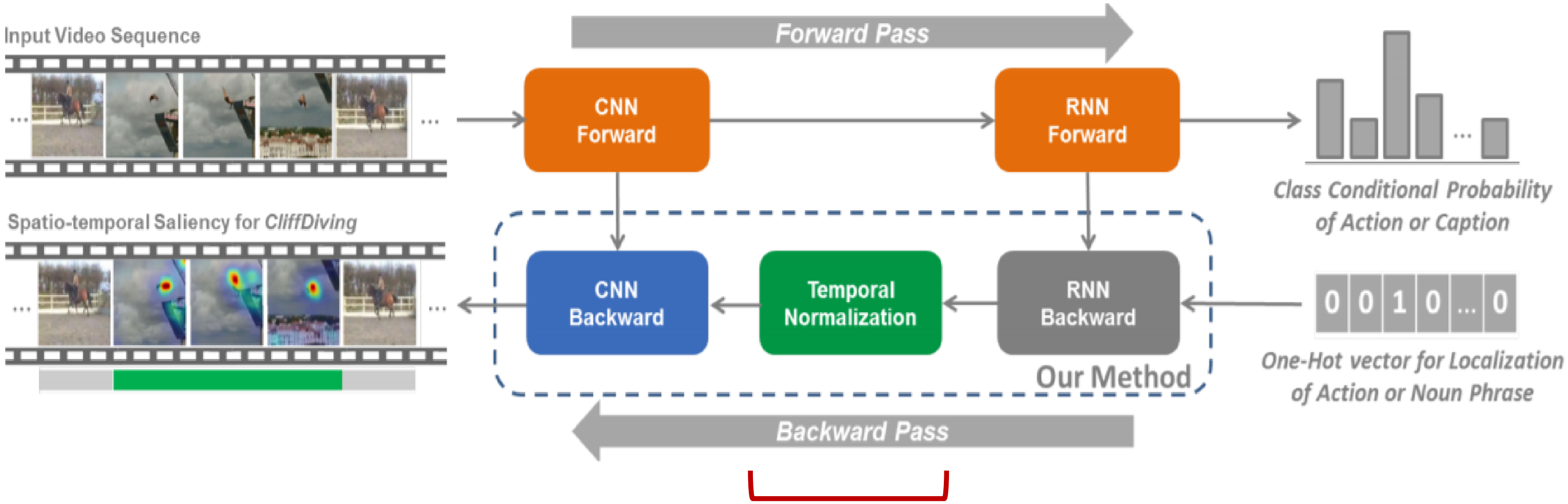
- For every time-step t:



$$P^t(a_i) = \sum_{a_j \in \mathcal{P}_i} P^t(a_i|a_j)P^t(a_j)$$

RNN Backward

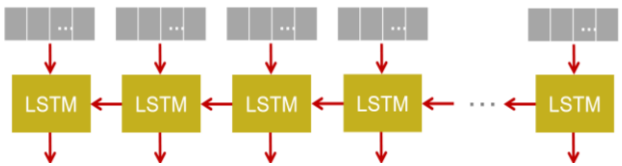
- For every time-step t:



$$P_N^t(a_i) = P^t(a_i) / \sum_{t=1}^{T} P^t(a_i)$$

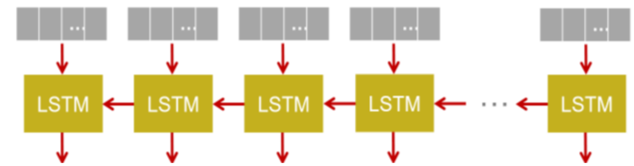
Contrastive Evidence

Original weights



−

Multiply top layer weights by -1



diving − non diving

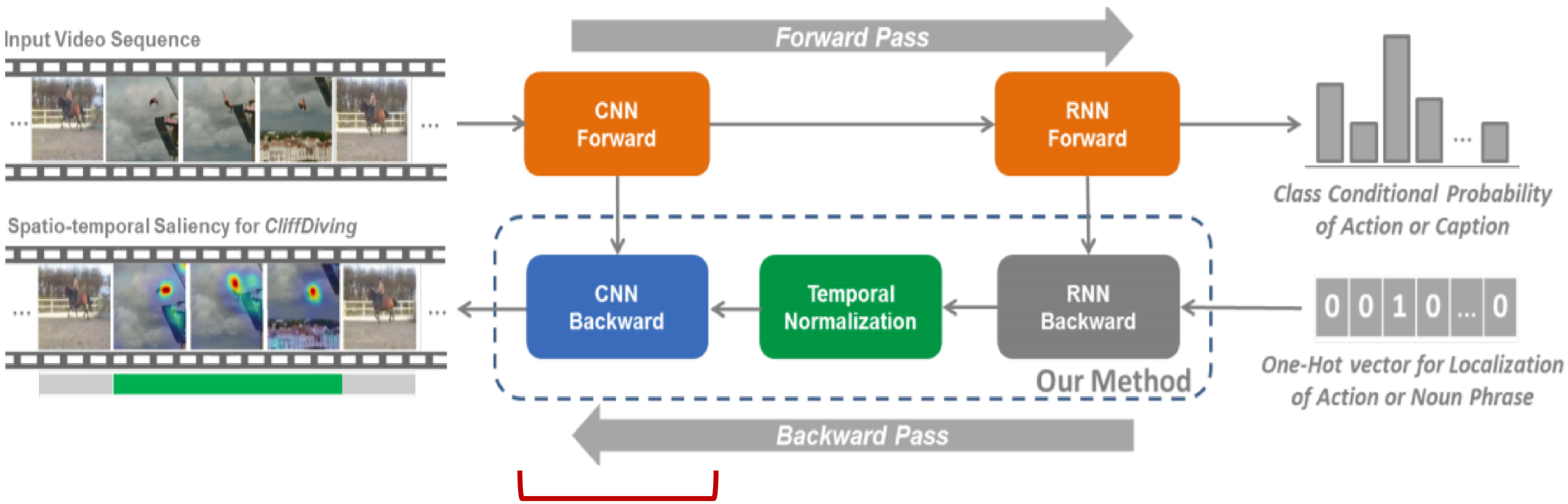
$P_N^t(a_i)$ $\overline{P}_N^t(a_i)$

Maps discriminative for diving

$$Map^t(a_i) = P_N^t(a_i) - \overline{P}_N^t(a_i)$$

RNN Backward

- For every time-step t:



$$Map^t(a_i) = \sum_{a_j \in P_i} P^t(a_i|a_j) Map^t(a_j)$$

Applications

- Action Detection (*videos*)
- Caption Grounding (*images, videos*)
- Reflecting the Abstraction Capability of Models

Applications

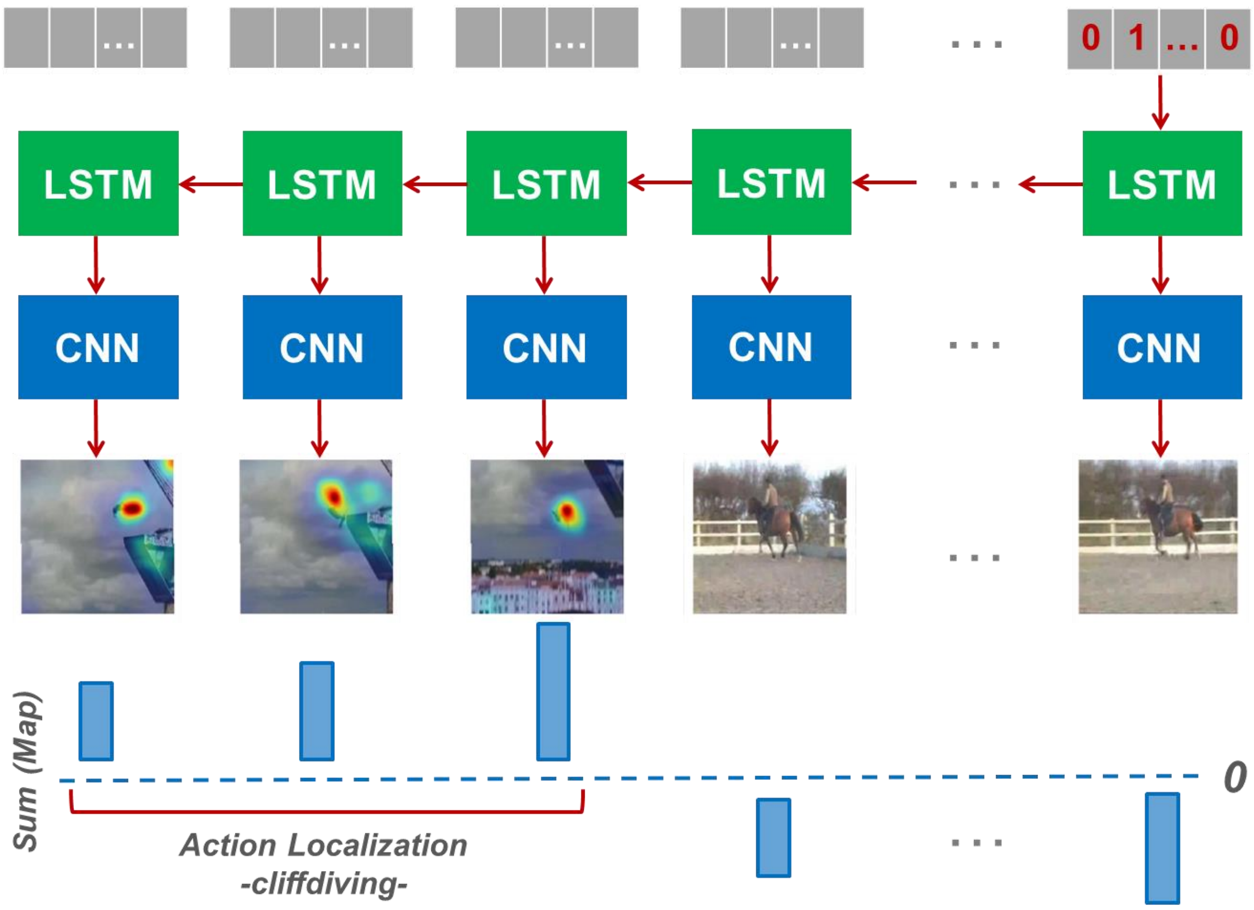
- Action Detection (*videos*)
- Caption Grounding (*images, videos*)
- Reflecting the Abstraction Capability of Models

UCF101 Dataset: Spatiotemporal Grounding

Handstand Walking

Ice Dancing

Spatiotemporal Action Detection



THUMOS'14 Dataset: Action Detection

Method	mAP ($\alpha = 0.1$)
Karaman *et al.* [6]	4.6
Wang *et al.* [23]	18.2
Oneata *et al.* [10]	36.6
Richard *et al.* [14]	39.7
Shou *et al.* [17]	47.7
Yeung *et al.* [28]	48.9
Yuan *et al.* [29]	51.4
Xu *et al.* [24]	54.5
Zhao *et al.* [32]	60.3
Kaufman *et al.* [8]	61.1
Ours [2]	57.9

Our weakly supervised approach *vs.* fully supervised approaches for action detection on THUMOS'14, measured by mAP at IoU threshold $\alpha = 0.1$.

Applications

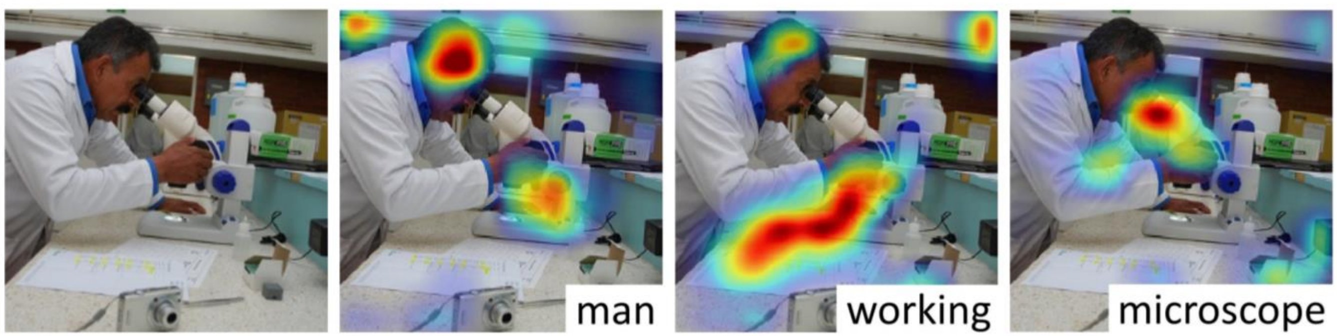
- Action Detection (*videos*)
- Caption Grounding (*images, videos*)
- Reflecting the Abstraction Capability of Models

Flicker30kEntities Dataset: Grounding Words of an Image Caption

image caption: *A man in a lab coat is working on a microscope.*

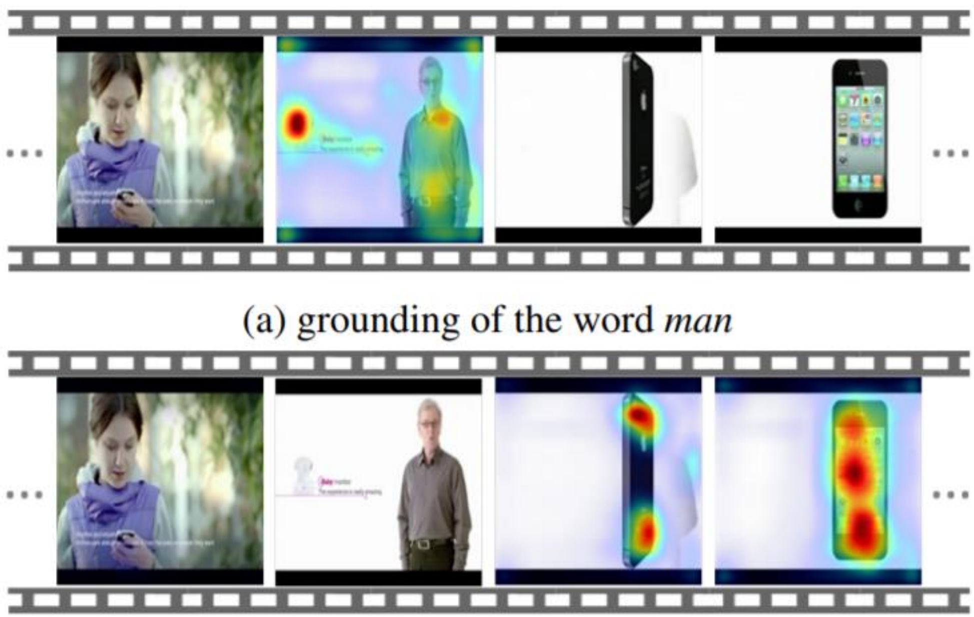
Flicker30kEntities Dataset: Grounding Words of an Image Caption

image caption: *A man in a lab coat is working on a microscope.*



MSRVTT Dataset: Grounding Words of a Video Caption

video caption: *"A man is talking about a phone"*



(a) grounding of the word *man*

(b) grounding of the word *phone*

Applications

- Action Detection (*videos*)

- Caption Grounding (*images, videos*)

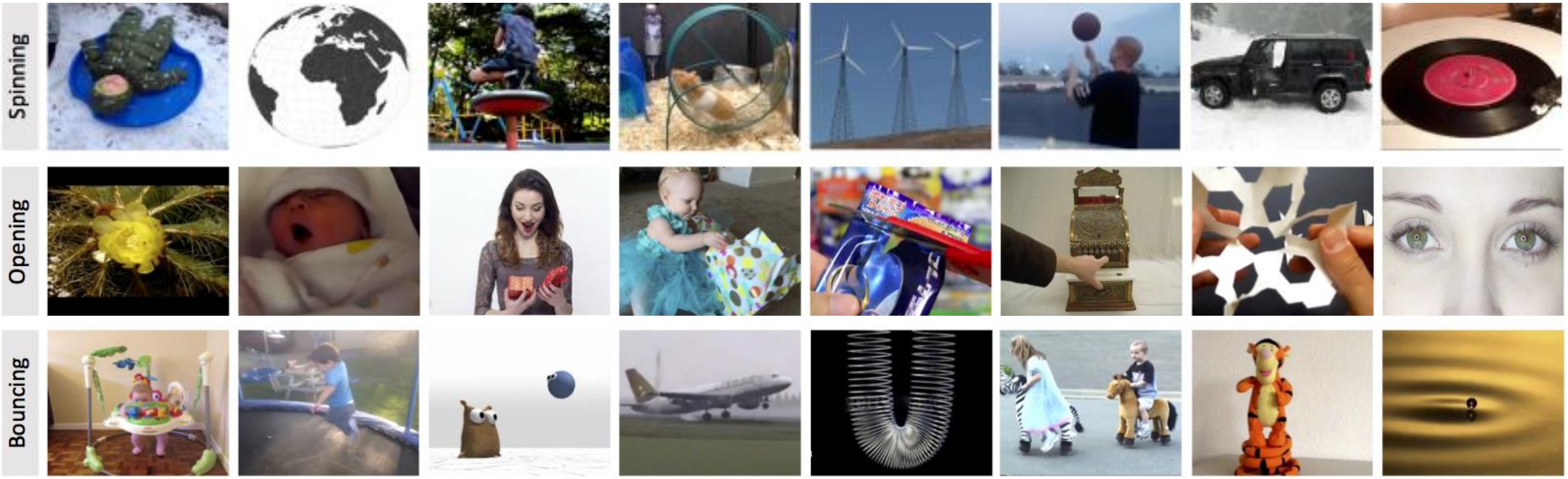
- Reflecting the Abstraction Capability of Models

Reflecting the Abstraction Capability of Models

- Moments in Time Dataset

M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfruend, C. Vondrick, A. Oliva. "Moments in Time Dataset: one million videos for event understanding." *TPAMI*, 2019.

- Videos of abstract dynamical events performed by various actors.

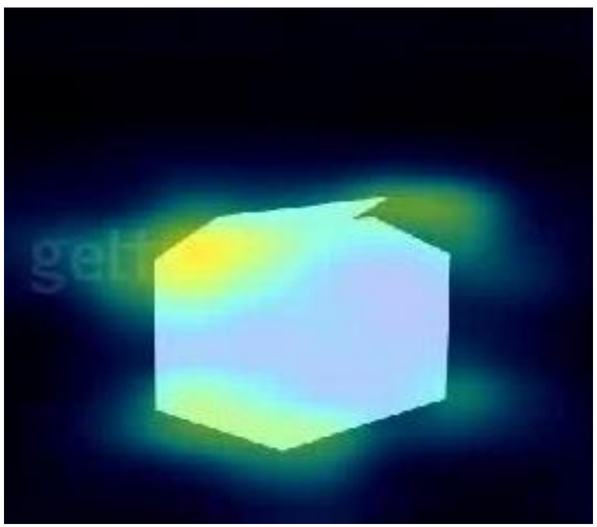


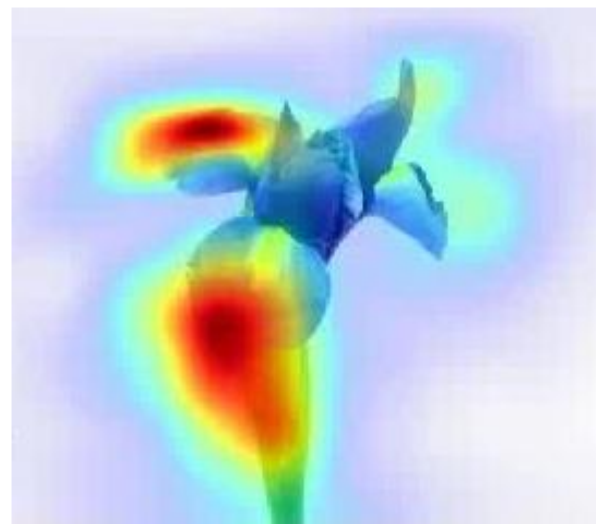
Moments in Time Dataset

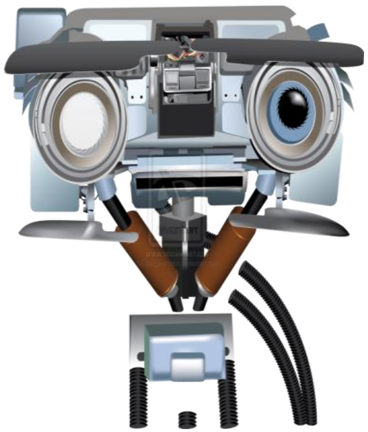
- Typically, classification accuracy is reported to summarize the recognition capability of models.
- However, classification accuracy alone is not representative as to whether the models are really modeling this diversity of actors.
- A classifier may be incorrectly classifying a whole subset of cases/actors.

Moments in Time Dataset

- Class: *Opening*



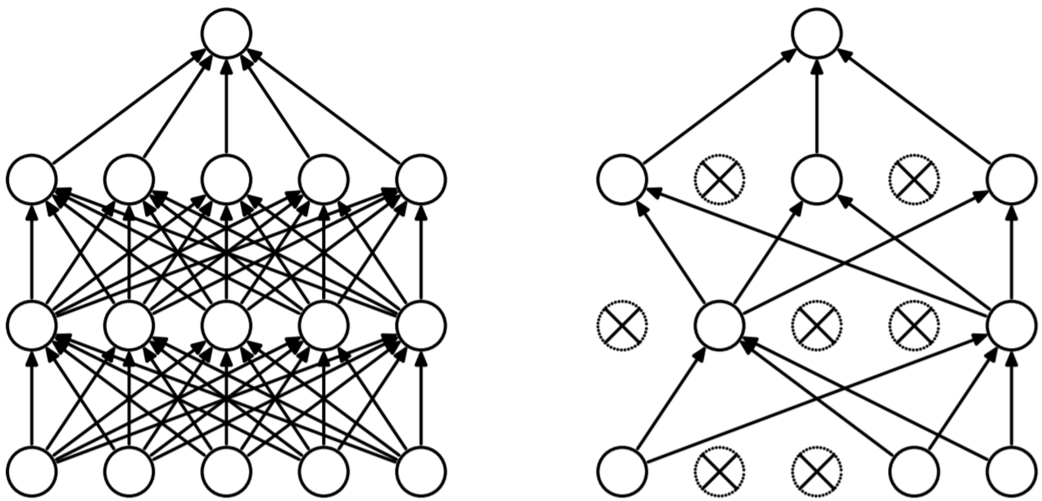




Explainability for Better Models

Dropout: A Classical Regularization Technique

- Many Deep Models employ dropout at training time to avoid overfitting, allowing



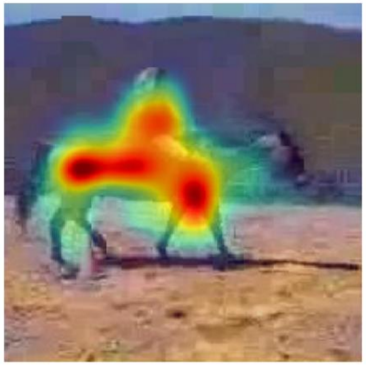
(a) Standard Neural Net (b) After applying dropout. [Srivastava *et al.*]

Excitation Dropout

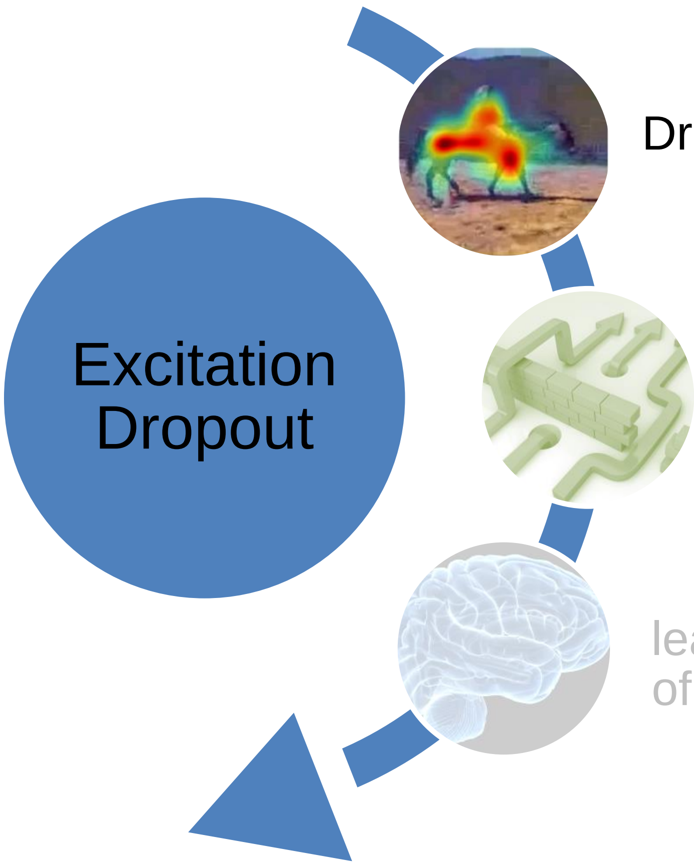
- We target answering the question: *Which neurons to drop out?*
  - *Neurons that have a higher contribution to the ground-truth prediction.*
  - *Example for ground-truth class HorseRiding:*

$image$

$evidence$: p_{EB}



Our Approach

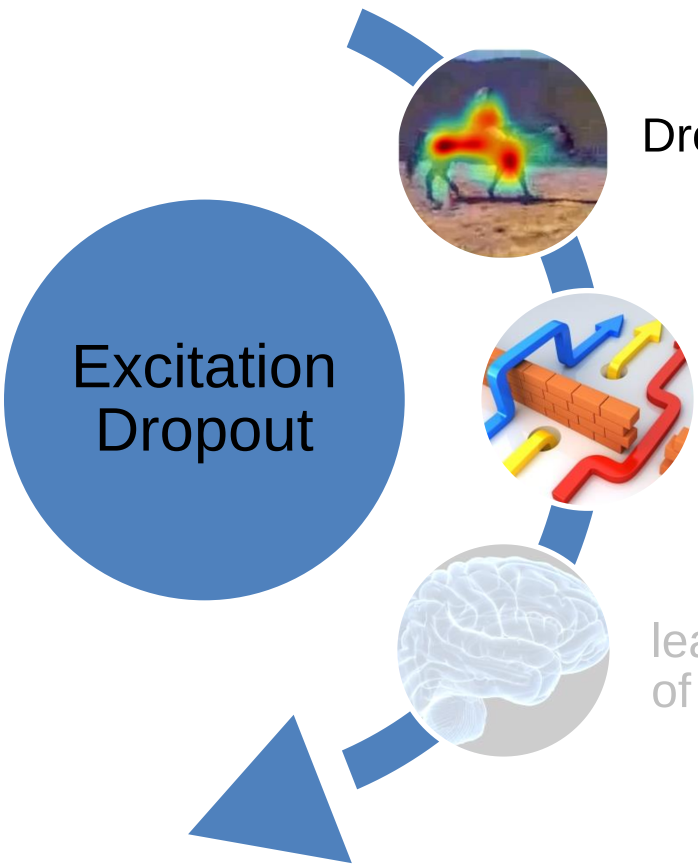


Dropping neurons having stronger evidence,

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



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

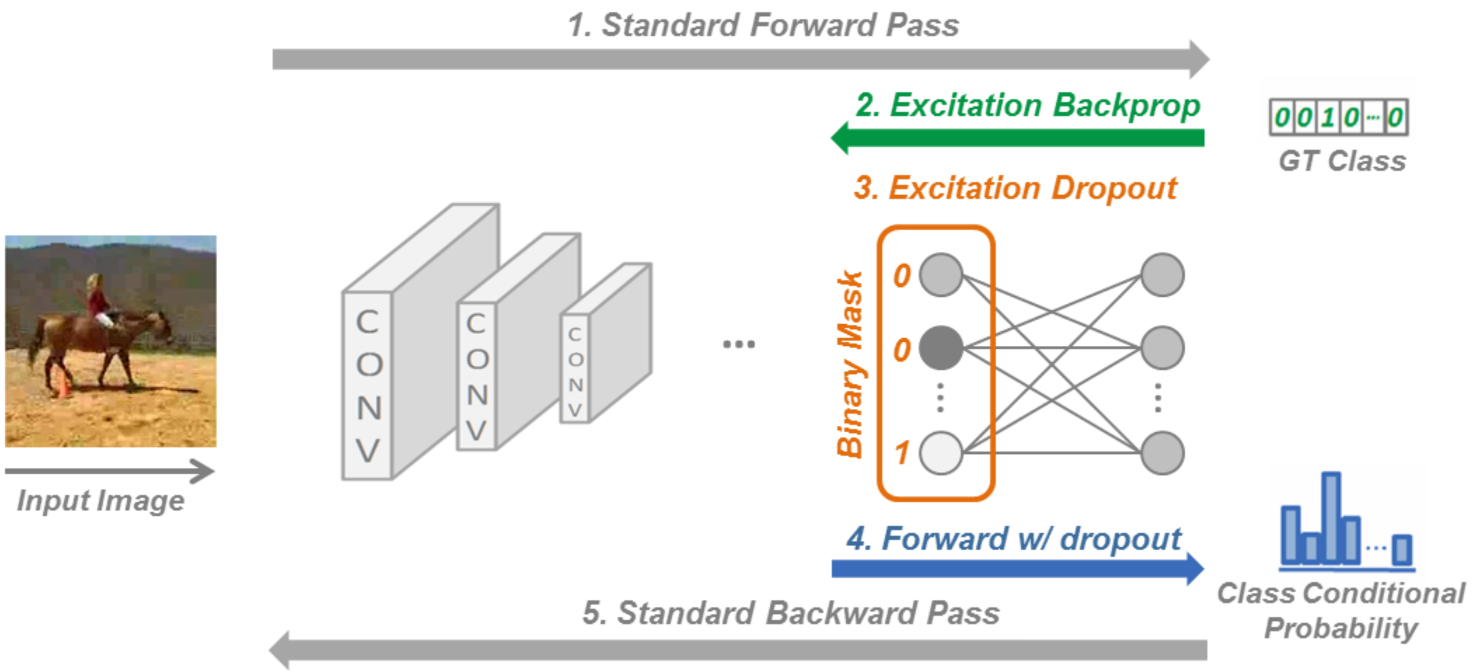
Excitation Dropout

Dropping neurons having stronger evidence,

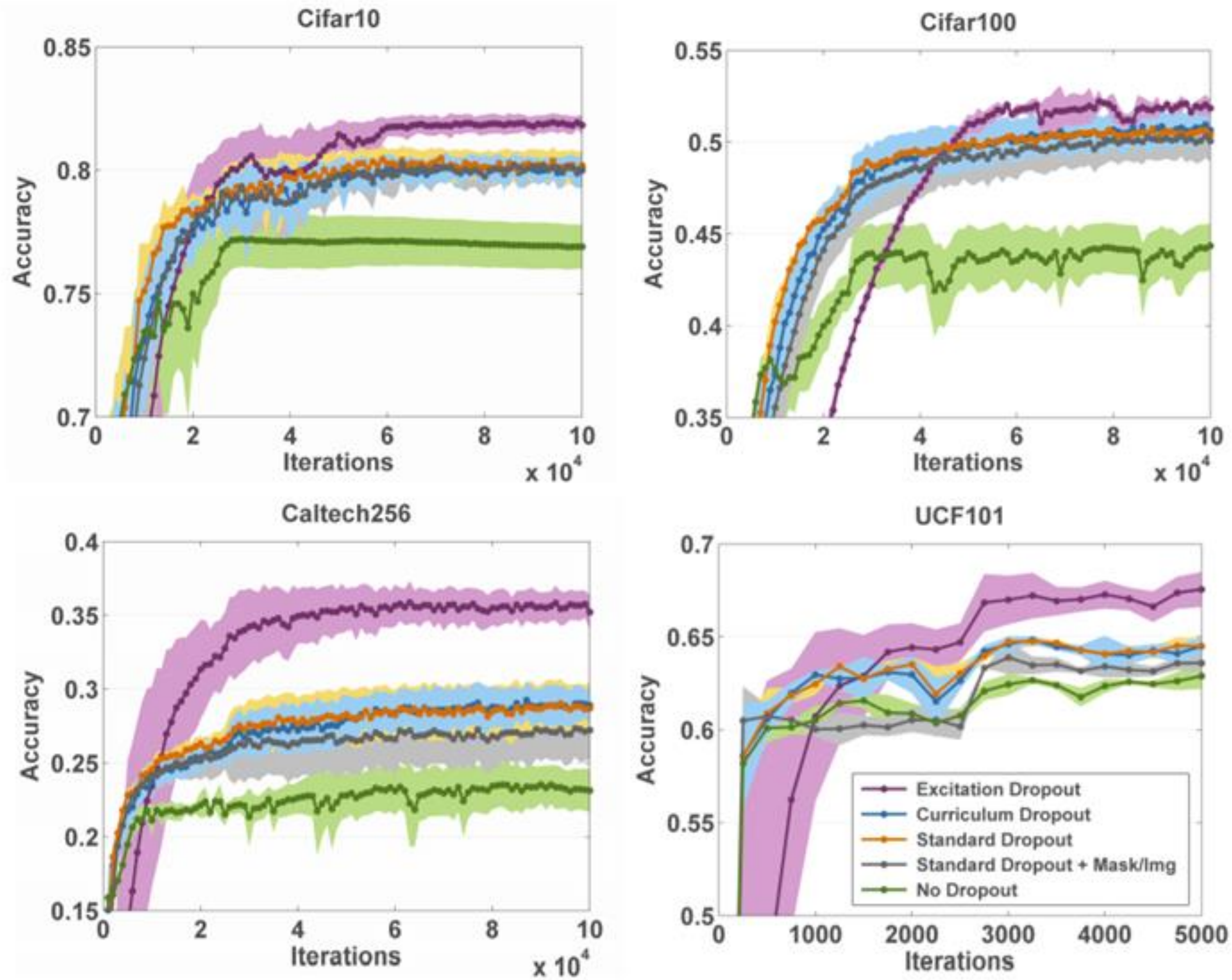
forces a network to learn alternative paths in order to maintain loss minimization,

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

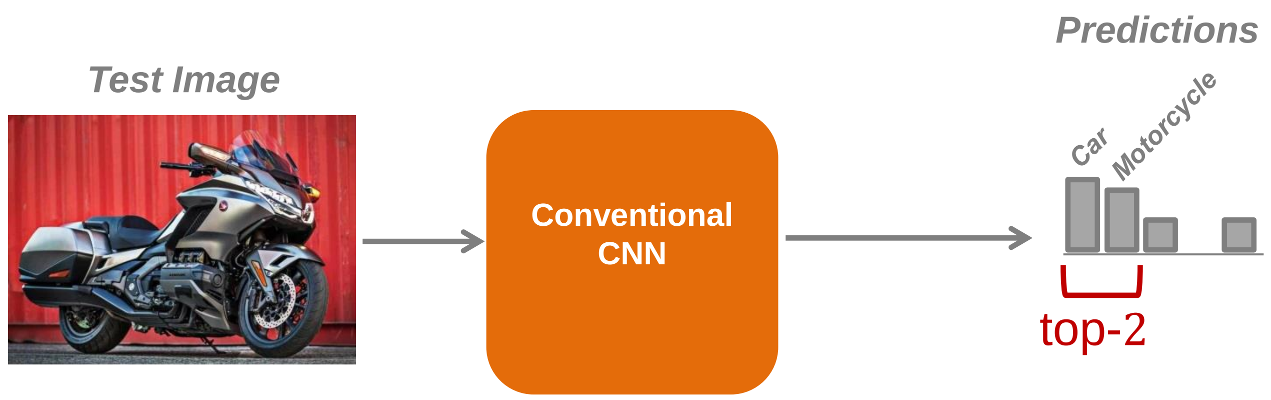
Excitation Dropout Pipeline



Improved Generalization



Conventional Deep Classification

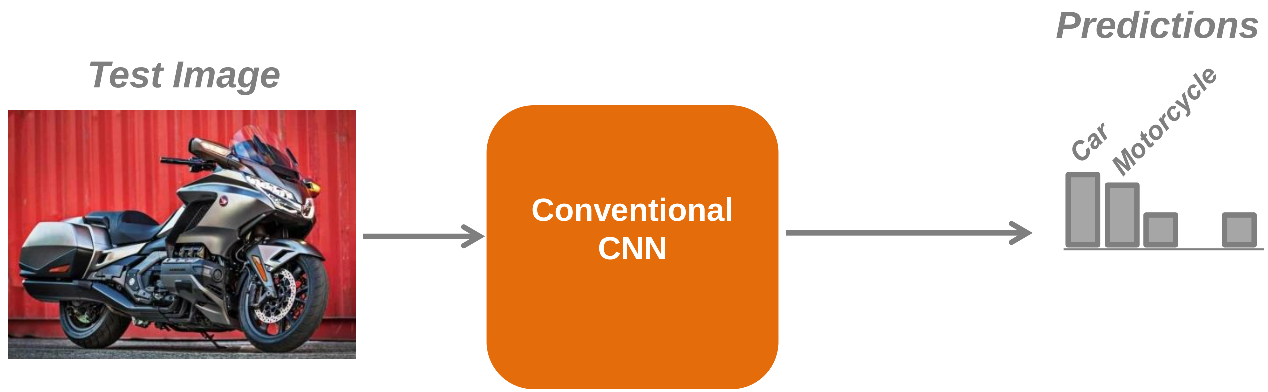


- The top-k $(k = 2{,}3{,}4, \ldots)$ classification accuracy is usually significantly higher than the top-1 accuracy.

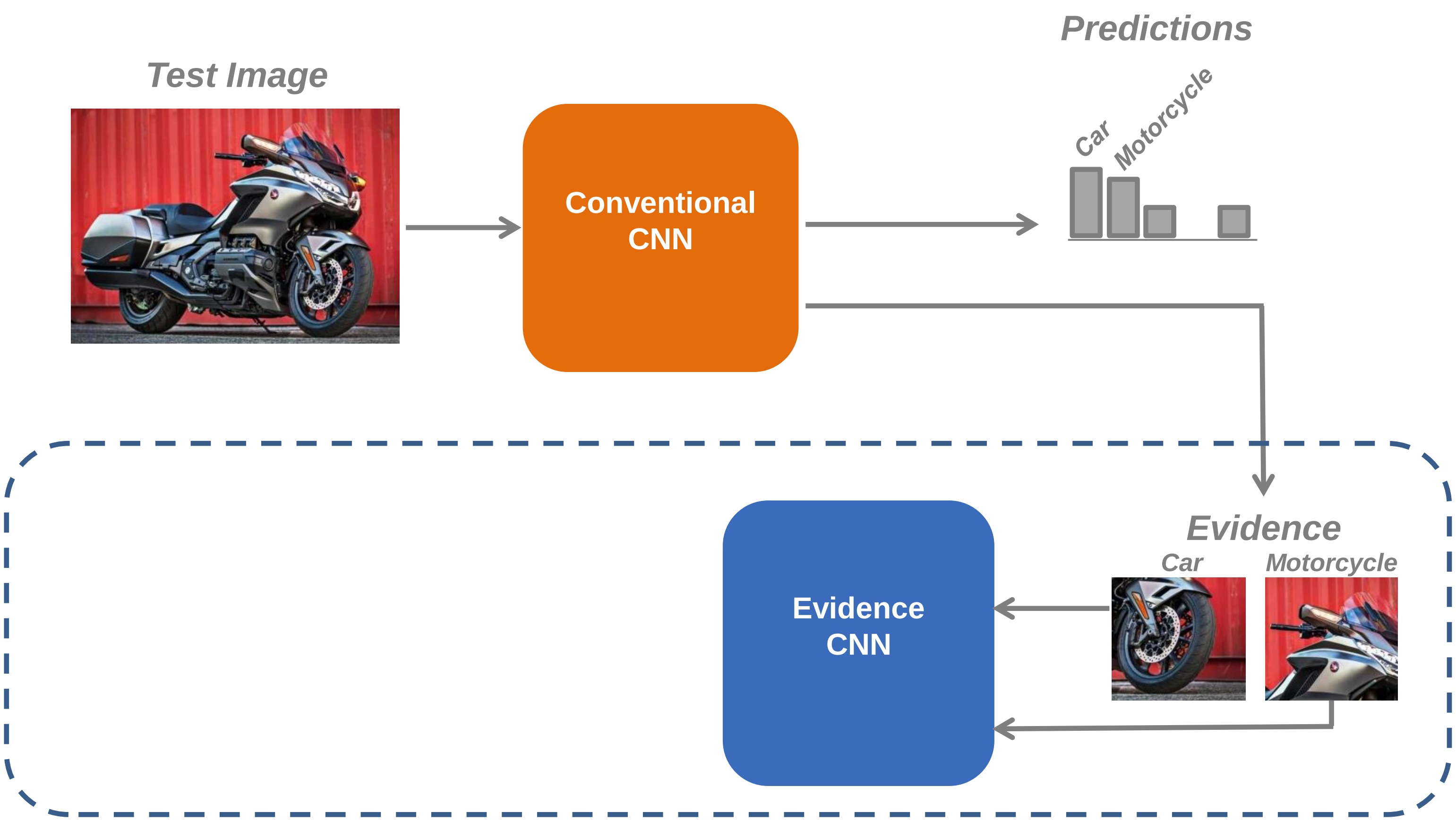
- This is more evident in fine-grained datasets, where differences between classes are quite subtle.

  Stanford Dogs: top-1: 86.9%, top-5: 98.9%

Guided Zoom: Pipeline

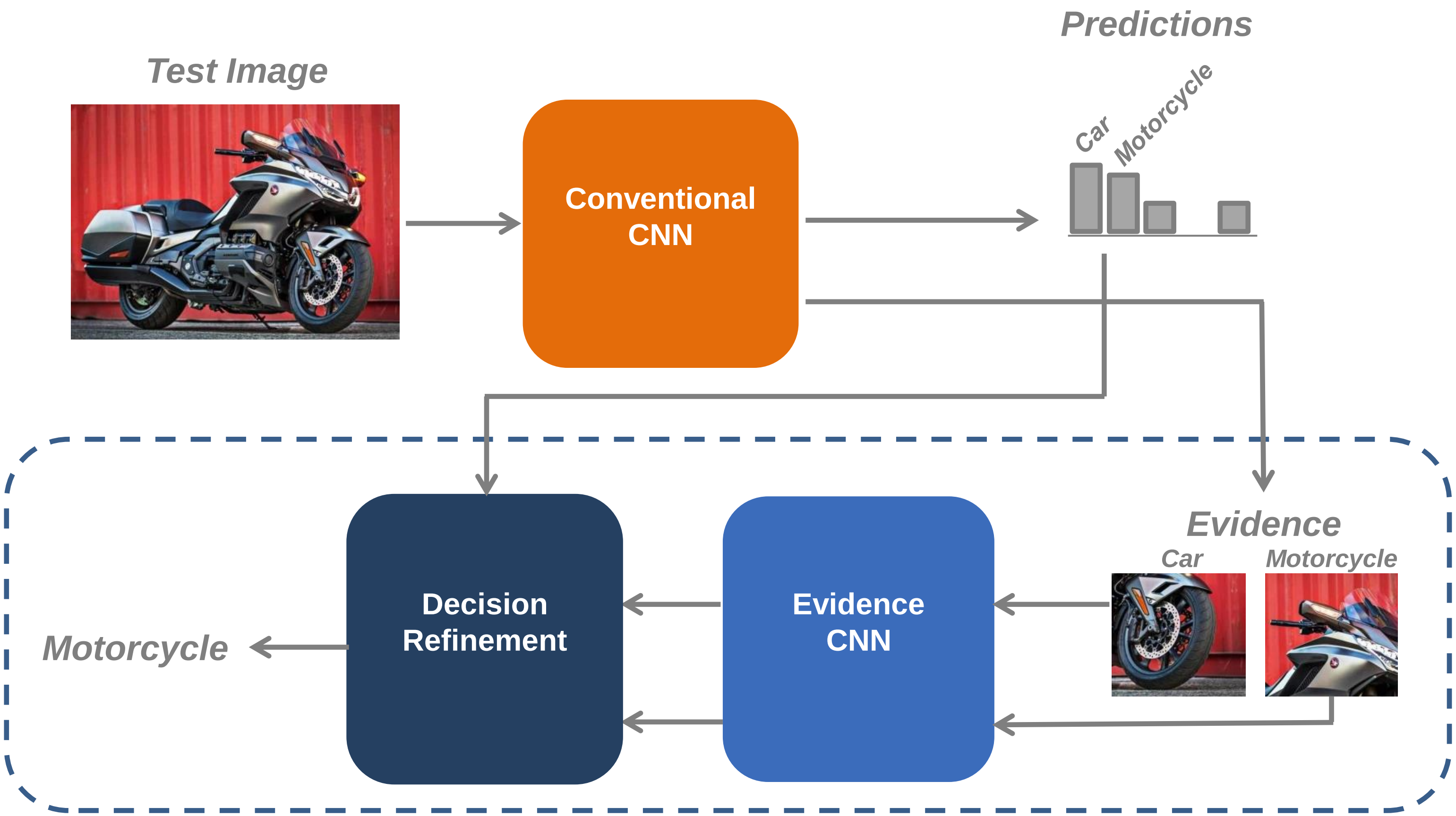


Guided Zoom: Pipeline



- **Evidence CNN** is trained to classify an evidence pool
- Generation of an Evidence Pool P

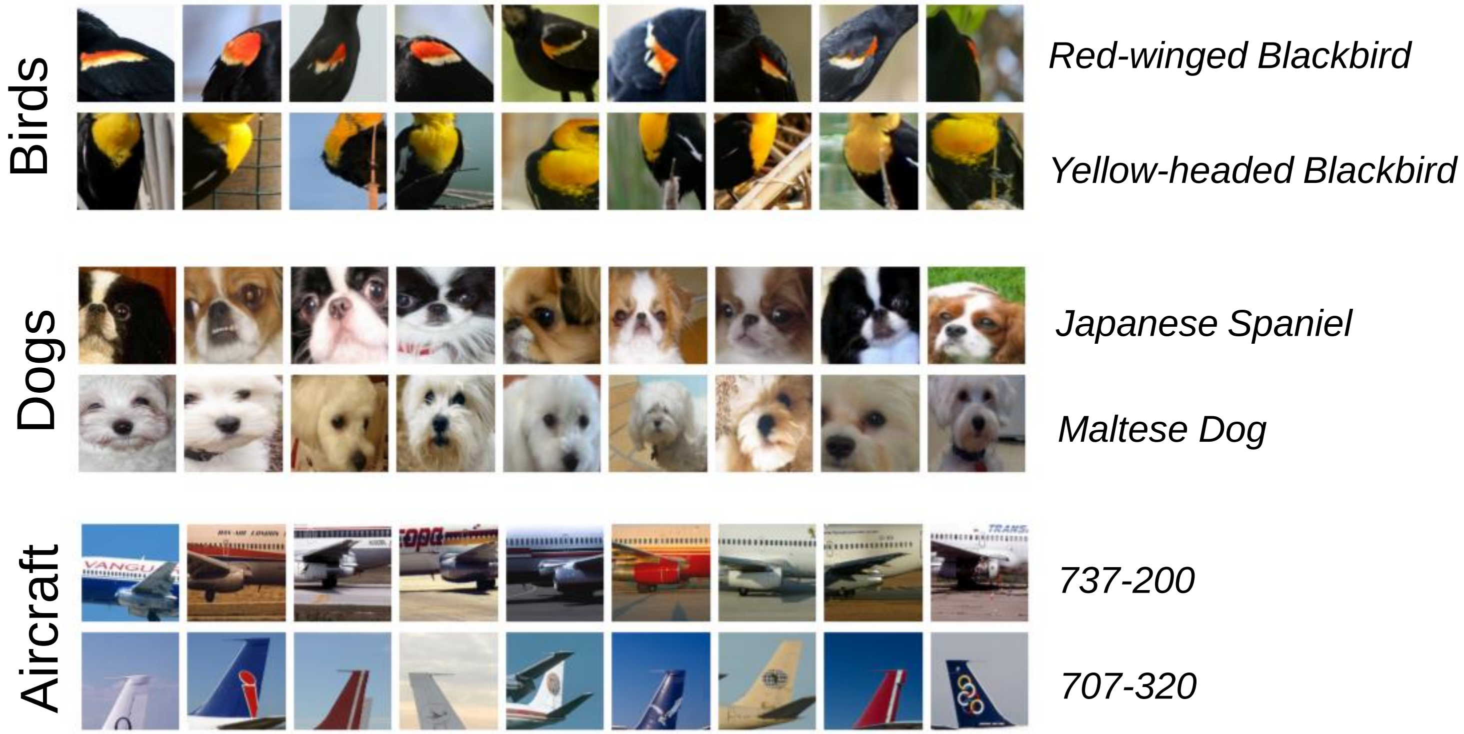
Guided Zoom: Pipeline



- **Conventional CNN** Prediction
- **Evidence CNN** Prediction

Evidence Pool P

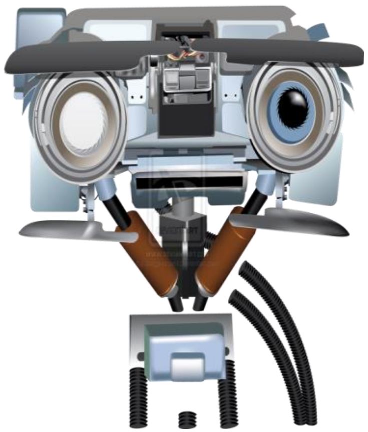
- We extract evidence patches from original training images around the peak saliency.



Results

- Classification accuracy of three fine-grained datasets:

Method	CUB-200-2011 Birds Dataset	Stanford Dogs Dataset	FGVC-Aircraft Dataset
Conventional CNN (ResNet-101)	82.3%	86.9%	87.5%
Guided Zoom (ResNet-101)	85.4%	88.5%	89.0%



Domain Adaptation

Kate Saenko

Has deep learning solved vision?

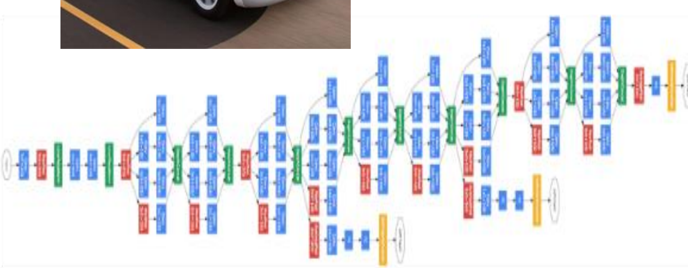
pedestrian detection FAIL

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

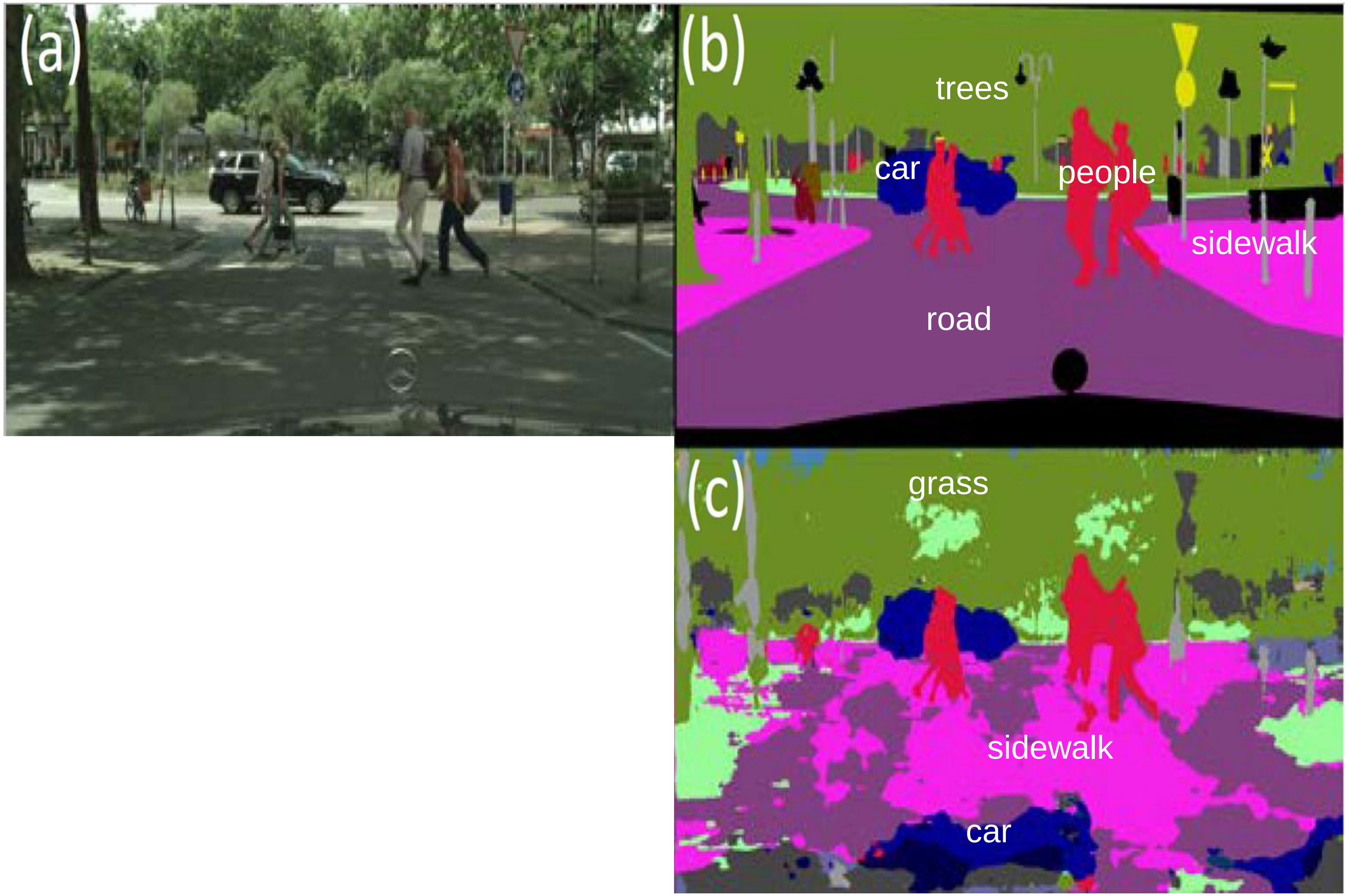
True Segmentation

Model Output

Solution: Domain Adaptation

Input Image

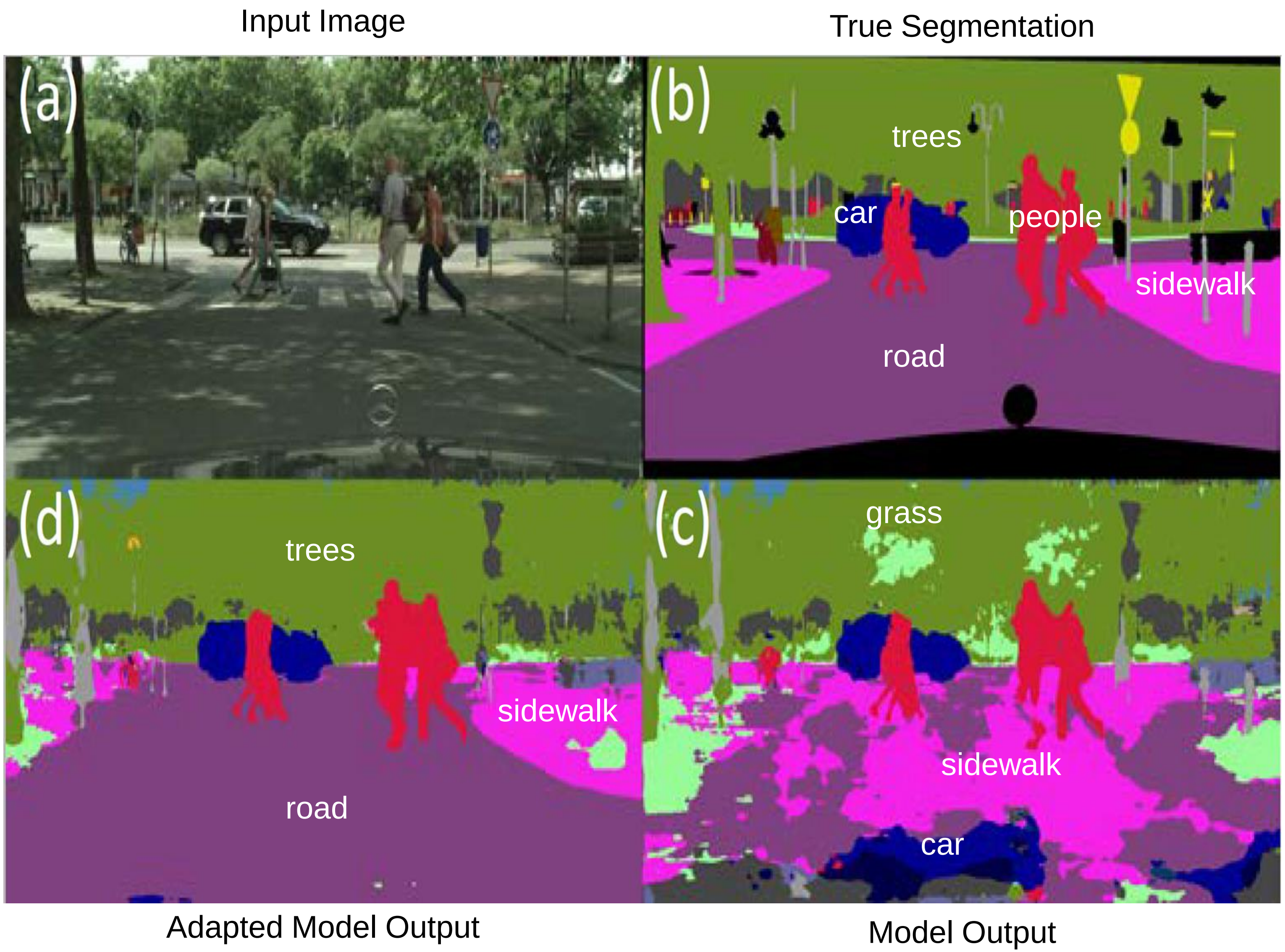
True Segmentation



Adapted Model Output

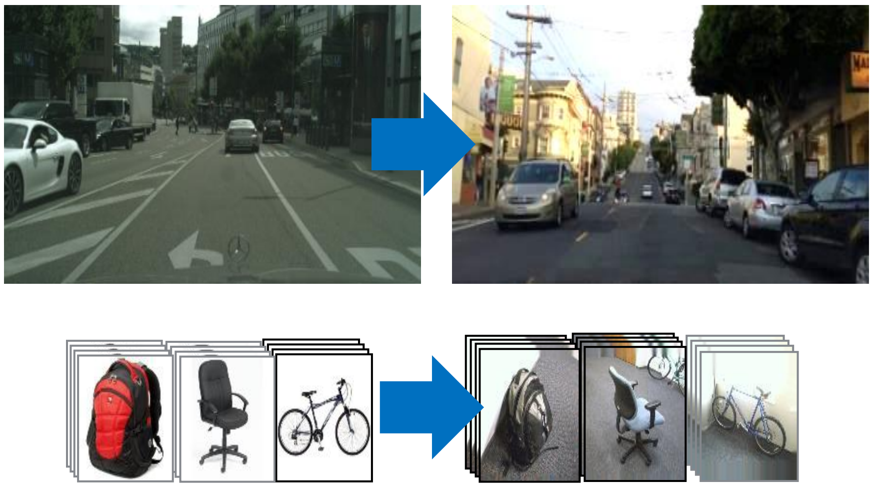
Model Output

Solution: Domain Adaptation



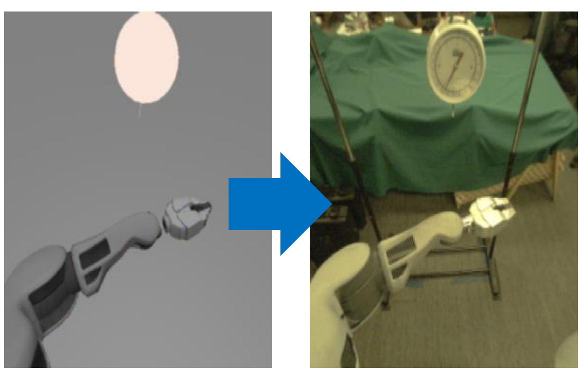
Applications of Domain Adaptation

From dataset to dataset

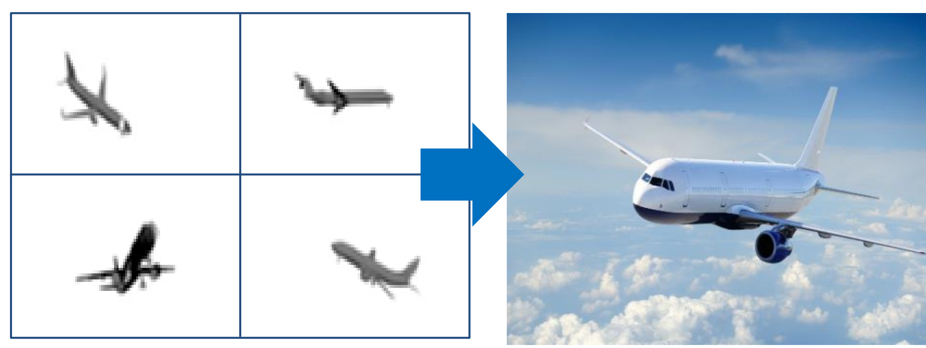


From RGB to depth

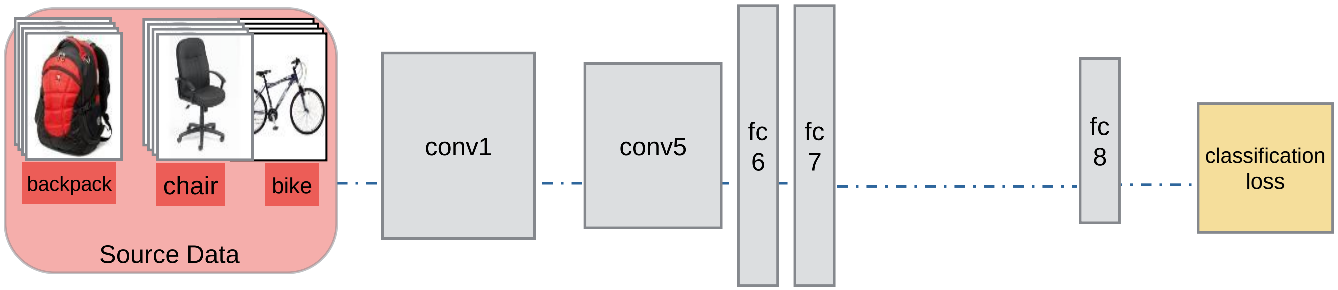
From simulated to real control



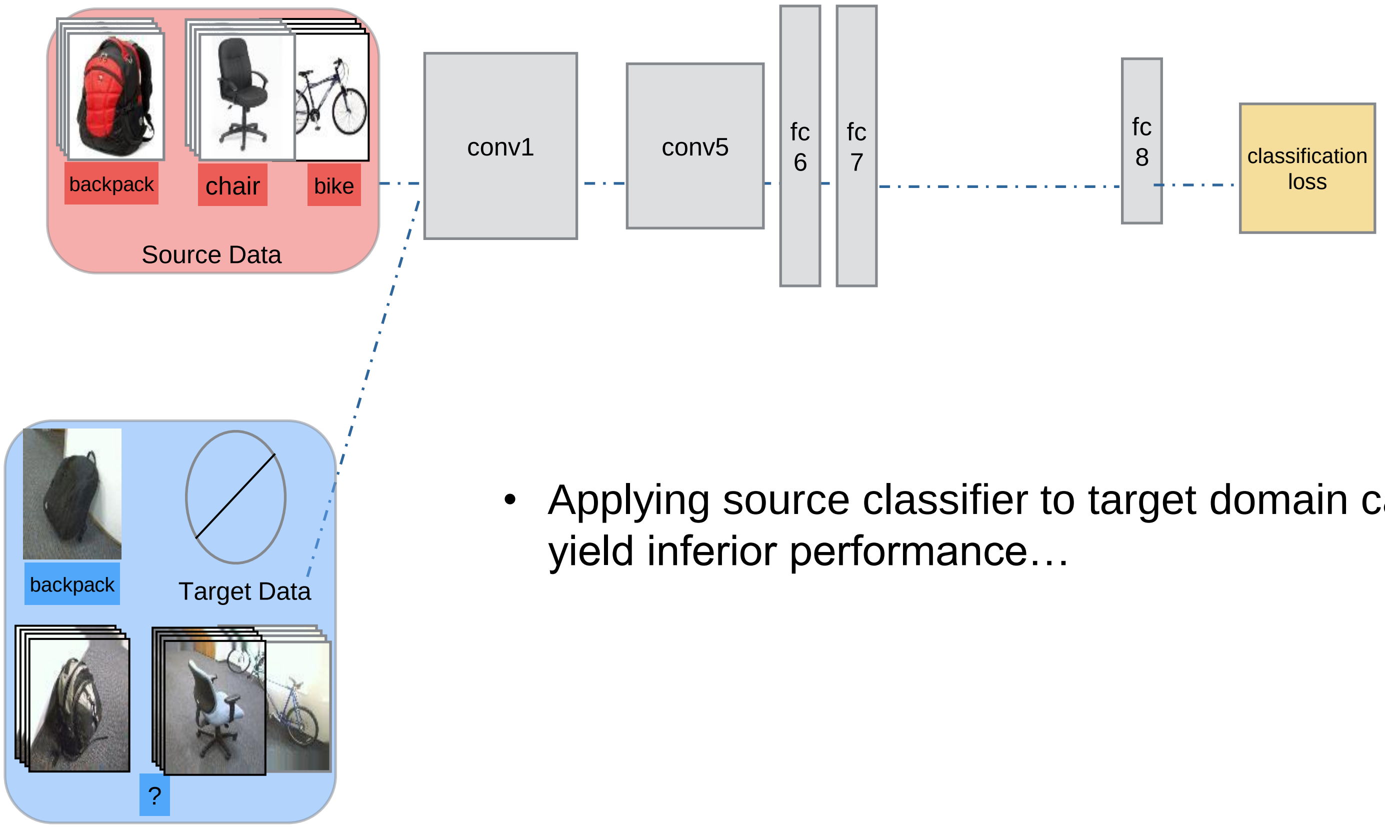
From CAD models to real images



How to adapt a deep network?



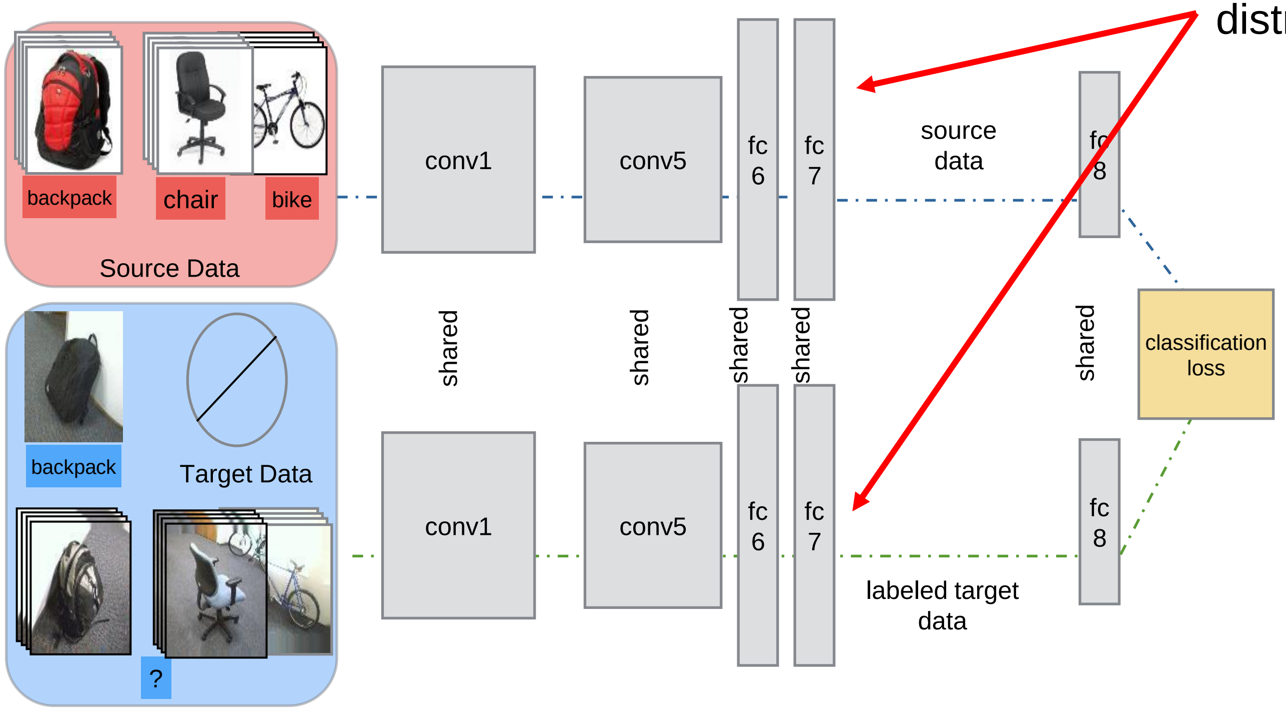
How to adapt a deep network?



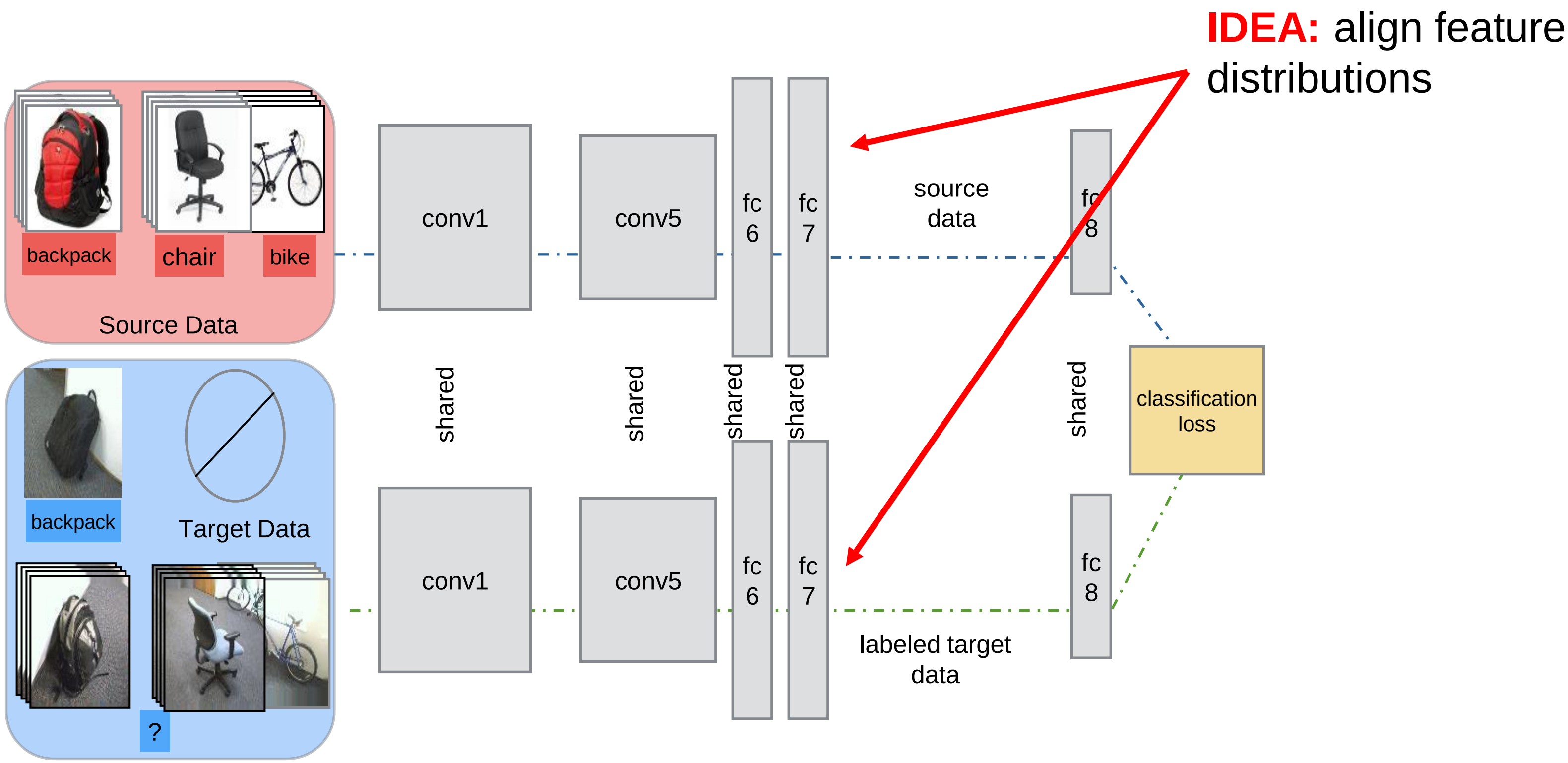
- Applying source classifier to target domain can yield inferior performance…

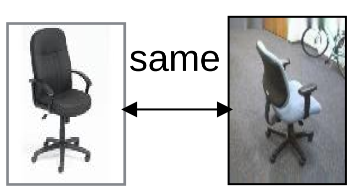
How to adapt a deep network?

IDEA: align feature distributions



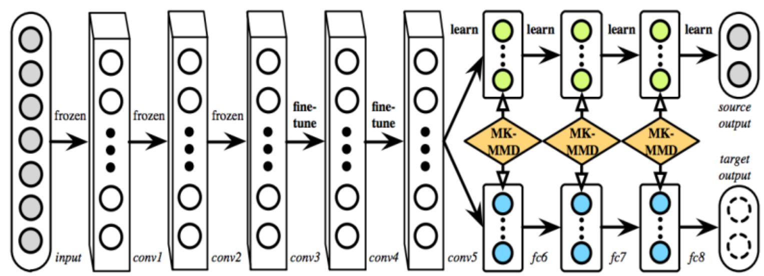
How to adapt a deep network?



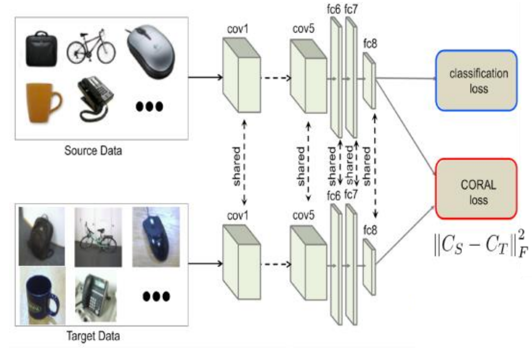


Solution: align deep feature distributions

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

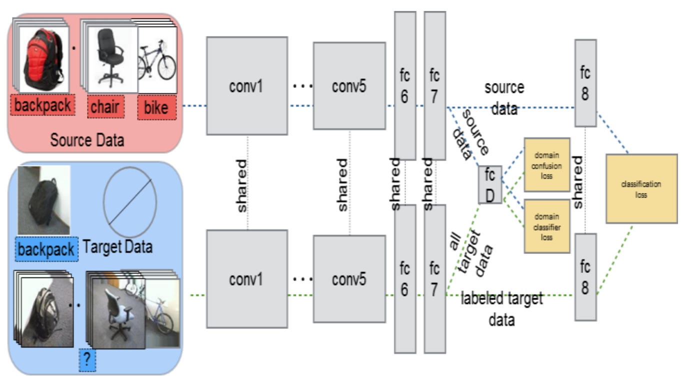


Maximum Mean Discrepancy M. Long, et al. ICML 2015

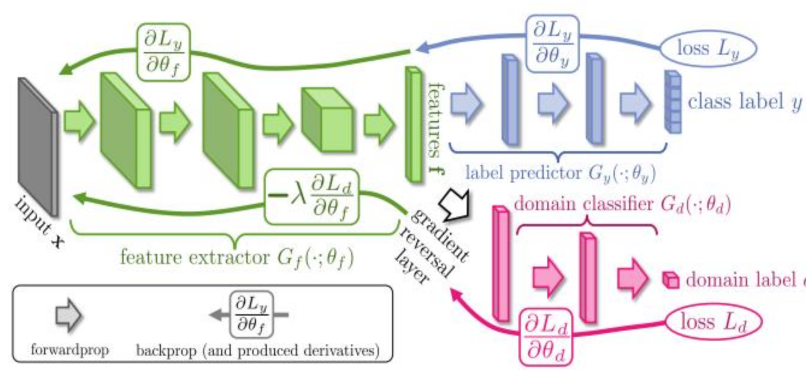


CORrelation ALignment Sun and Saenko, AAAI 2016

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



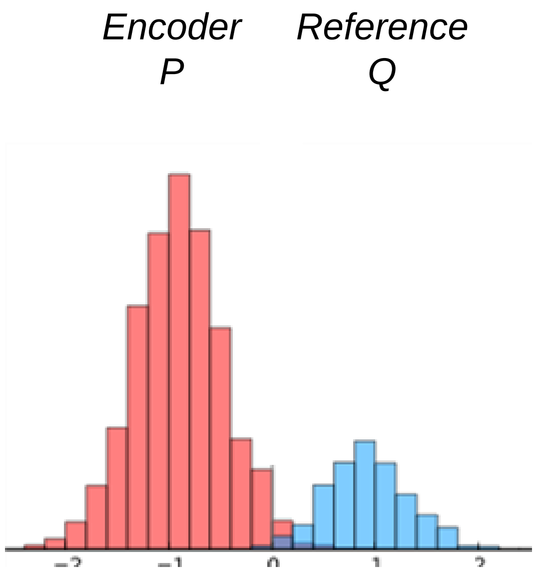
Domain Confusion E. Tzeng et al. ICCV 2015

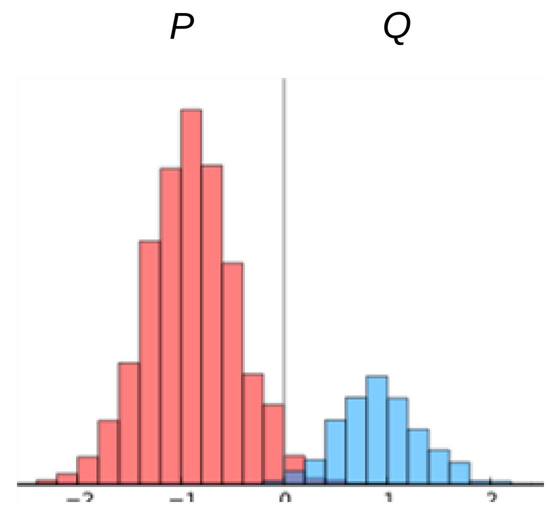


Reverse Gradient Y. Ganin and V. Lempitsky ICML 2015

Adversarial Feature Alignment

Adversarial networks



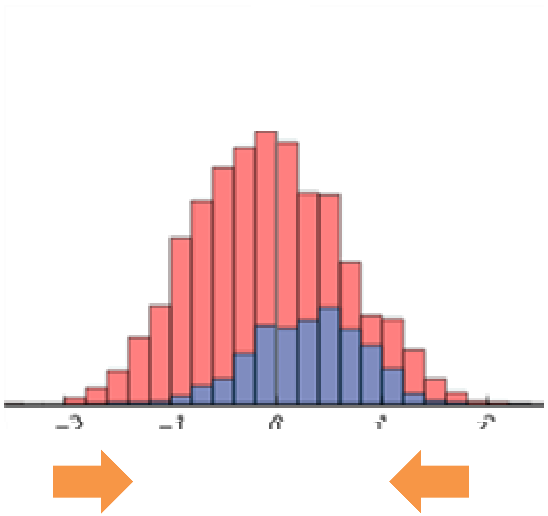


Encoder
Generates features such that their distribution P matches reference distribution Q

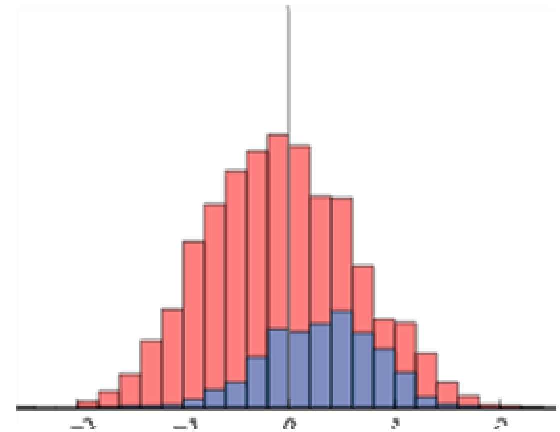
Adversary
Tries to discriminate between samples from P and samples from Q

Adversarial networks

Encoder *Reference*
P *Q*



P *Q*



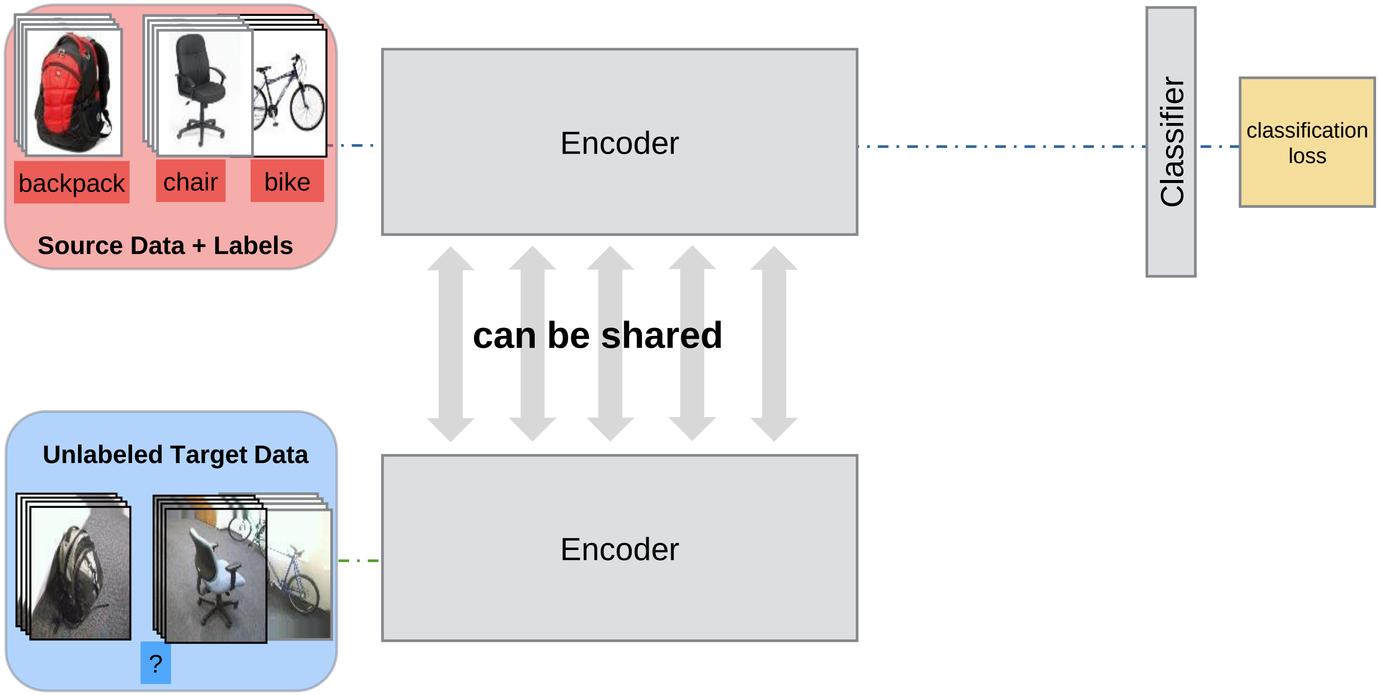
Encoder
Generates features such that their distribution P matches reference distribution Q

fools adversary

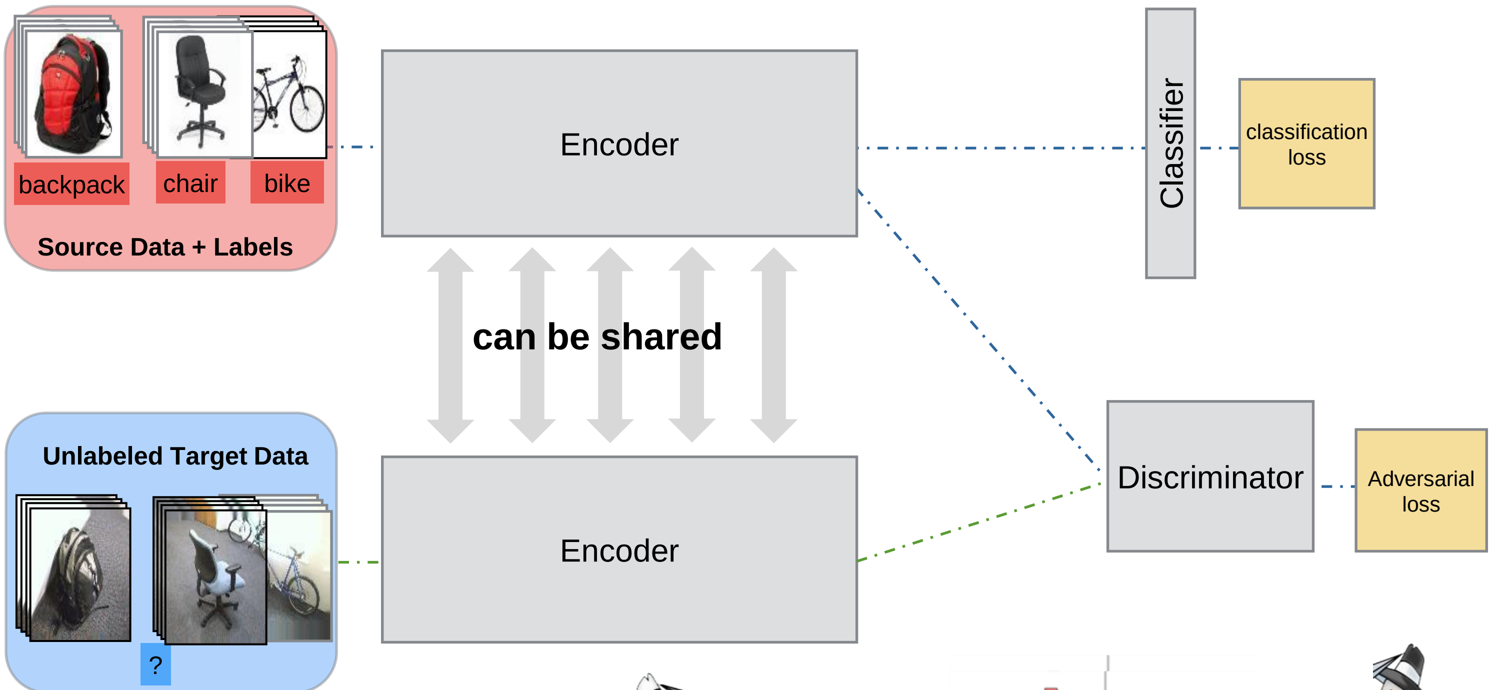
Adversary
Tries to discriminate between samples from P and samples from Q

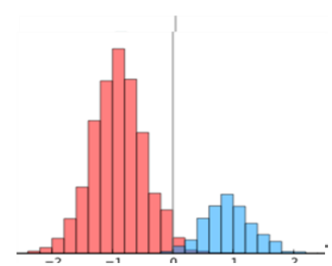
tries harder

Adversarial domain adaptation

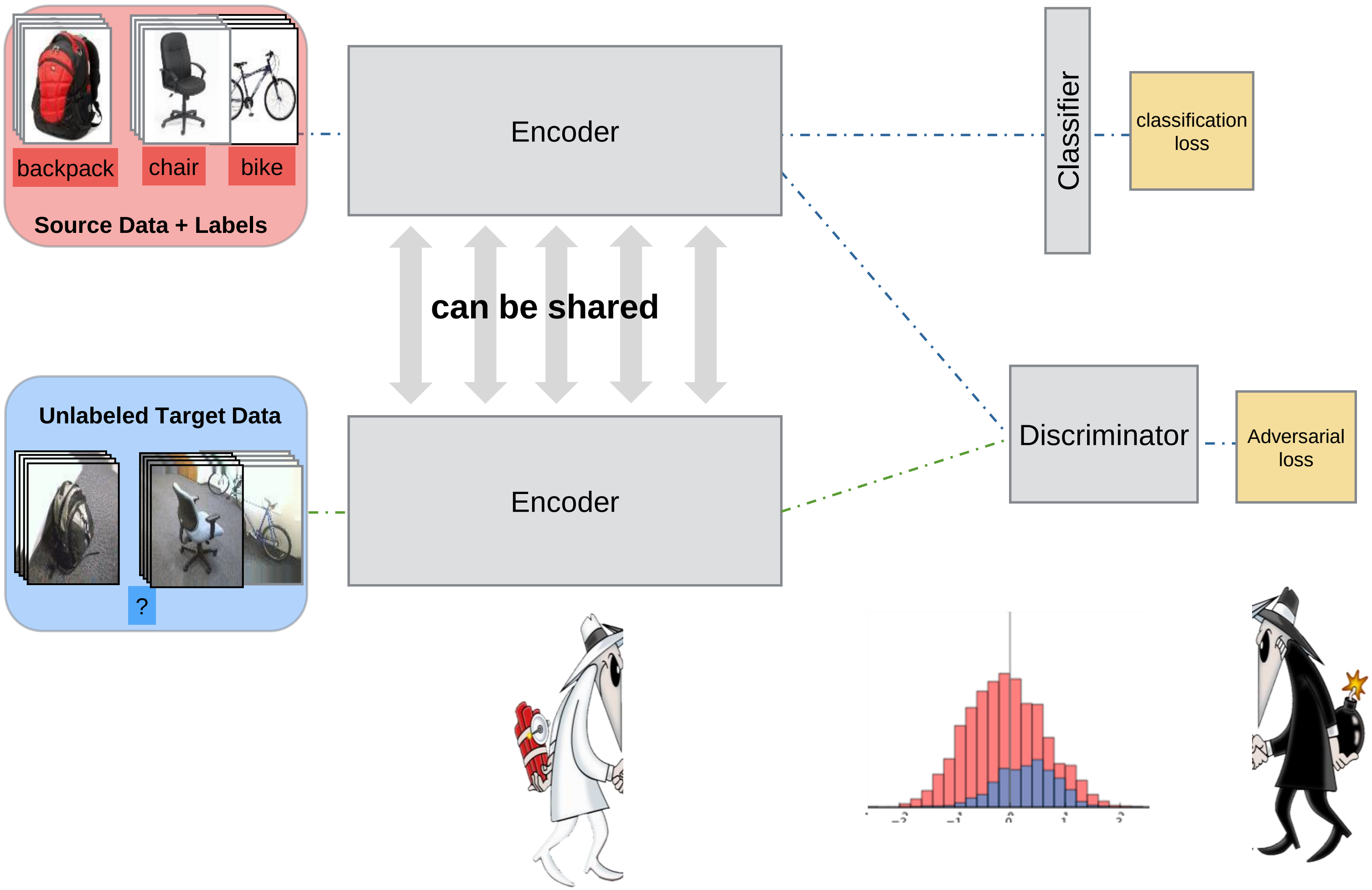


Adversarial domain adaptation

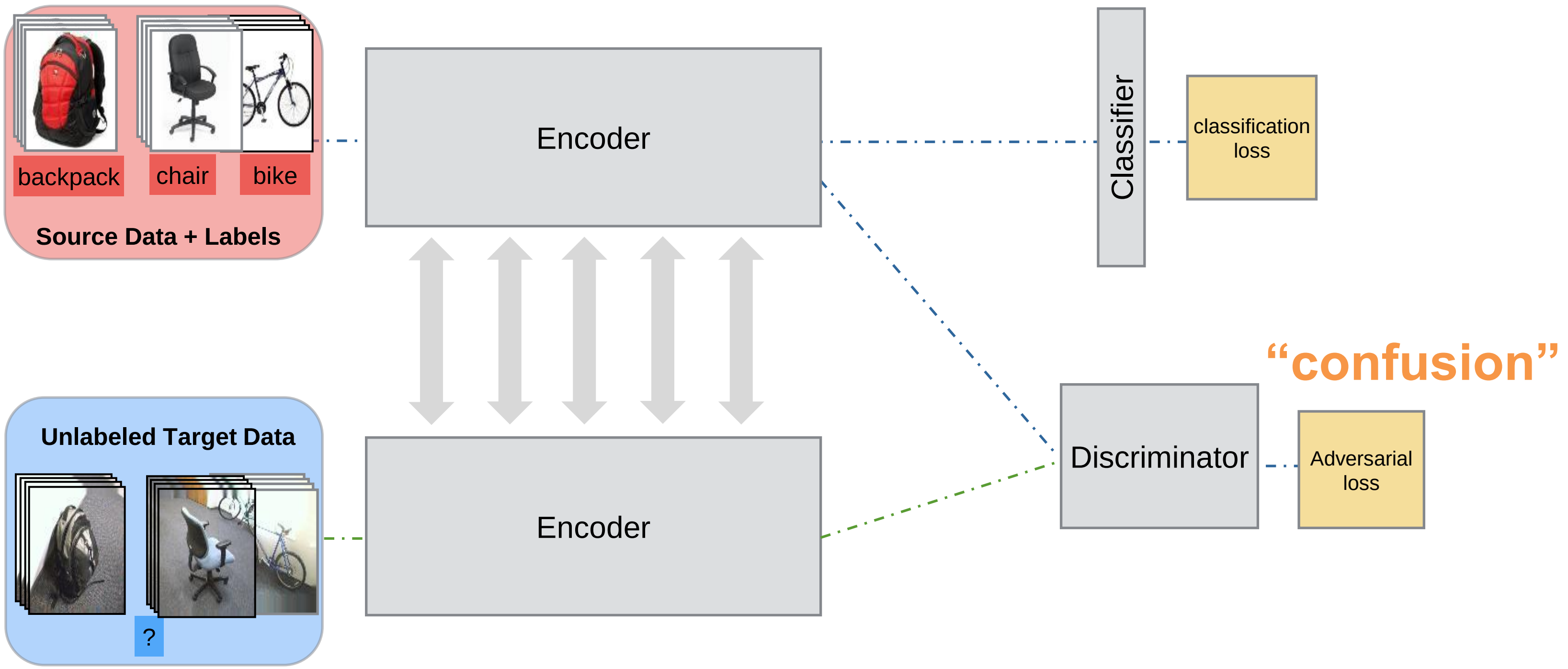




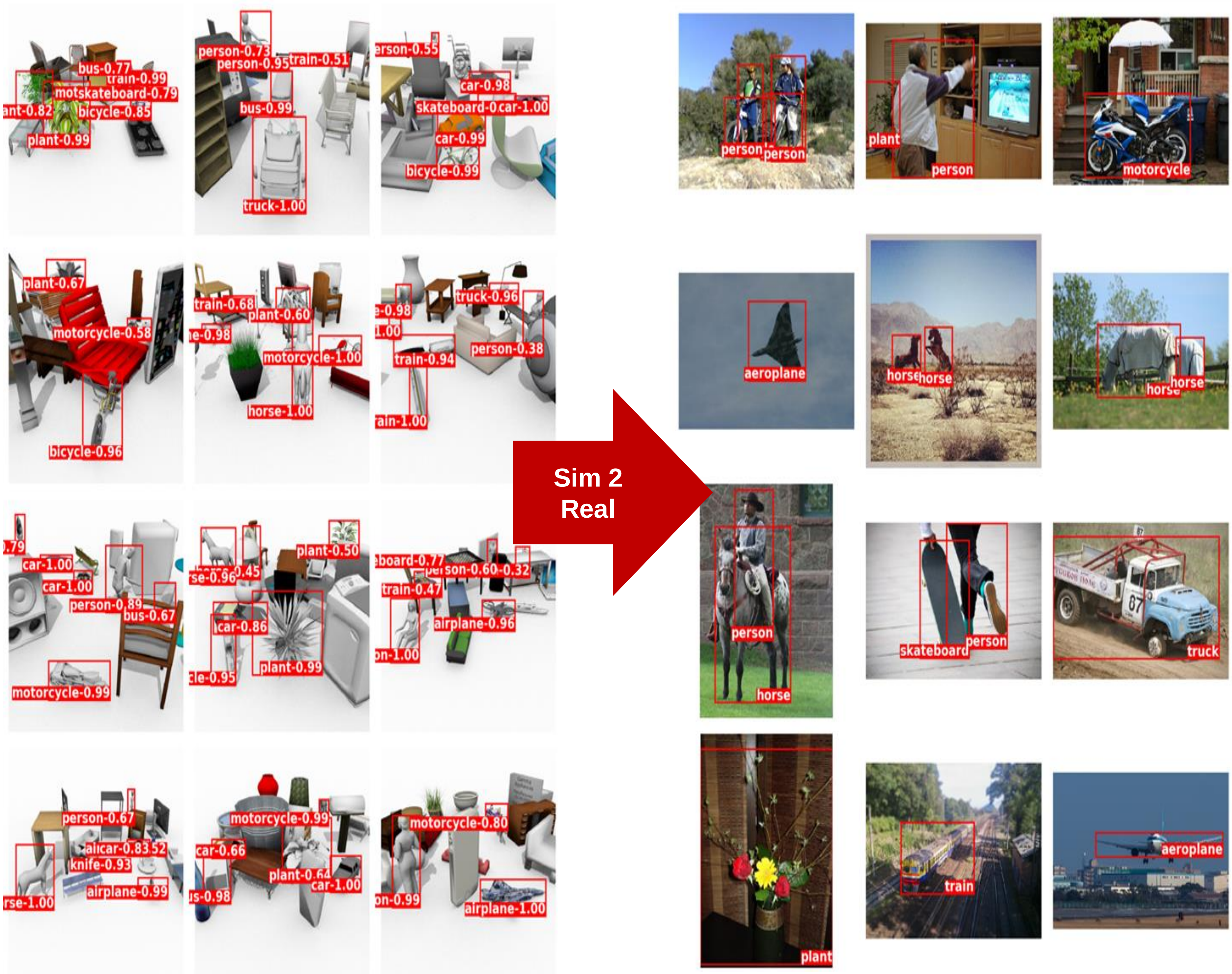
Adversarial domain adaptation

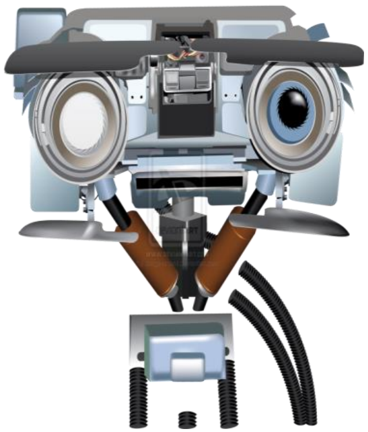


Design choices in adversarial adaptation



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


bus-0.77
train-0.99
motskateboard-0.79
ant-0.82
bicycle-0.85
plant-0.99
person-0.73
person-0.95
train-0.51
bus-0.99
truck-1.00
erson-0.55
car-0.98
skateboard-0
car-1.00
car-0.99
bicycle-0.99
plant-0.67
motorcycle-0.58
bicycle-0.96
train-0.68
plant-0.60
e-0.98
motorcycle-1.00
horse-1.00
truck-0.96
train-0.94
person-0.38
ain-1.00
.79
car-1.00
car-1.00
person-0.89
bus-0.67
motorcycle-0.99
plant-0.50
se-0.96
car-0.86
plant-0.99
cle-0.95
board-0.77
train-0.47
airplane-0.96
on-1.00
person-0.67
knife-0.93
airplane-0.99
rse-1.00
motorcycle-0.99
car-0.66
plant-0.64
car-1.00
us-0.98
motorcycle-0.80
on-0.99
airplane-1.00
Sim 2
Real
person
person
plant
person
motorcycle
aeroplane
horse
horse
horse
horse
person
horse
skateboard
person
truck
plant
train
aeroplane

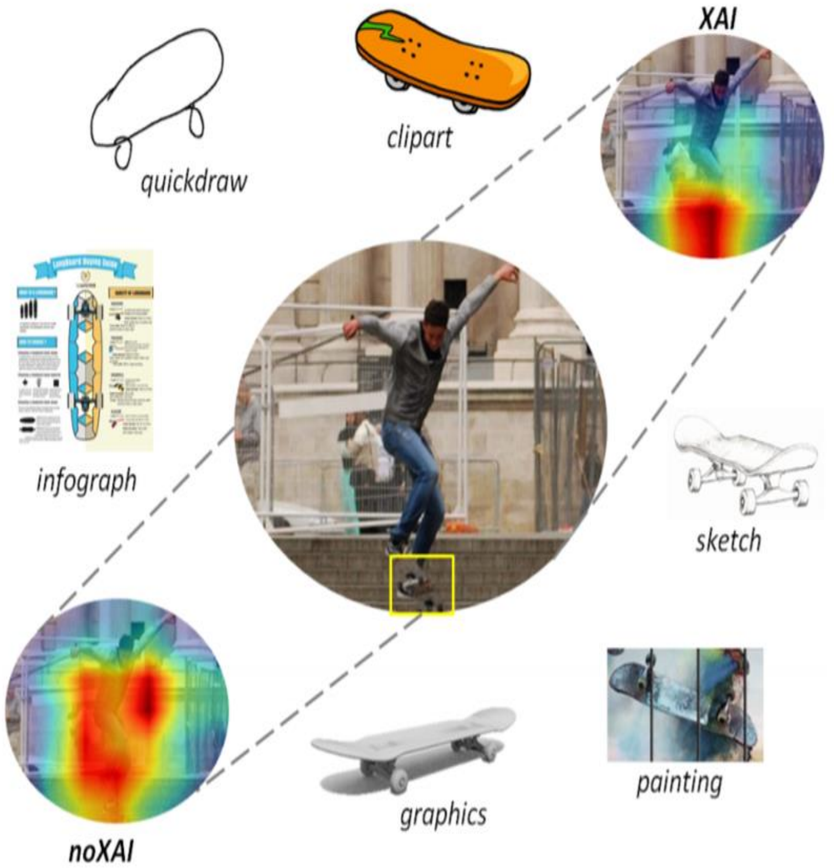


Explainability and Domain Generalization

Bargal & Saenko

Explainable AI (XAI) for Domain Generalization

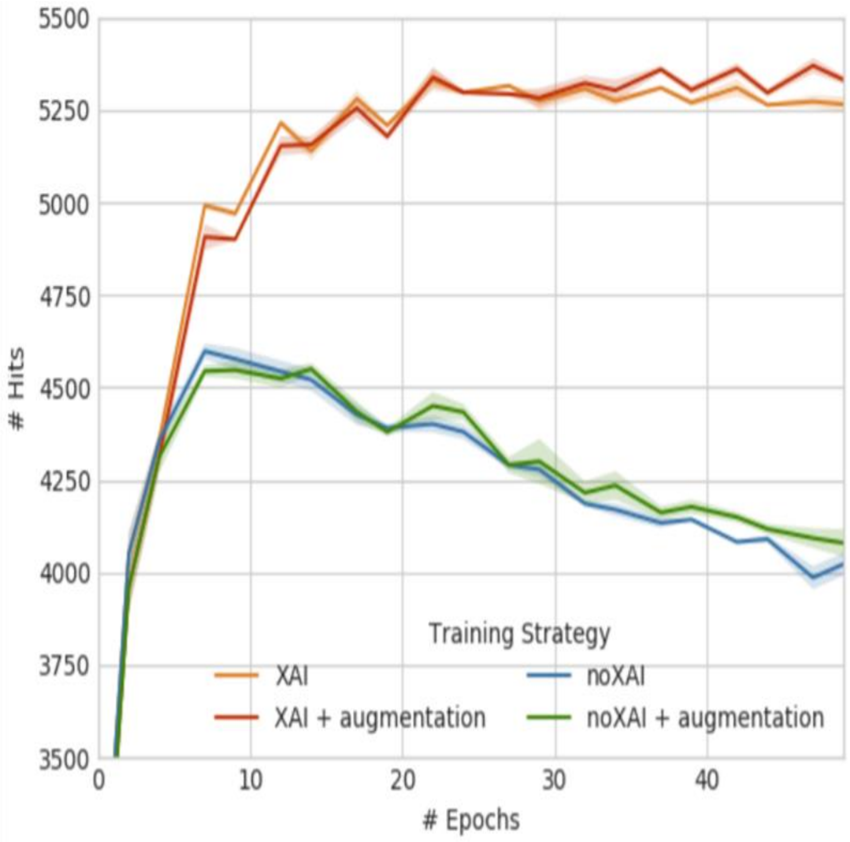
Training a deep neural network model to enforce explainability, *e.g.* focusing on the skateboard region (red is most salient, and blue is least salient) for the ground-truth class skateboard in the central training image, enables improved generalization to other domains where the background is not necessarily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



Explainability Results: Quantitative *[Automated]*

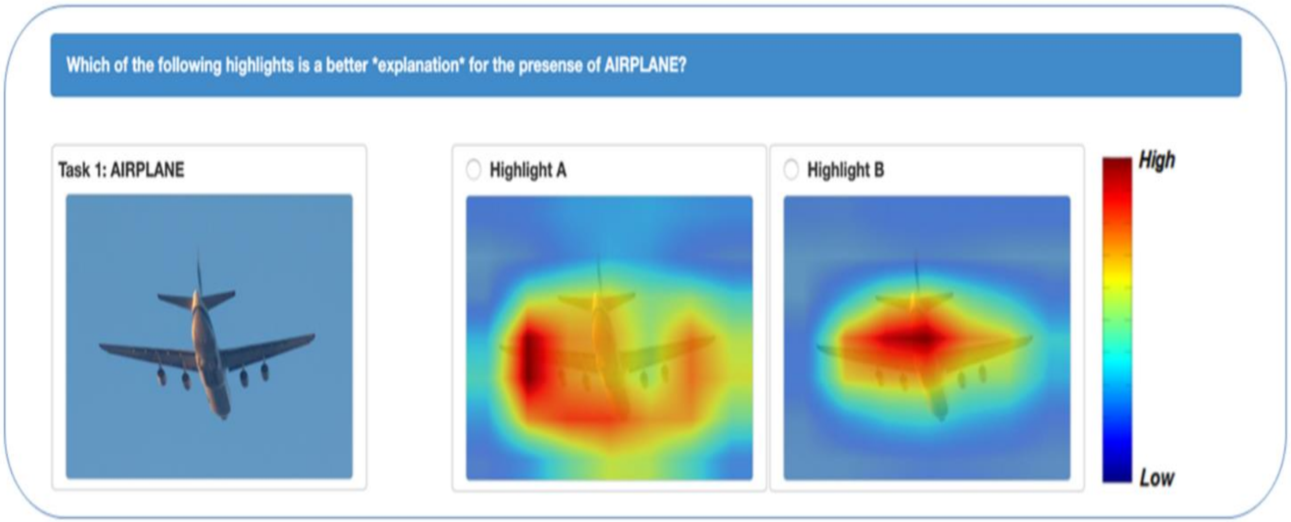
- The number of unseen MSCOCO images, among the 16K validation set, where the model is able to provide an accurate explanation for, among the correctly classified ones during training.

- 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.



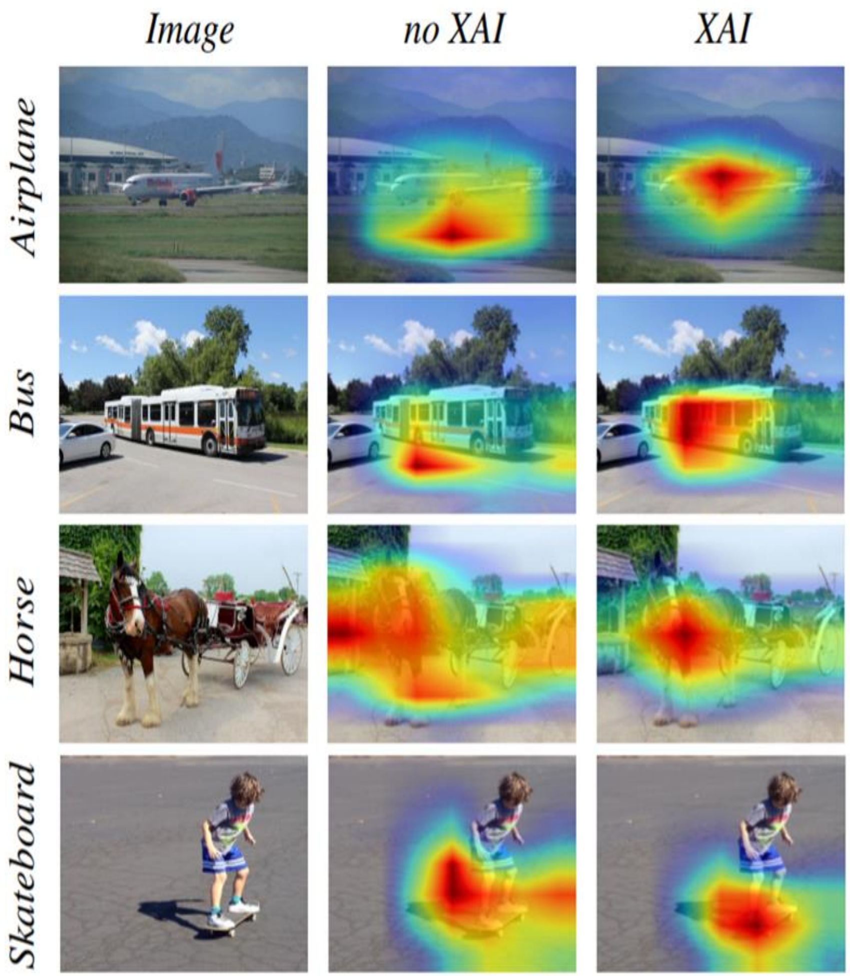
Explainability Results: Quantitative *[Human Judgment]*

- 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.



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.

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.