Towards Adaptive and Explainable Artificial Intelligence

Prof. Trevor Darrell
UC Berkeley

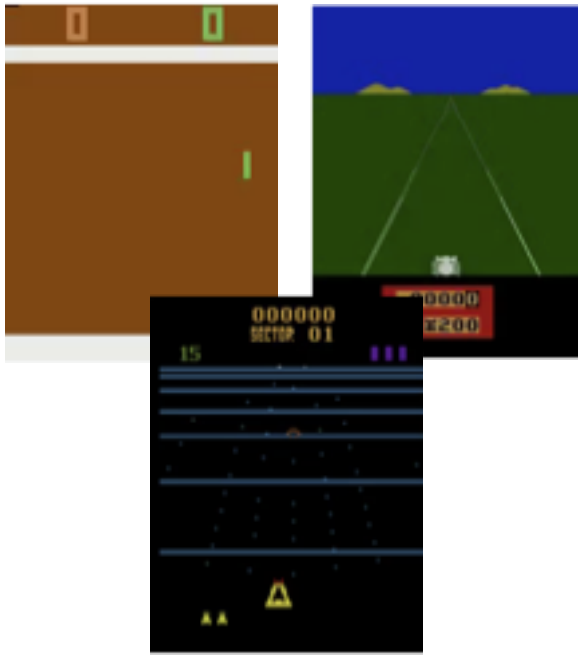
AI seems to come of age

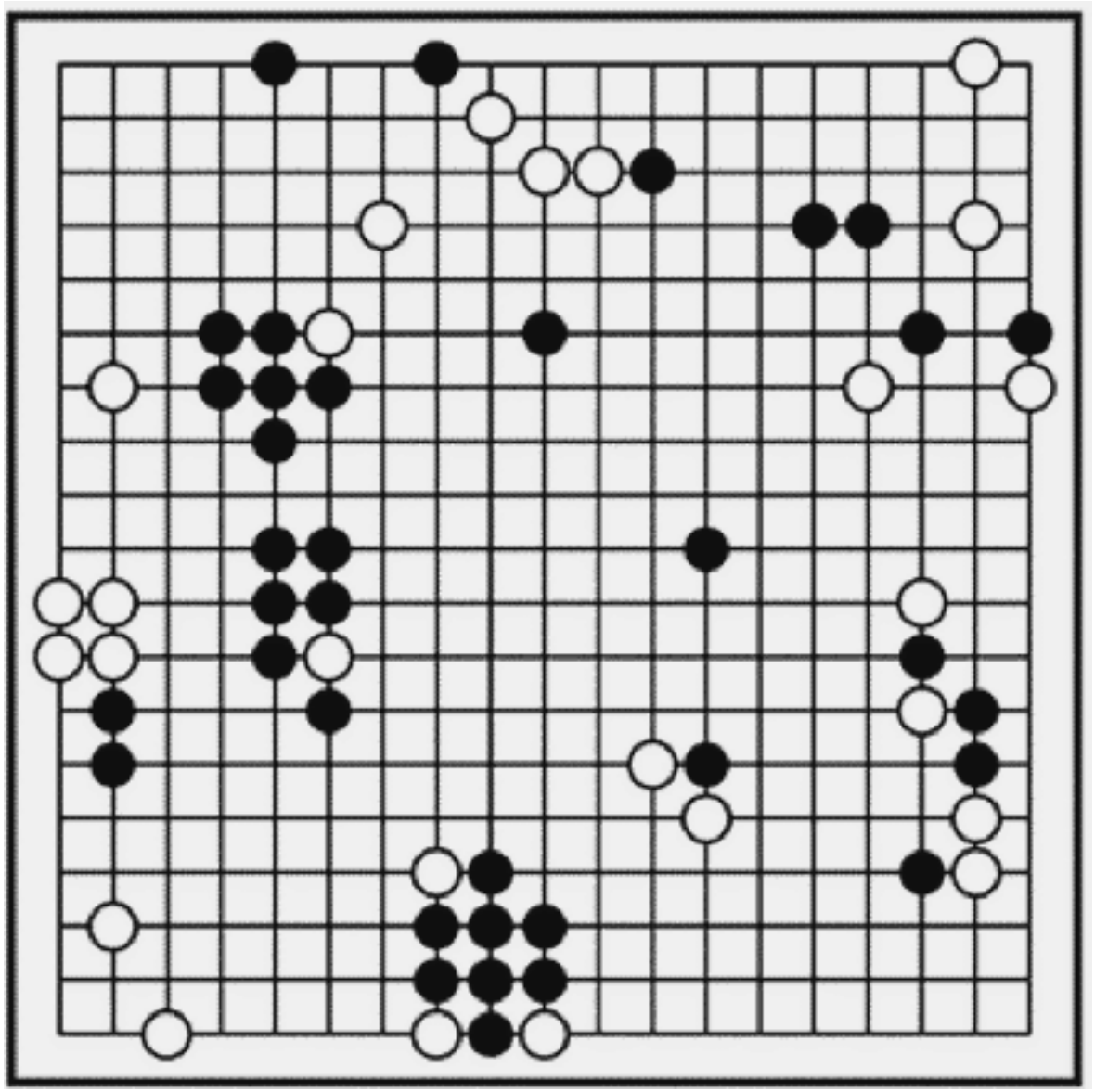
massive data sets

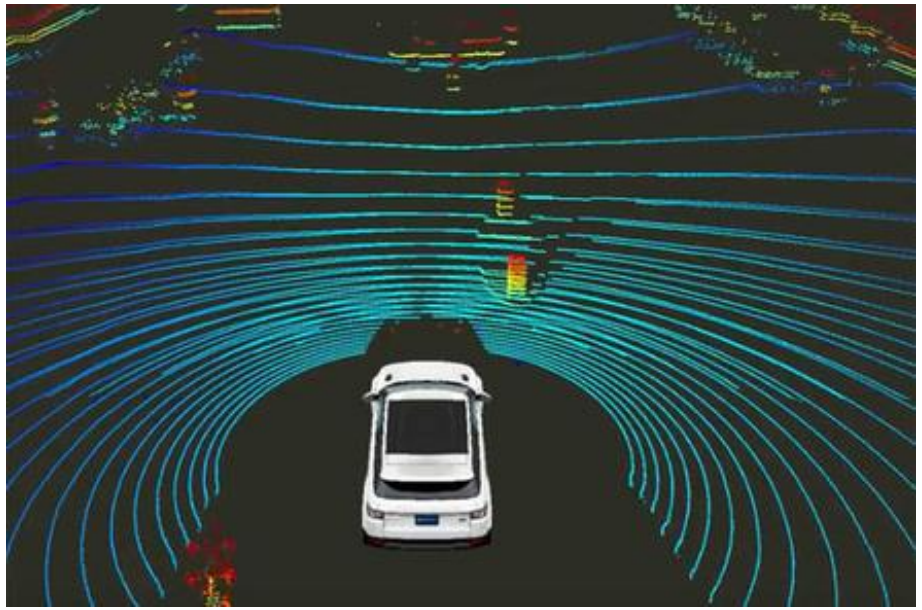
new computational capabilities

new sensors and actuators capbilities

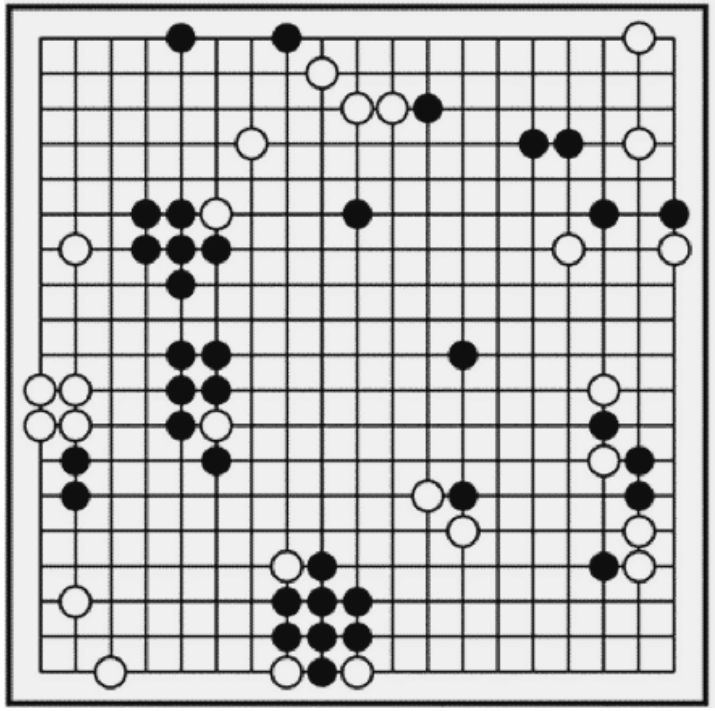
new algorithms







AI seems to come of age…..



simulation

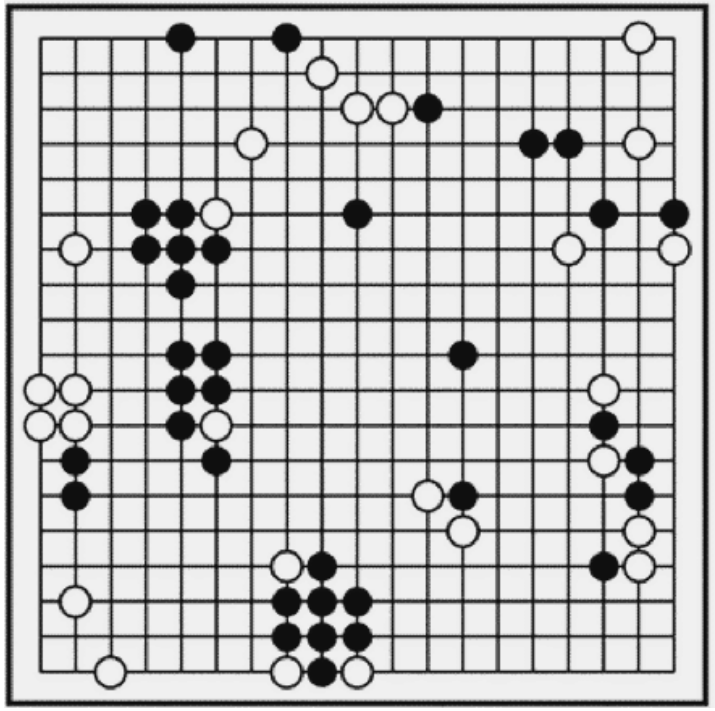
fully observable

short horizon

single agent

game scenarios

… but we've just scratched the surface



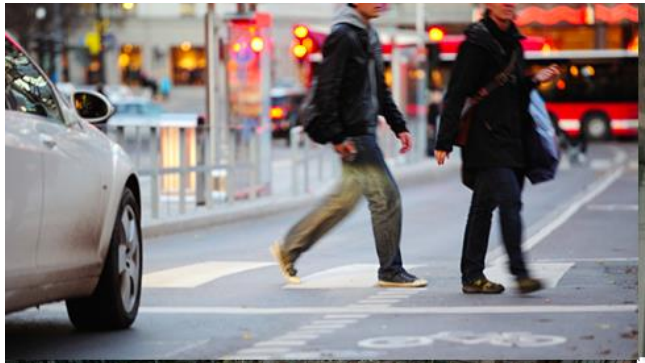
simulation

short horizon

fully observable

single agent

game scenarios



real world

extended timescale

uncertainty

massively distributed

really helping **people**

Limitations of Contemporary Deep Methods

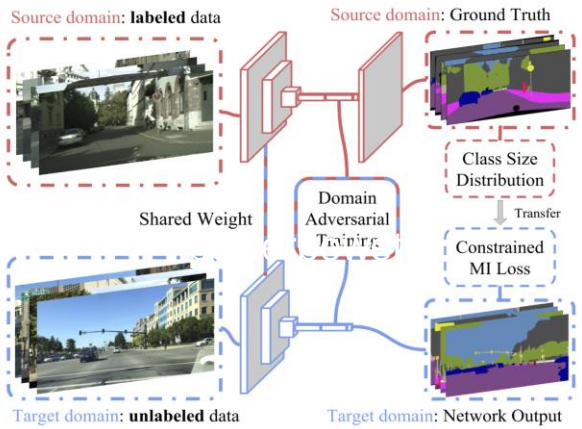
- Generalization to new domains
- Learning new concepts from few examples
- Estimating uncertainty
- Avoiding bias, explaining and trusting models
- Learning tasks without clear supervision
- …

These are the focus of current research by many / most faculty in BAIR, and especially in my lab….

Today I'll touch on **adaptation and explanation.** (And maybe **fairness).**

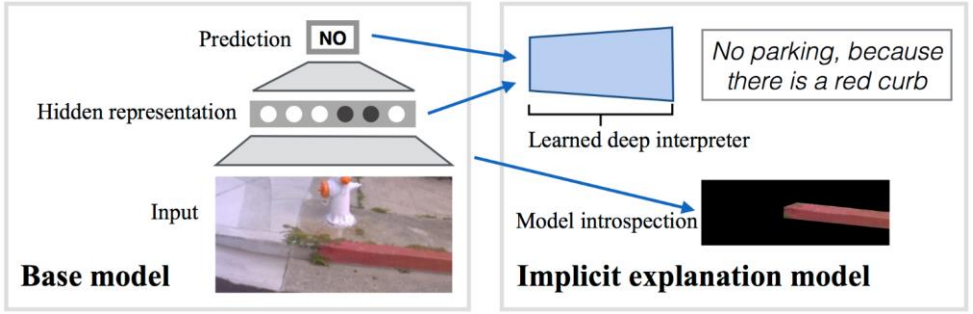
Deep *learning to learn*

Major current research theme in BAIR: beyond supervised learning

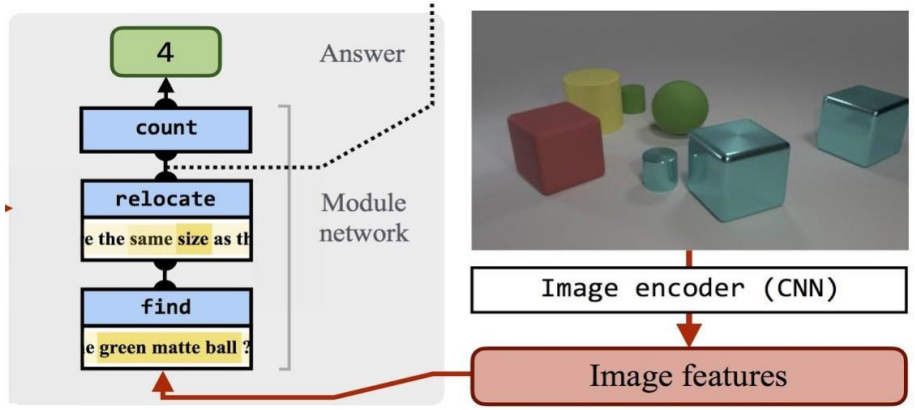


Adapting to new domains and tasks

Self-supervised / Curious learning



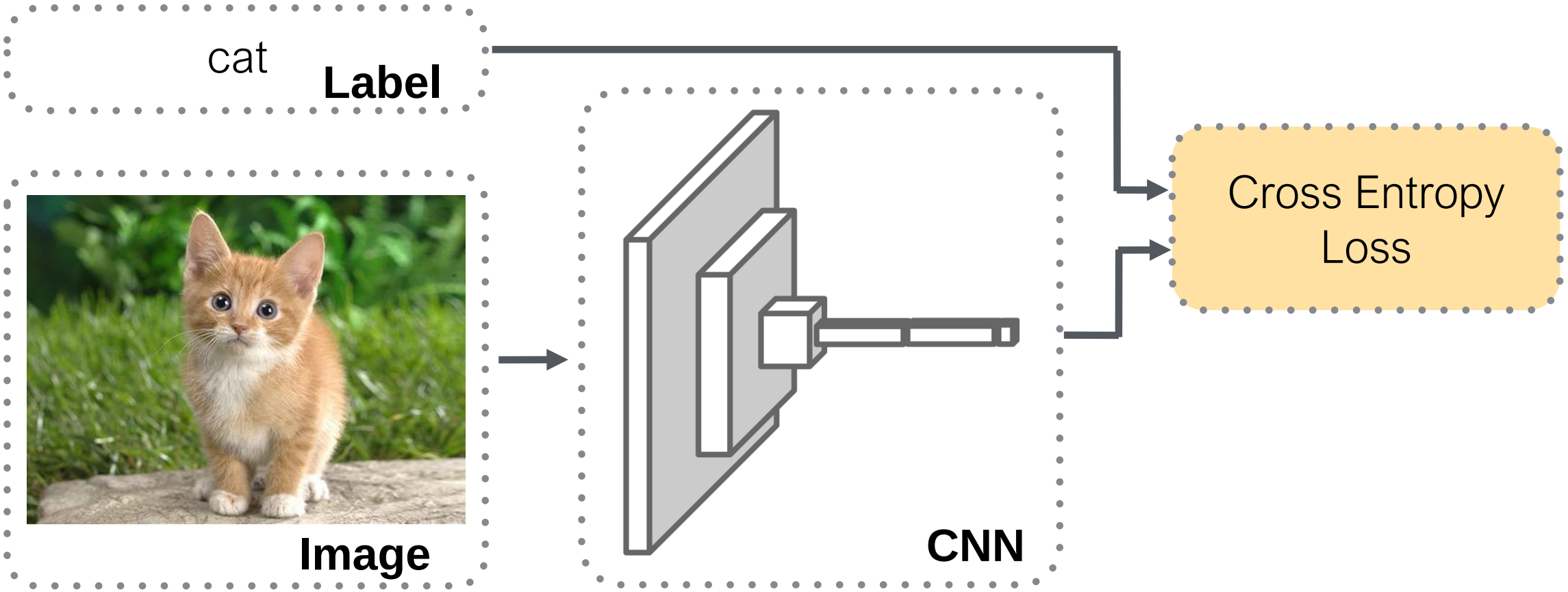
Learning to explain deep learning models



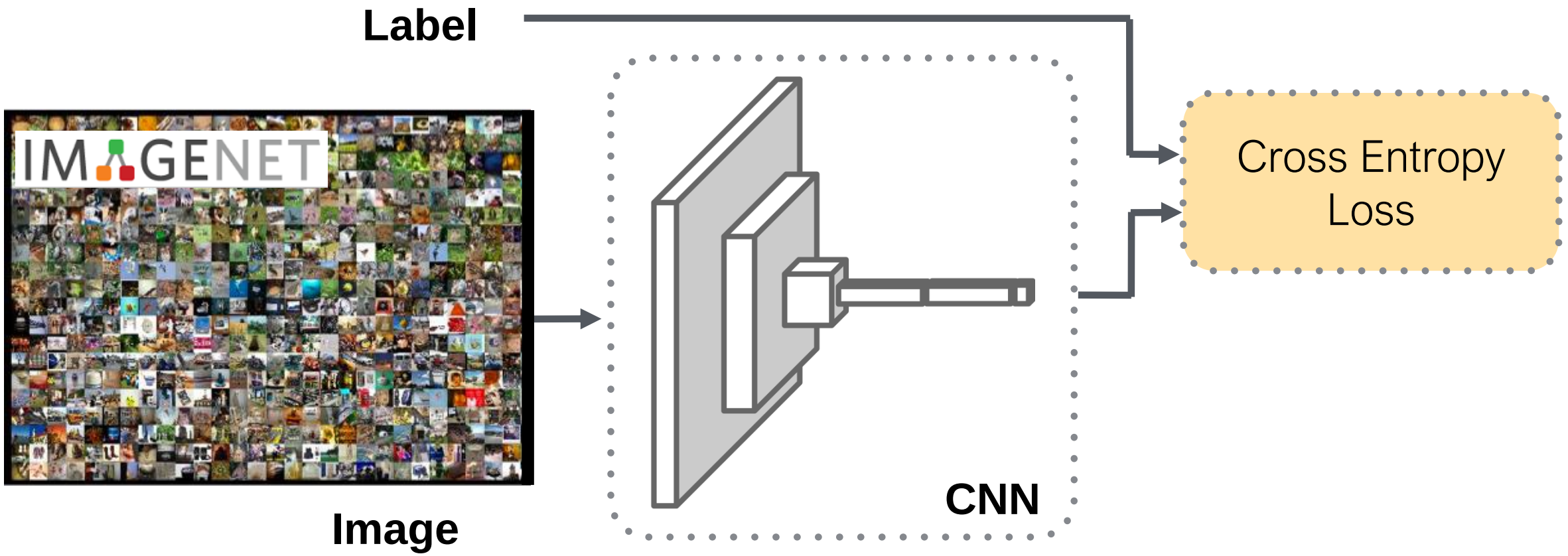
Reasoning over network structures

Adaptation

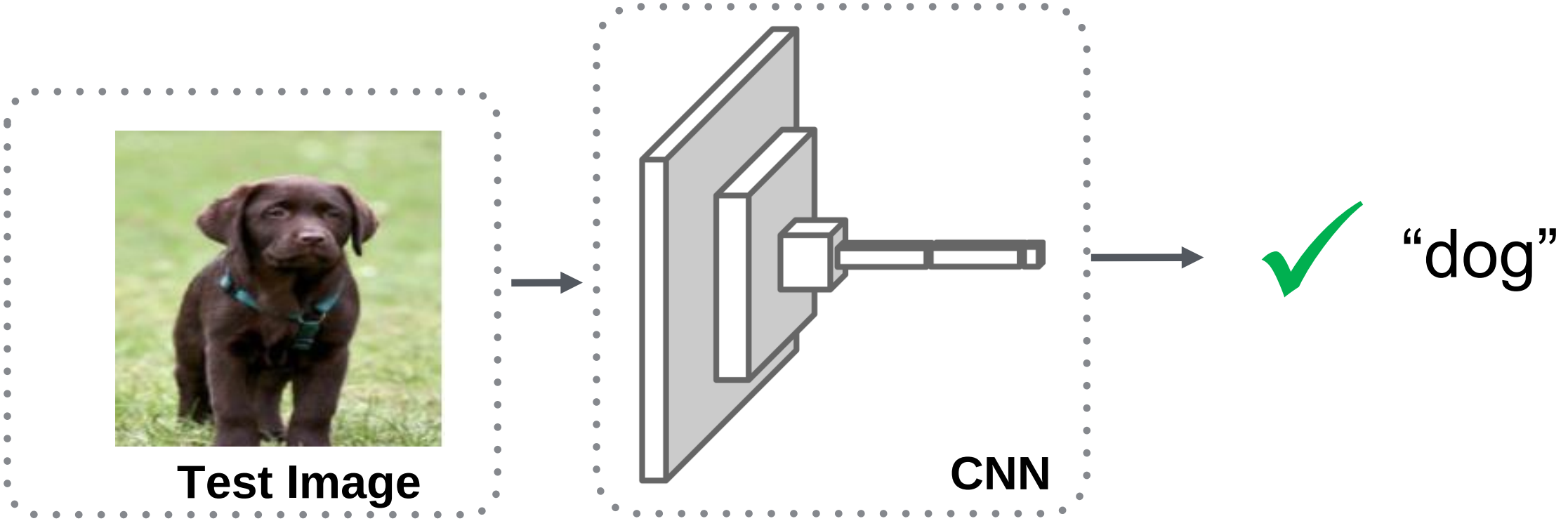
Supervised Learning



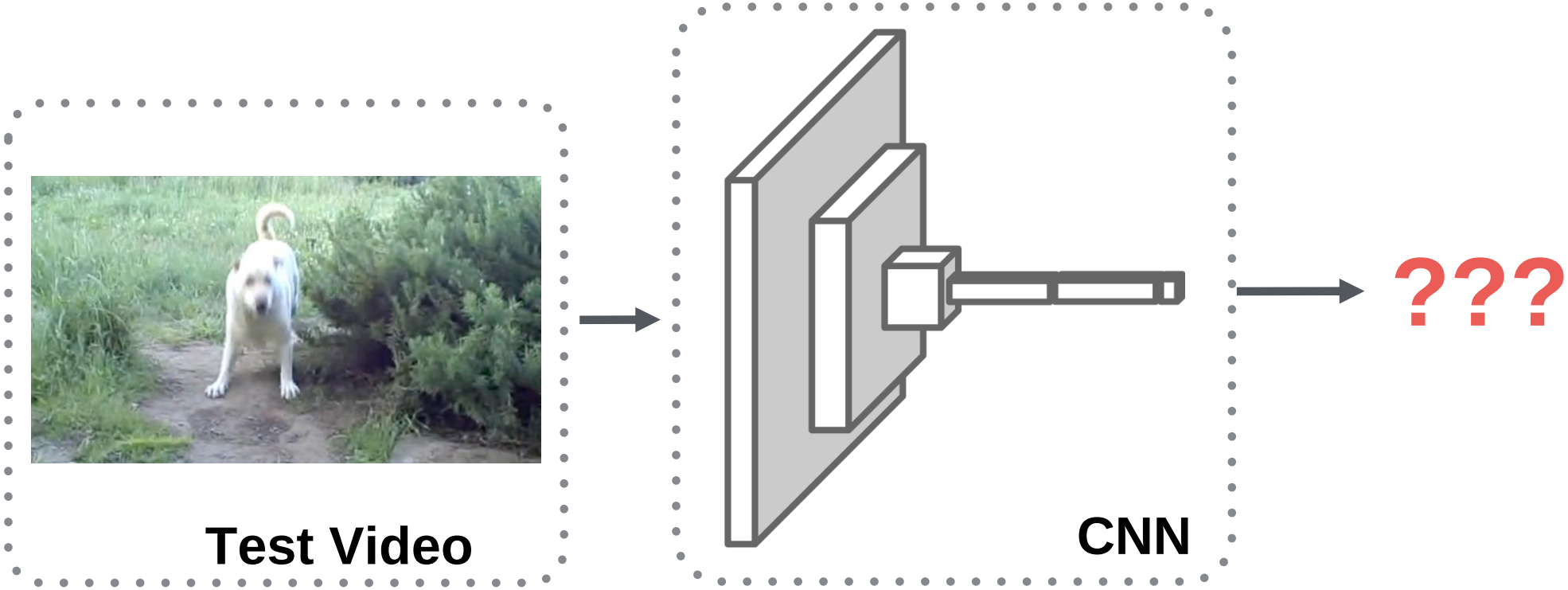
Supervised Learning



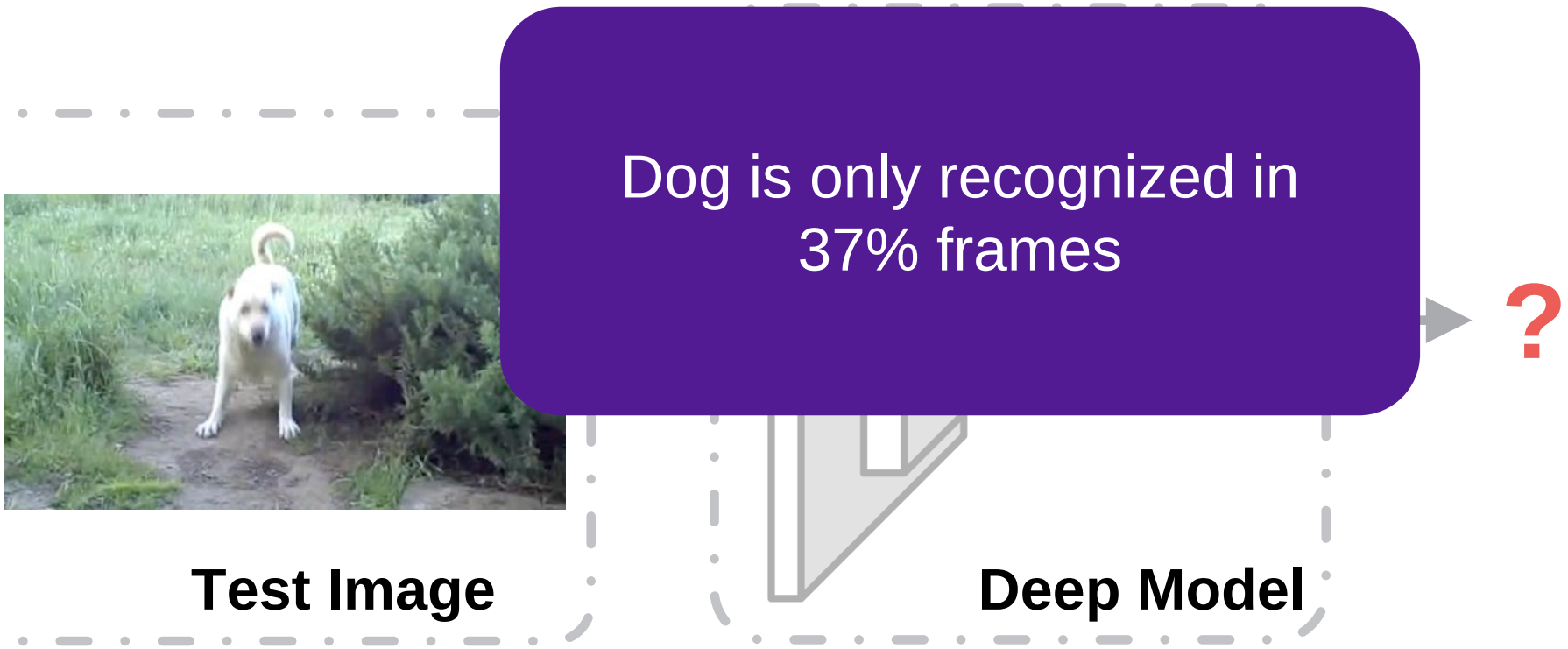
Supervised Learning



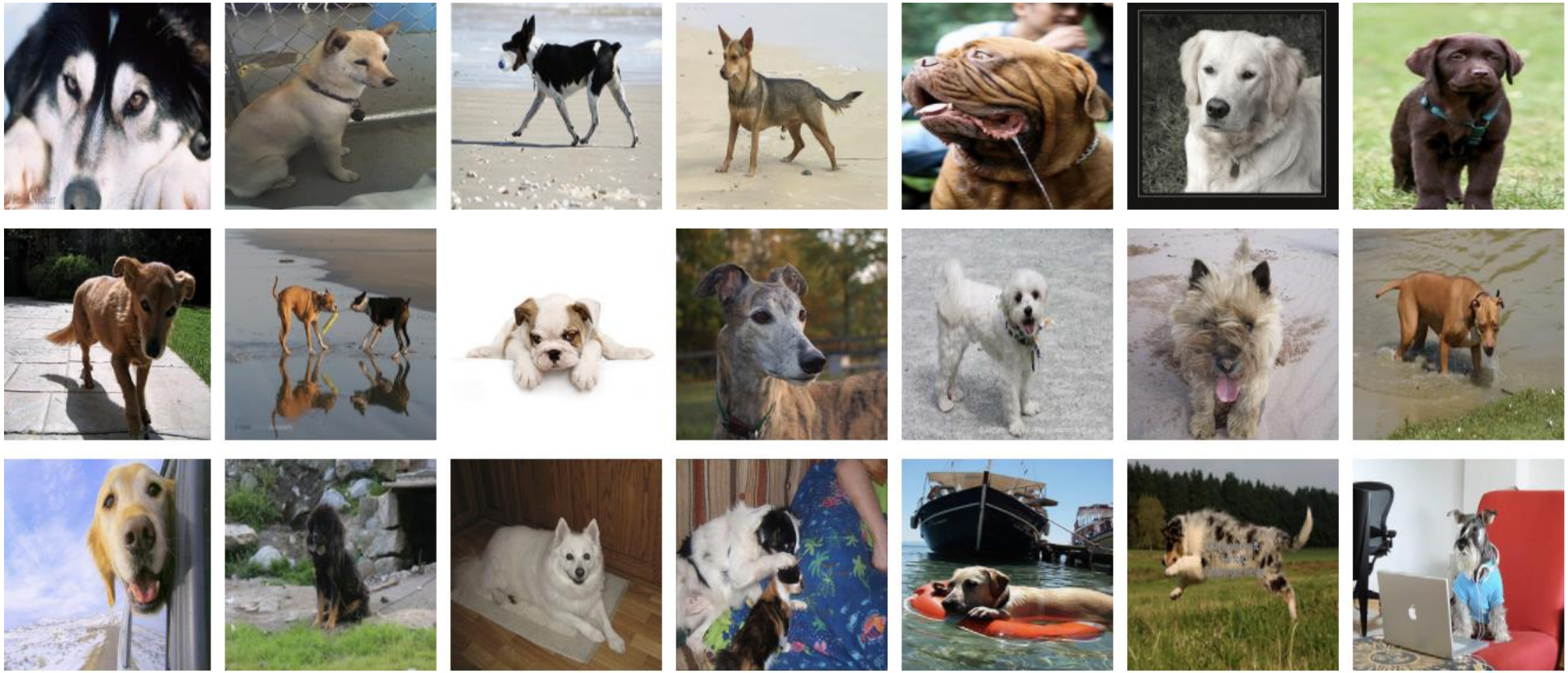
Poor Performance on New Domains



Dataset Bias



Dataset Bias



IMAGENET

Dataset Bias

Low resolution

Motion Blur

Pose Variety

How can we learn to generalize?

Domain Adaptation: Train on Source Test on Target

Source Domain $\sim P_S(X_S, Y_S)$

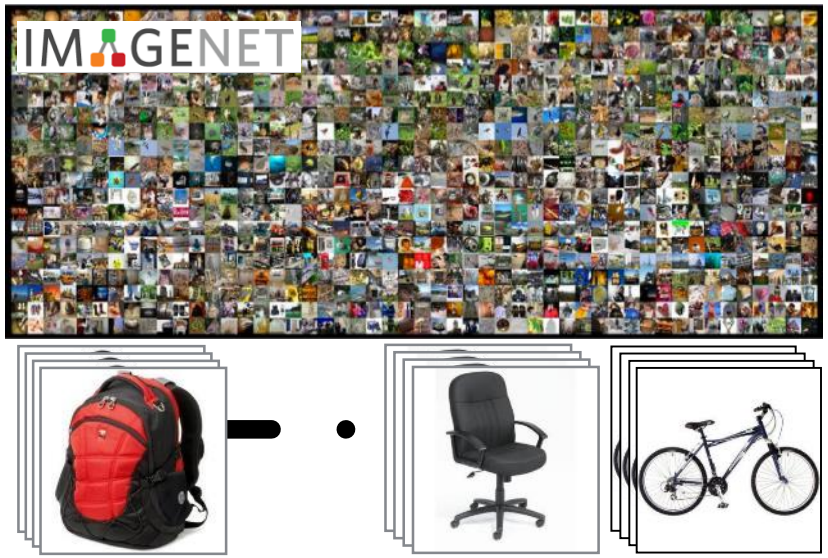
lots of **labeled** data

$\neq$

Target Domain $\sim P_T(X_T, Y_T)$

unlabeled or limited labels

Domain Adaptation: Train on Source Test on Target



≠

Source Domain $\sim P_S(X_S, Y_S)$

lots of **labeled** data

Target Domain $\sim P_T(X_T, Y_T)$

unlabeled or limited labels

Domain Adaptation: Train on Source Test on Target

≠

Source Domain $\sim P_S(X_S, Y_S)$

lots of **labeled** data

Target Domain $\sim P_T(X_T, Y_T)$

unlabeled or limited labels

Domain Adaptation: Train on Source Test on Target

≠

Source Domain $\sim P_S(X_S, Y_S)$

lots of **labeled** data

Target Domain $\sim P_T(X_T, Y_T)$

unlabeled or limited labels

Domain Adaptation: Train on Source Test on Target

Adapt

Source Domain $\sim P_S(X_S, Y_S)$

lots of **labeled** data

Target Domain $\sim P_T(X_T, Y_T)$

unlabeled or limited labels

Domain Adaptation Paradigms

- Feature Augmentation – add training data transformed based on knowledge of domain.

  (A great and old idea; used at least since face detectors of 1990's…but does presume prior external knowledge of what domain shift/variation is…)

- Bootstrapping, a.k.a. "self-ensembling", ... – take high confidence predictions of source-only model and add them to training data, iterate….

  (Also classic….but only works for small shifts….risk of concept drift; usually applied with a constraint to not diverge too far from source model.)

- **Distribution alignment or transformation: domain adversarial learning / domain confusion using GAN-like models….**

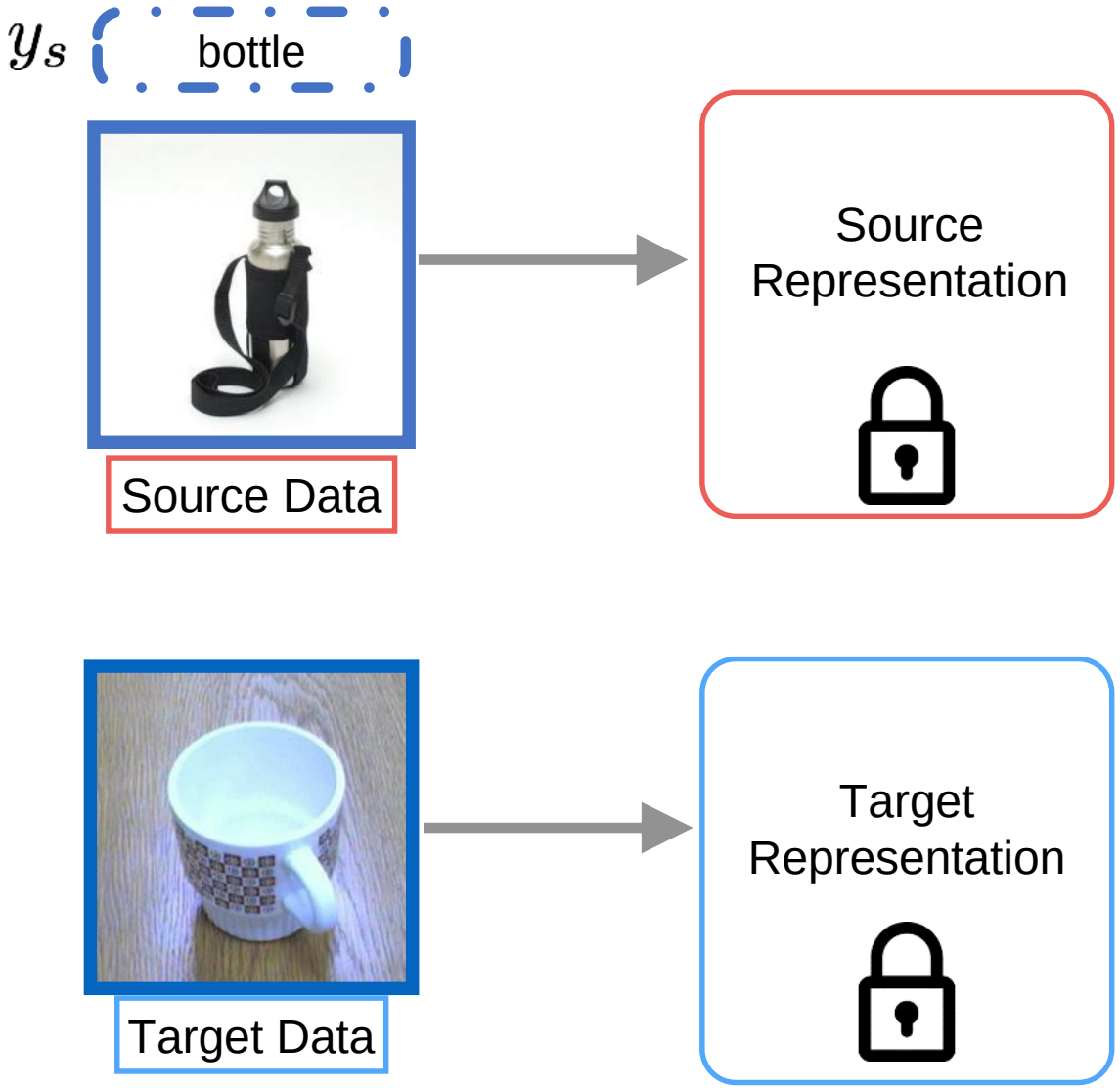
Classic Domain Adaptation

y_s bottle

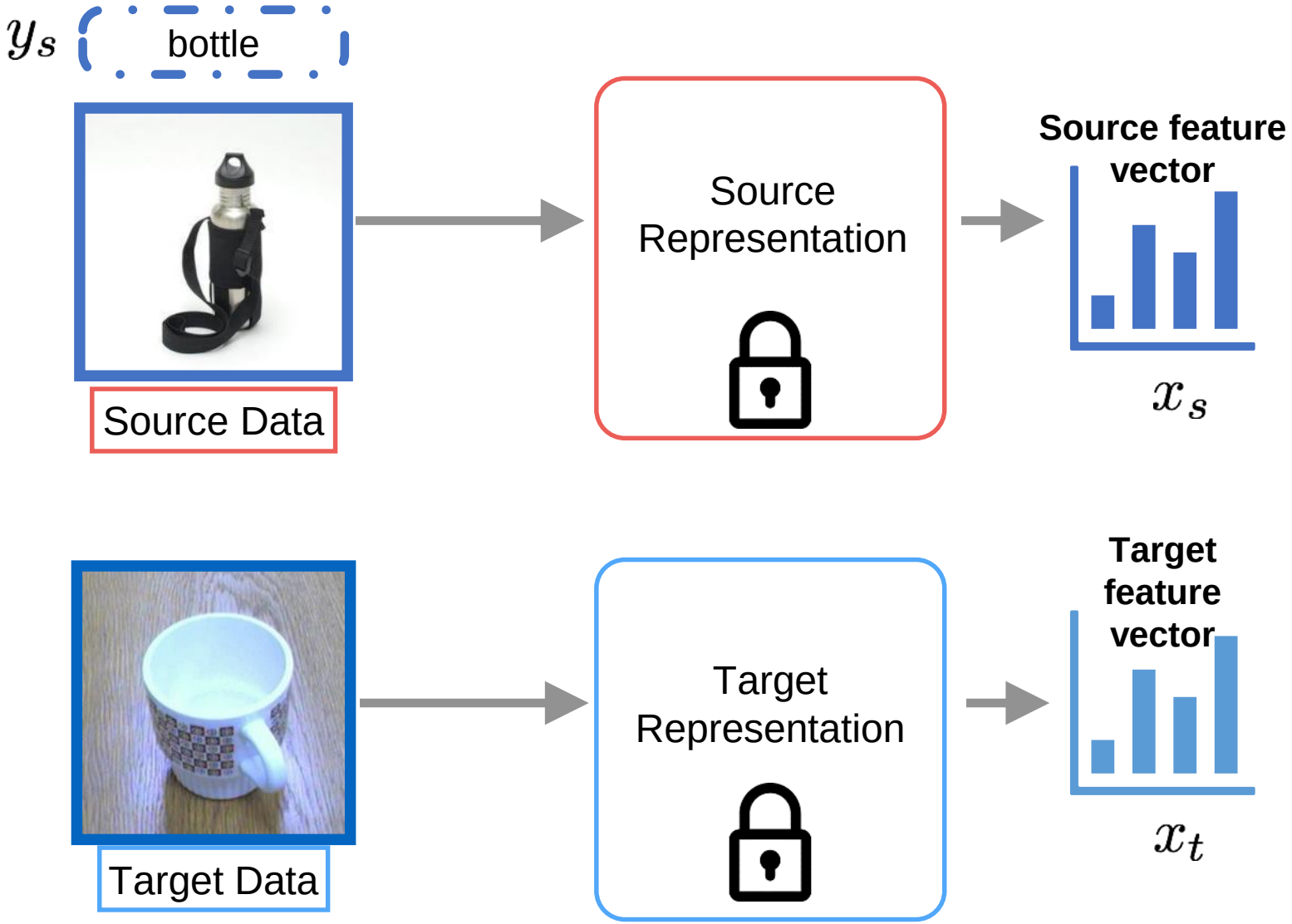
Source Data

Target Data

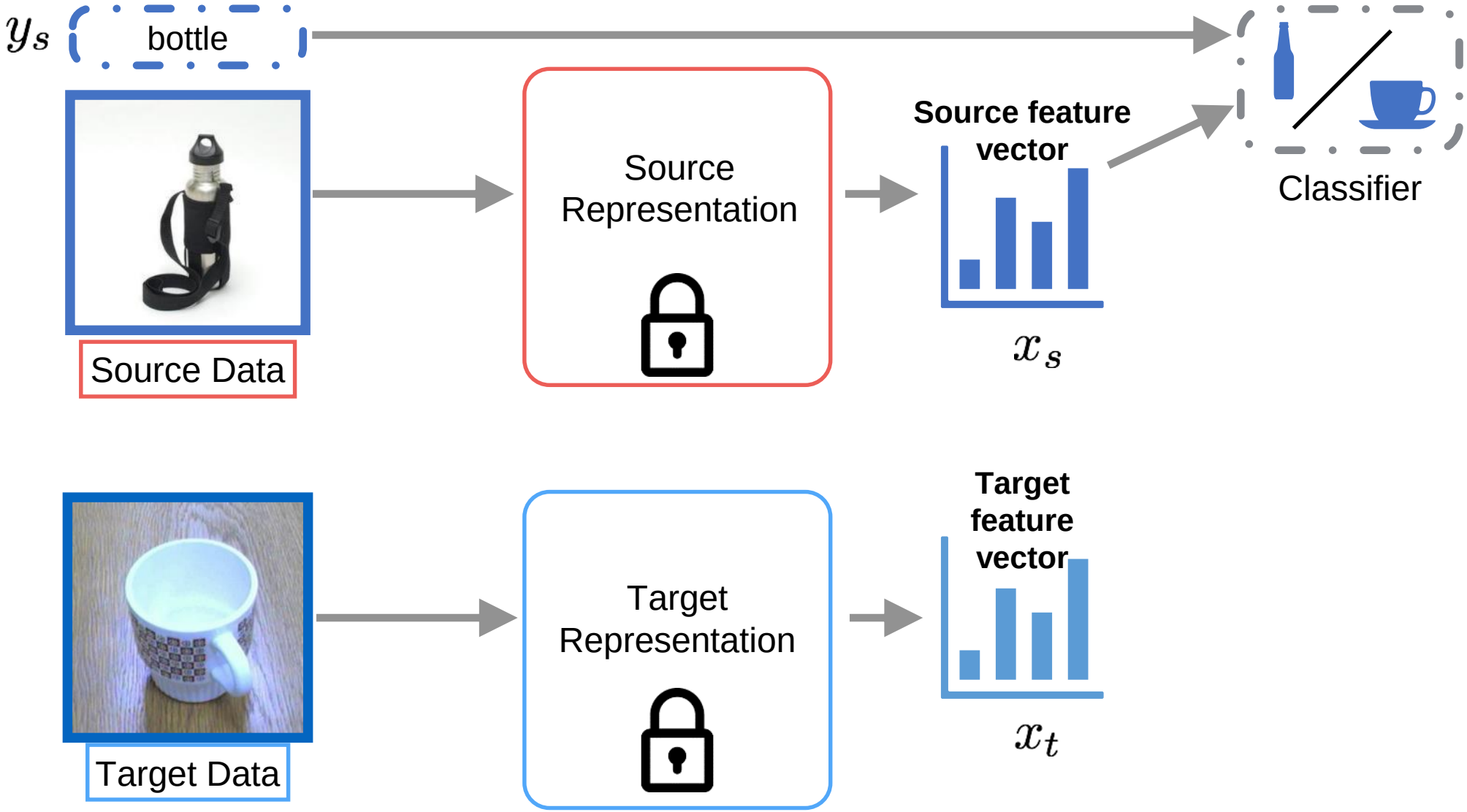
Classic Domain Adaptation



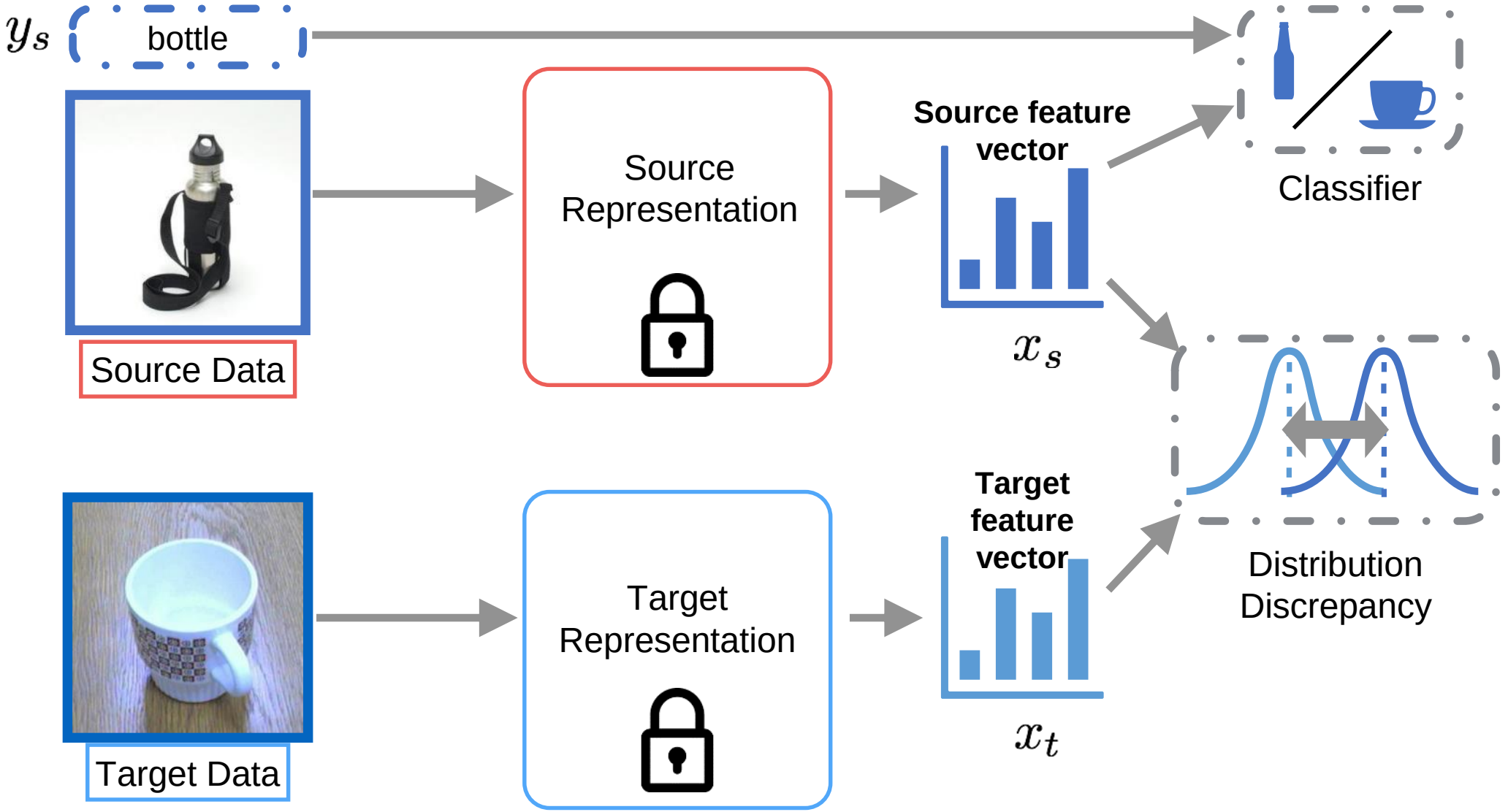
Classic Domain Adaptation



Classic Domain Adaptation



Classic Domain Adaptation

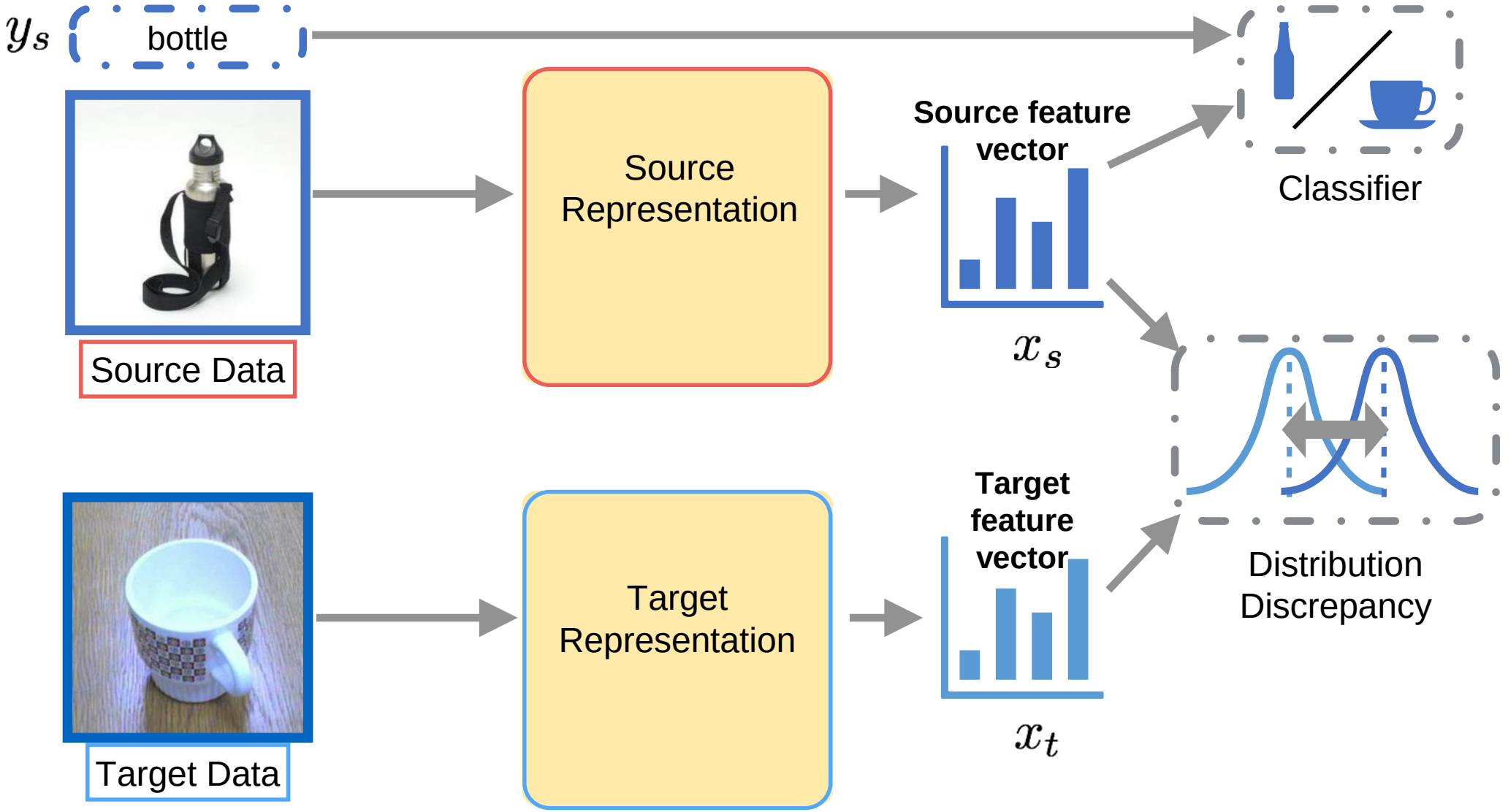


Saenko et al. ECCV 10. Torralba & Efros CVPR 11. Gong et al. CVPR 12. Fernando et al. ICCV 13. **Hoffman** et al. ICLR 13

Idea:

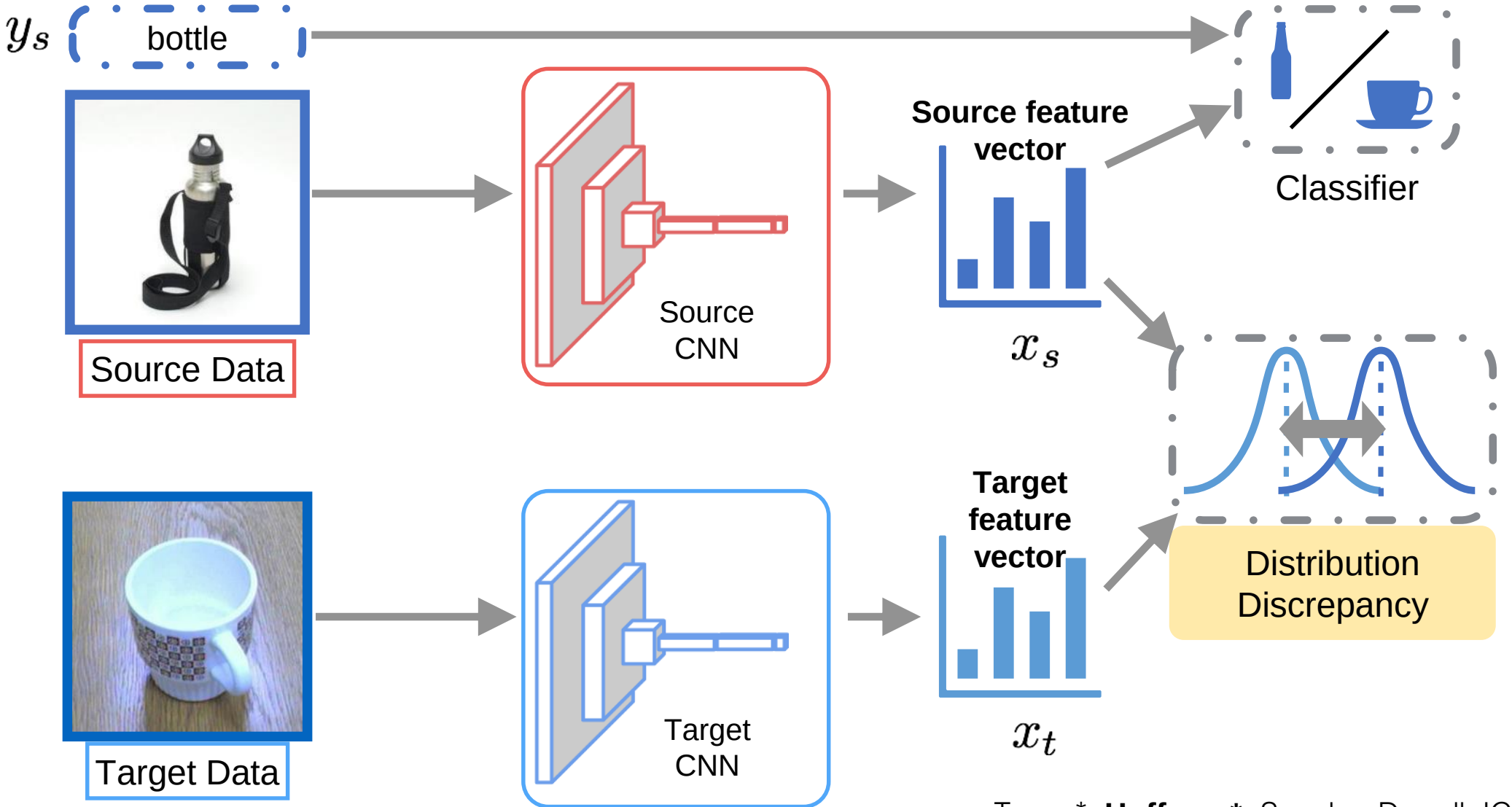
Learn a representation that cannot distinguish domains

Deep Domain Adaptation



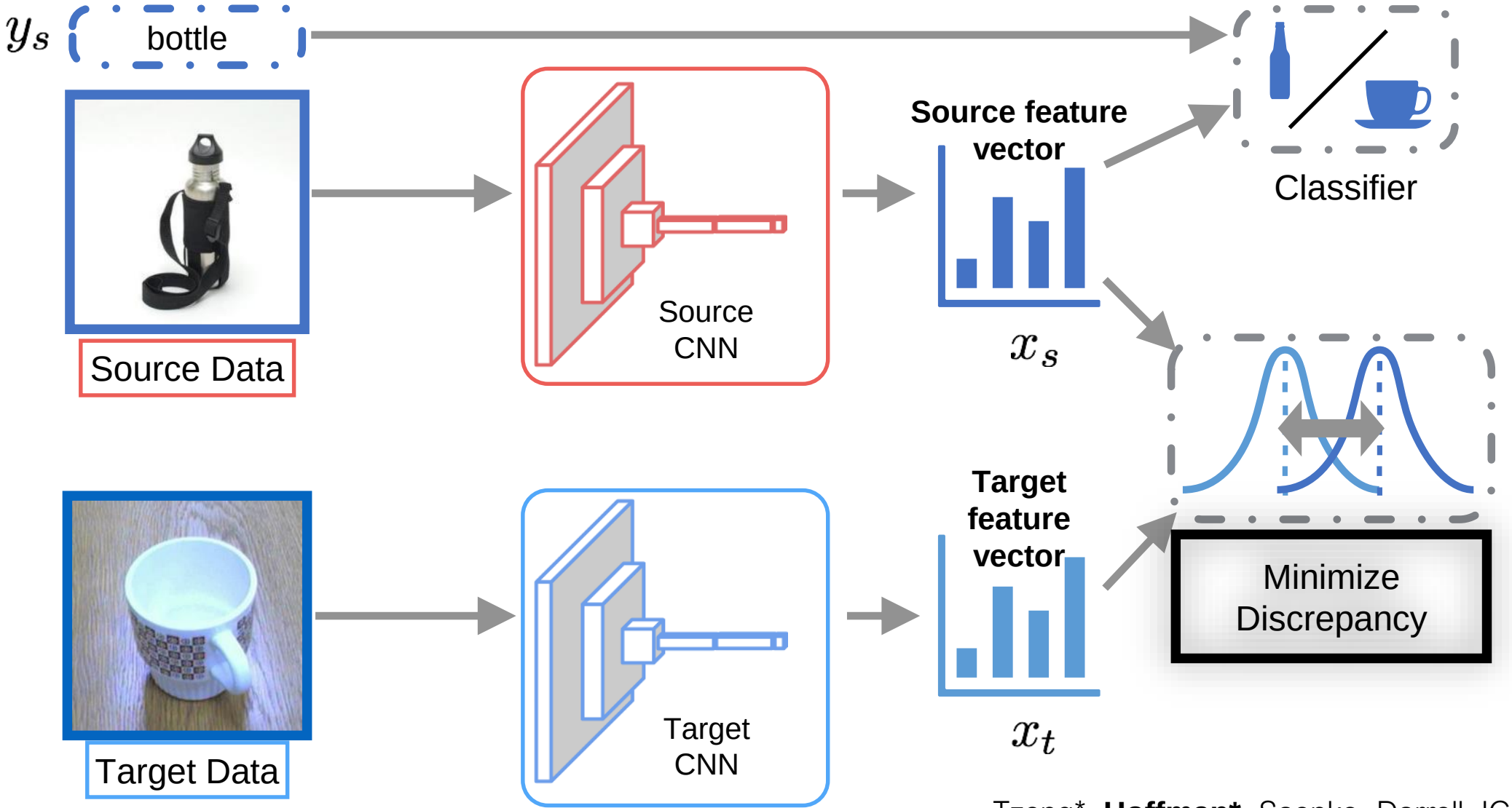
Saenko et al. ECCV 10. Torralba & Efros CVPR 11. Gong et al. CVPR 12. Fernando et al. ICCV 13. **Hoffman** et al. ICLR 13

Deep Domain Adaptation



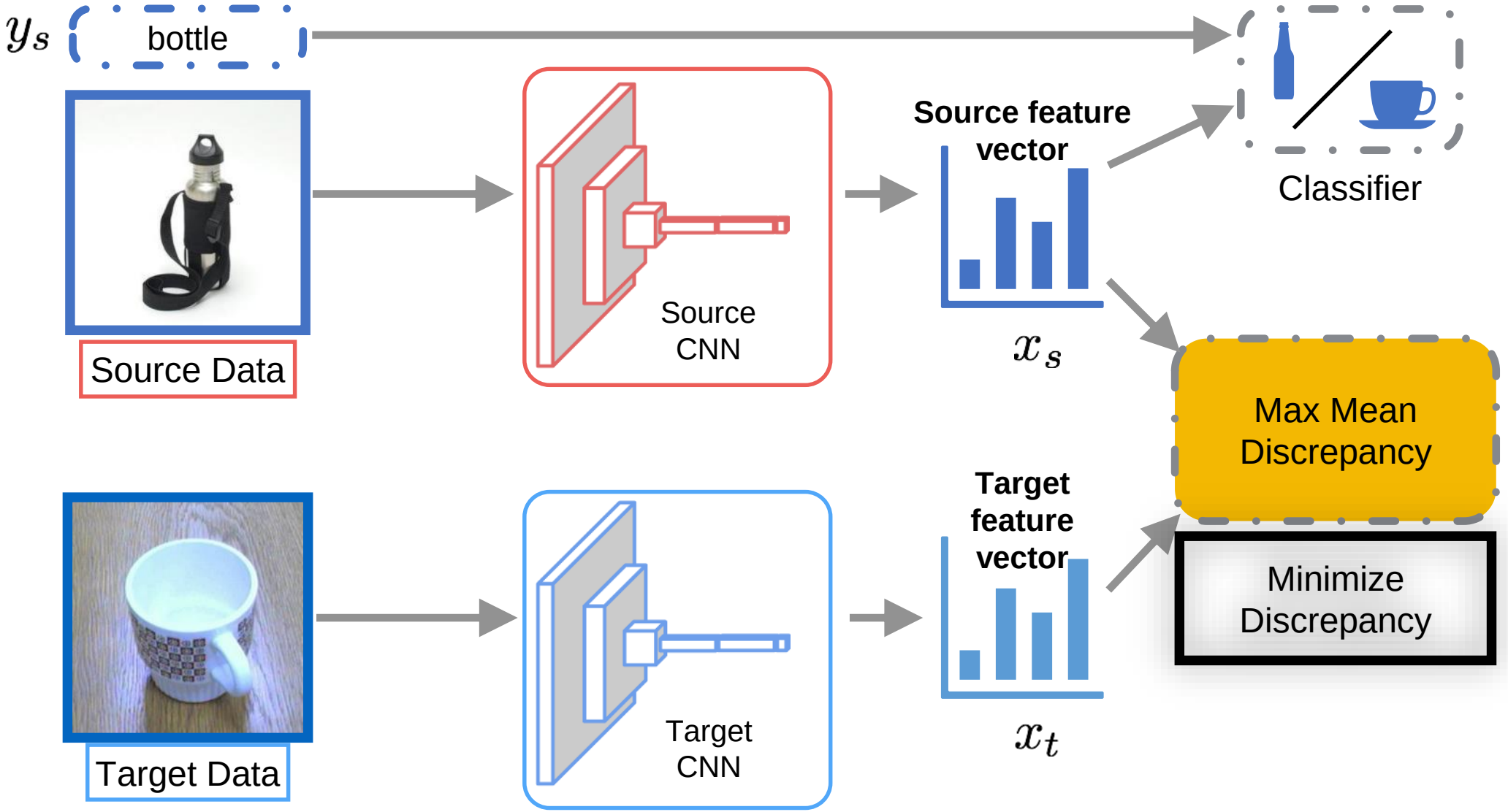
Tzeng*, **Hoffman***, Saenko, Darrell, ICCV 2015.
Tzeng, **Hoffman**, Saenko, Darrell. *CVPR* 2017.

Deep Domain Adaptation



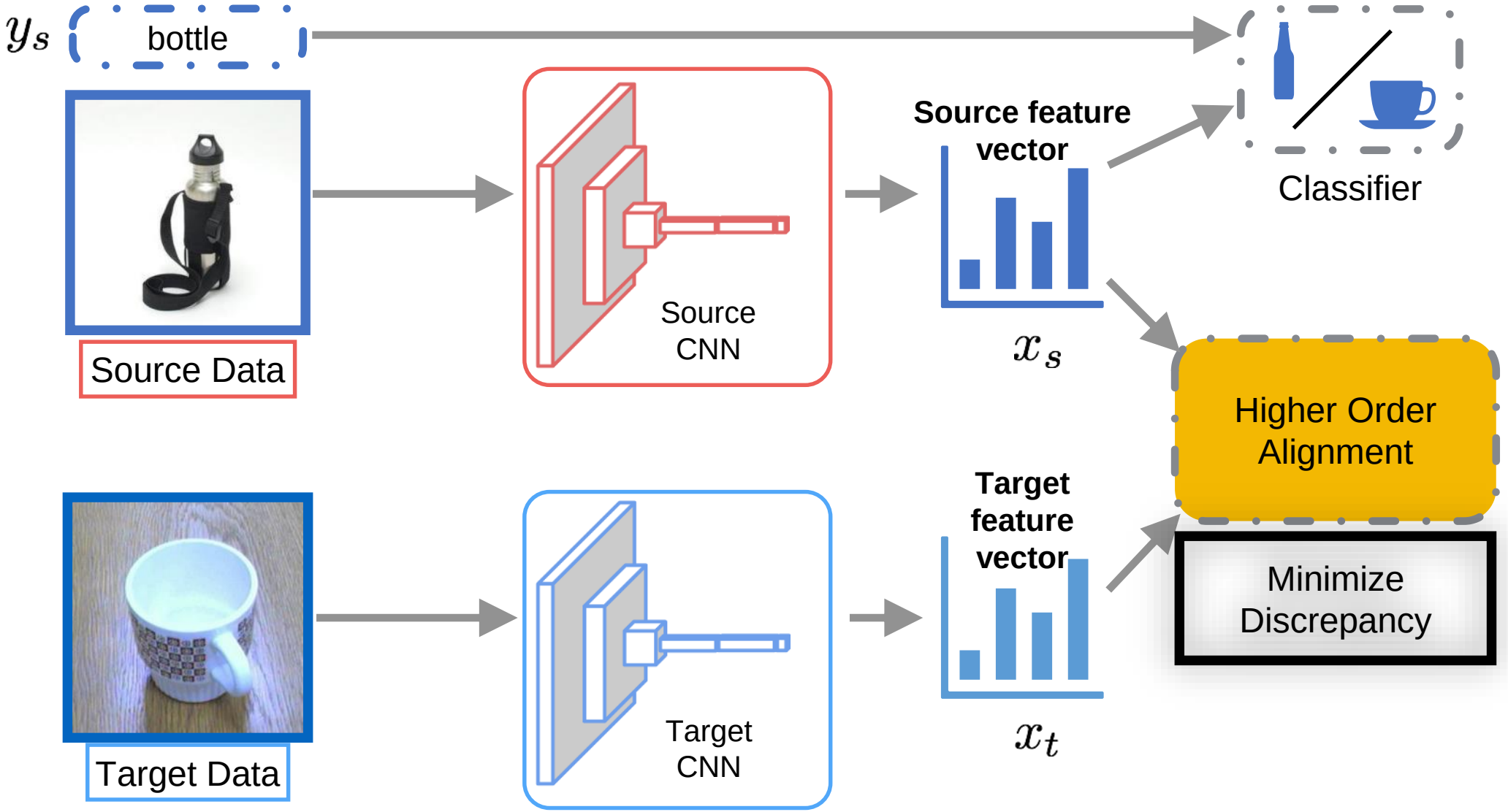
Tzeng*, **Hoffman***, Saenko, Darrell, ICCV 2015.
Tzeng, **Hoffman**, Saenko, Darrell. *CVPR* 2017.

Deep Domain Adaptation



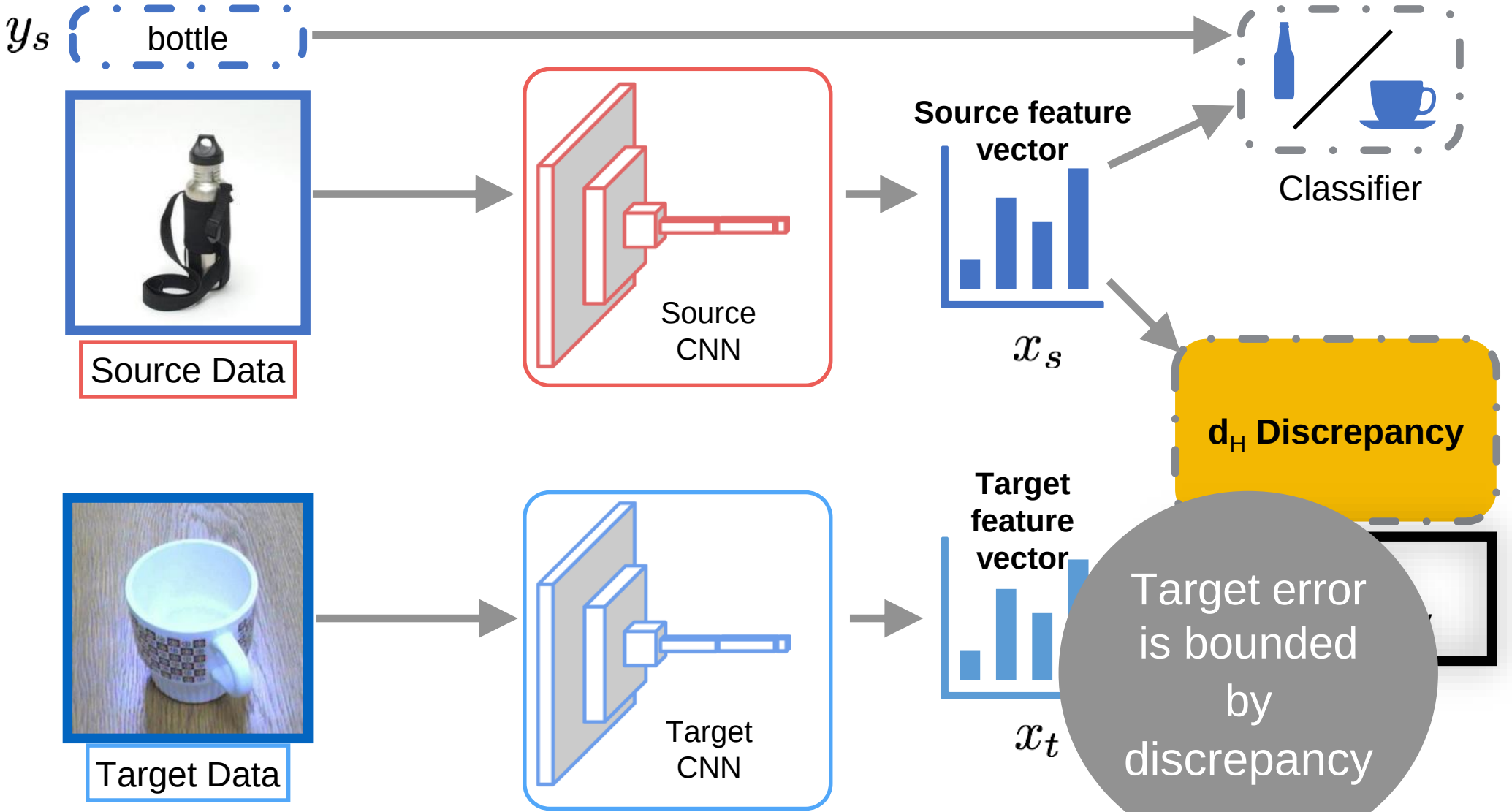
Tzeng, **Hoffman**, Zhang, Saenko, Darrell. *arXiv 2014.*

Deep Domain Adaptation



Long, Cao, Wang, Jordan, ICML 2015. Sun, Saenko, ECCV 2016.

Deep Domain Adaptation



Tzeng, **Hoffman**, Saenko, Darrell. *CVPR* 2017.

Discrepancy Between Source and Target

$$\mathcal{A} \subseteq 2^{|X|}$$

$$d_{\mathcal{H}}(D_S, D_T) = 2 \sup_{A \in \mathcal{A}_{\mathcal{H}}} |Pr_{D_S}[A] - Pr_{D_T}[A]|$$

discrepancy

Ben David et al., NIPS 2006
Blitzer et al. NIPS 2007

Discrepancy Between Source and Target

$$\mathcal{A} \subseteq 2^{|X|} \qquad d_{\mathcal{H}}(D_S, D_T) = 2 \sup_{A \in \mathcal{A}_{\mathcal{H}}} |Pr_{D_S}[A] - Pr_{D_T}[A]|$$

$$\epsilon_T(h) \leq \hat{\epsilon}_S(h) + d_{\mathcal{H}}(D_S, D_T) + \sqrt{\frac{4}{M}\left(d \log \frac{2eM}{d} + \log \frac{4}{\delta}\right)} + \lambda$$

target error

Ben David et al., NIPS 2006
Blitzer et al. NIPS 2007

Discrepancy Between Source and Target

$$\mathcal{A} \subseteq 2^{|X|} \qquad d_{\mathcal{H}}(D_S, D_T) = 2 \sup_{A \in \mathcal{A}_{\mathcal{H}}} |Pr_{D_S}[A] - Pr_{D_T}[A]|$$

$$\epsilon_T(h) \leq \hat{\epsilon}_S(h) + d_{\mathcal{H}}(D_S, D_T) + \sqrt{\frac{4}{M}\left(d \log \frac{2eM}{d} + \log \frac{4}{\delta}\right)} + \lambda$$

target error

source error

Ben David et al., NIPS 2006
Blitzer et al. NIPS 2007

Discrepancy Between Source and Target

$$\mathcal{A} \subseteq 2^{|X|} \qquad d_{\mathcal{H}}(D_S, D_T) = 2 \sup_{A \in \mathcal{A}_{\mathcal{H}}} |Pr_{D_S}[A] - Pr_{D_T}[A]|$$

$$\epsilon_T(h) \leq \hat{\epsilon}_S(h) + d_{\mathcal{H}}(D_S, D_T) + \sqrt{\frac{4}{M}\left(d \log \frac{2eM}{d} + \log \frac{4}{\delta}\right)} + \lambda$$

target error

source error

discrepancy

Ben David et al., NIPS 2006
Blitzer et al. NIPS 2007

Discrepancy Between Source and Target

$$\mathcal{A} \subseteq 2^{|X|} \qquad d_{\mathcal{H}}(D_S, D_T) = 2 \sup_{A \in \mathcal{A}_{\mathcal{H}}} |Pr_{D_S}[A] - Pr_{D_T}[A]|$$

$$\epsilon_T(h) \leq \hat{\epsilon}_S(h) + d_{\mathcal{H}}(D_S, D_T) + \sqrt{\frac{4}{M}\left(d \log \frac{2eM}{d} + \log \frac{4}{\delta}\right)} + \lambda$$

target error

source error

discrepancy

error of best model ++

Ben David et al., NIPS 2006
Blitzer et al. NIPS 2007

Discrepancy Between Source and Target

$$\mathcal{A} \subseteq 2^{|X|} \qquad d_{\mathcal{H}}(D_S, D_T) = 2 \sup_{A \in \mathcal{A}_{\mathcal{H}}} |Pr_{D_S}[A] - Pr_{D_T}[A]|$$

$$\epsilon_T(h) \leq \hat{\epsilon}_S(h) + d_{\mathcal{H}}(D_S, D_T) + \sqrt{\frac{4}{M}\left(d \log \frac{2eM}{d} + \log \frac{4}{\delta}\right)} + \lambda$$

target error

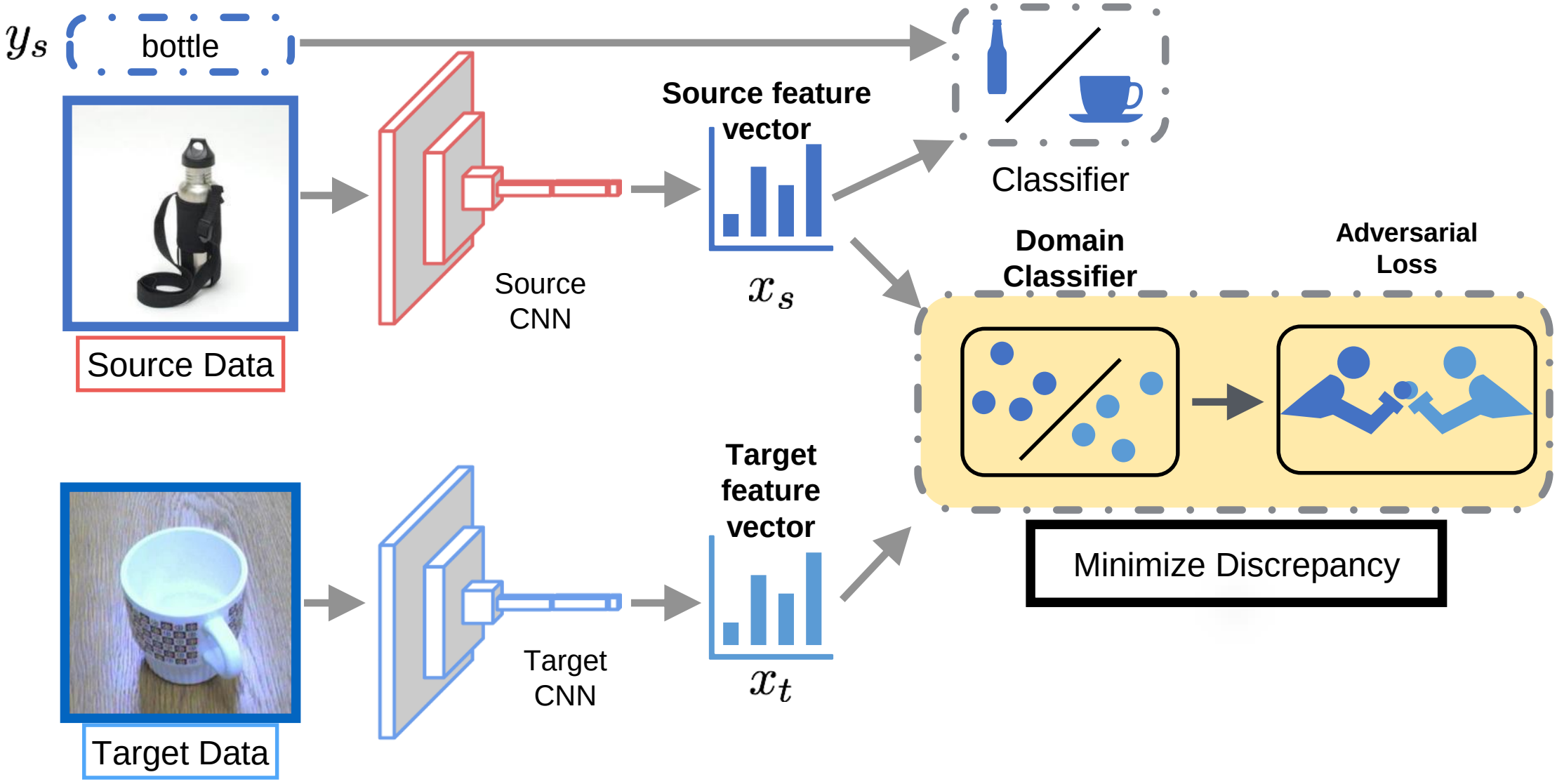
source error

discrepancy

error of best model ++

Ben David et al., NIPS 2006
Blitzer et al. NIPS 2007

Domain Adversarial Adaptation



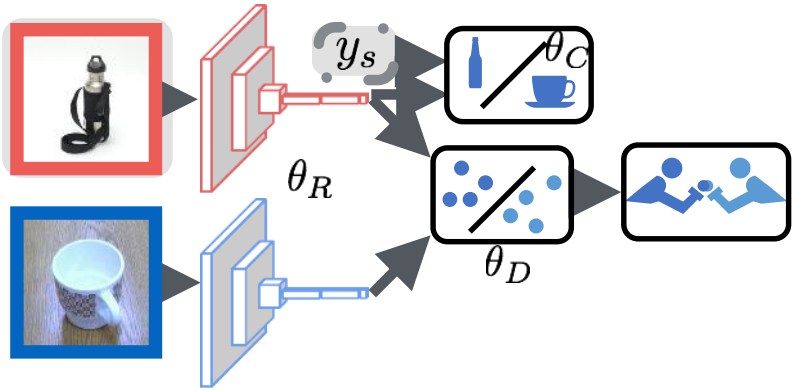
Tzeng, **Hoffman**, Saenko, Darrell. *CVPR* 2017.

Domain Adversarial Optimization

Tzeng, **Hoffman**, Saenko, Darrell. *CVPR* 2017.

Domain Adversarial Optimization

$$\min_{\theta_C, \theta_R} \mathcal{L}_{\text{cls}}(\mathbf{X}_s, \mathbf{Y}_s; \theta_C, \theta_R)$$



bottle

mug

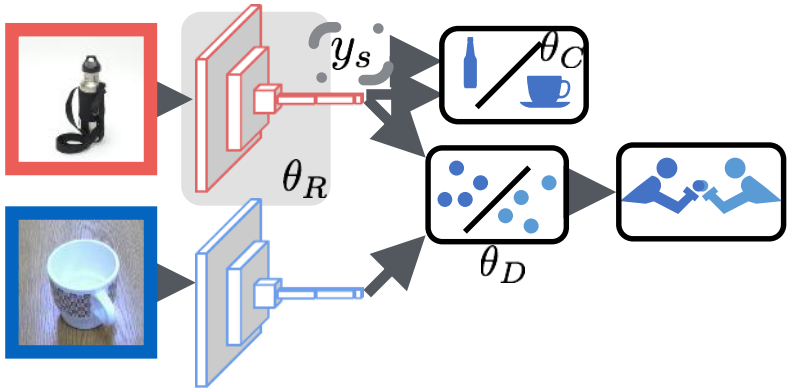
$\mathbf{Y}_s$

$\mathbf{X}_s$

Tzeng, **Hoffman**, Saenko, Darrell. *CVPR* 2017.

Domain Adversarial Optimization

$$\min_{\theta_C, \theta_R} \mathcal{L}_{\text{cls}}(\mathbf{X}_s, \mathbf{Y}_s; \theta_C, \theta_R)$$



bottle

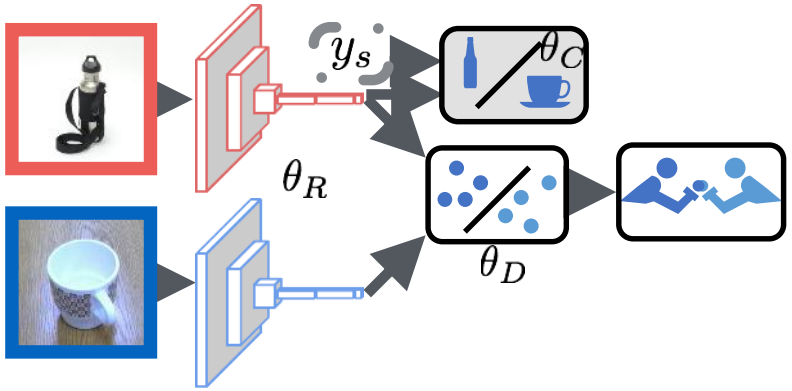
mug

$\mathbf{Y}_s$ $\mathbf{X}_s$

Tzeng, **Hoffman**, Saenko, Darrell. *CVPR* 2017.

Domain Adversarial Optimization

$$\min_{\theta_C, \theta_R} \mathcal{L}_{\text{cls}}(\mathbf{X}_s, \mathbf{Y}_s; \theta_C, \theta_R)$$



bottle

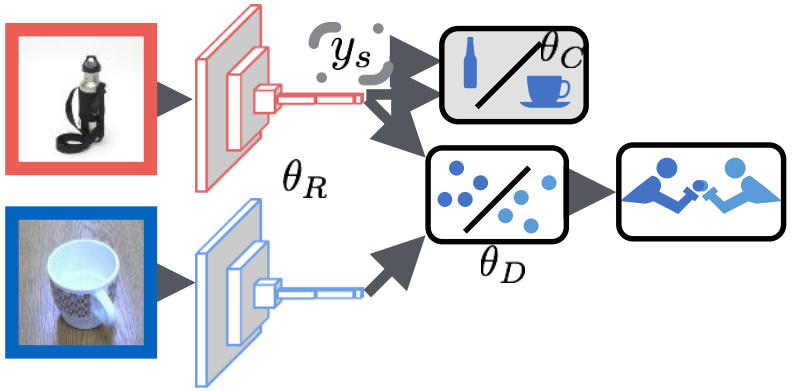
mug

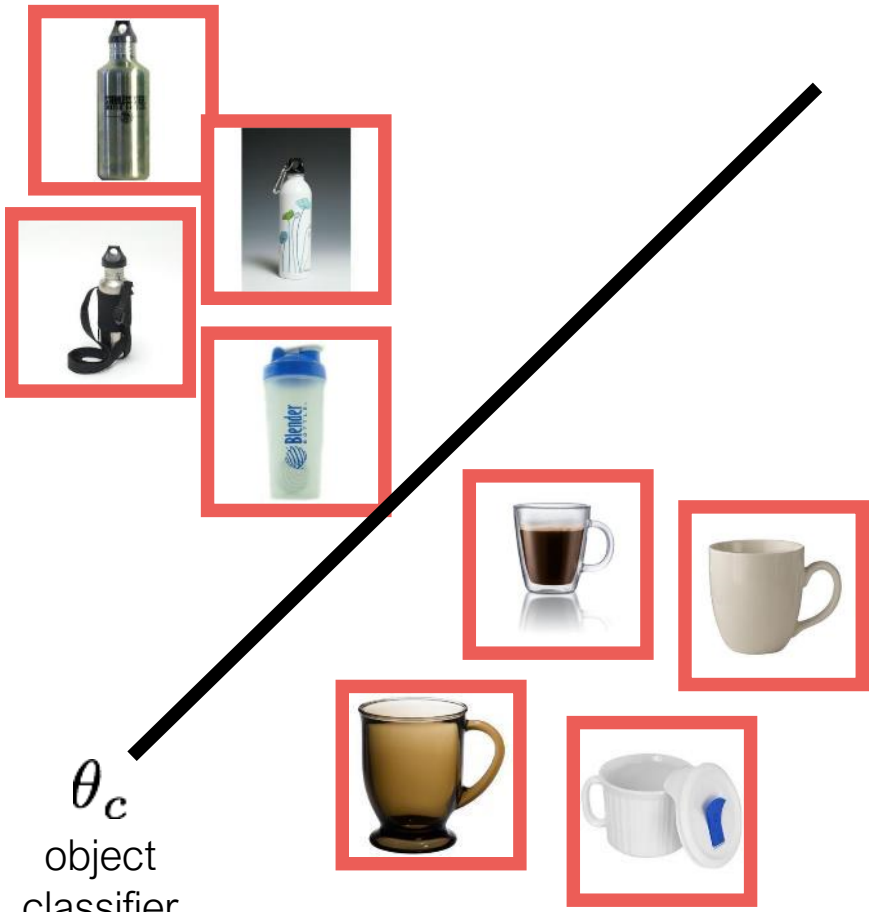
$\mathbf{Y}_s$ $\mathbf{X}_s$

Tzeng, **Hoffman**, Saenko, Darrell. *CVPR* 2017.

Domain Adversarial Optimization

$$\min_{\theta_C, \theta_R} \mathcal{L}_{\text{cls}}(\mathbf{X}_s, \mathbf{Y}_s; \theta_C, \theta_R)$$



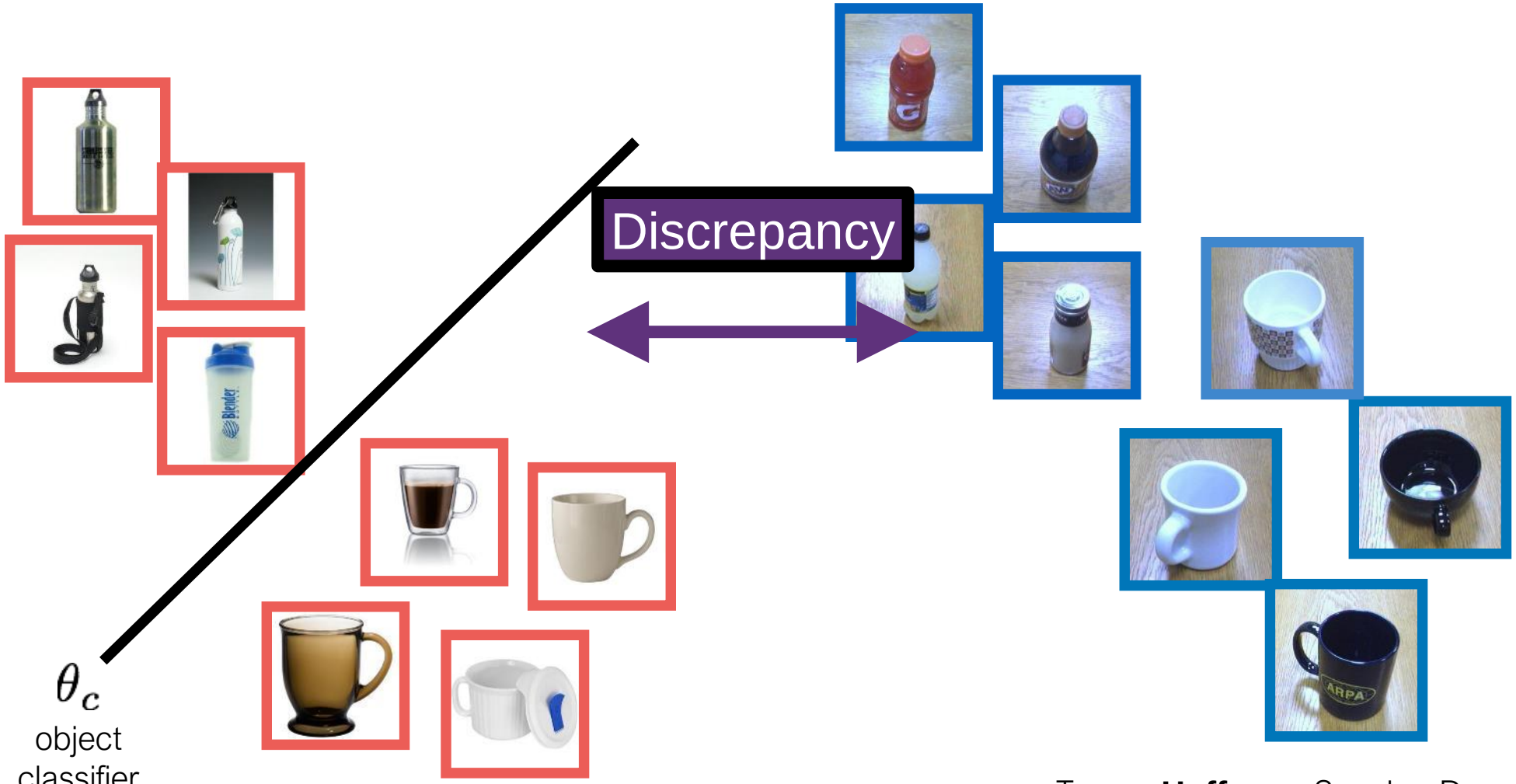


Tzeng, **Hoffman**, Saenko, Darrell. *CVPR* 2017.

Domain Adversarial Optimization

$$\min_{\theta_D} \mathcal{L}_{\text{dom}}(\mathbf{X}_s, \mathbf{X}_t, \theta_R; \theta_D)$$

$$\min_{\theta_R} \mathcal{L}_{\text{rep}}(\mathbf{X}_s, \mathbf{X}_t, \theta_D; \theta_R)$$

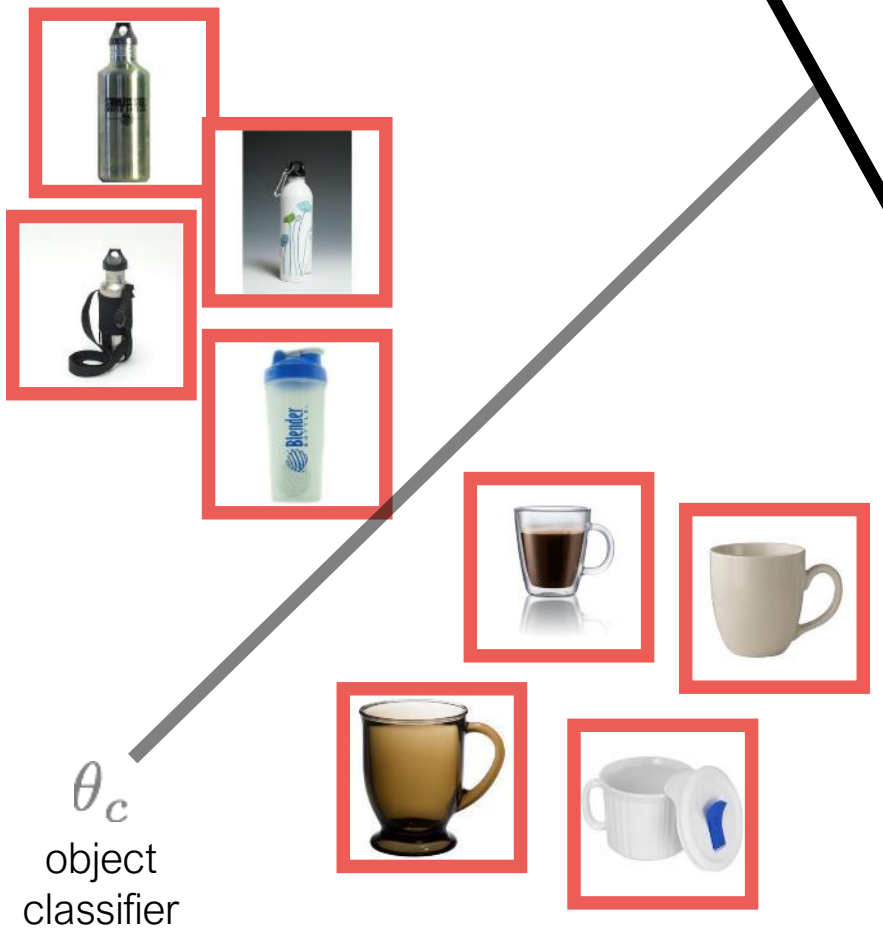


Tzeng, **Hoffman**, Saenko, Darrell. *CVPR* 2017.

Domain Adversarial Optimization

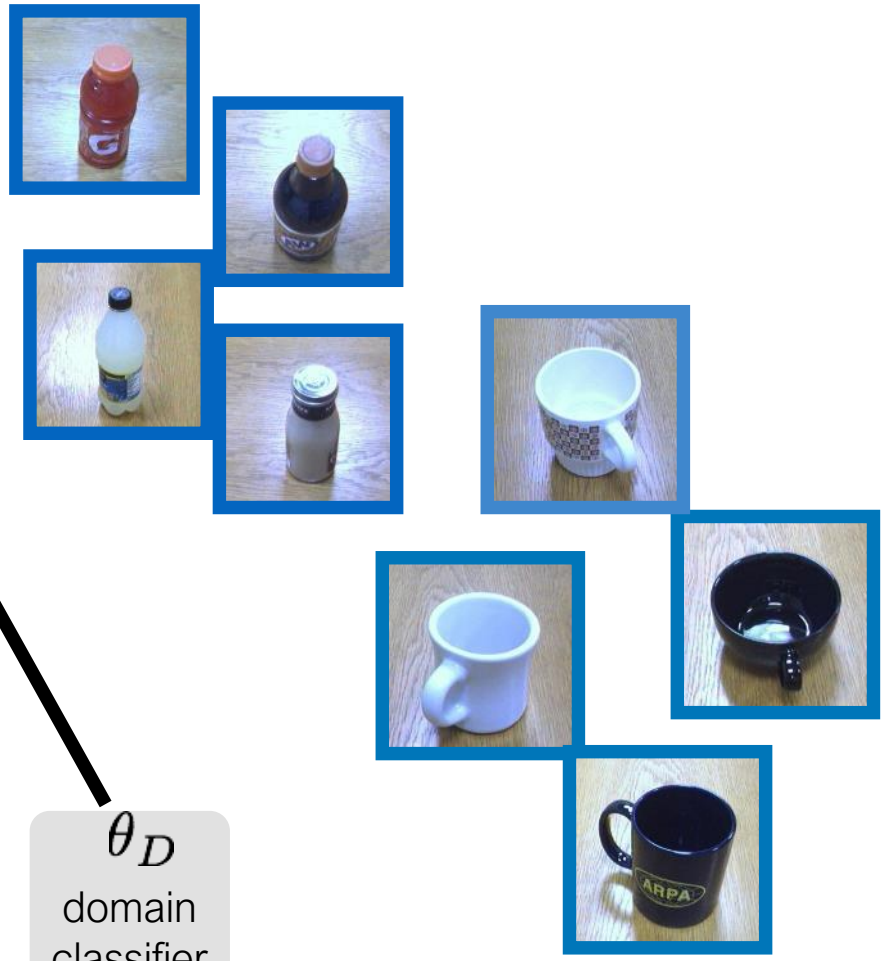
$$\min_{\theta_D} \mathcal{L}_{\text{dom}}(\mathbf{X}_s, \mathbf{X}_t, \theta_R; \theta_D)$$

$$\min_{\theta_R} \mathcal{L}_{\text{rep}}(\mathbf{X}_s, \mathbf{X}_t, \theta_D; \theta_R)$$



θ_c
object classifier

θ_D
domain classifier



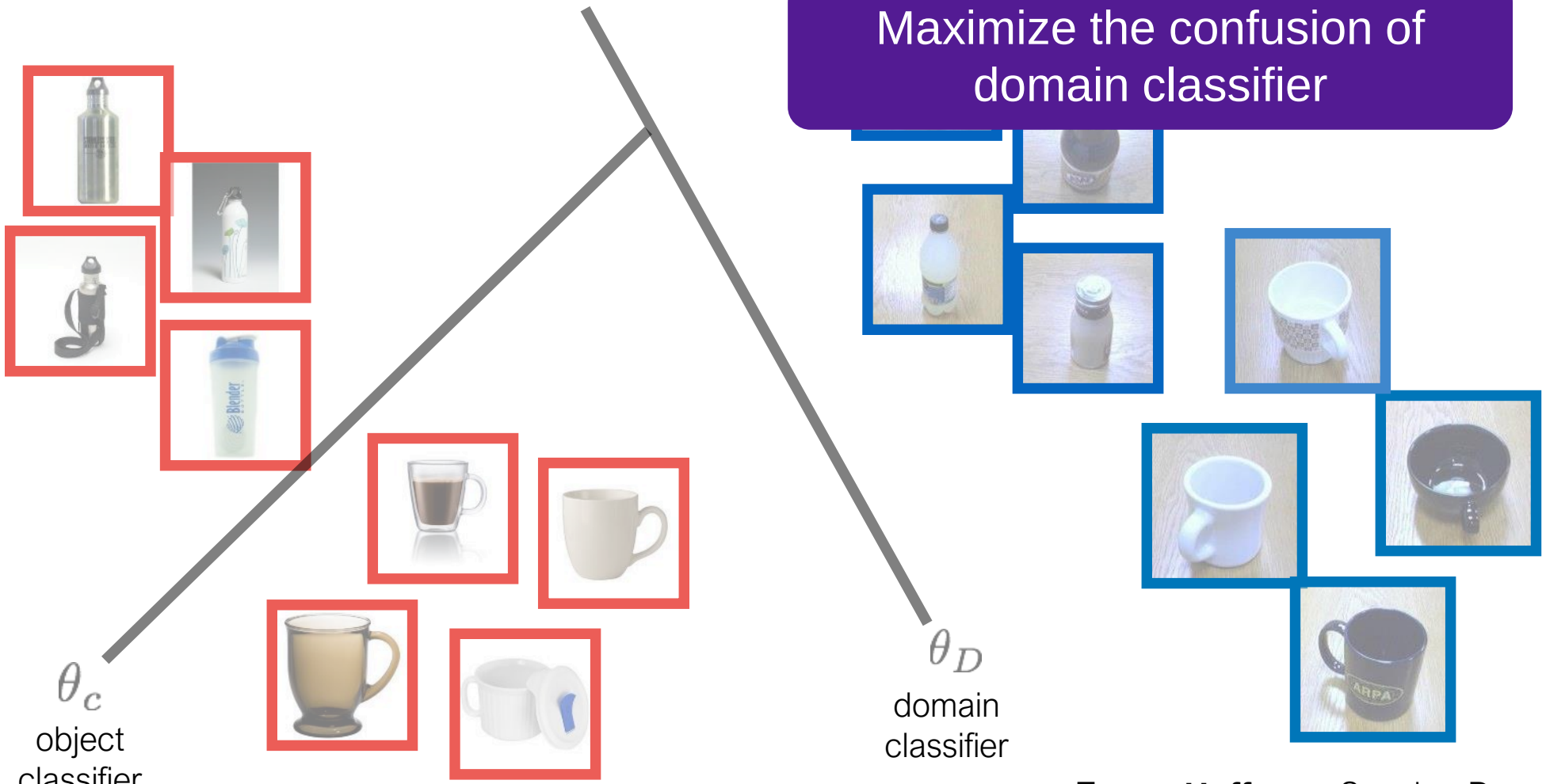
Tzeng, **Hoffman**, Saenko, Darrell. *CVPR* 2017.

Domain Adversarial Optimization

$$\min_{\theta_D} \mathcal{L}_{\text{dom}}(\mathbf{X}_s, \mathbf{X}_t, \theta_R; \theta_D)$$

$$\min_{\theta_R} \mathcal{L}_{\text{rep}}(\mathbf{X}_s, \mathbf{X}_t, \theta_D; \theta_R)$$

Maximize the confusion of domain classifier

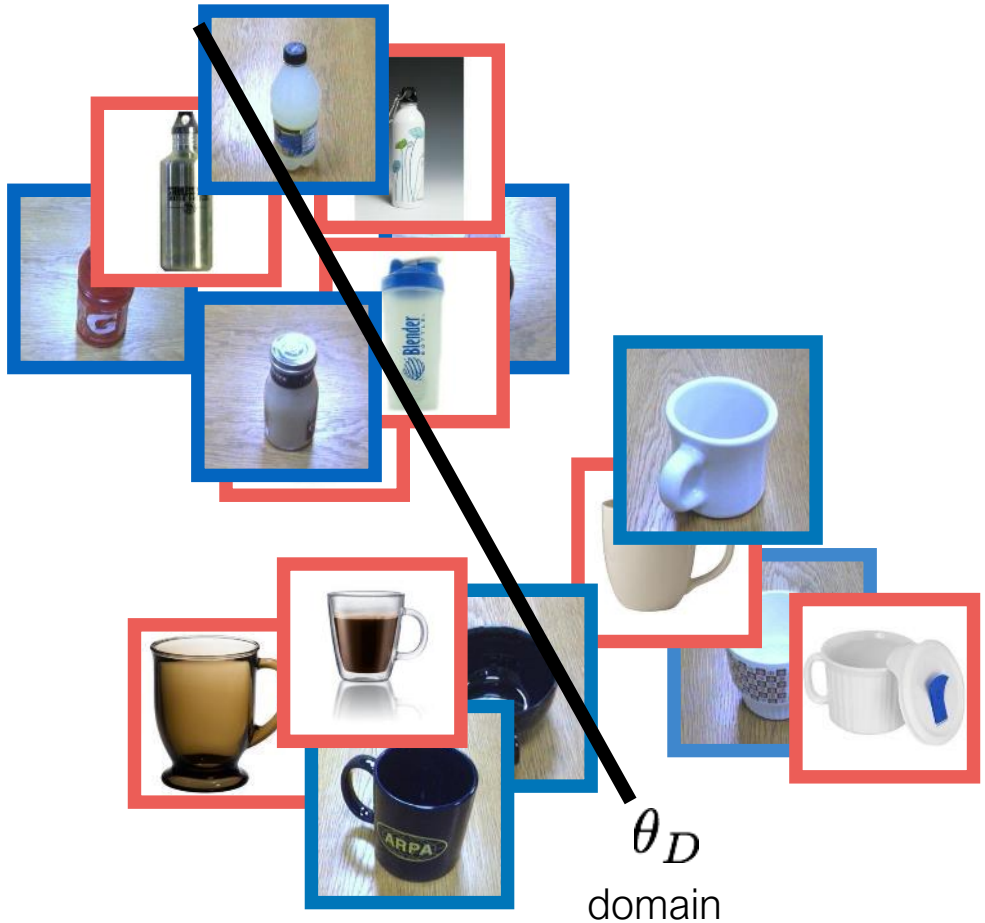


Tzeng, **Hoffman**, Saenko, Darrell. *CVPR* 2017.

Domain Adversarial Optimization

$$\min_{\theta_D} \mathcal{L}_{\text{dom}}(\mathbf{X}_s, \mathbf{X}_t, \theta_R; \theta_D)$$

$$\min_{\theta_R} \mathcal{L}_{\text{rep}}(\mathbf{X}_s, \mathbf{X}_t, \theta_D; \theta_R)$$



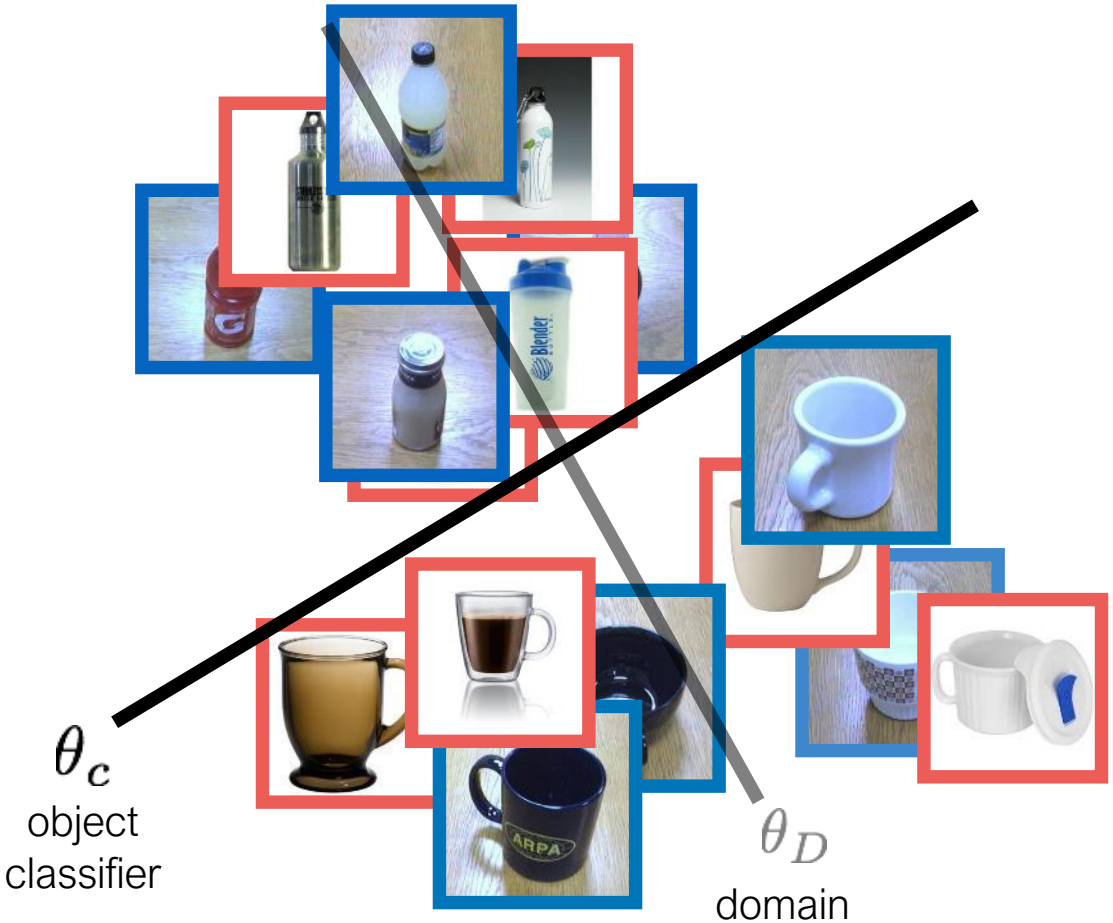
Tzeng, **Hoffman**, Saenko, Darrell. *CVPR* 2017.

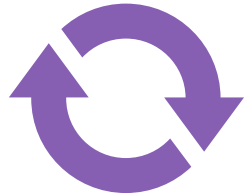
Domain Adversarial Optimization

$$\min_{\theta_D} \mathcal{L}_{\text{dom}}(\mathbf{X}_s, \mathbf{X}_t, \theta_R; \theta_D)$$

$$\min_{\theta_R} \mathcal{L}_{\text{rep}}(\mathbf{X}_s, \mathbf{X}_t, \theta_D; \theta_R)$$

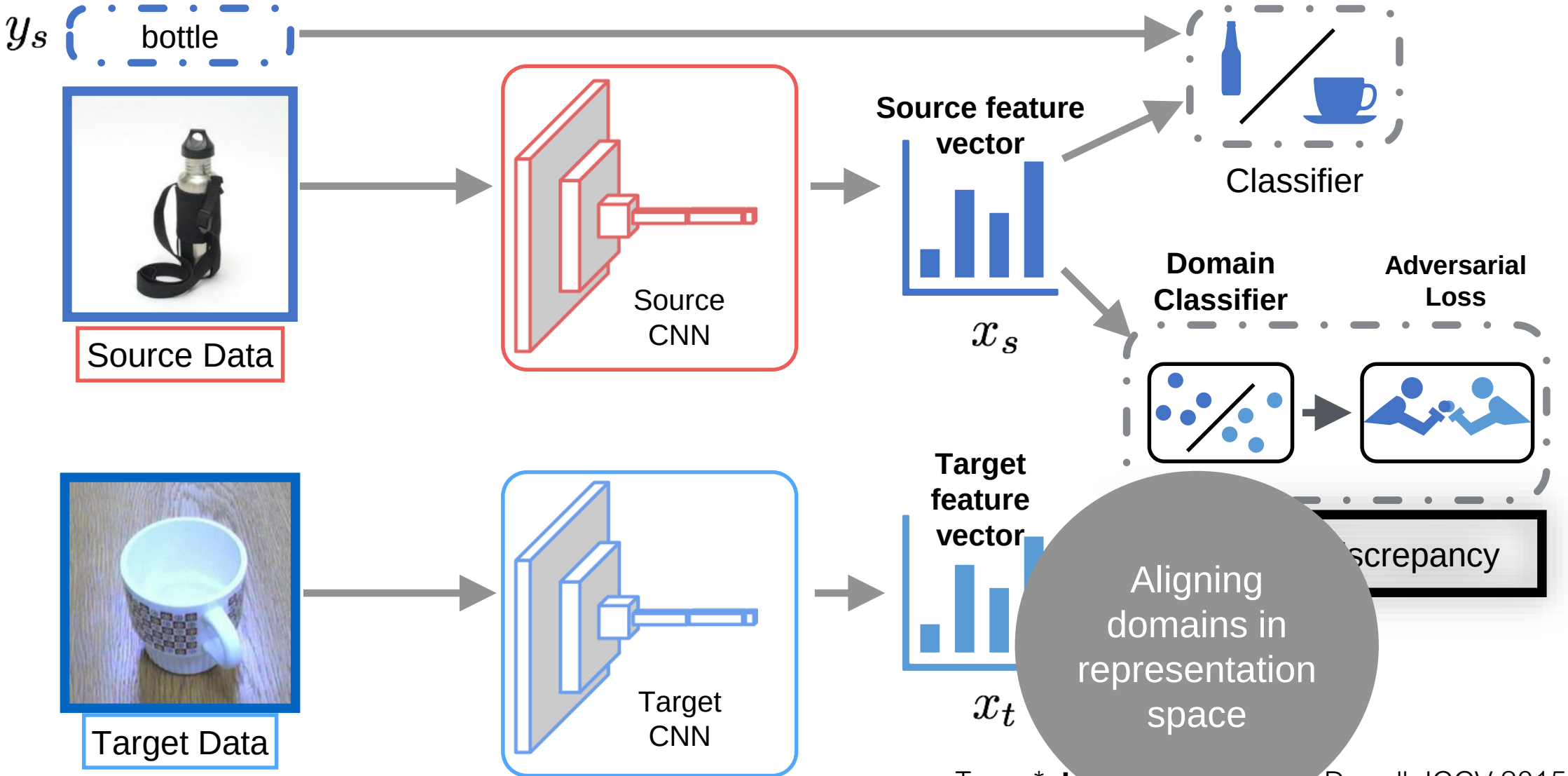
$$\min_{\theta_C, \theta_R} \mathcal{L}_{\text{cls}}(\mathbf{X}_s, \mathbf{Y}_s; \theta_C, \theta_R)$$





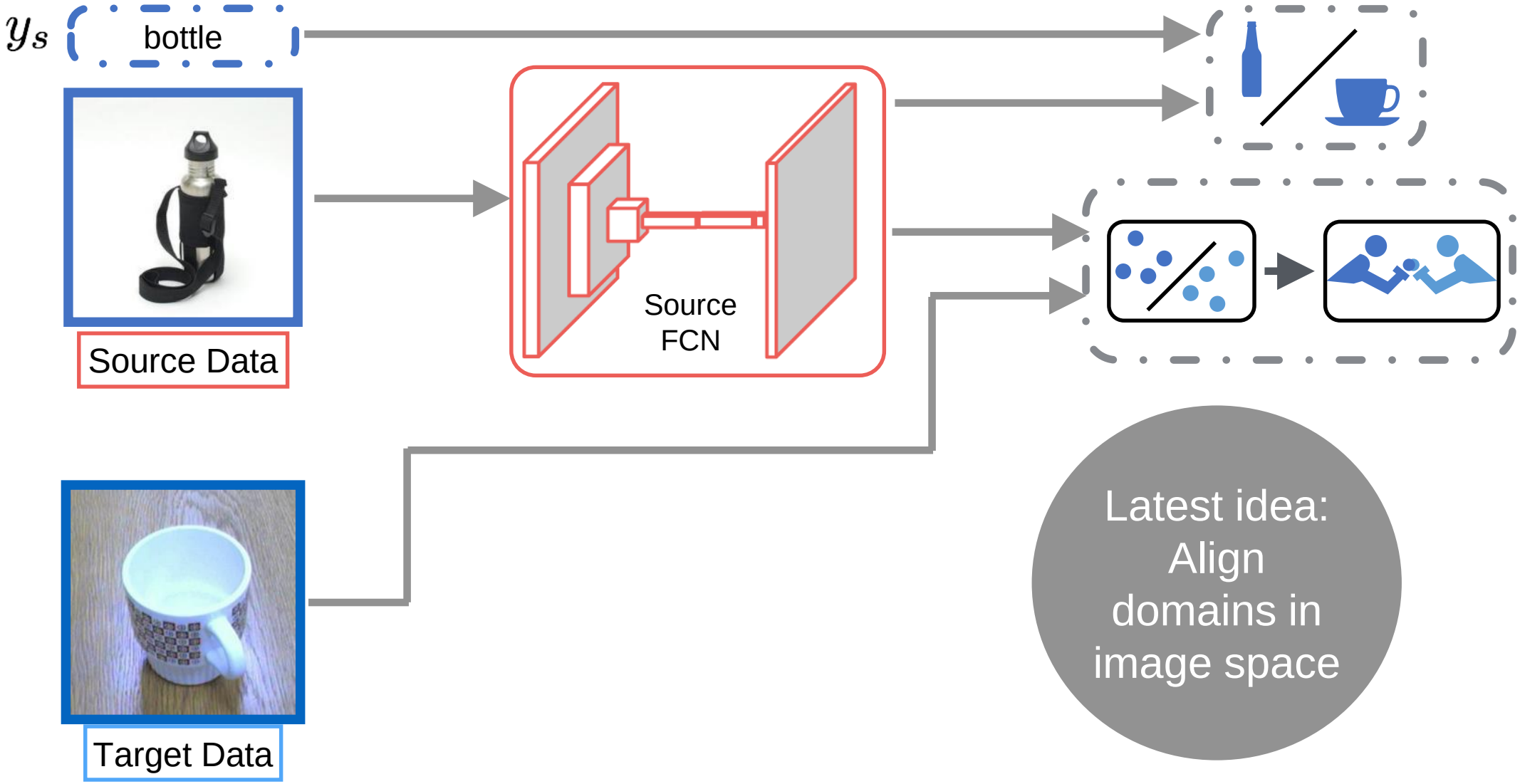
Tzeng, **Hoffman**, Saenko, Darrell. *CVPR* 2017.

Domain Adversarial Adaptation



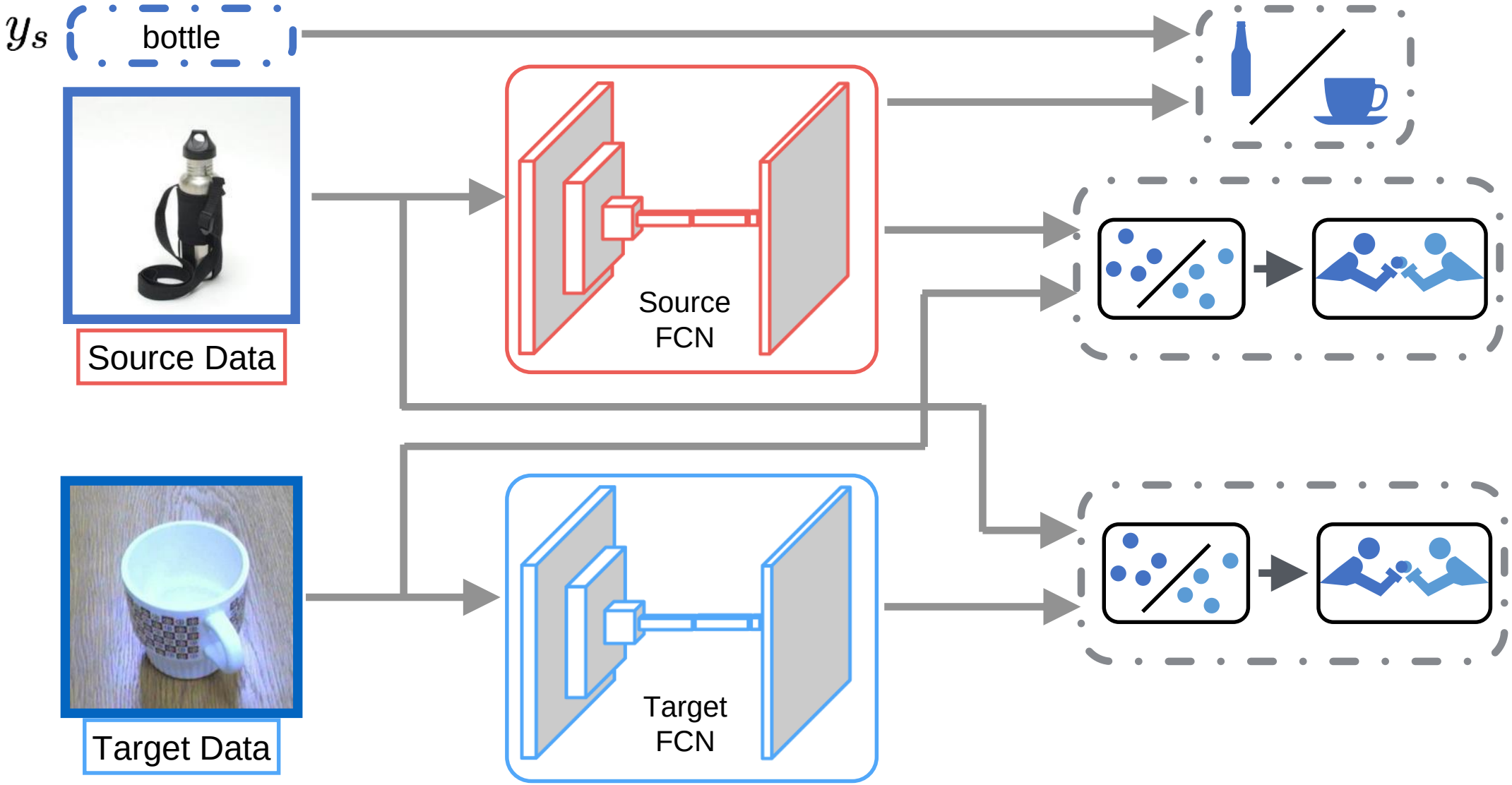
Tzeng*, **Ho**, Darrell. ICCV 2015.
Tzeng, **Hoffman**, Saenko, Darrell. *CVPR* 2017.

Domain Adversarial Adaptation



Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, arXiv 2017.

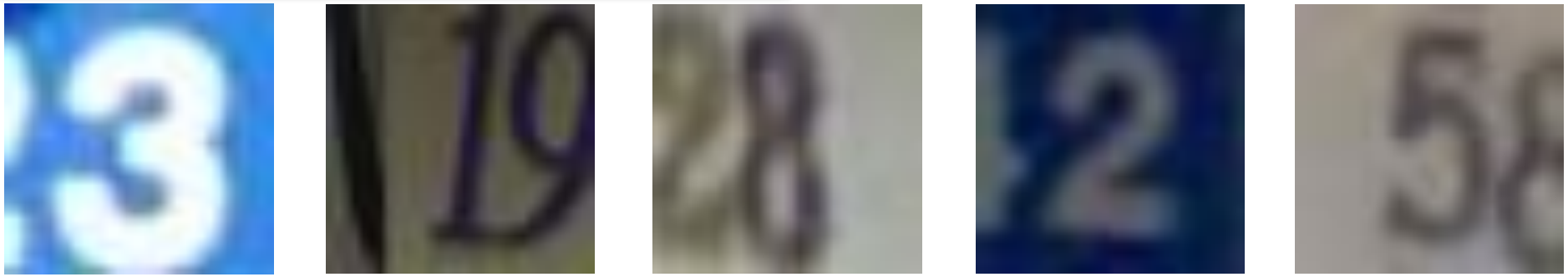
Domain Adversarial Adaptation



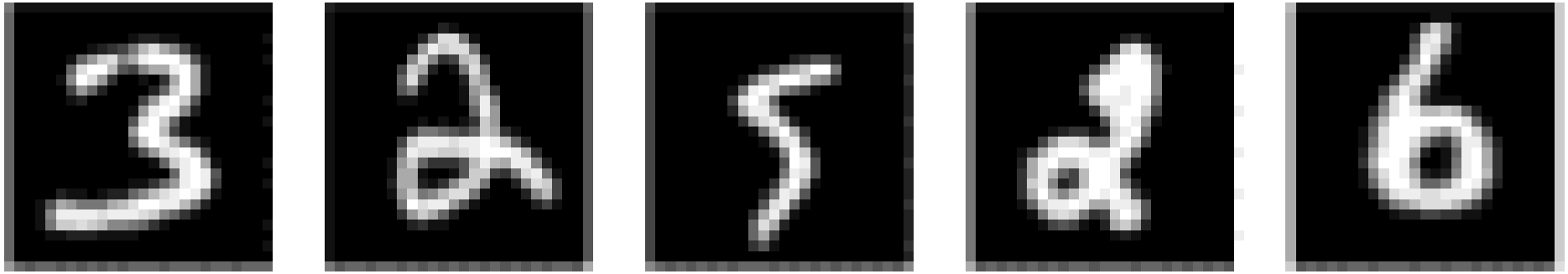
Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, arXiv 2017.

Adaptation Results: Digit Recognition

Google Street View House Numbers



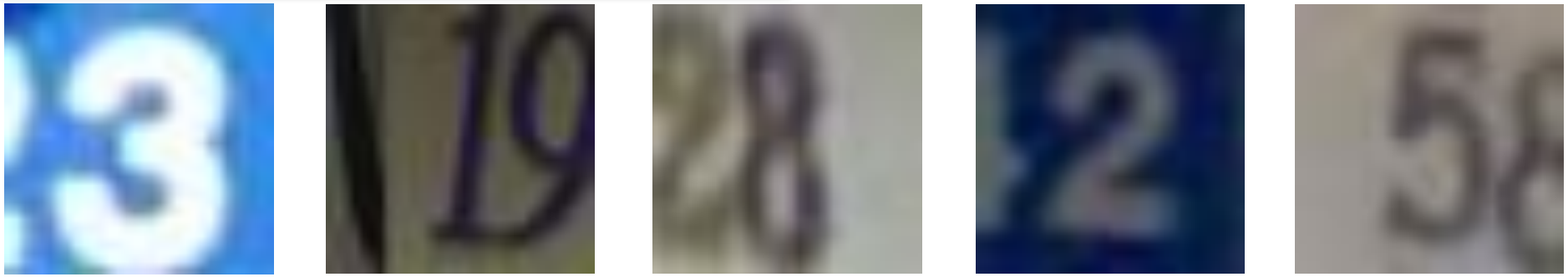
MNIST (handwritten digits)



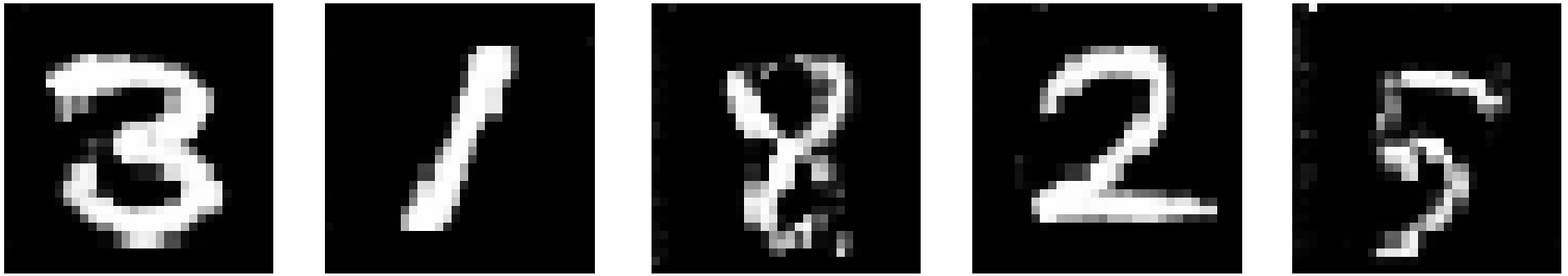
Tzeng, **Hoffman**, Saenko, Darrell. *CVPR* 2017.
Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, arXiv 2017.

Adaptation Results: Digit Recognition

Google Street View House Numbers

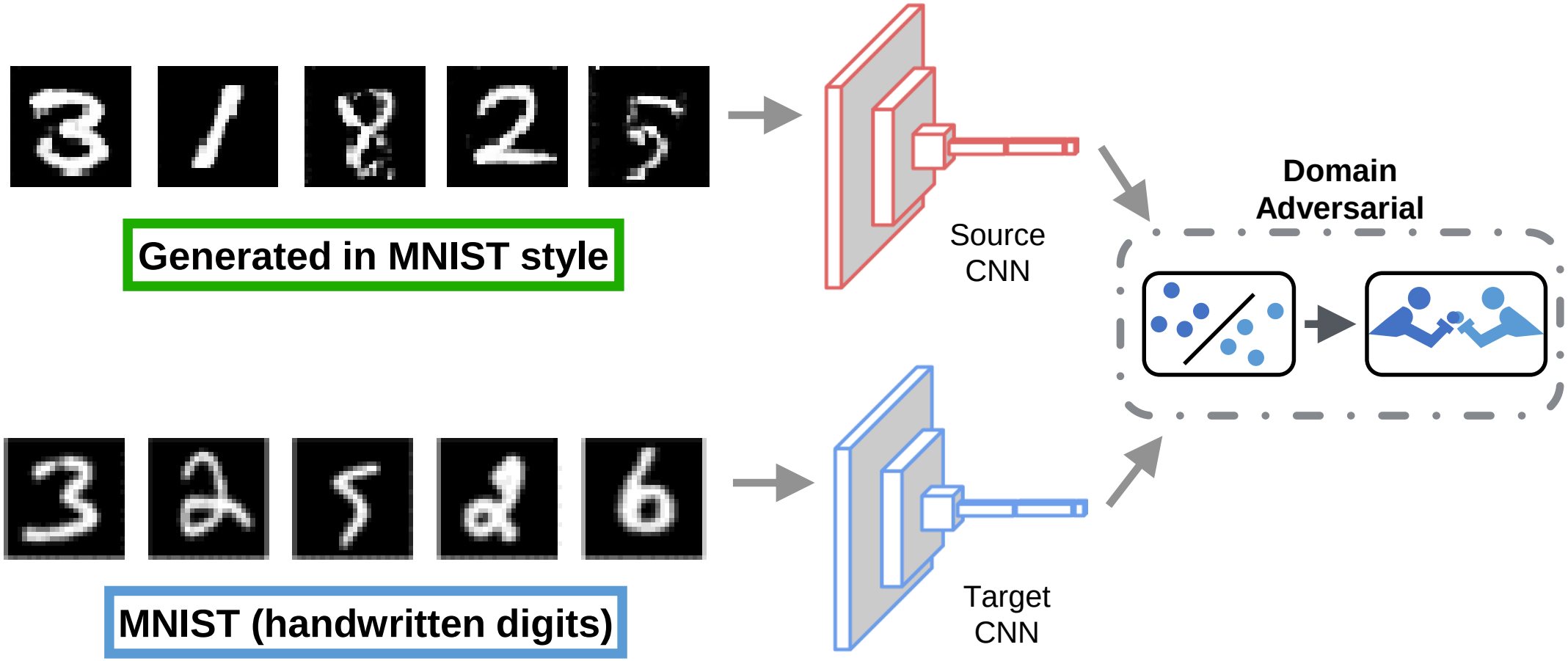


Generated in MNIST style



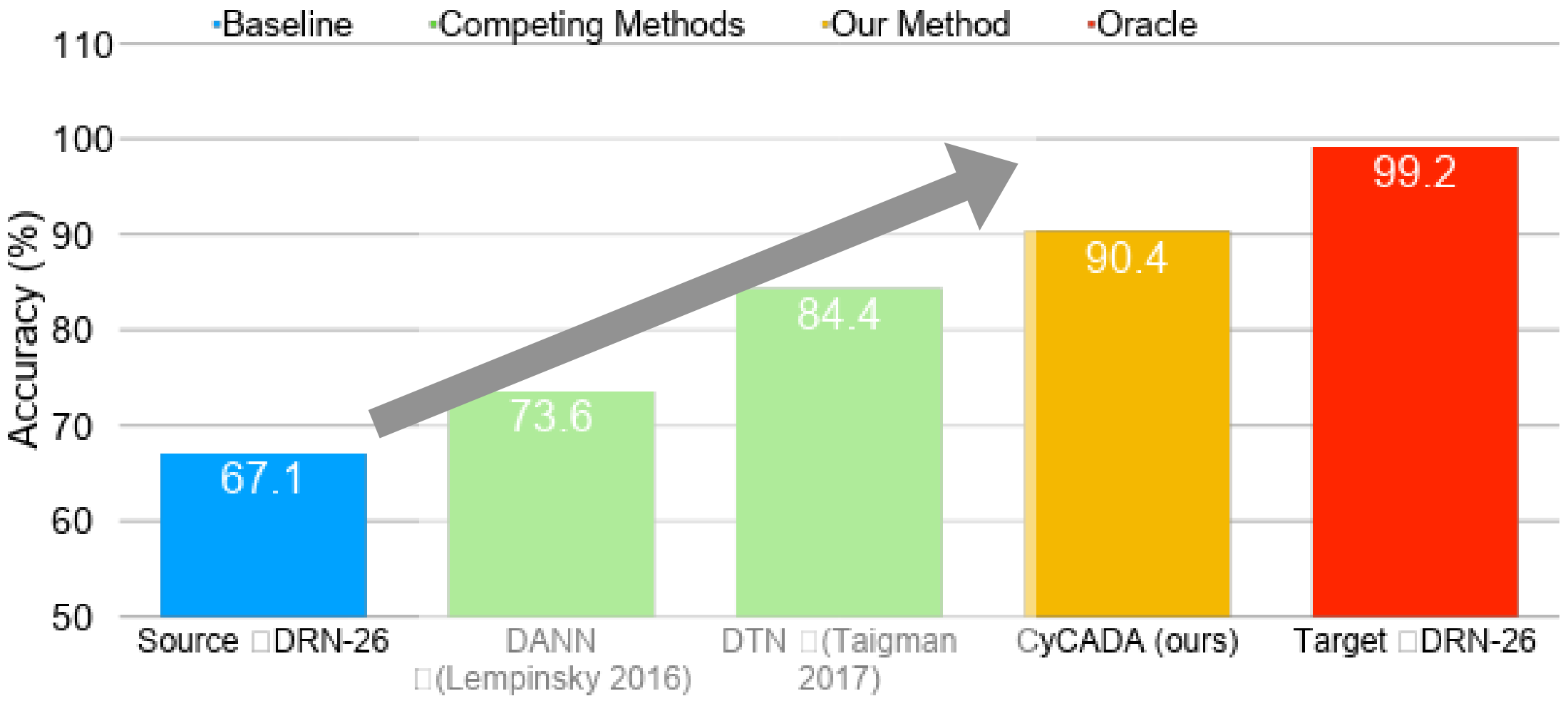
Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, arXiv 2017.

Adaptation Results: Digit Recognition



Tzeng, **Hoffman**, Saenko, Darrell. *CVPR* 2017.
Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, arXiv 2017.

Adaptation Results: Digit Recognition



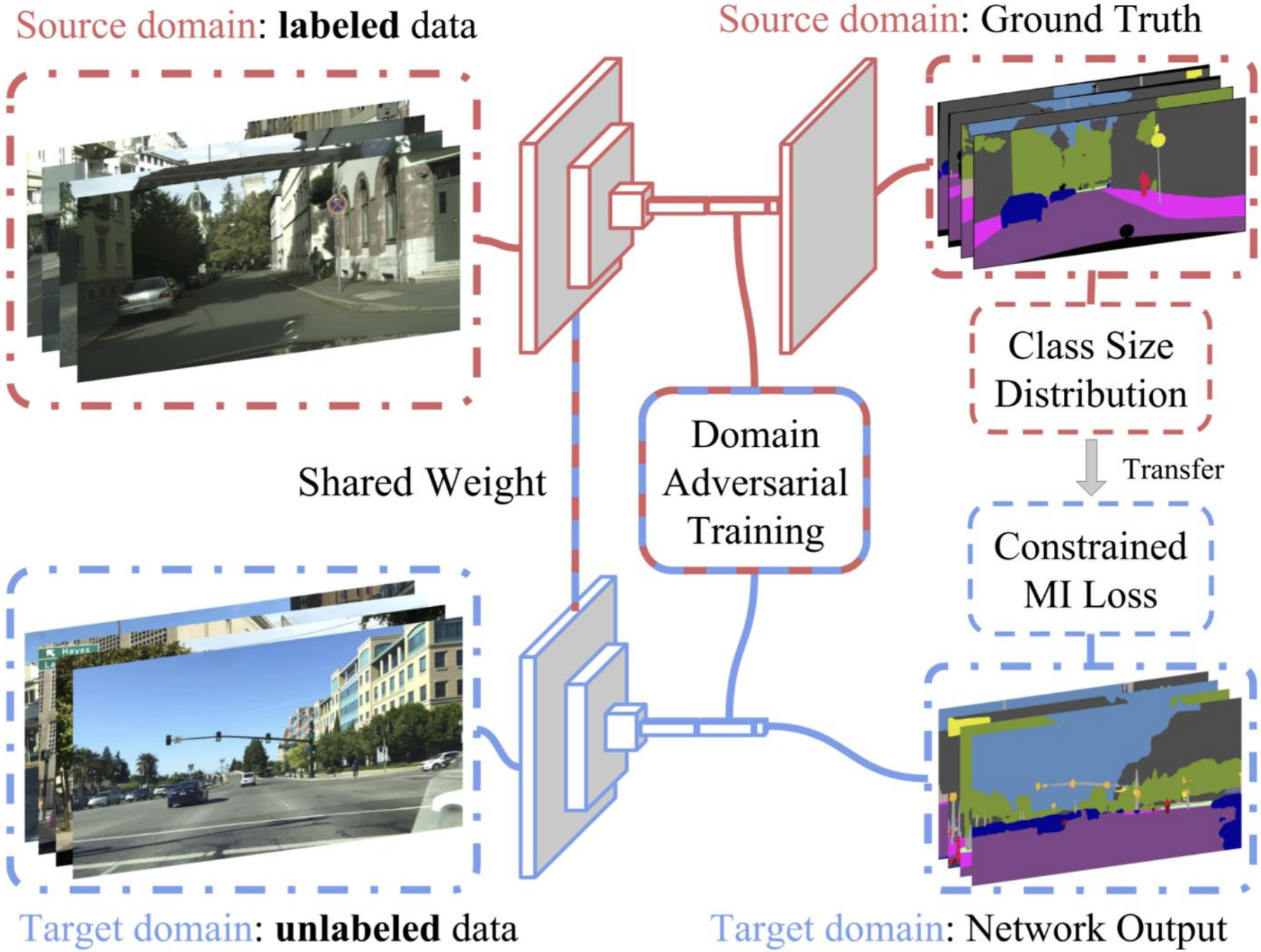
Tzeng, **Hoffman**, Saenko, Darrell. *CVPR* 2017.
Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, arXiv 2017.

In-domain fully supervised FCN

Train on Cityscapes, Test on Cityscapes

Domain shift: Cityscapes to SF

Train on Cityscapes, Test on **San Francisco Dashcam**

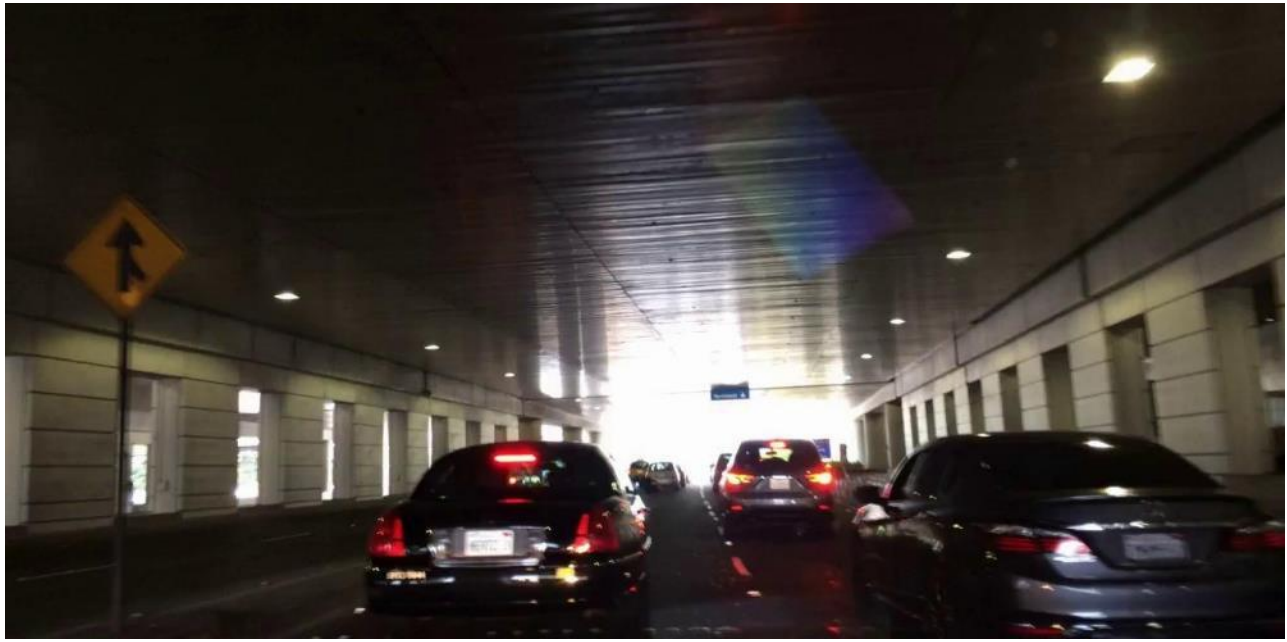

Source domain: labeled data
Source domain: Ground Truth
Class Size Distribution
Transfer
Shared Weight
Domain Adversarial Training
Constrained MI Loss
Target domain: unlabeled data
Target domain: Network Output

Cross-city Adaptation

Train

CityScapes (Germany)

Test

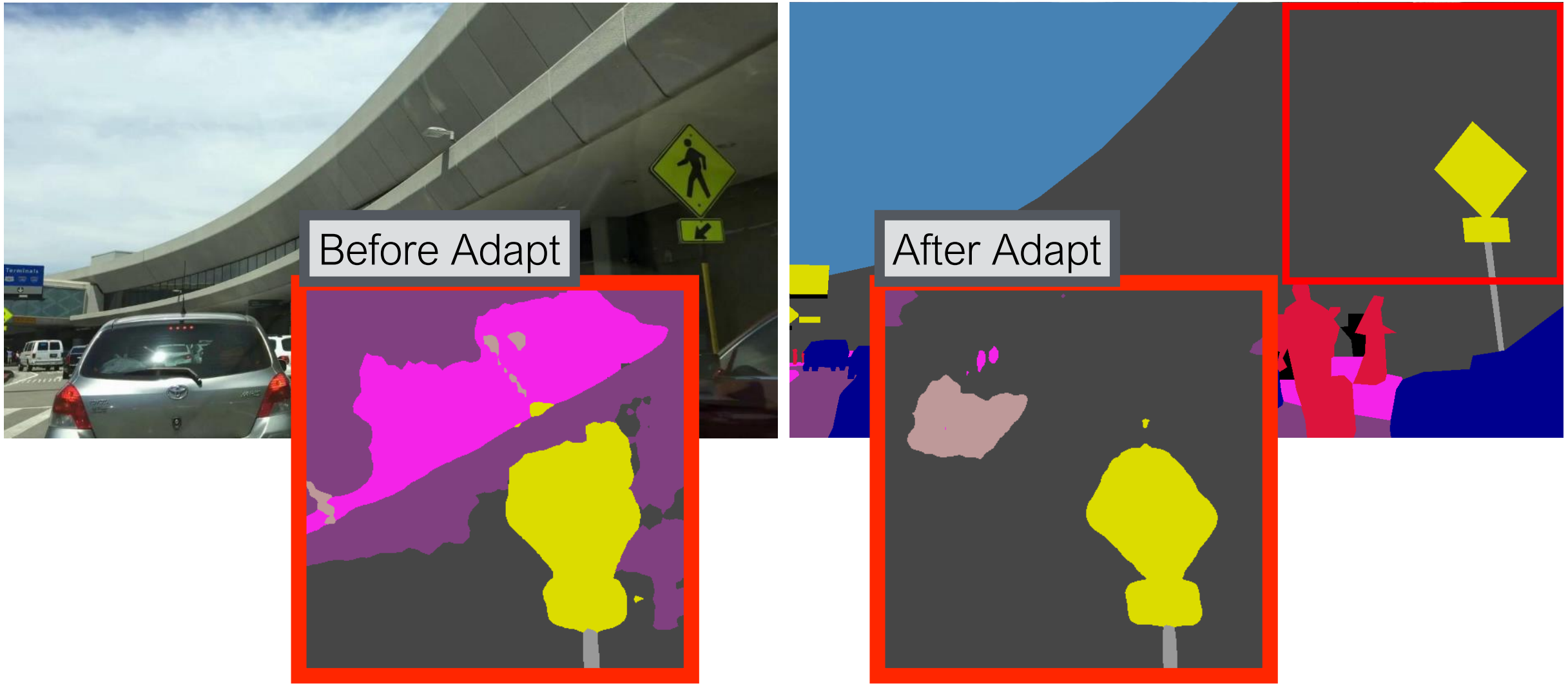


San Francisco

***Differences*: signs, tunnels, size of roads**

Hoffman, Wang, Yu, Darrell, arXiv 2017.

Cross-city Adaptation



CityScapes (Germany) to San Francisco

Cross Season Adaptation

Train

Fall Image

Test

Winter Image

SYNTHIA Dataset

Hoffman, Wang, Yu, Darrell, arXiv 2017

Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, arXiv 2017

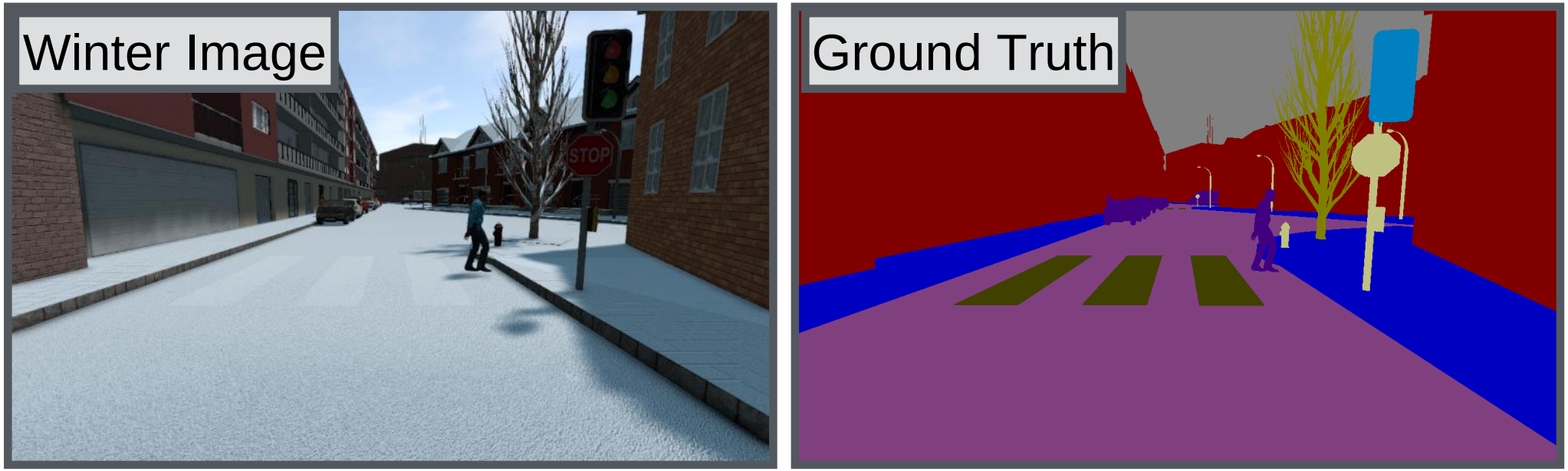
Cross Season Pixel Adaptation

Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, arXiv 2017.

Cross Season Pixel Adaptation

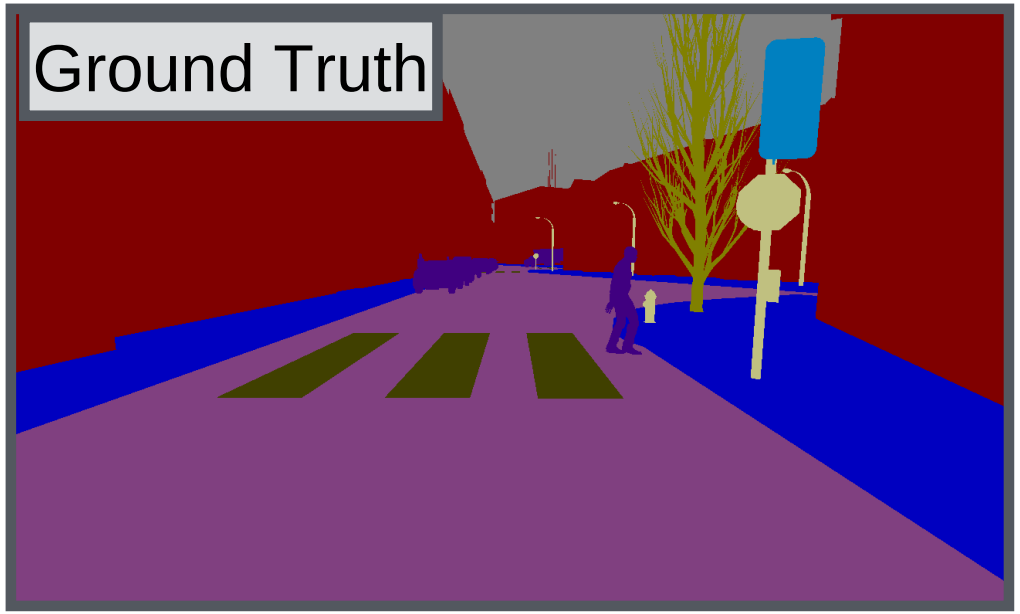
Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, arXiv 2017.

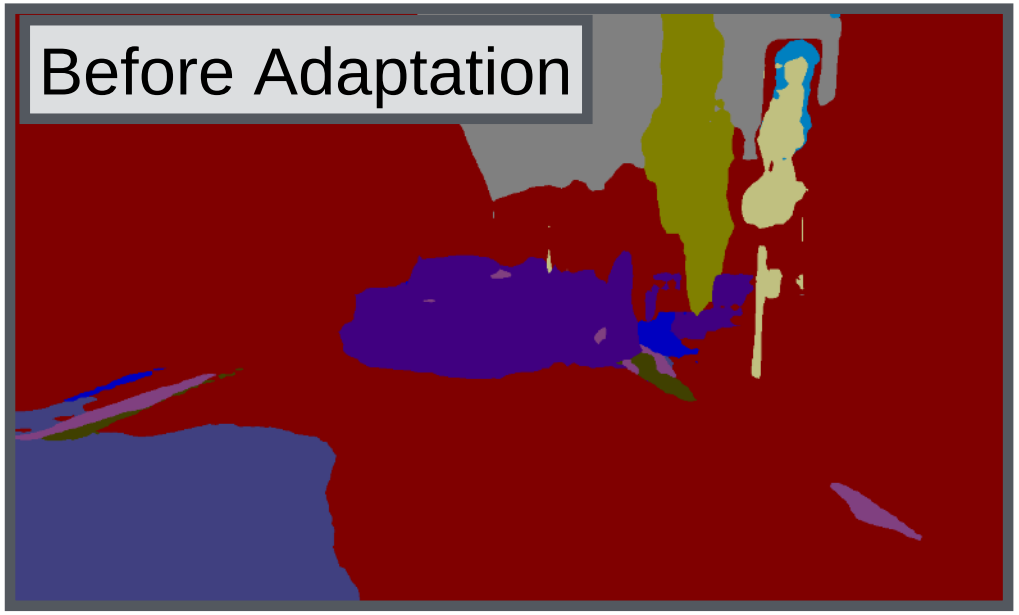
Cross Season Adaptation



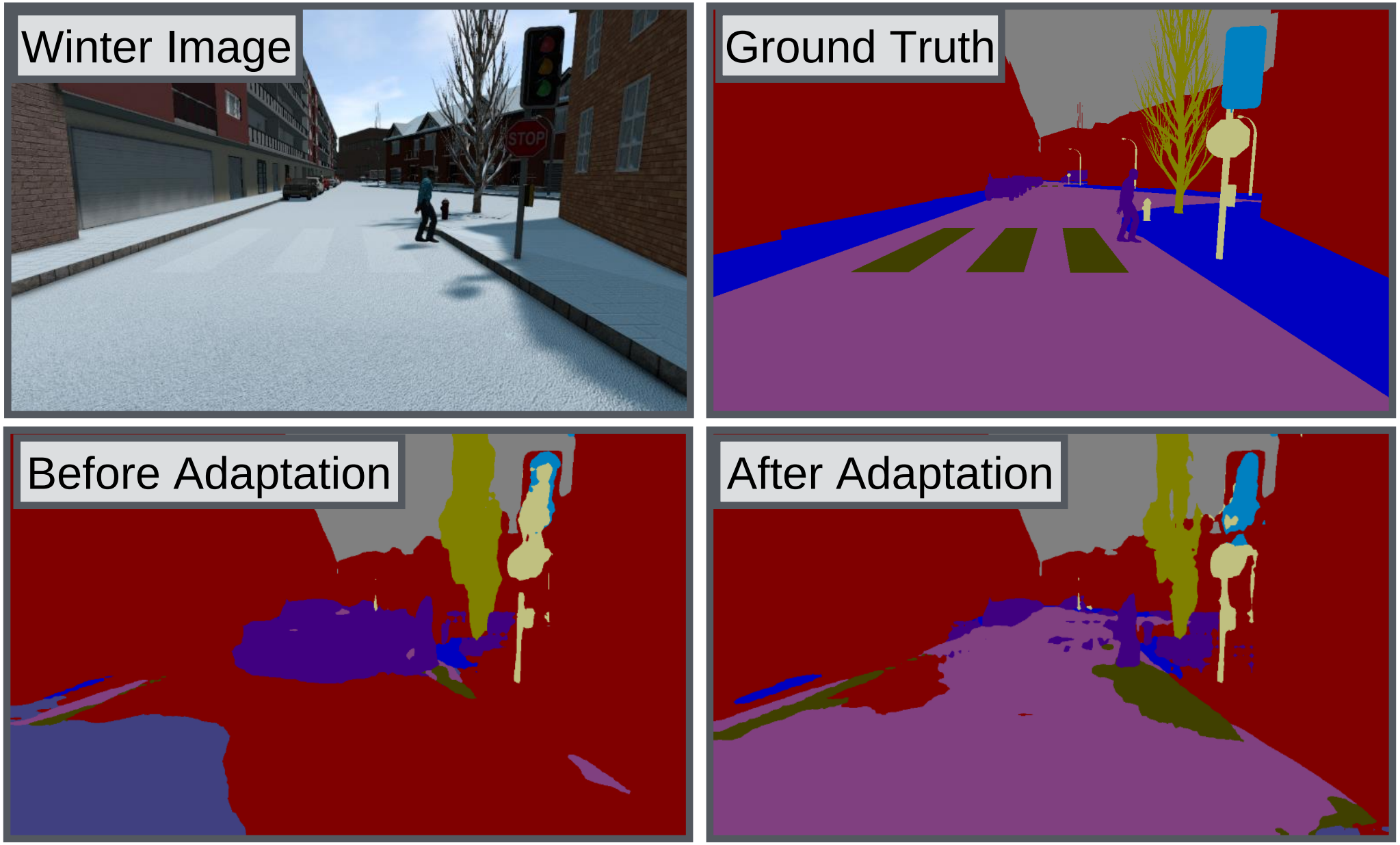
Hoffman, Wang, Yu, Darrell, arXiv 2017.

Cross Season Adaptation





Cross Season Adaptation



Synthetic to Real Pixel Adaptation

Train

GTA (synthetic)

Test

CityScapes (Germany)

Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, arXiv 2017.

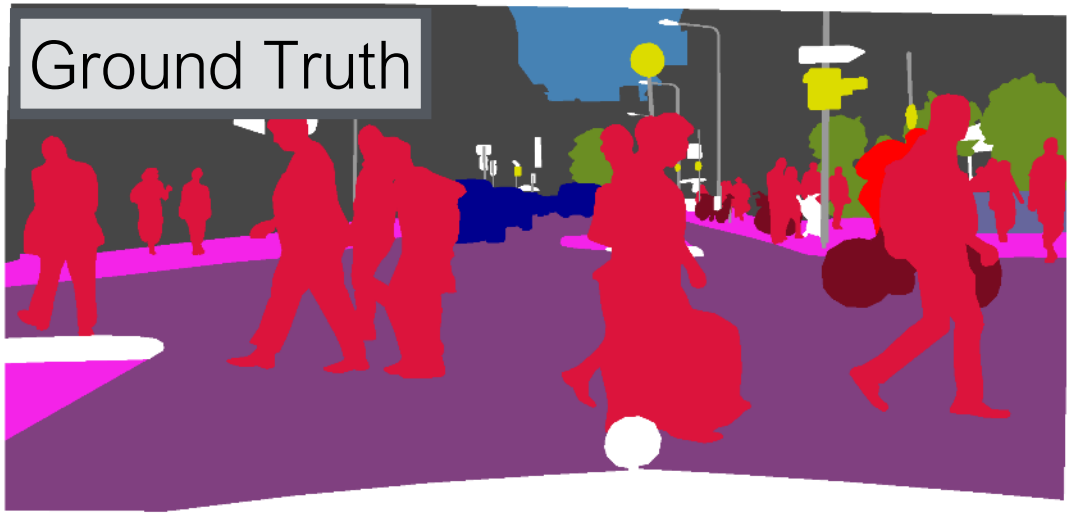
Synthetic to Real Pixel Adaptation

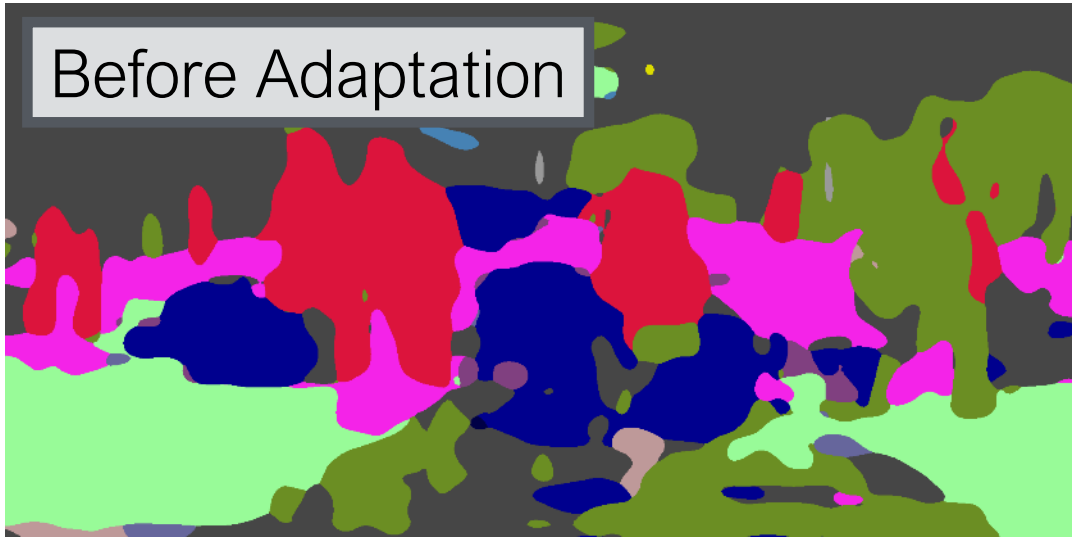
Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, arXiv 2017.

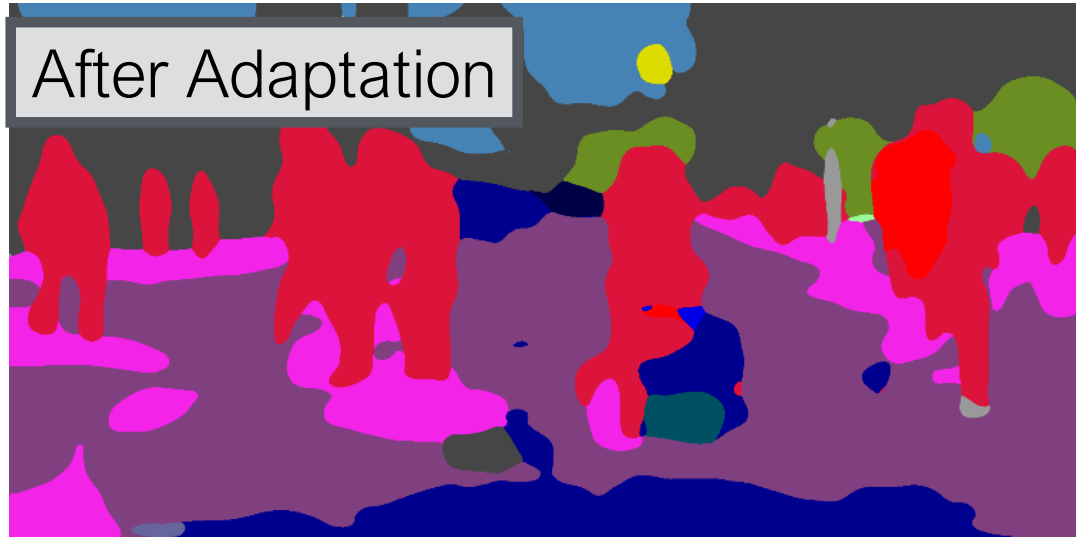
Real to Synthetic Pixel Adaptation

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

Results: Train on GTA, test on Cityscapes

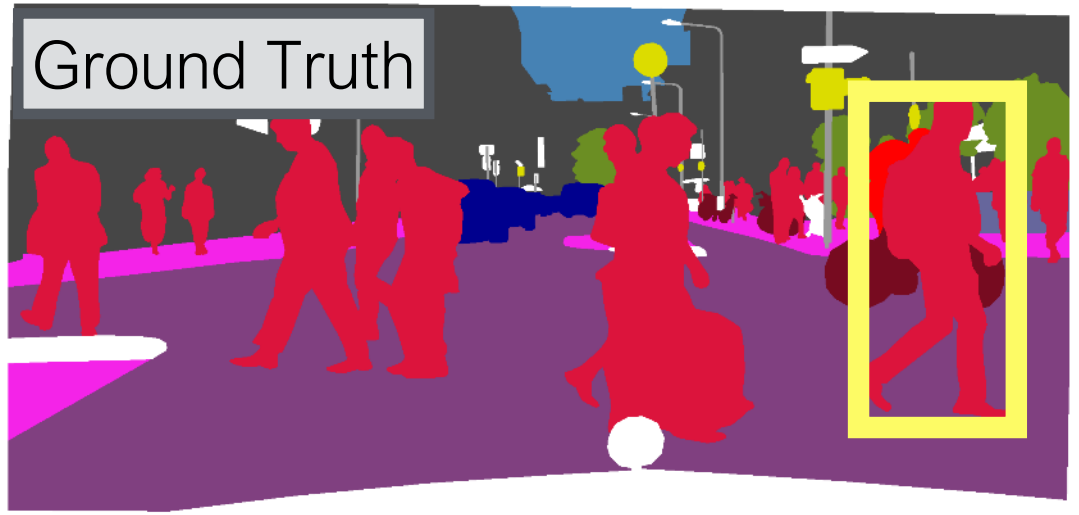


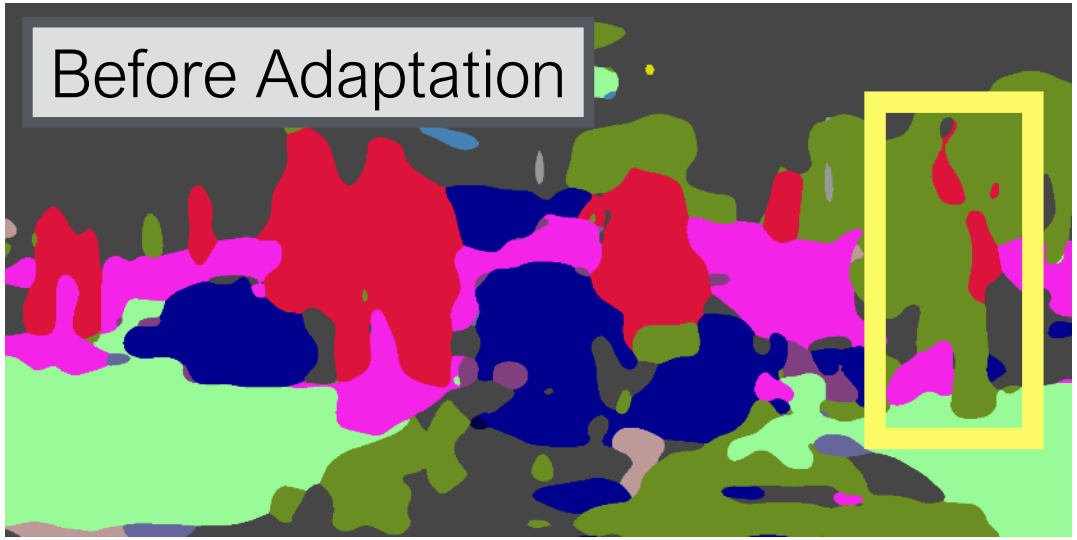


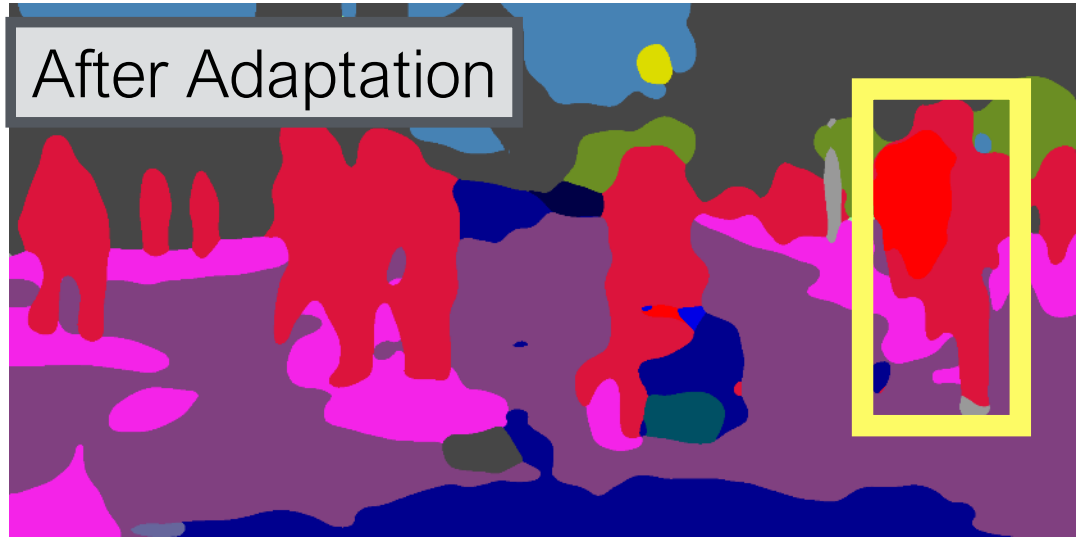


Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, arXiv 2017.

Results: Train on GTA, test on Cityscapes

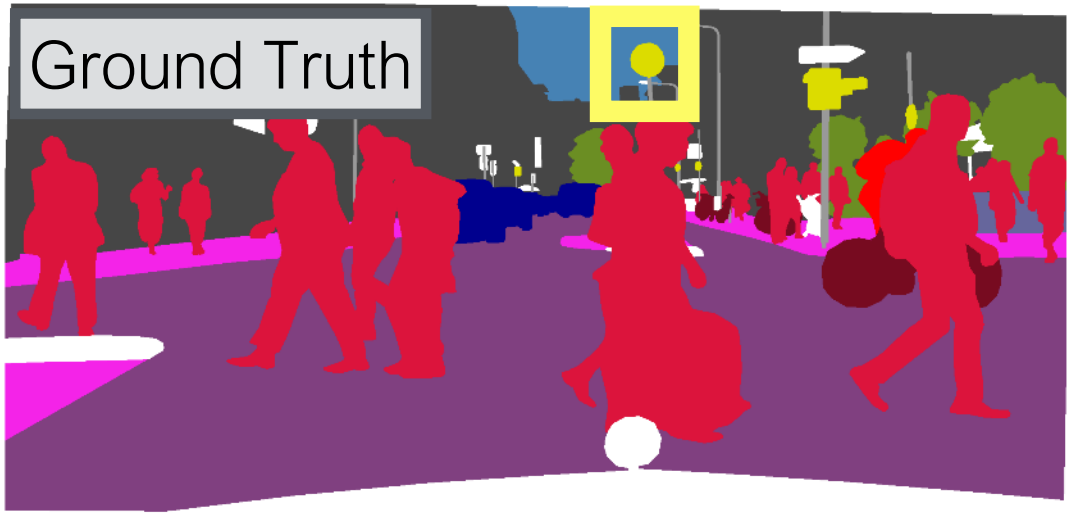


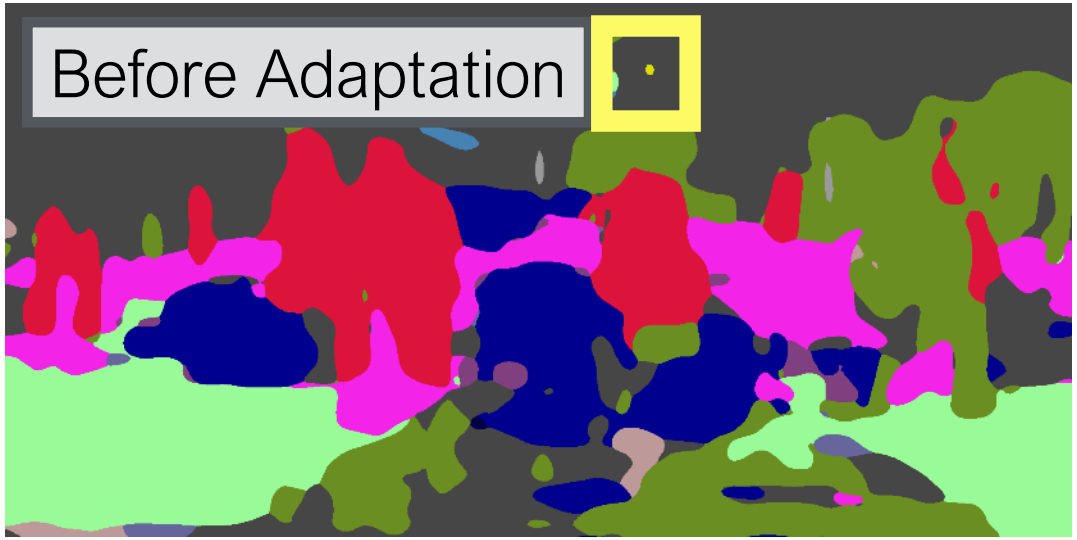


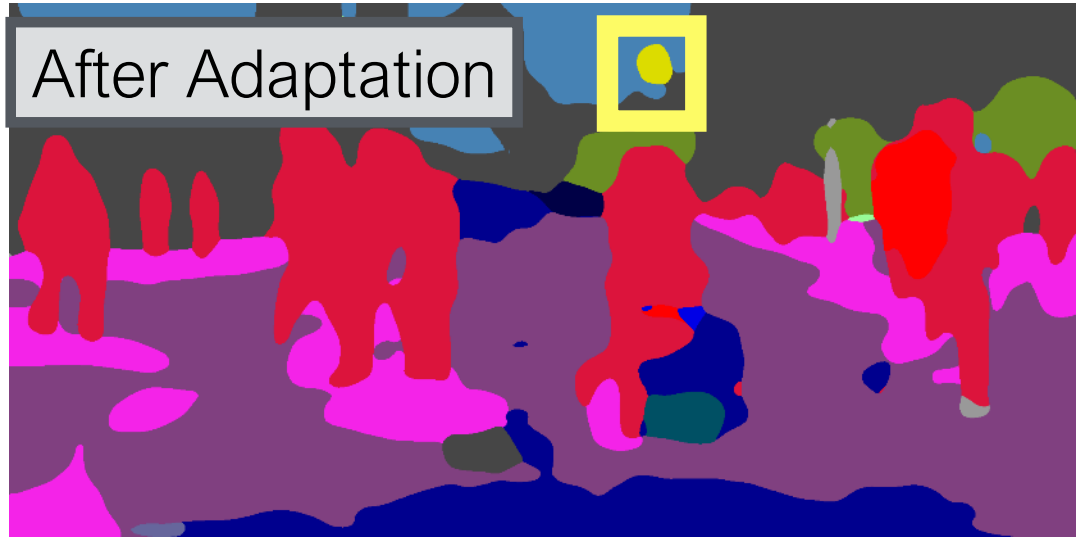


Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, arXiv 2017.

Results: Train on GTA, test on Cityscapes

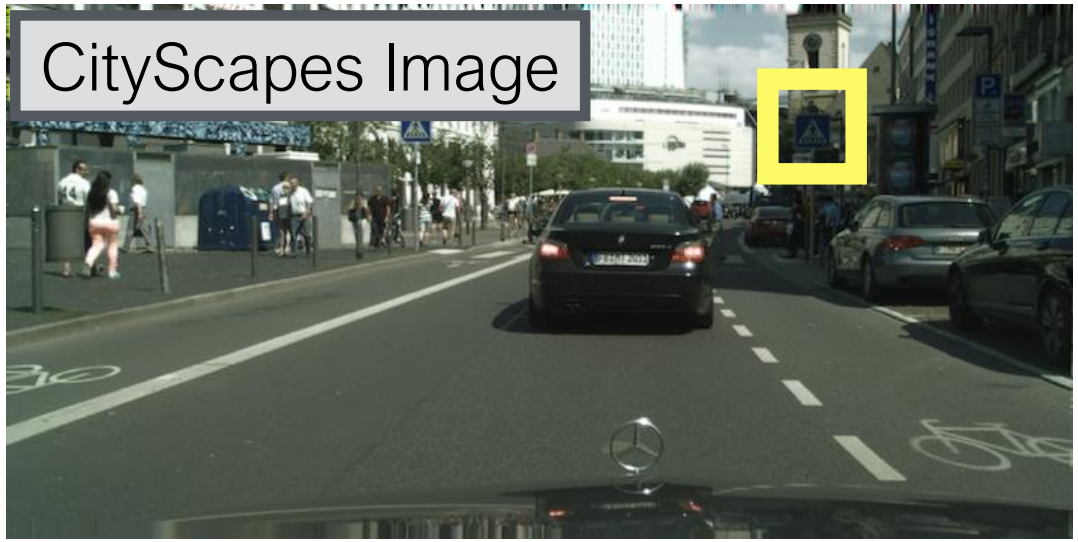


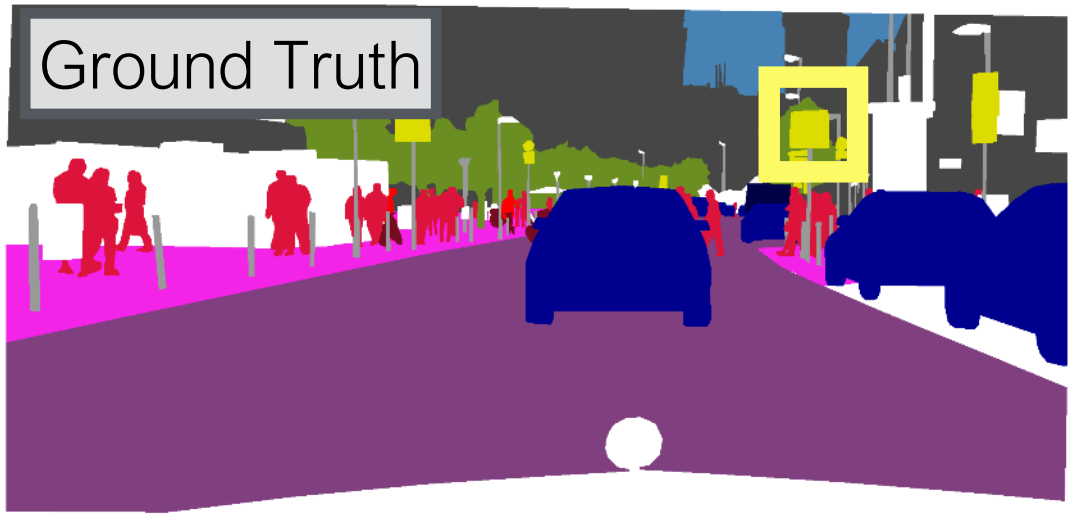


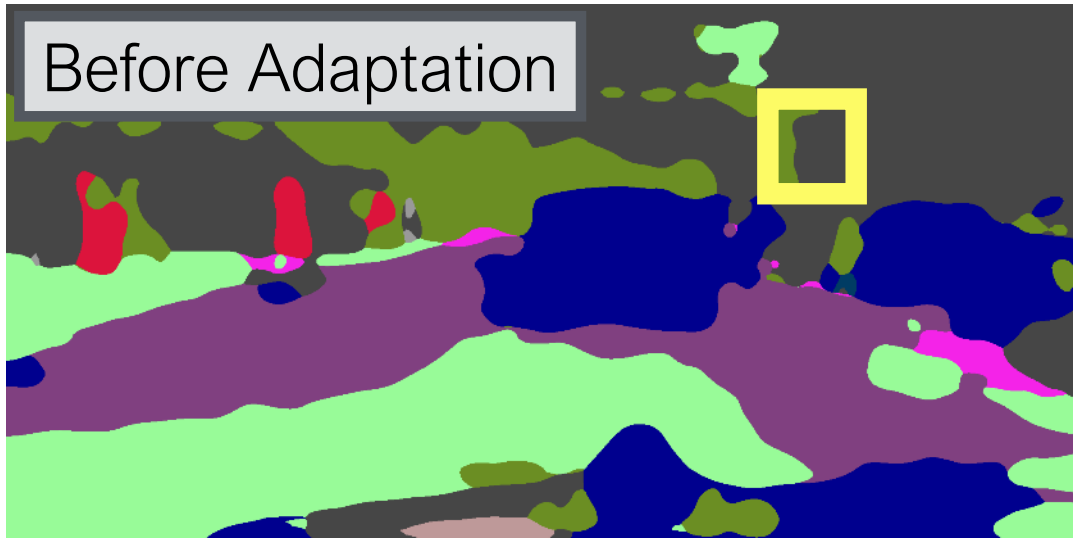


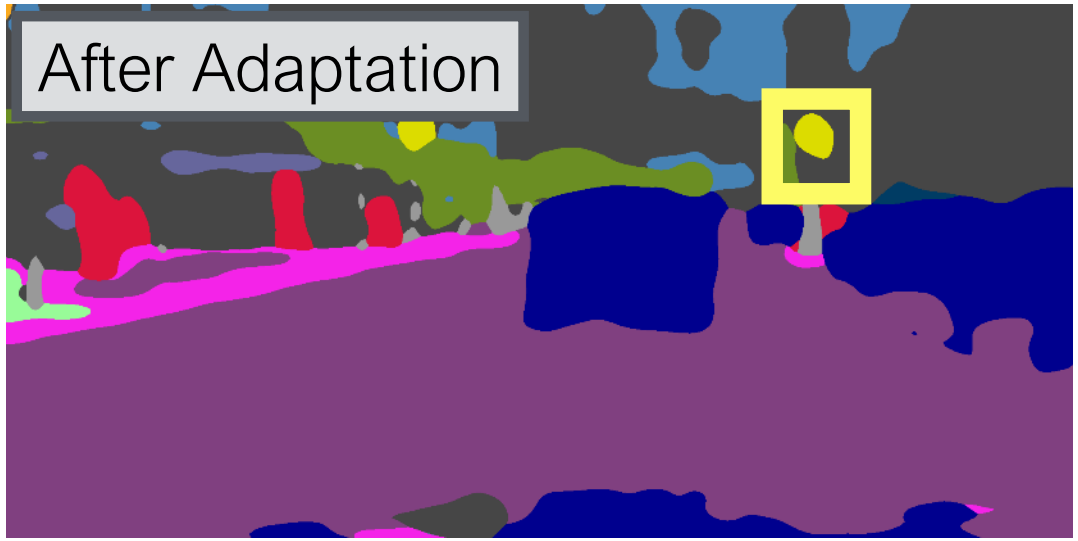
Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, arXiv 2017.

Results: Train on GTA, test on Cityscapes



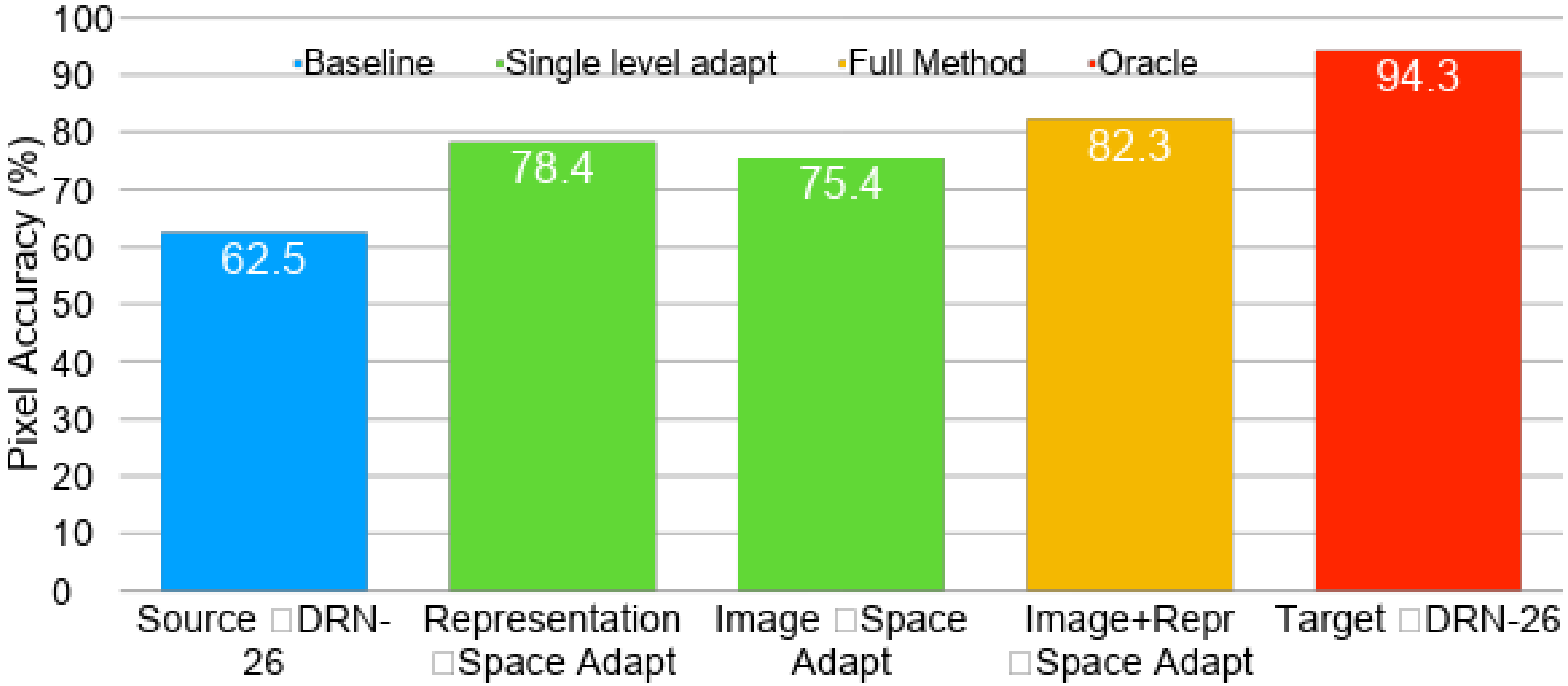






Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, arXiv 2017.

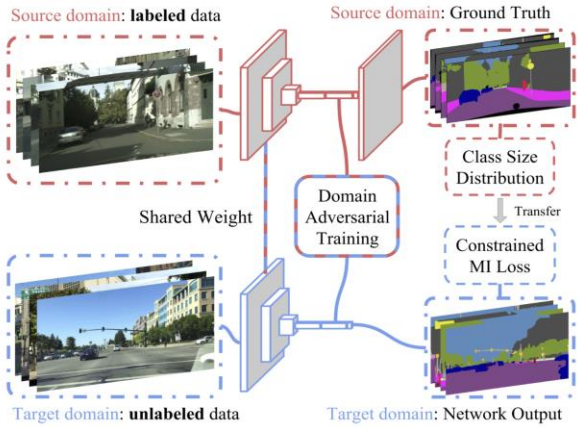
Results: Joint Image and Representation Adapt



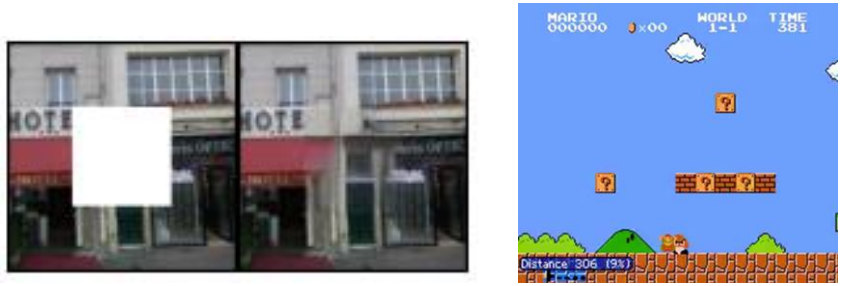
J. Hoffman, E. Tzeng, T. Park, J. Zhu, P. Isola, K. Saenko, A. Efros, T. Darrell, arXiv 2017.

Deep *learning to learn*

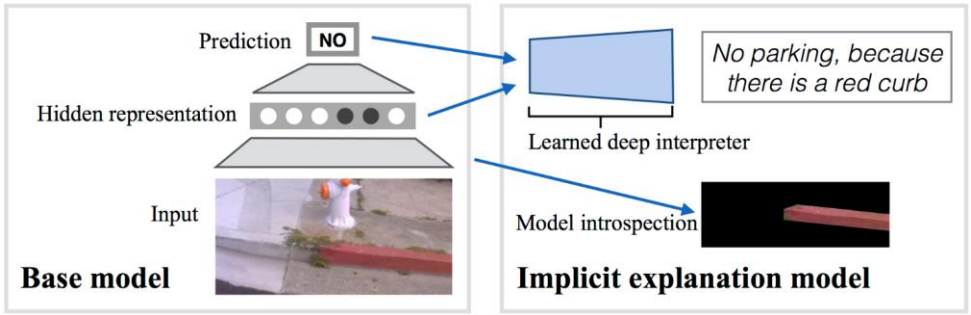
Major current research theme in BAIR: beyond supervised learning



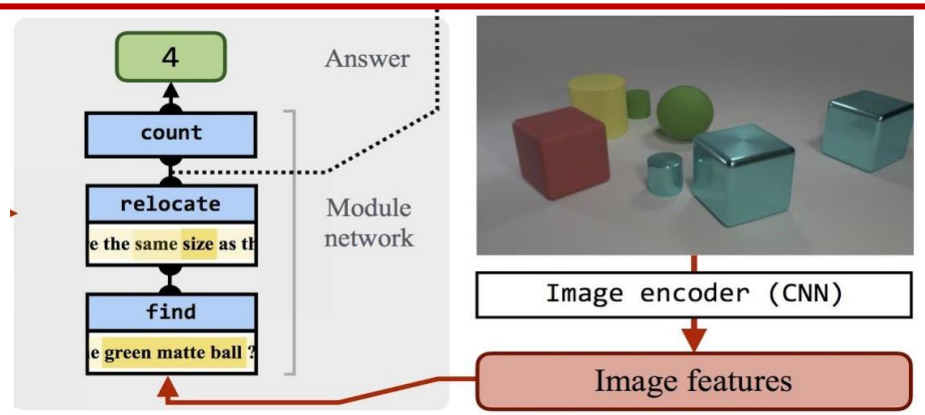
Adapting to new domains and tasks



Self-supervised / Curious learning



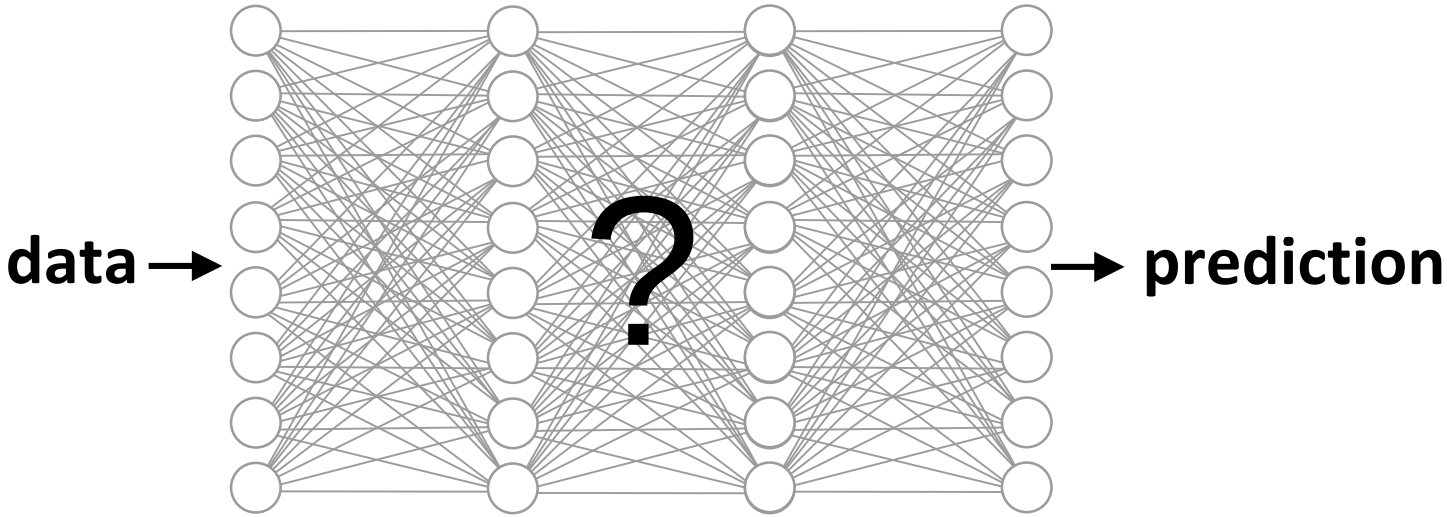
Learning to explain deep learning models



Reasoning over network structures

Explainable AI

Current deep neural networks (DNNs) are "**black boxes**"



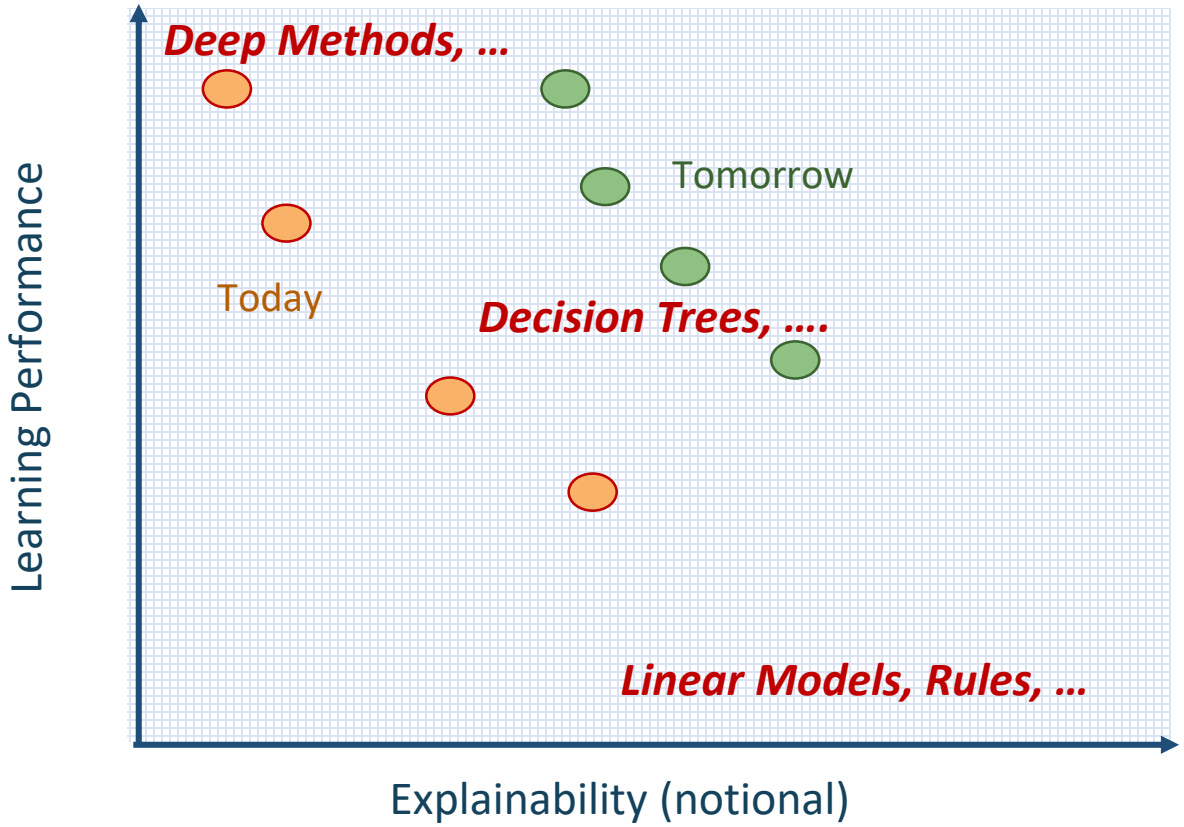
- Do not expose their **decision making** process
- Do not provide their **confidence** in their predictions
- Not clear whether they can be **trusted** and/or **corrected**

How do we make

Deep Learning

more **explainable** and **trustworthy**?

Performance vs Explainability / Interpretability?



(source: DARPA XAI)

XAI Perspectives and Goals

- **All-in for Deep Models**: “The solution to Interpretable Deep Learning is more Deep Learning”

data → ? → prediction

Post-hoc rationalization
“story-telling” models

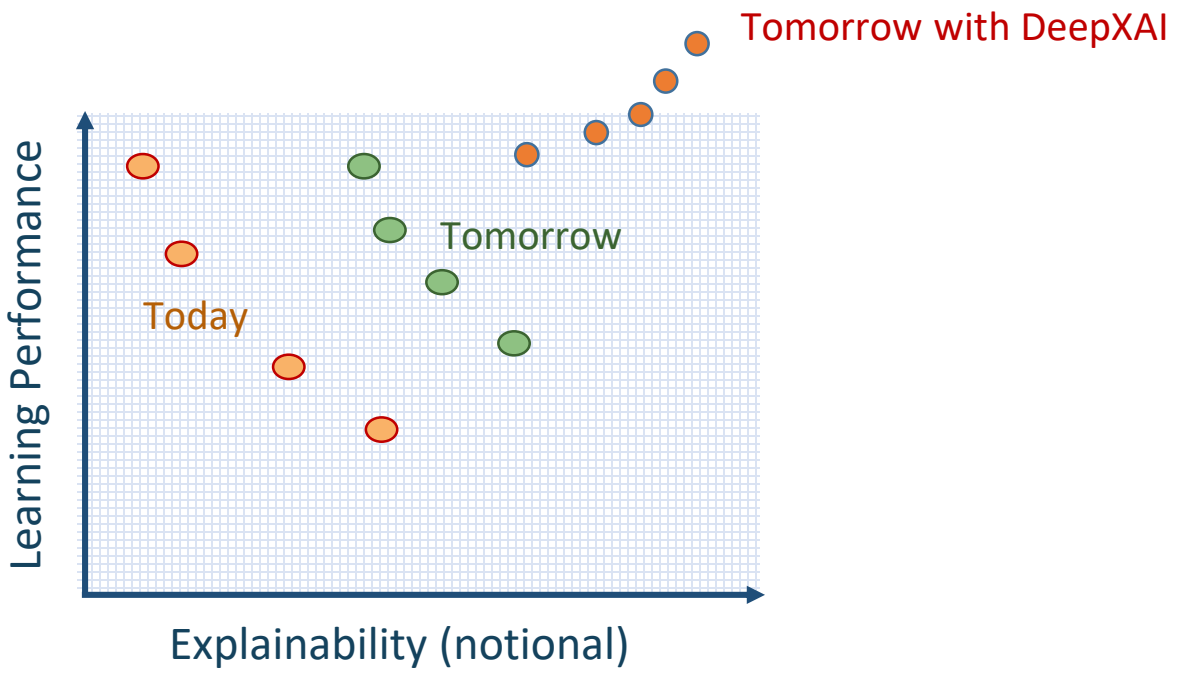
MODULARITY

Introspective /
causal models

JUSTIFICATION

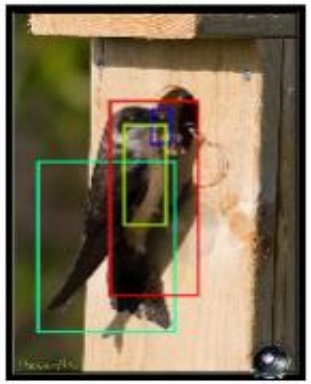
XAI Perspectives and Goals

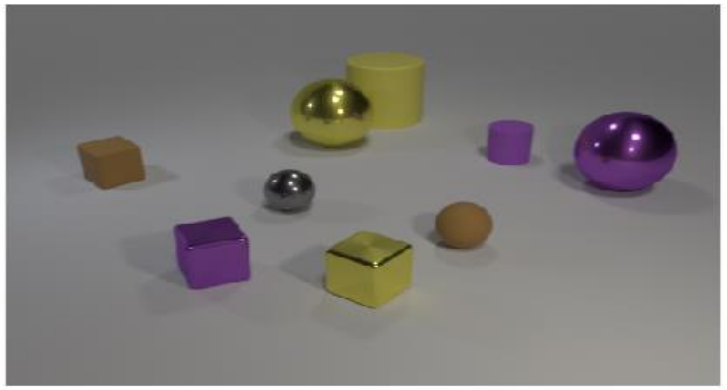
- **All-in for Deep Models**: “The solution to Interpretable Deep Learning is more Deep Learning”; via both **Modularity** and **Justification** paradigms…
- Invert the curve: Deep models may perform **better** when augmented with XAI losses

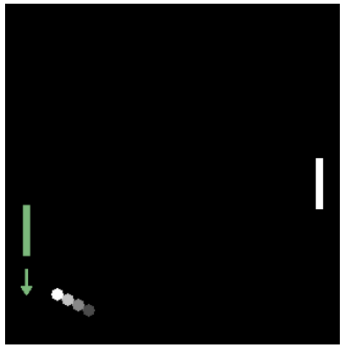


XAI Perspectives and Goals

- **All-in for Deep Models**: “The solution to Interpretable Deep Learning is more Deep Learning”; via both **Modularity** and **Justification** paradigms
- Invert the curve: Deep models may perform **better** when augmented with XAI losses
- Target domains include **image / video analysis, game/vehicle control, strategic games**

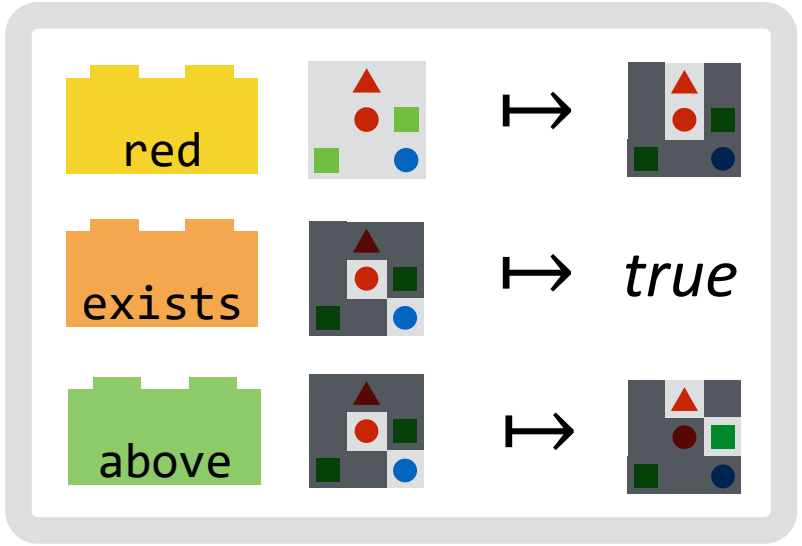


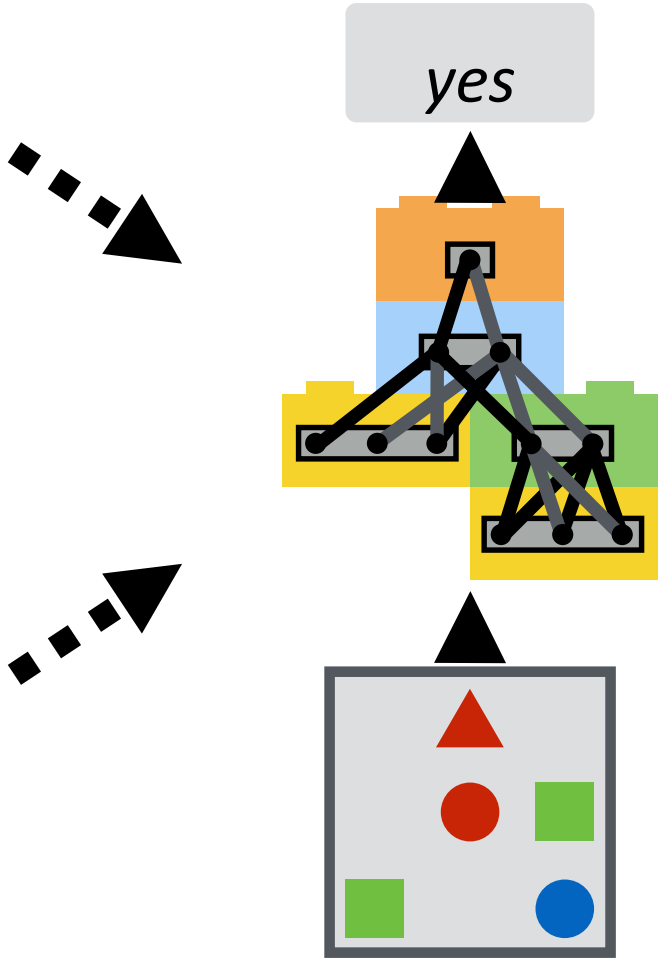




Neural module networks

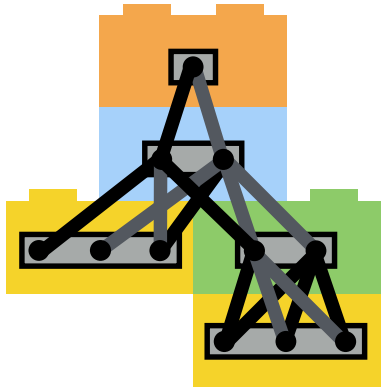
Is there a red shape above a circle?





Neural module networks

Linguistic or Task structure dynamically generates model structure

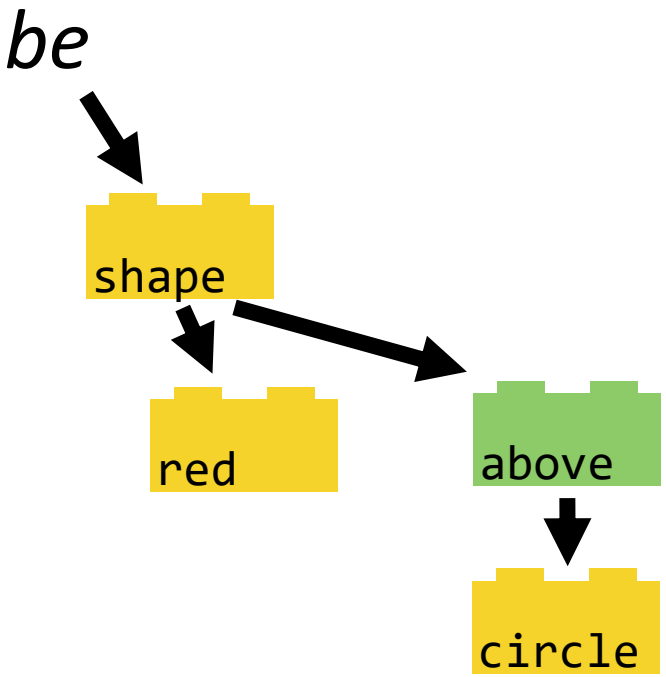


Combines advantages of:

- Representation learning (like a neural net)
- Compositionality (like a semantic parser or planner / reasoning component)

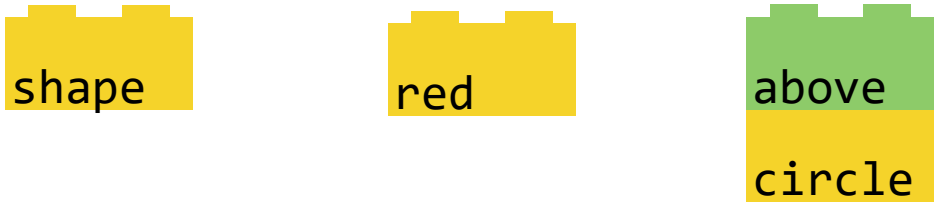
Where do layouts come from?

Is there a red shape above a circle?



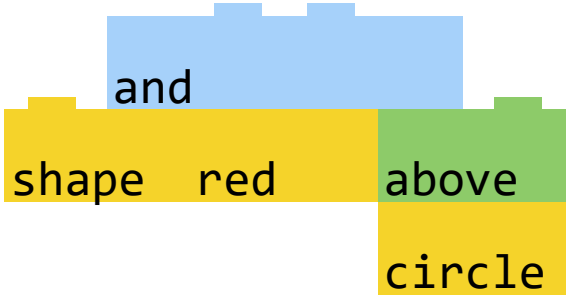
Where do layouts come from?

Is there a red shape above a circle?



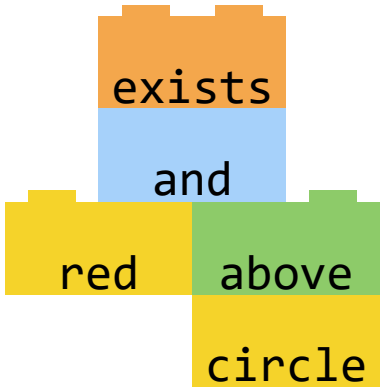
Where do layouts come from?

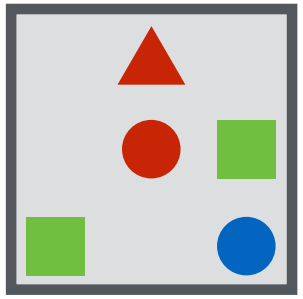
Is there a red shape above a circle?



Learning

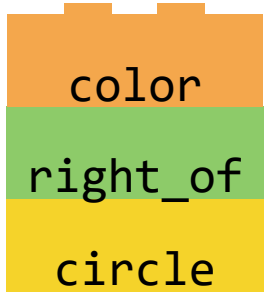
yes

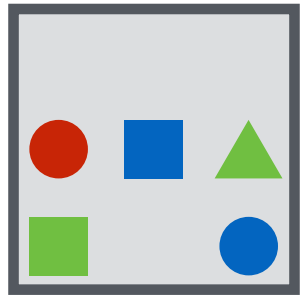




Is there a red shape above a circle?

blue

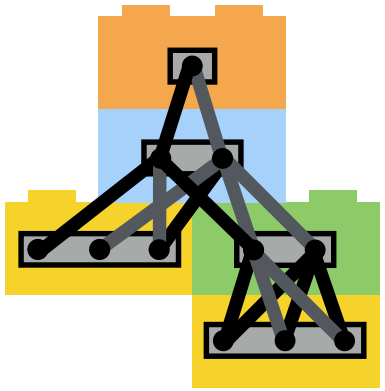


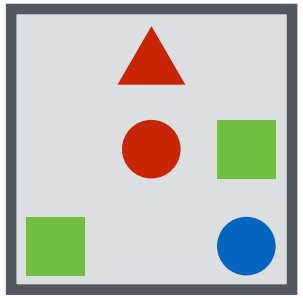


What color is the shape right of a circle?

Learning

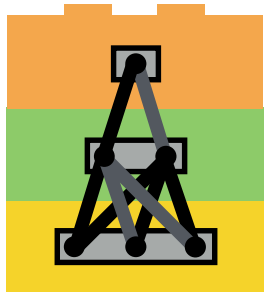
yes

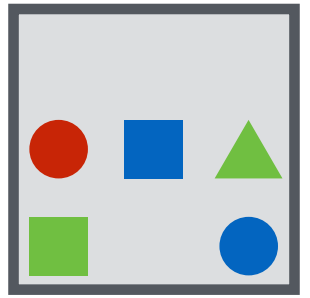




Is there a red shape above a circle?

blue

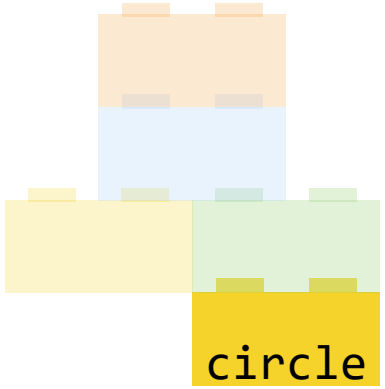


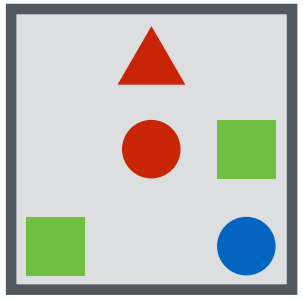


What color is the shape right of a circle?

Parameter tying

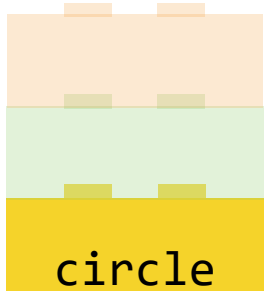
yes

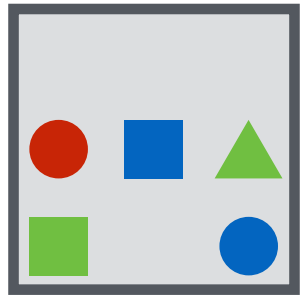




Is there a red shape above a circle?

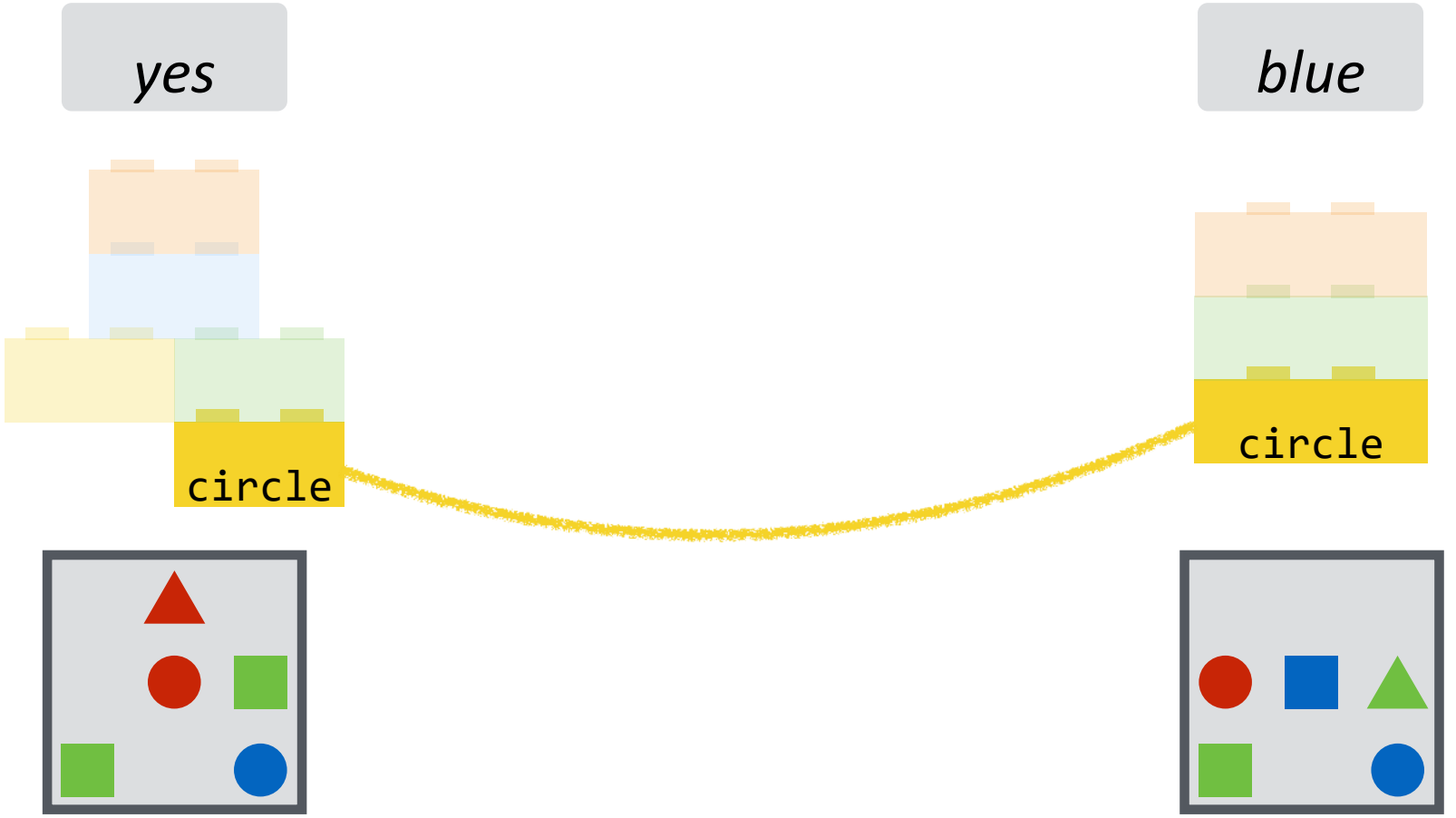
blue





What color is the shape right of a circle?

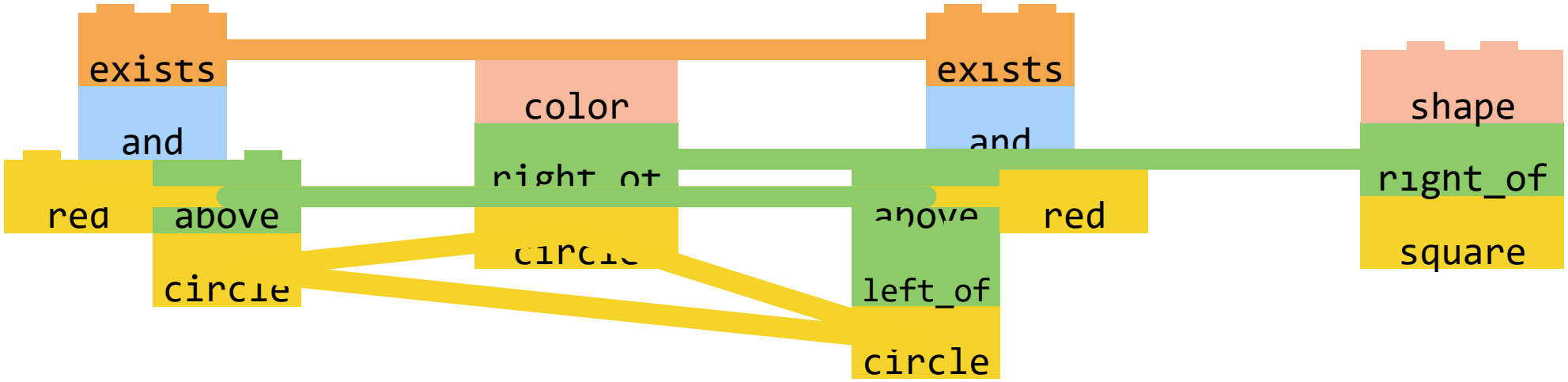
Parameter tying



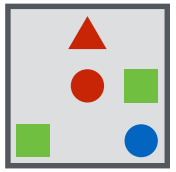
Is there a red shape above a circle?

What color is the shape right of a circle?

Extreme parameter tying



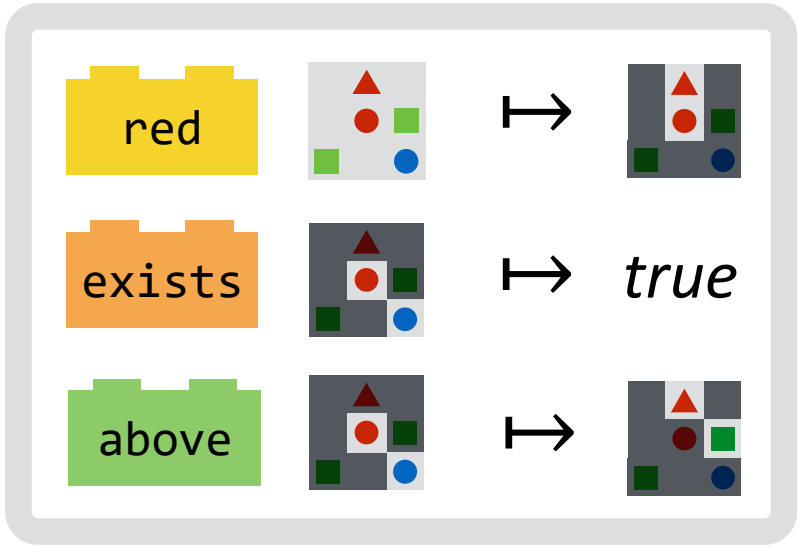
Learning with fixed layouts is easy!

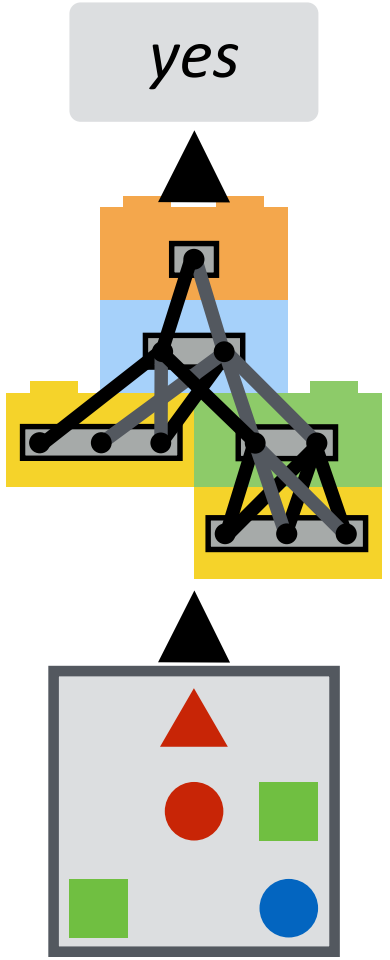
$$\arg\max_{W} \sum \mathrm{p}(\textit{yes} \mid \text{  }, \text{  }; W)$$

(where every root module outputs a distribution over answers and W is the set of all module parameters)

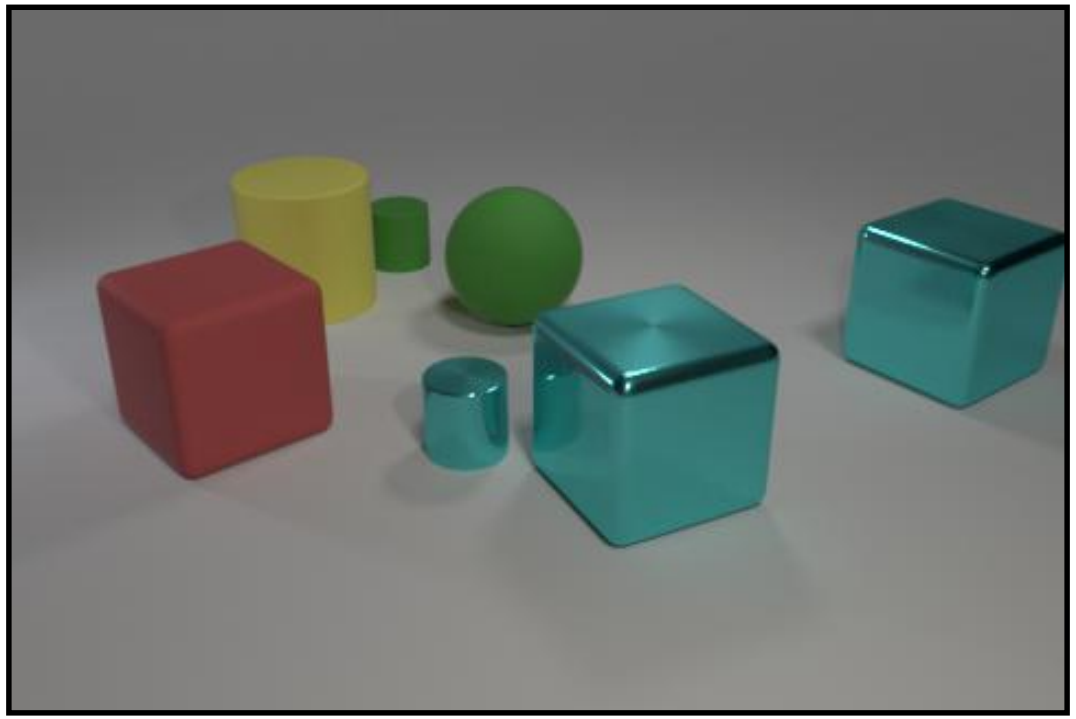
Neural module networks

Is there a red shape above a circle?



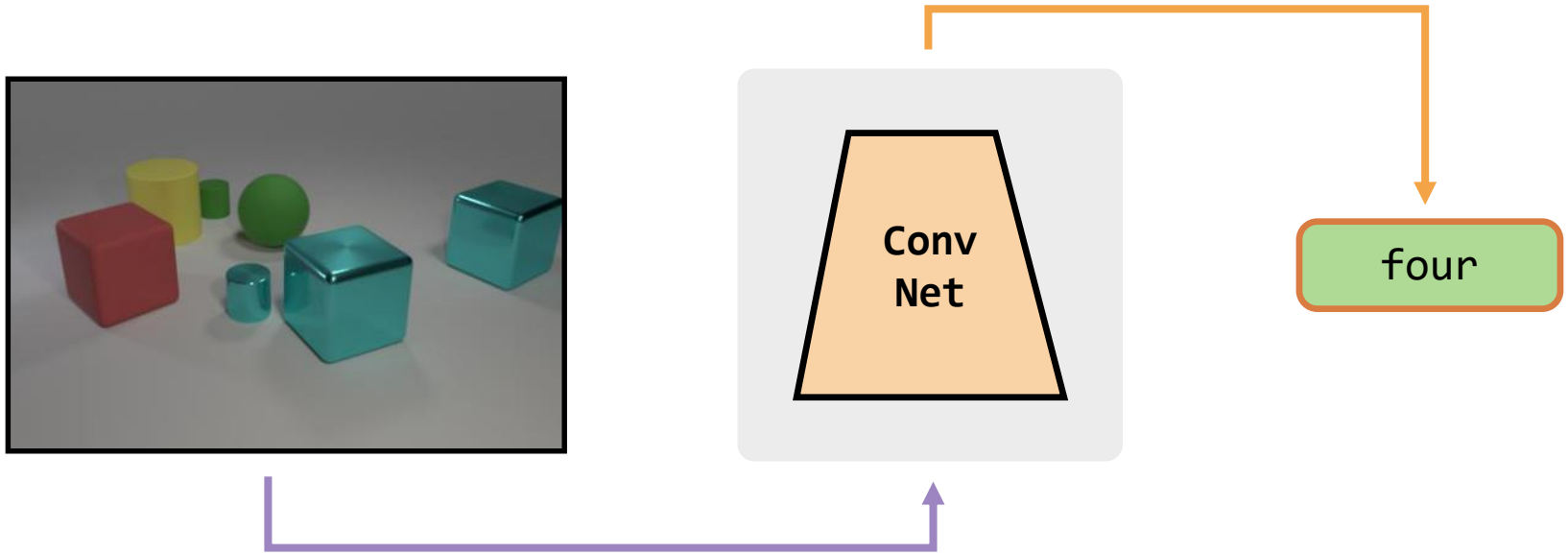


How many other things are of the same size as the green matte ball?



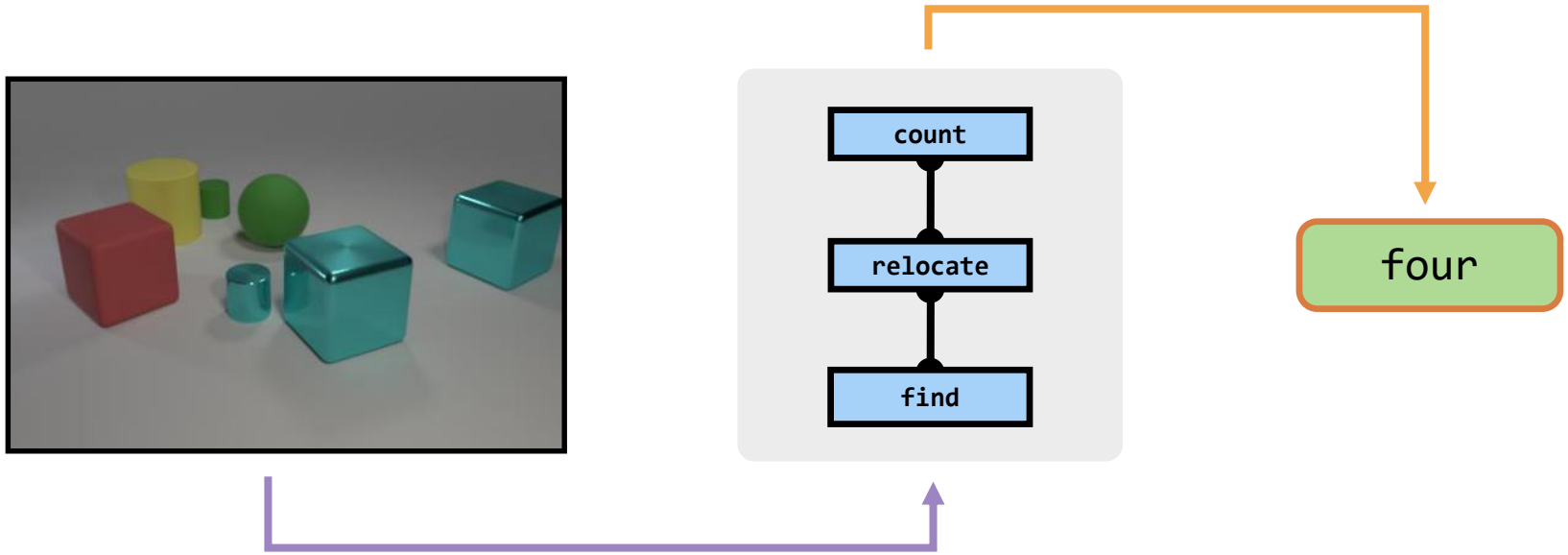
This video demonstrates end-to-end module networks,
which jointly learn to construct and execute deep networks
for visual problem-solving tasks.

How many other things are of the same size as the green matte ball?



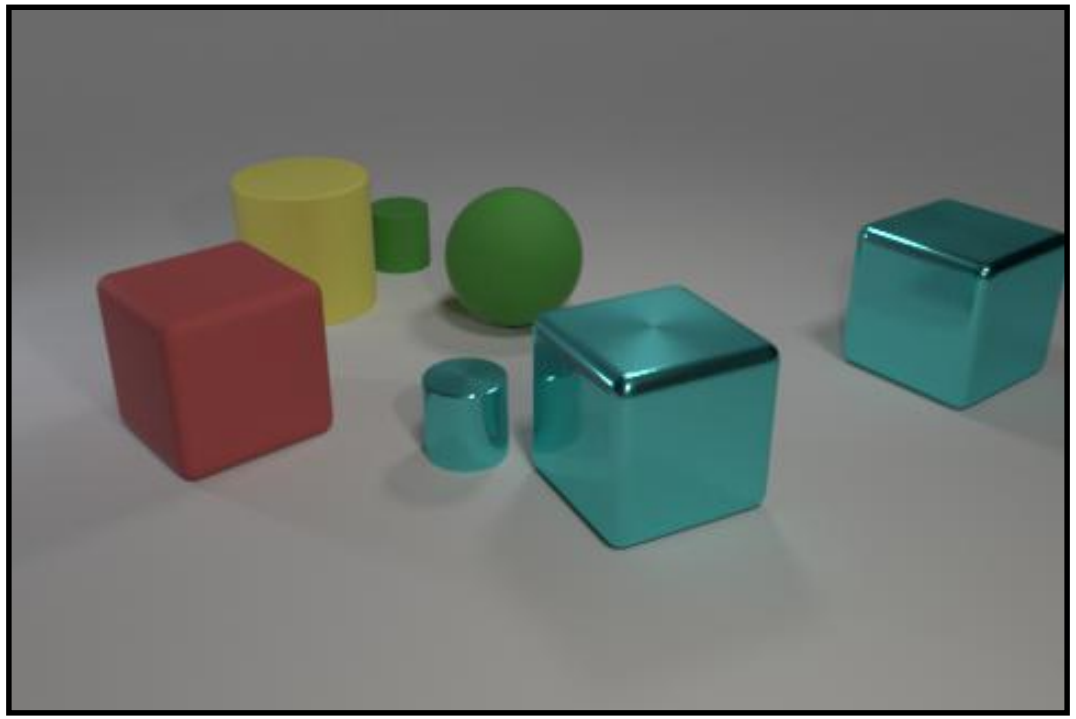
Ordinary neural nets struggle with these kinds of problems because they have a hard time representing the many distinct computational structures required by different kinds of questions.

How many other things are of the same size as the green matte ball?

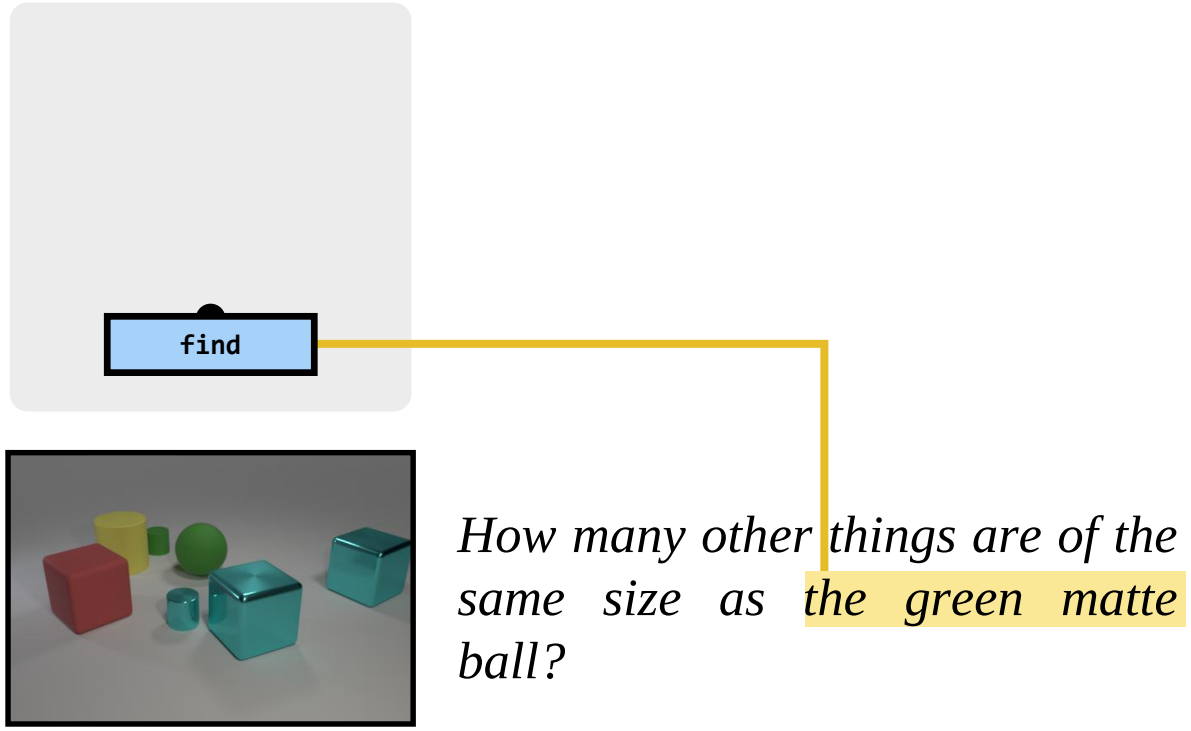


At a high-level, end-to-end module networks work
by building up question-specific neural networks on the fly
and executing them to obtain an answer.

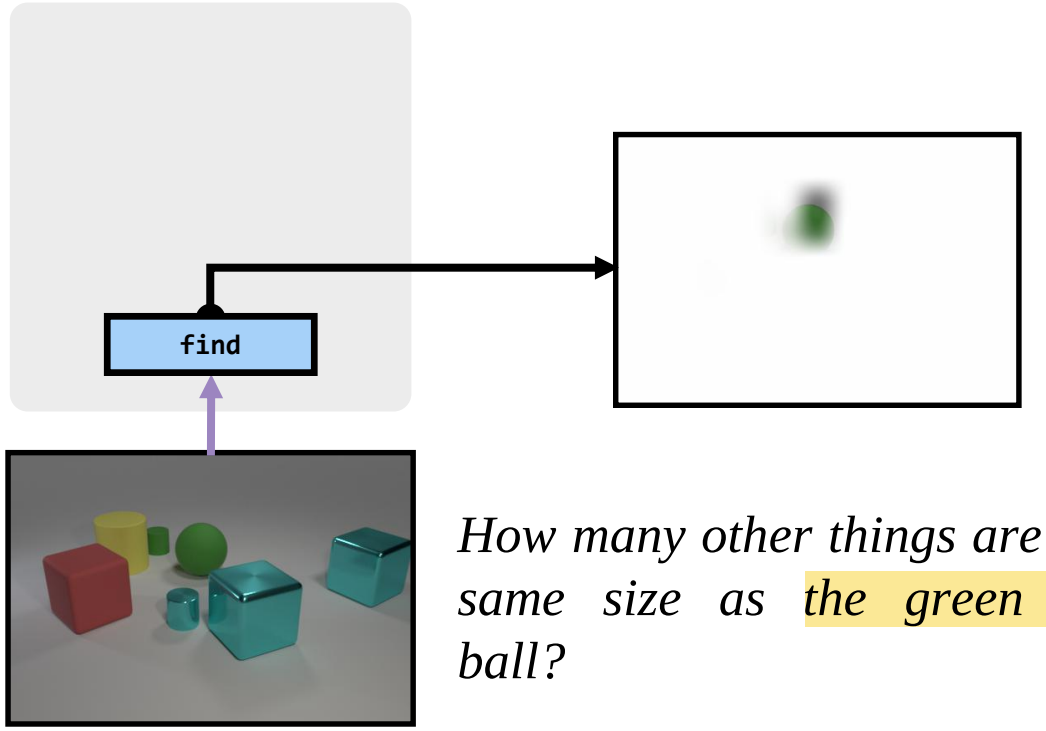
How many other things are of the same size as the green matte ball?



Let's see how a module network handles the problem above.

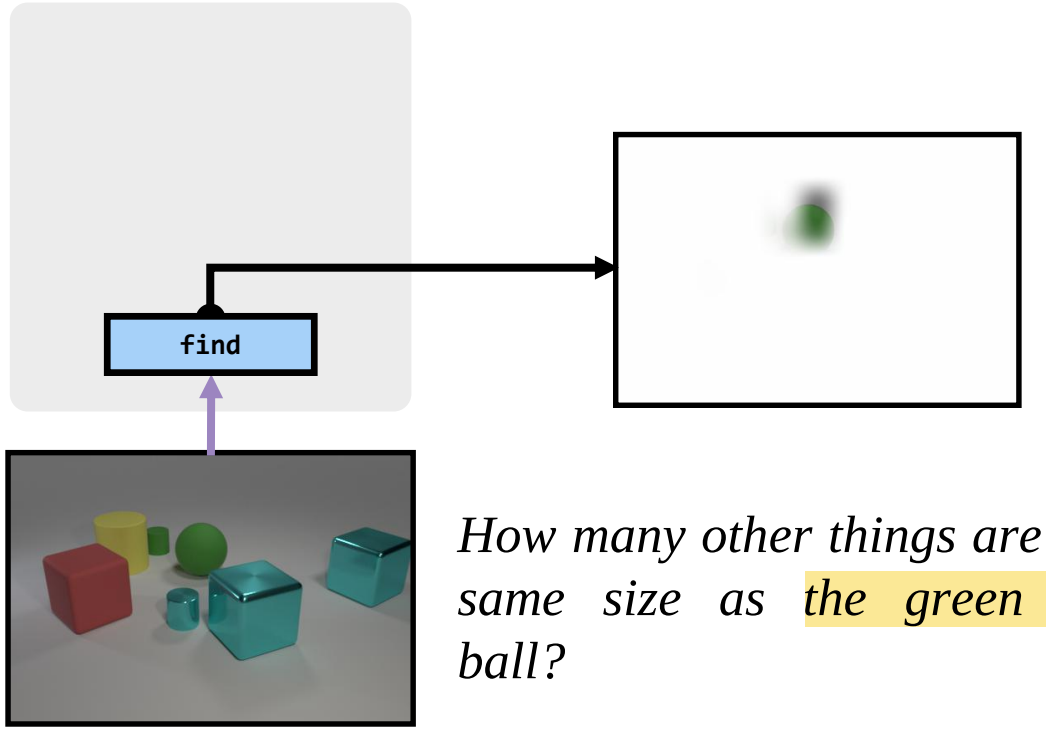


First, it constructs a **`find`** module
and parameterizes it to search for a green matte ball.



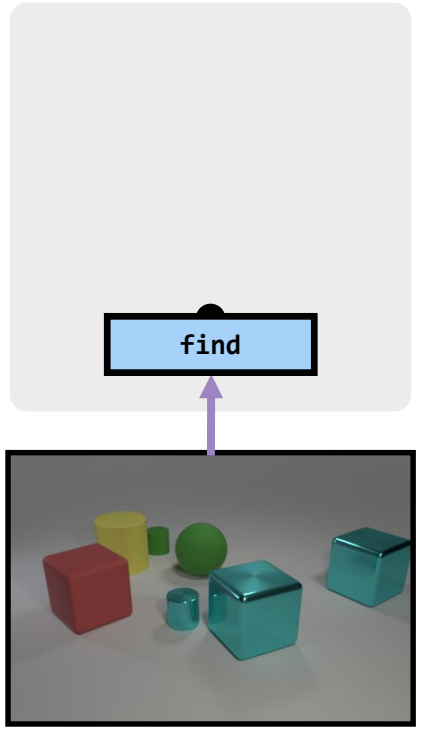
How many other things are of the same size as the green matte ball?

The output of this **`find`** module is an attention



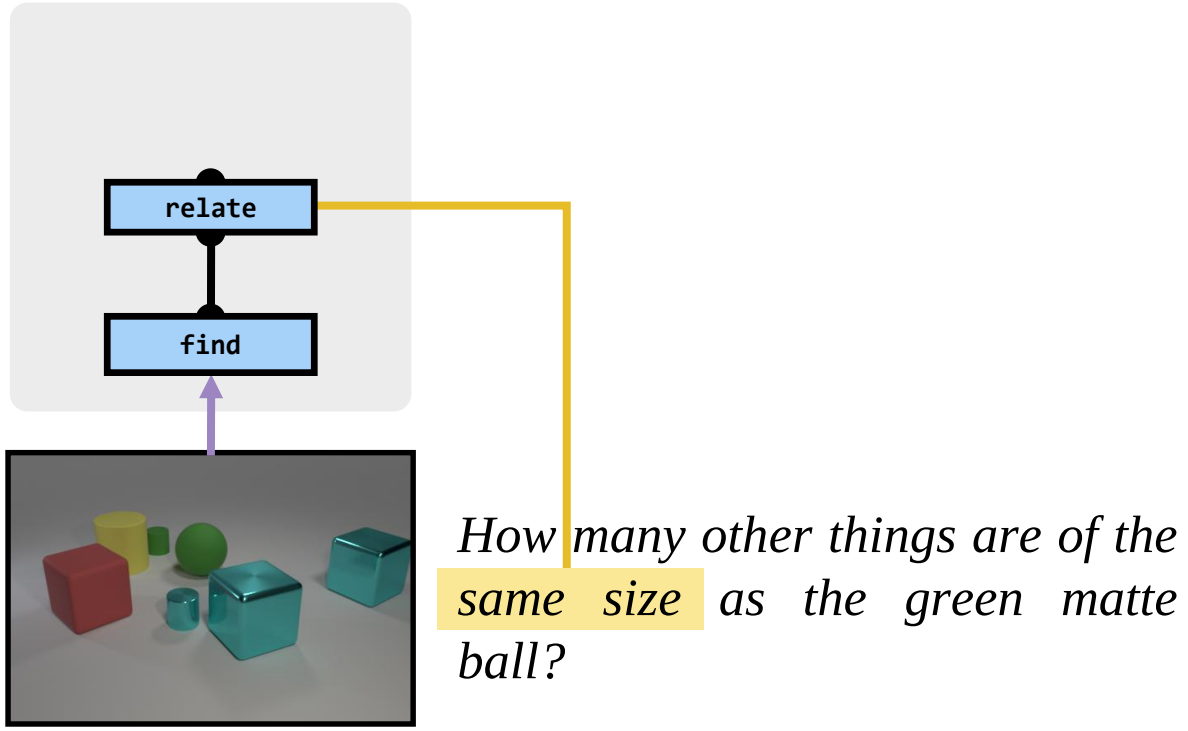
How many other things are of the same size as the green matte ball?

Here we can see that the module focuses on the right object.

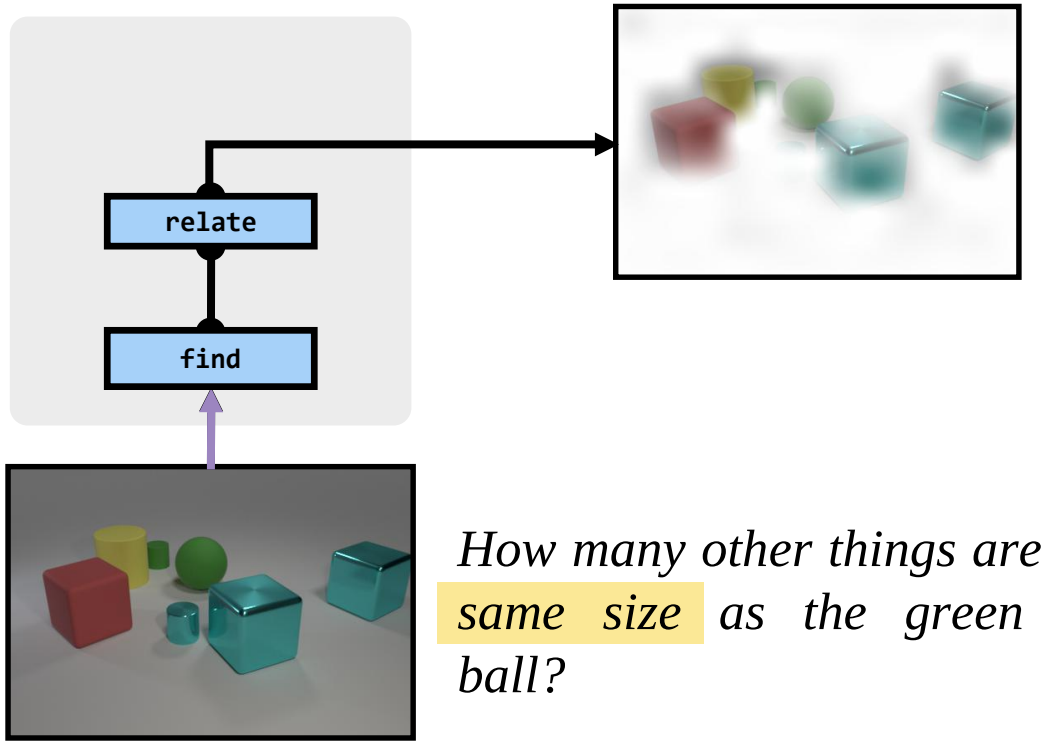


How many other things are of the same size as the green matte ball?

The output of the **`find`** module is stored in memory.

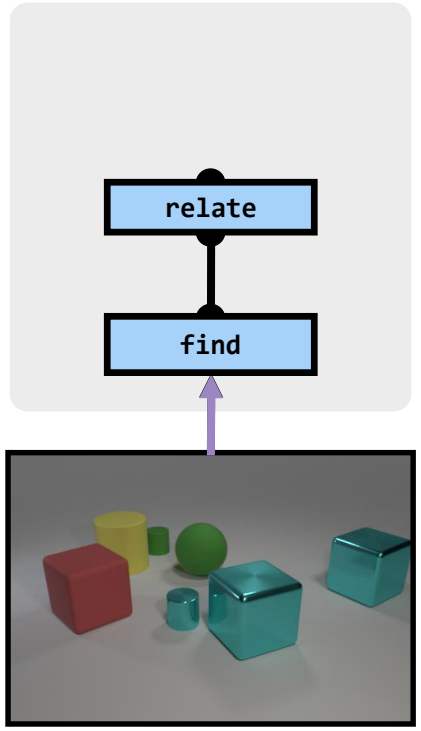


Next, the model constructs a **`relate`** module, which uses the output of the **`find`** module to identify related regions of the image:



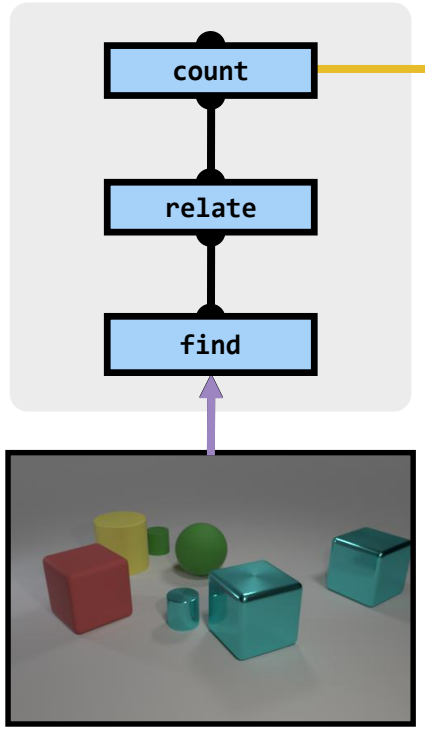
How many other things are of the same size as the green matte ball?

in this case, objects with the same size.



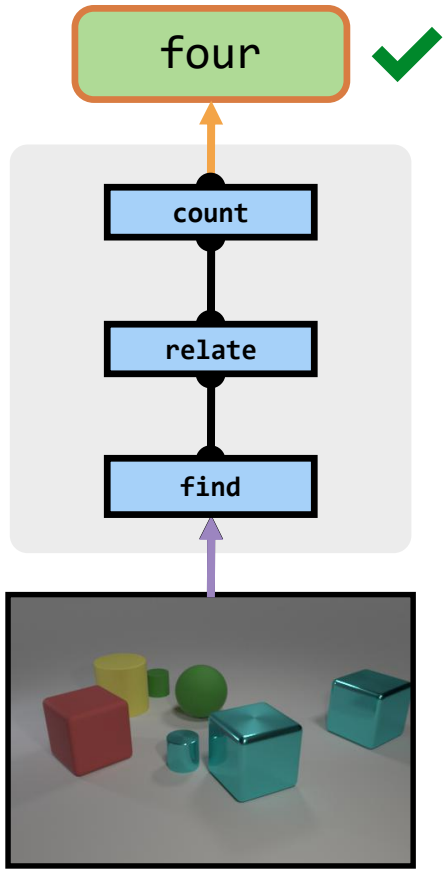
How many other things are of the same size as the green matte ball?

in this case, objects with the same size.



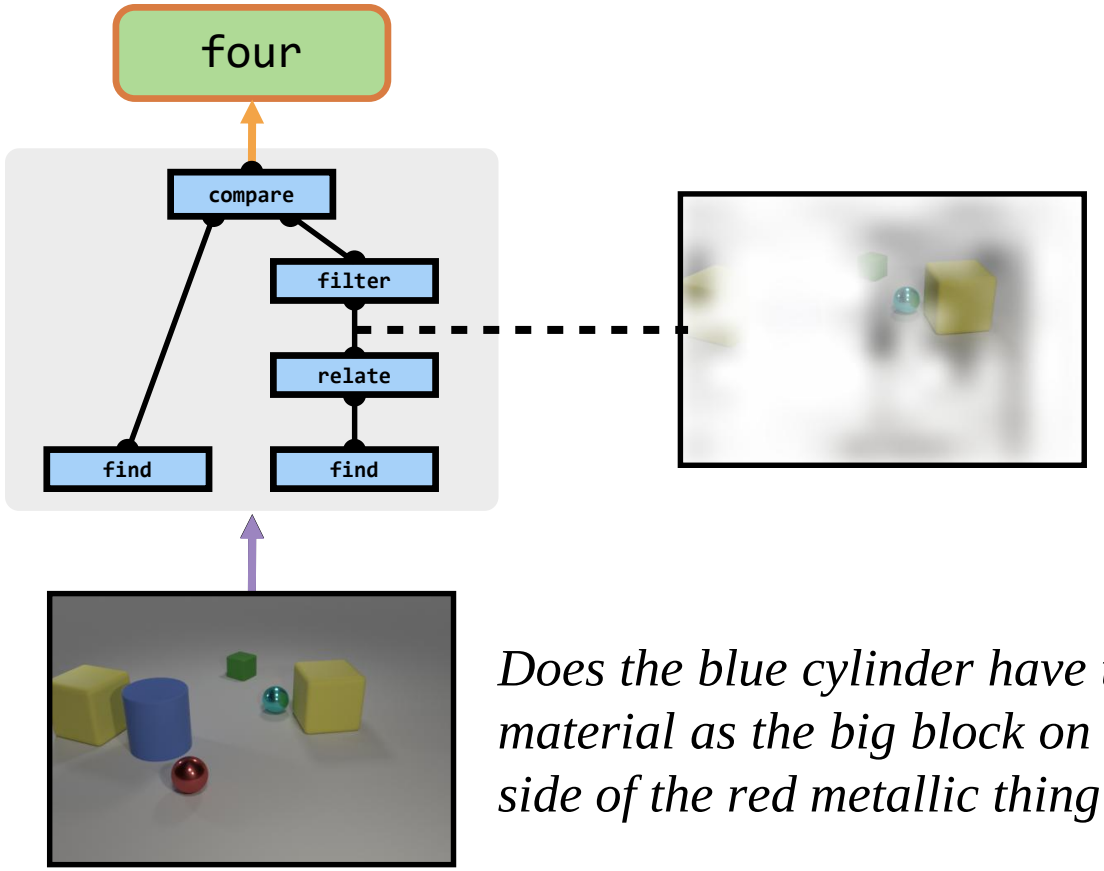
How many other things are of the same size as the green matte ball?

Finally, the model constructs a **`count`** module,
which counts the number of other objects attended to.



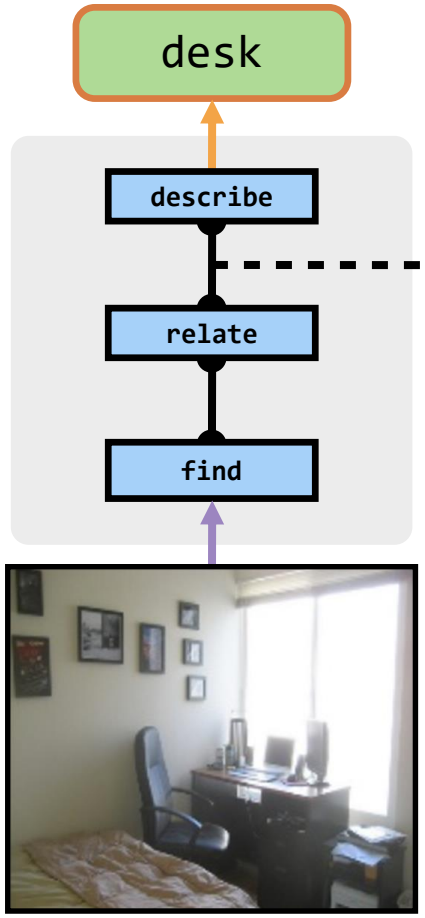
How many other things are of the same size as the green matte ball?

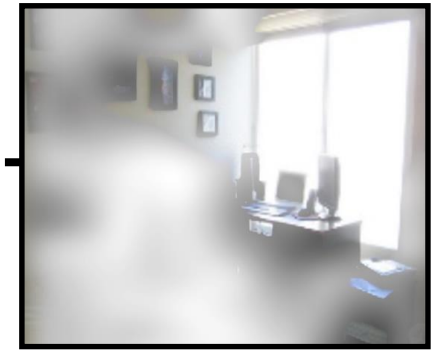
Here it has correctly answered the question.



Does the blue cylinder have the same material as the big block on the right side of the red metallic thing?

Our approach also supports more complex questions,





What is behind the foot of the bed?

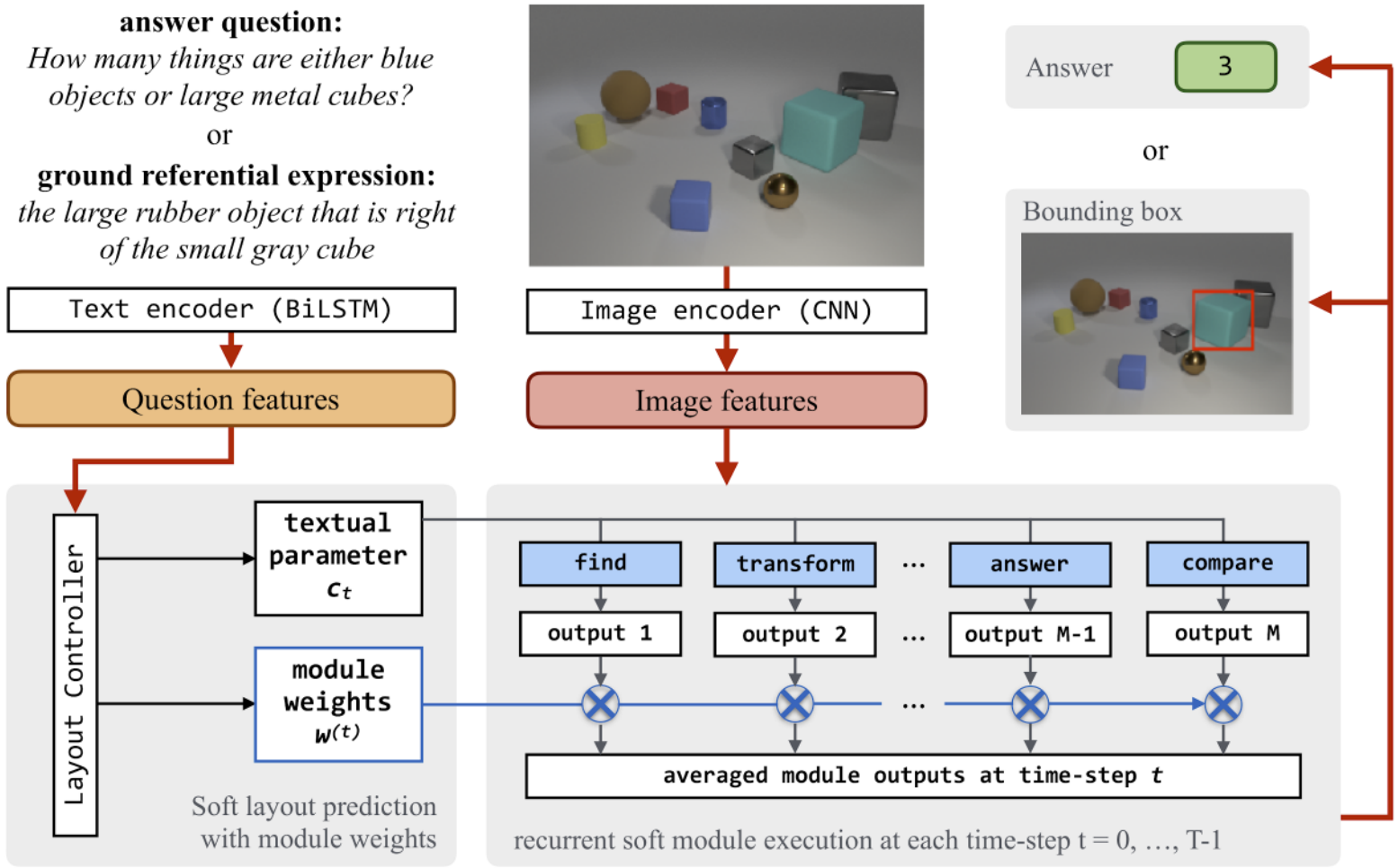
as well as natural images.

Neural module networks with soft layout - Model

Differentiable: replacing previous *discrete* execution graph with *continuous* soft layout (via module weights), not requiring "expert layout" supervision or RL.

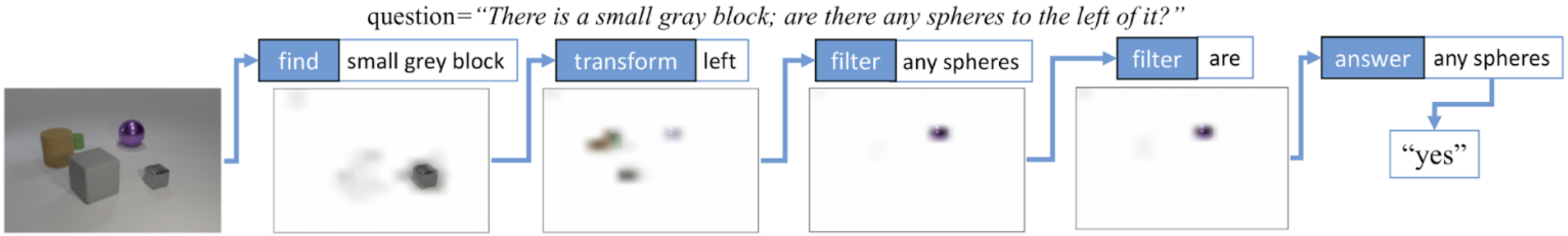
Interpretable as humans can understand its reasoning steps and detect its failure.

Multi-task by sharing a common set of sub-tasks (modules).

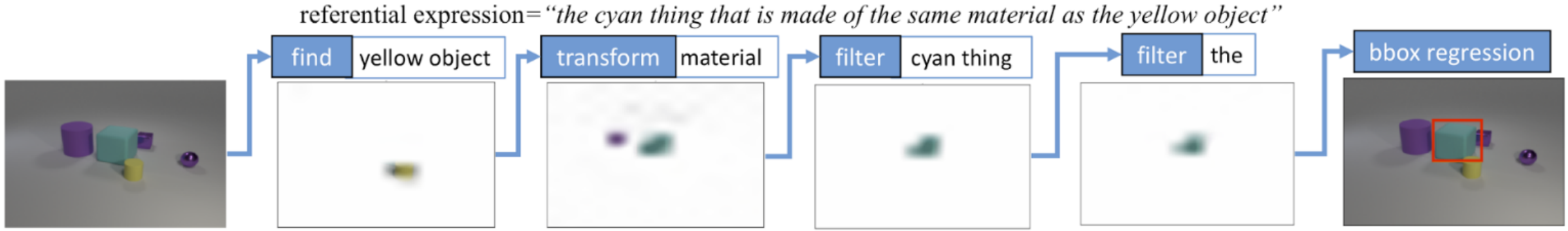


Neural module networks with soft layout - Examples

Example predictions on Visual Question Answering (VQA)



Example predictions on Referential Expression Grounding (REF)



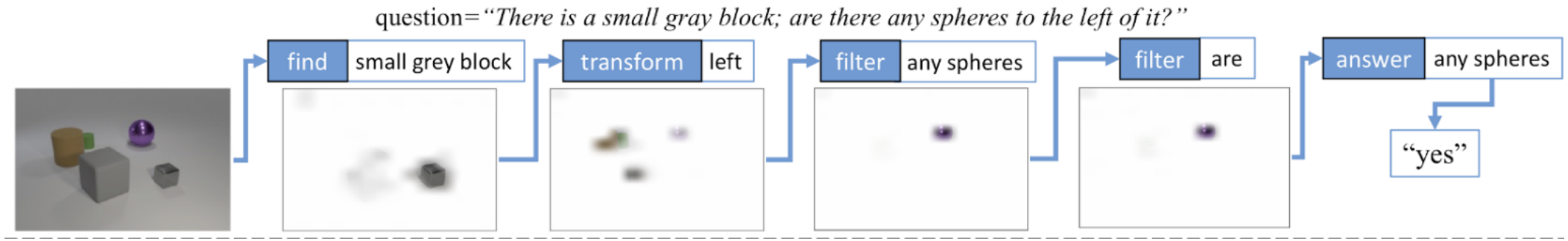
Accuracy evaluation

Our method does not require “expert layout” supervision.
It almost closes the gap of between whether having “expert layout”.

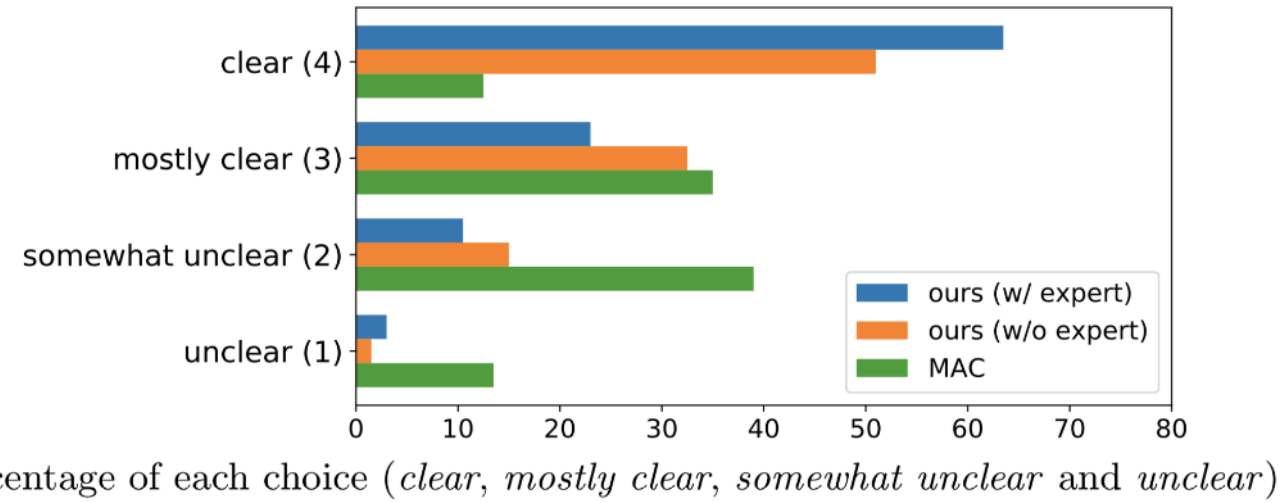
Method	expert layout	VQA accuracy
N2NMN [12]	yes	83.7
PG+EE [18]	yes	**96.9**
Ours	yes	96.0
Ours (+REF)	yes	96.1
N2NMN [12]	no	69.0
PG+EE [18]	no	(does not converge)
Ours	no	93.8
Ours (+REF)	no	**94.1**

Interpretability evaluation - subjective understanding

We let human users judge (from the image and text attentions) whether the internal computation is clear to them.
Our model is much more often rated as "clear".

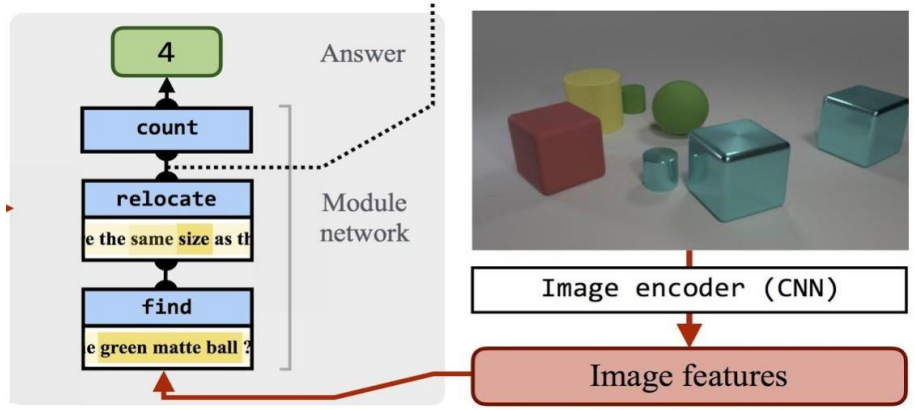


Question:
Are the internal reasoning steps above clear and understandable to you?

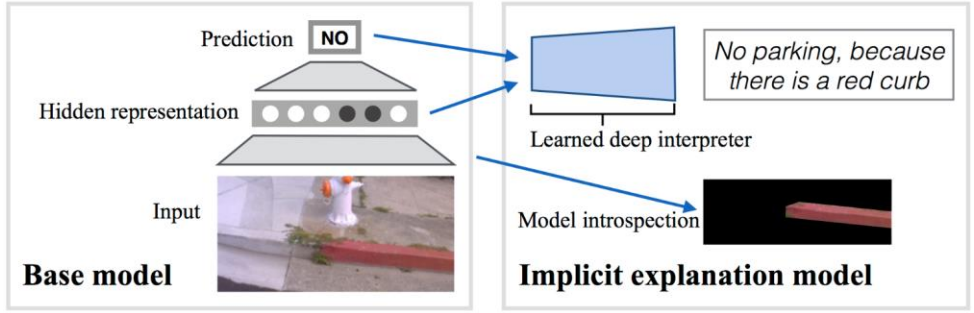


Deep Explanation Models

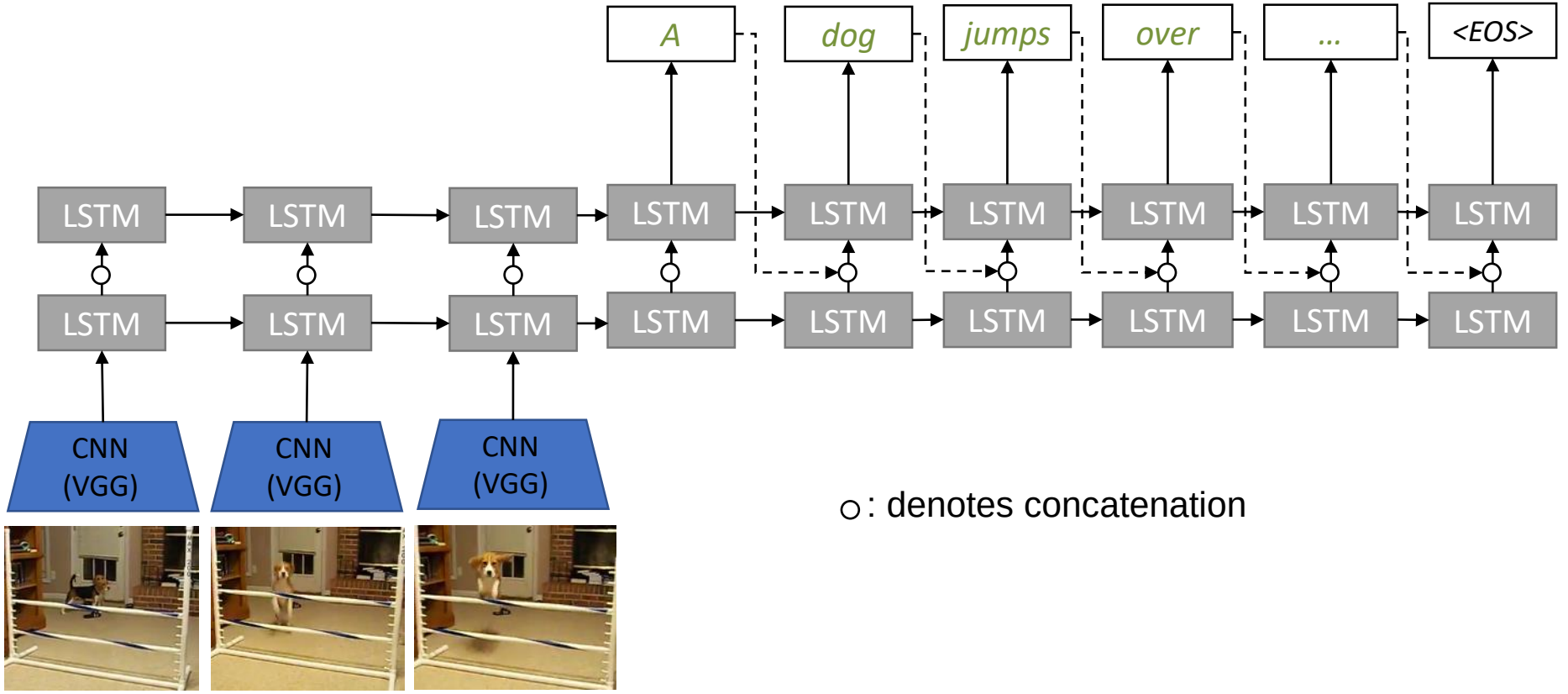
Explicit / Introspective models: interpretable internal visualization



Implicit / Justification models: post-hoc rationalization

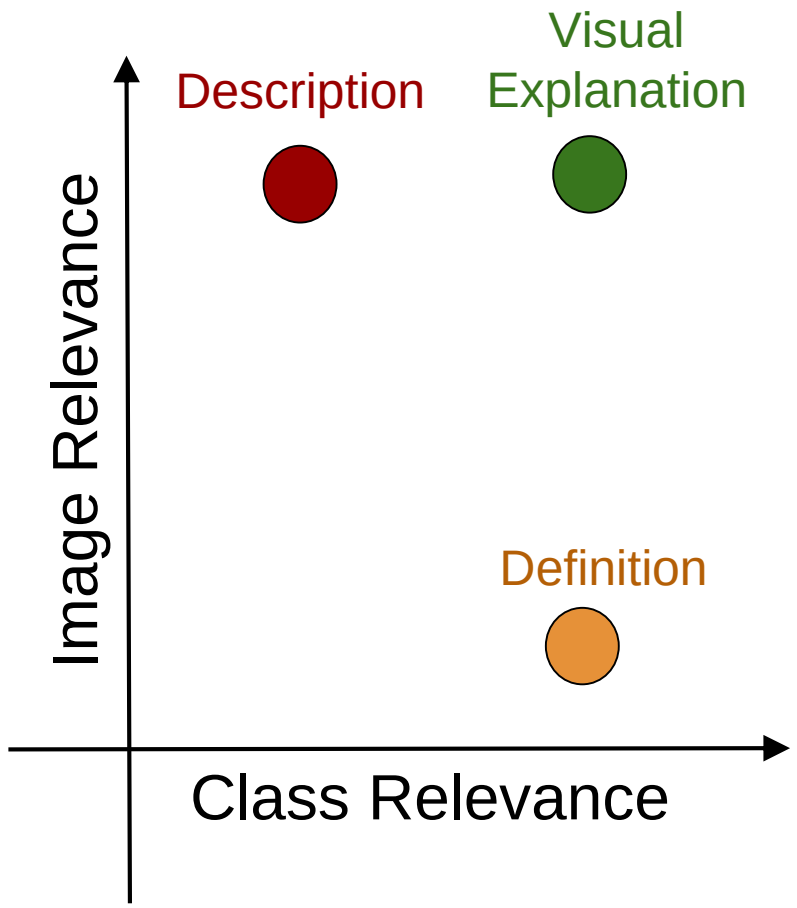


Past work: Image and Video Captions



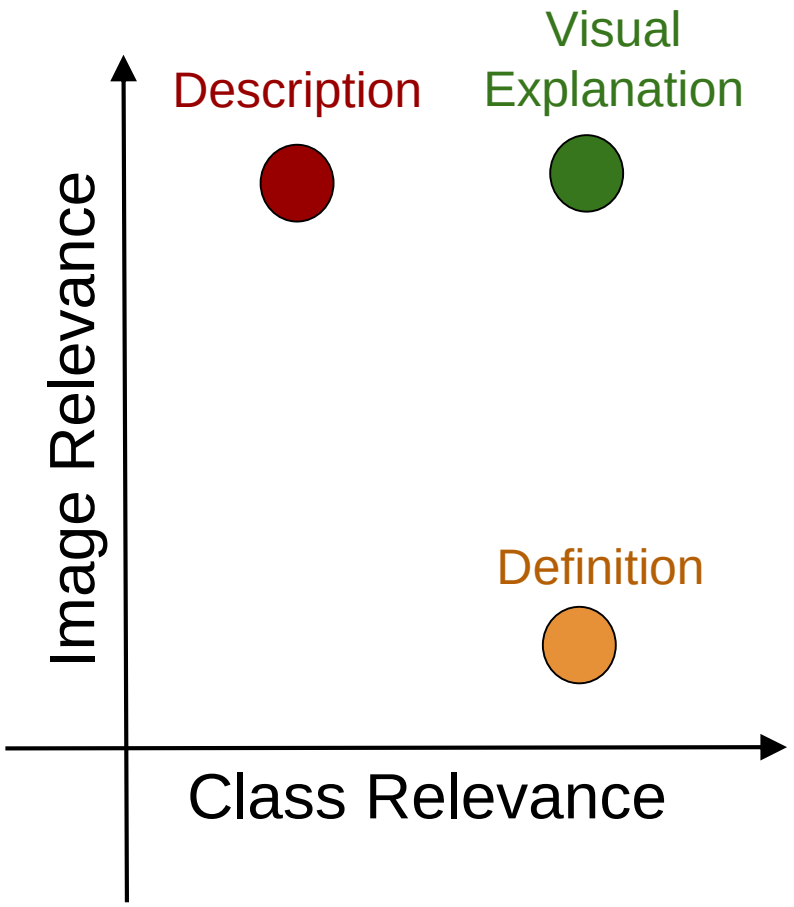
S. Venugopalan, M. Rohrbach, J Donahue, R Mooney, T Darrell, K Saenko;
Sequence to Sequence – Video to Text; **ICCV 2015**

From captions to explanations…



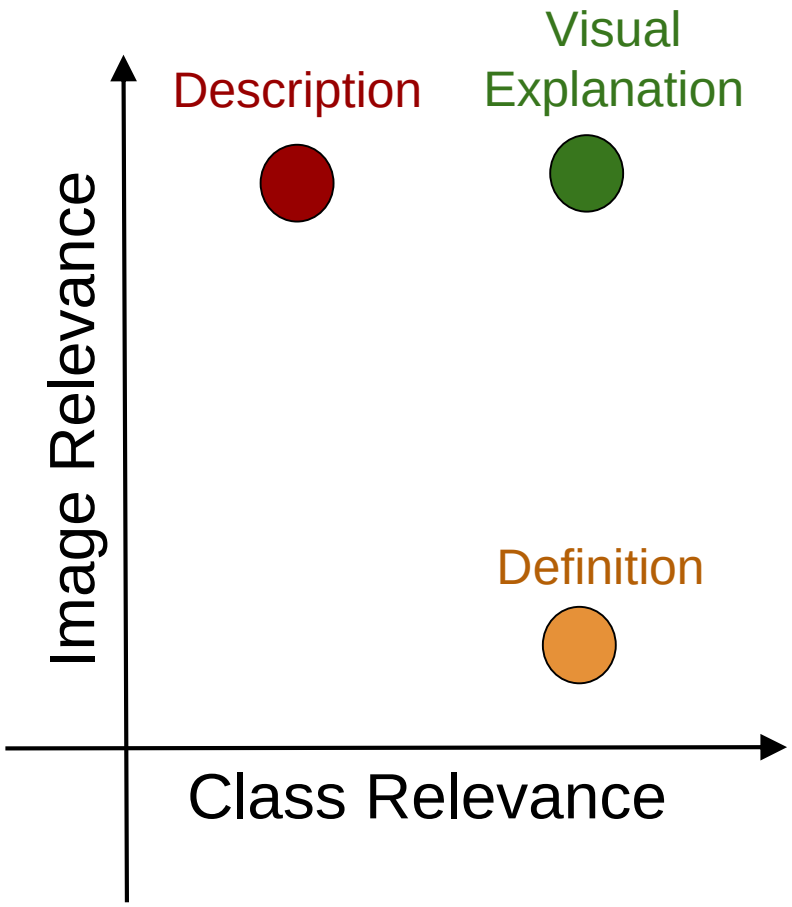
[Hendricks, Akata, Rohrbach, Donahue, Schiele, Darrell; ECCV 2016]

From captions to explanations…



[Hendricks, Akata, Rohrbach, Donahue, Schiele, Darrell; ECCV 2016]

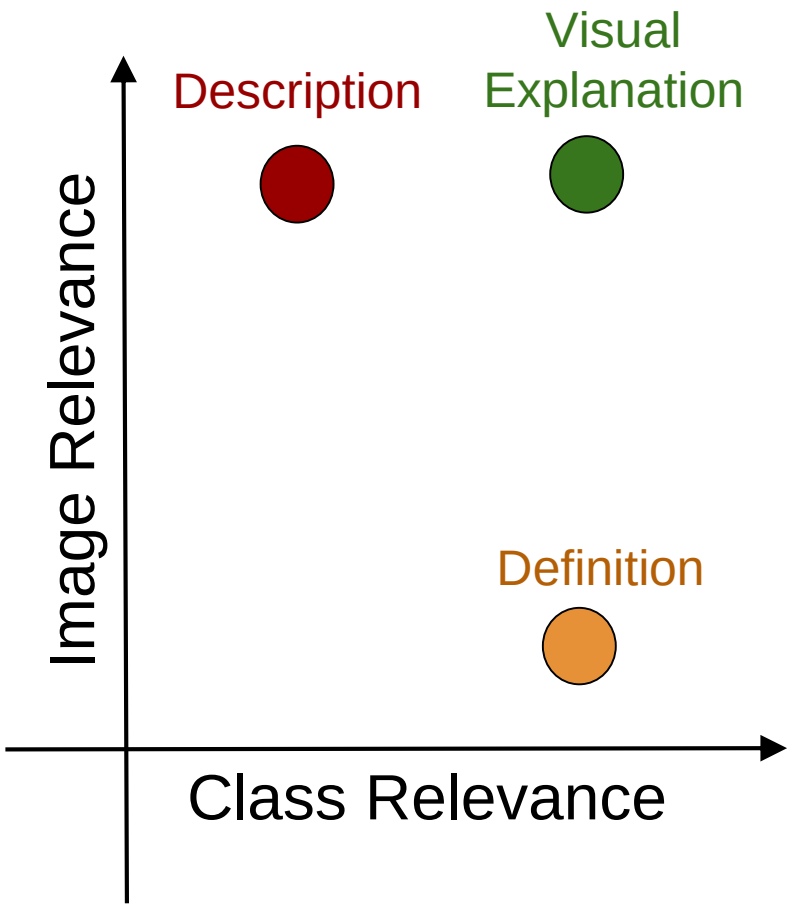
From captions to explanations…



A large bird with a **white neck** in the **water**.

[Hendricks, Akata, Rohrbach, Donahue, Schiele, Darrell; ECCV 2016]

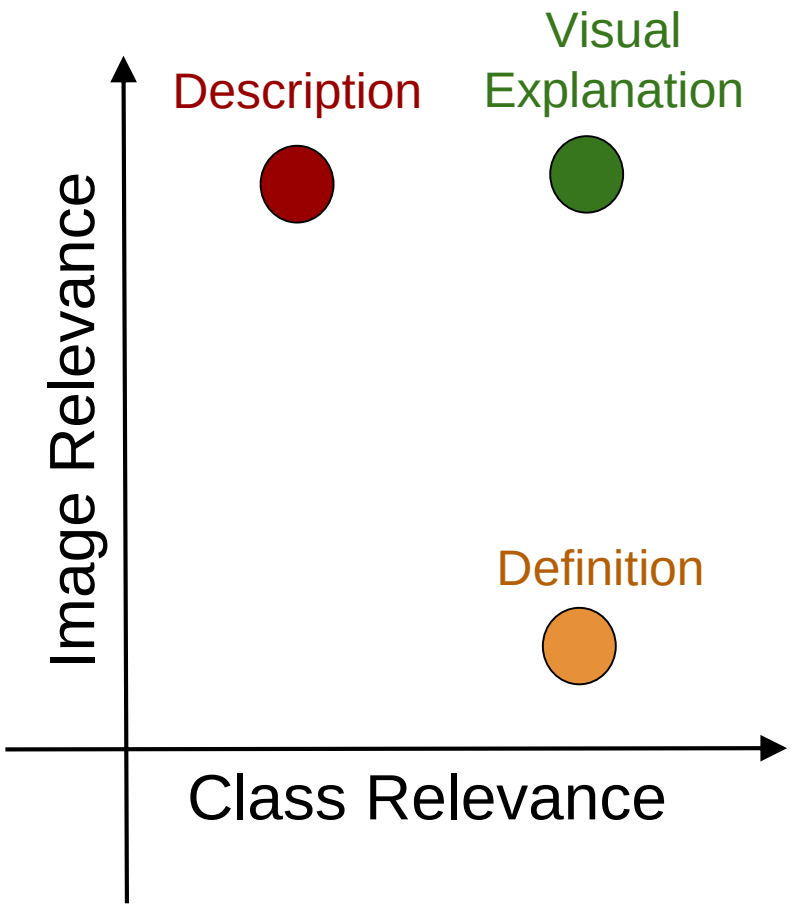
From captions to explanations…



A large bird with a **white neck** in the **water**.
Western Grebe has **yellow pointy beak**.

[Hendricks, Akata, Rohrbach, Donahue, Schiele, Darrell; ECCV 2016]

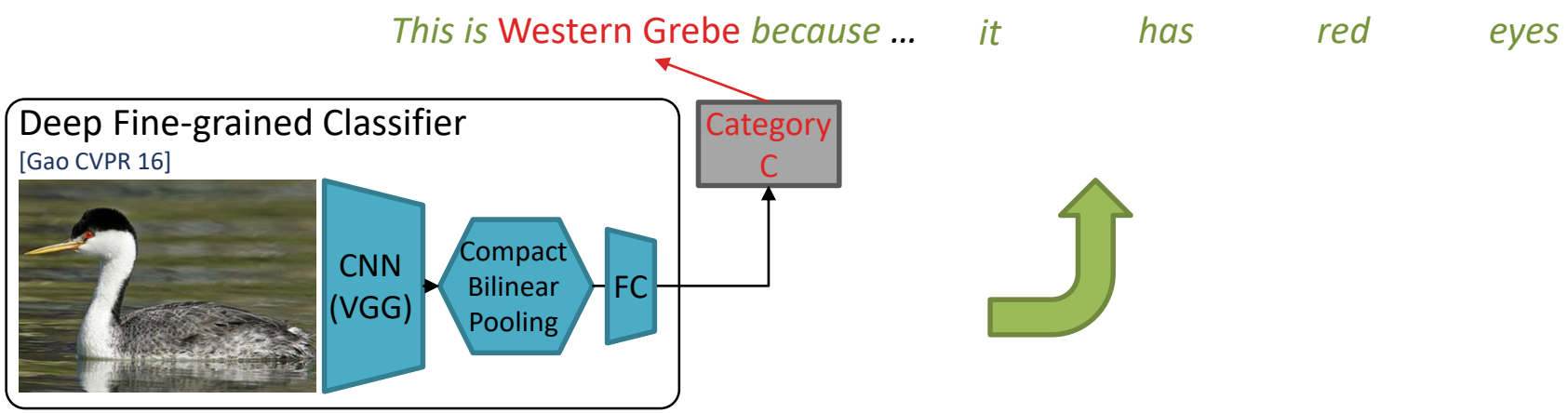
From captions to explanations…



A large bird with a **white neck** in the **water**.
Western Grebe has **yellow pointy beak**.
This is a *Western Grebe* because it has a **long white neck**, **pointy yellow beak** and **red eye**.

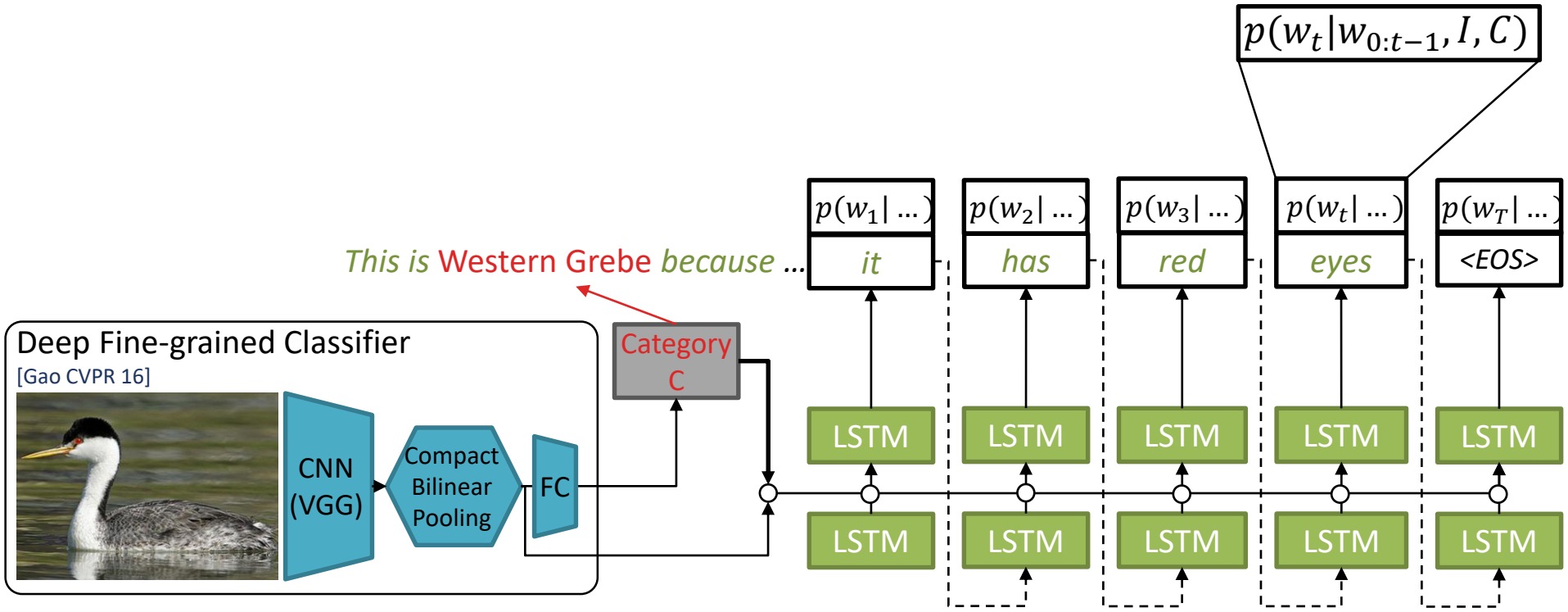
[Hendricks, Akata, Rohrbach, Donahue, Schiele, Darrell; ECCV 2016]

Visual Explanations – Test time



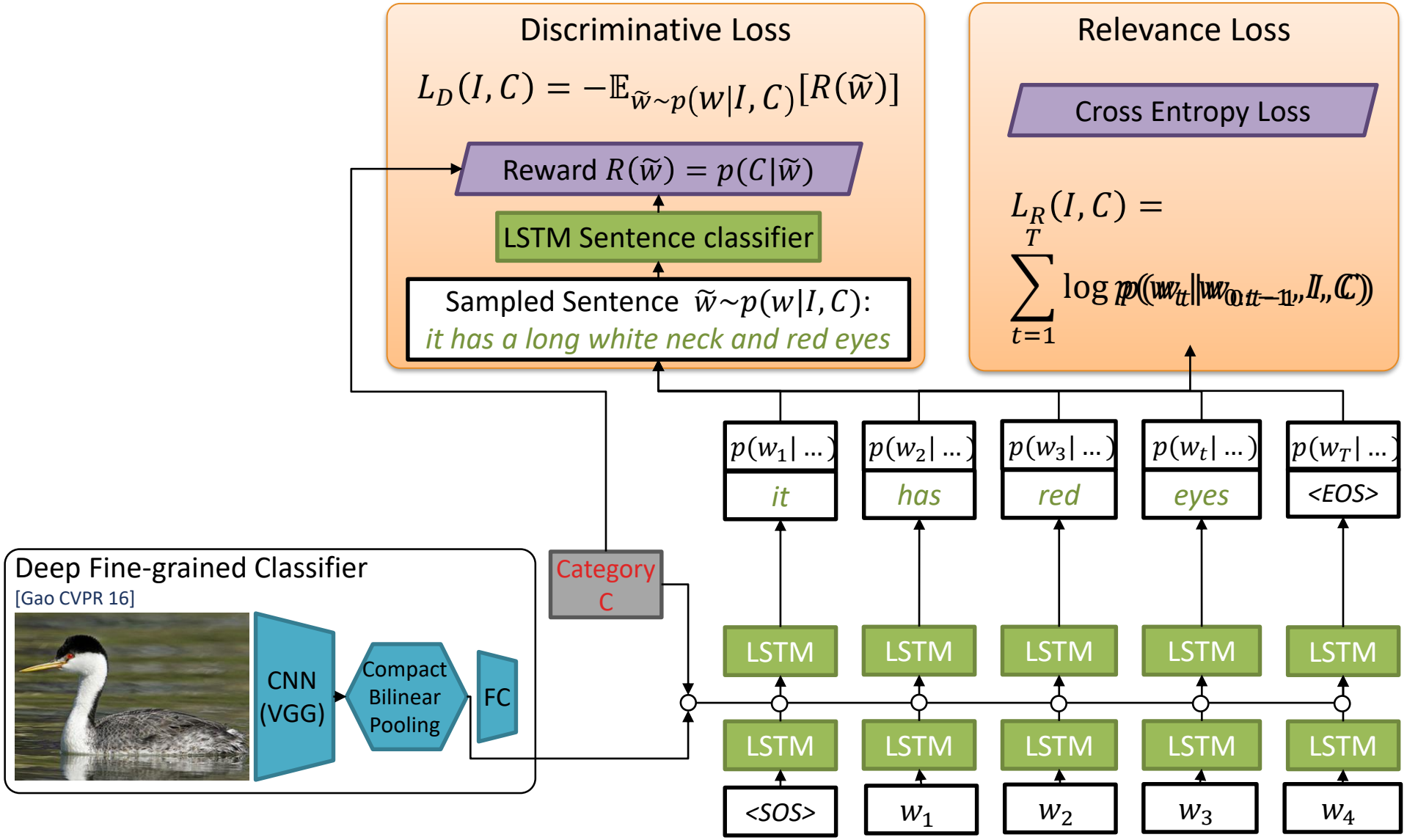
[Gao CVPR 16] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. CVPR 2016

Visual Explanations – Test time



[Gao CVPR 16] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. CVPR 2016

Visual Explanations – Training time



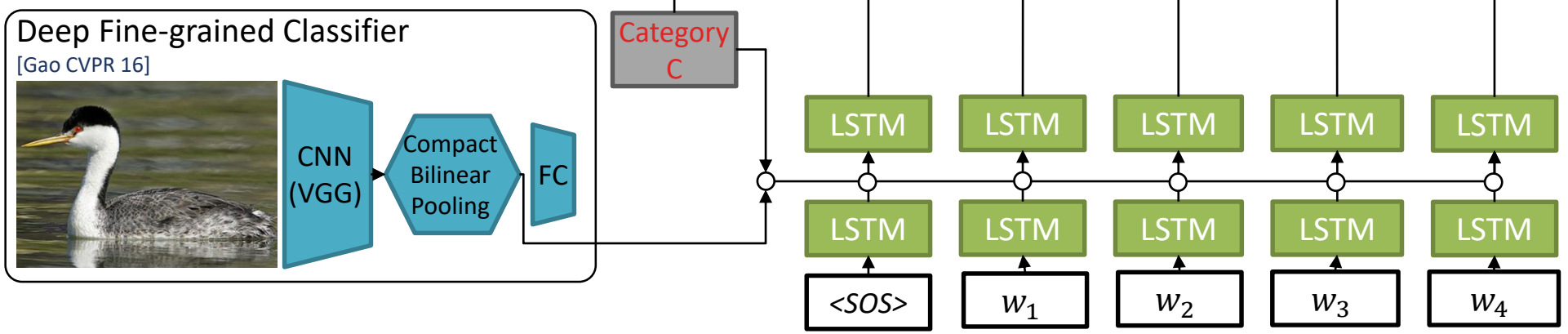
Visual Explanations – Training time

Backprop through sampling:

REINFORCE [Williams 92]:

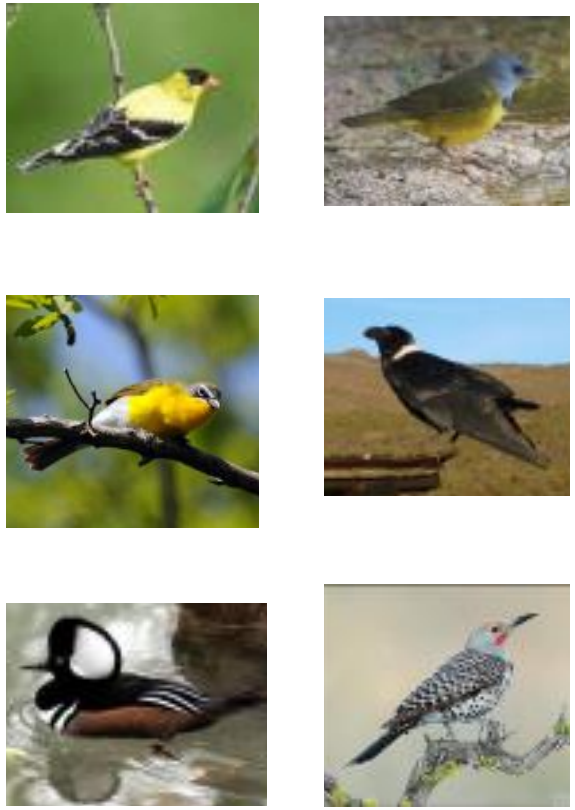
$$\Delta_\theta \mathbb{E}_{\tilde{w}\sim p(w|I,C)}[R(\tilde{w})]$$
$$= \mathbb{E}_{\tilde{w}\sim p(w|I,C)}[R(\tilde{w})\Delta_\theta \log p(\tilde{w})]$$

[Williams 92] Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning (1992)



Visual Explanations – Experimental Setup

- Bird images
  - Caltech UCSD Birds dataset (CUB) [Wah 2011]
  - 200 bird classes
  - 11,788 images
- Captions
  - 5 per image [Reed CVPR 2016]
    - **No** explanation



[Reed CVPR 2016] Reed, S., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine-grained visual descriptions. CVPR 2016
[Wah 2011] Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)

Visual Explanations – Qualitative Results

This is a **Bronzed Cowbird** *because ...*

Definition:	this bird is **black** with **blue** on its wings and has a long **pointy beak**.
Description:	this bird is **nearly all black** with a short **pointy bill**.
Explanation-Label:	this bird is **nearly all black** with **bright orange eyes**.
Explanation-Dis.:	this is a **black bird** with a **red eye** and a **white beak**.
Explanation:	this is a **black bird** with a **red eye** and a **pointy black beak**.

Definition vs. Explanation – Qualitative Results

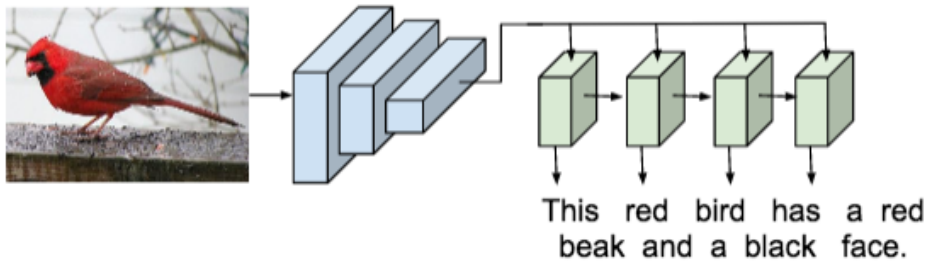
*This is a **Downy Woodpecker** because...*

Definition: this bird has a white breast black wings and a **red spot** on its head.
Explanation: this is a black and white bird with a **red spot** on its crown.

*This is a **Downy Woodpecker** because...*

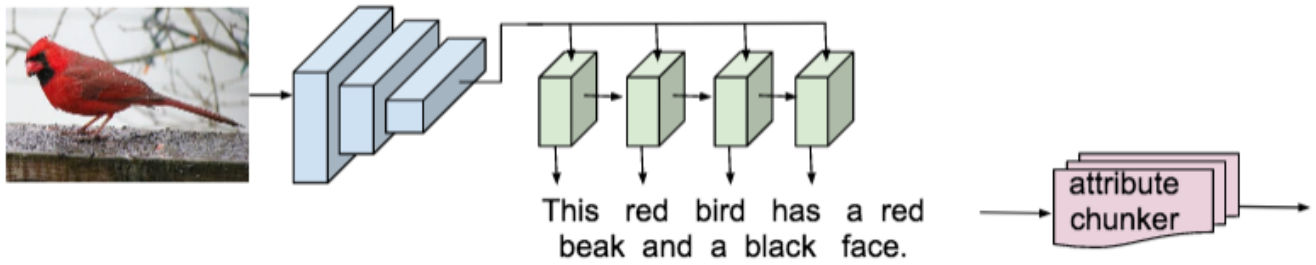
Definition: this bird has a white breast black wings and a **red spot** on its head.
Explanation: this is a white bird with a black wing and a black and white striped head.

Grounding Visual Explanations



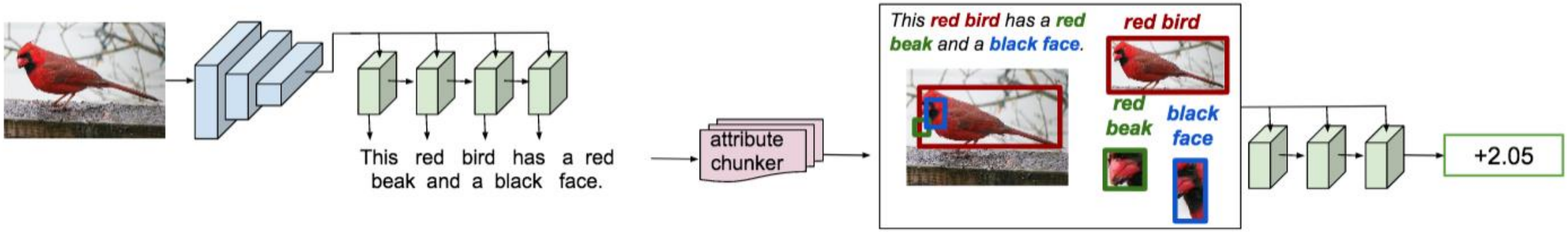
[Hendricks, Hu, Akata, Darrell; ECCV18]

Grounding Visual Explanations



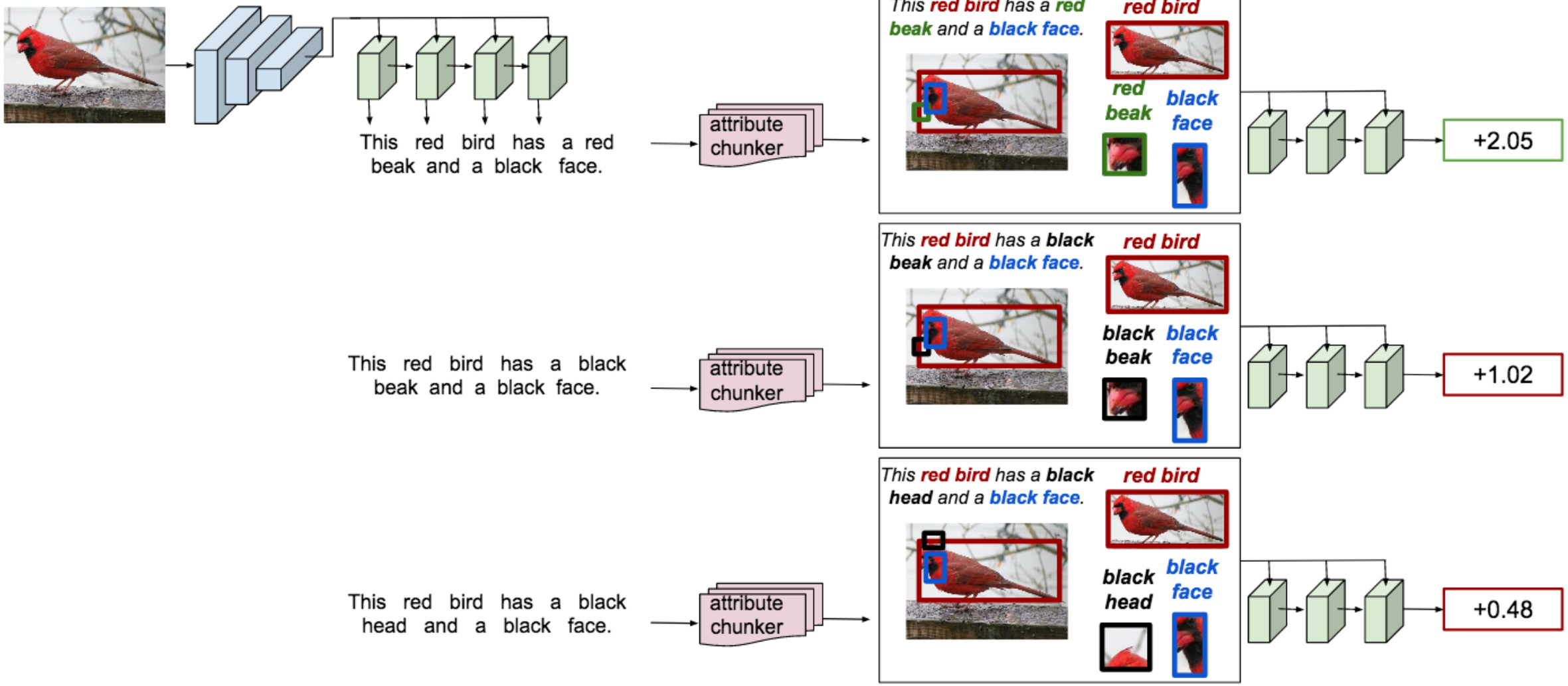
[Hendricks, Hu, Akata, Darrell; ECCV18]

Grounding Visual Explanations



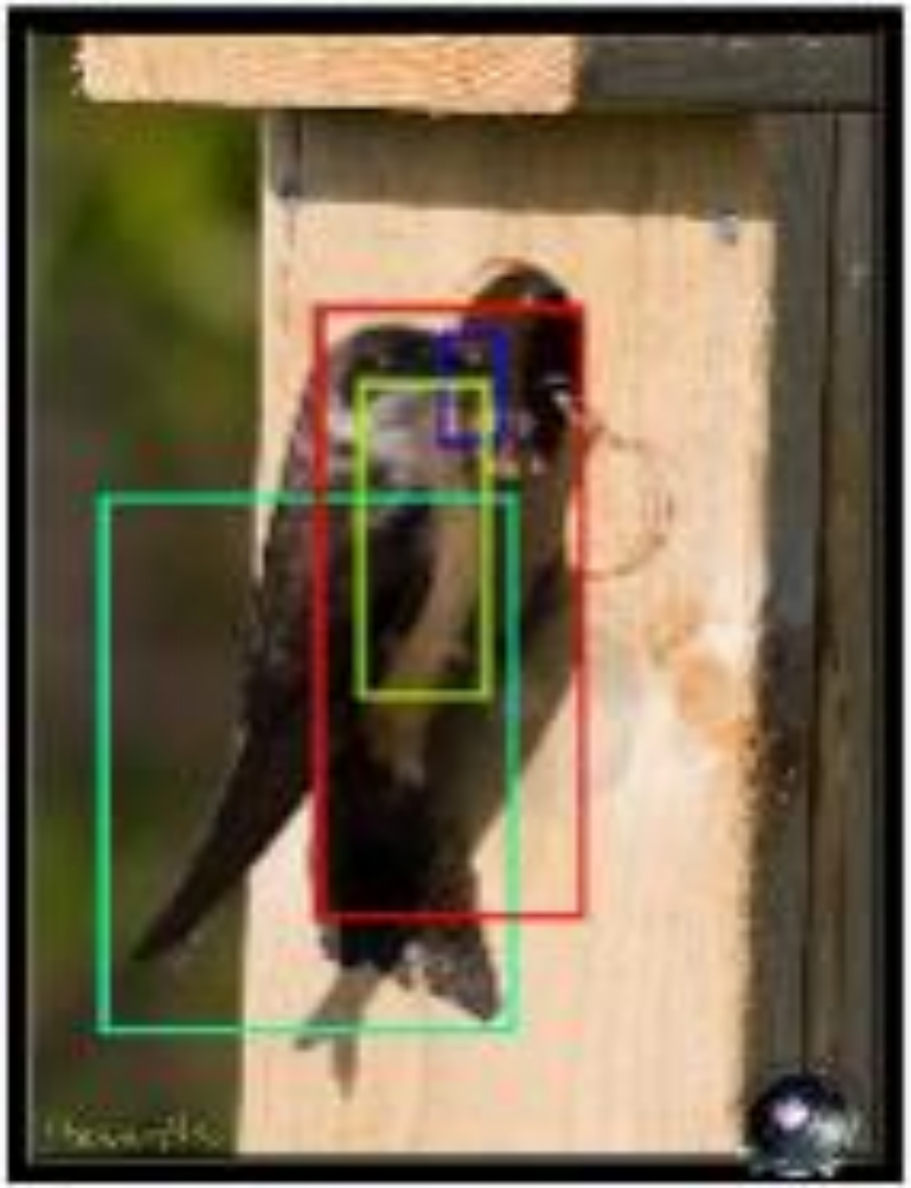
[Hendricks, Hu, Akata, Darrell; ECCV18]

Grounding Visual Explanations



[Hendricks, Hu, Akata, Darrell; ECCV18]

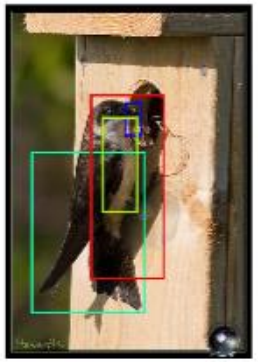
Multimodal Grounding



This is a *Bank Swallow* because this **small bird** has a **white belly and breast**, **brown wings**, and a short **black bill**.

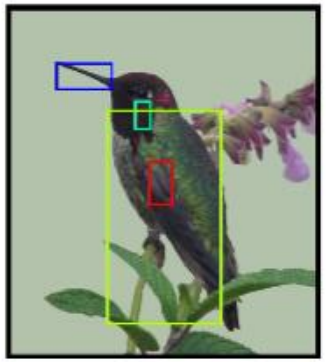
Qualitative Results: Ranking and Grounding

Score: 14.4901



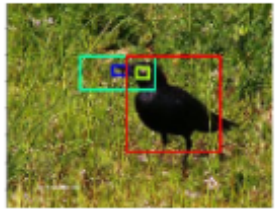
This is a *Bank Swallow* because this **small bird** has a **white belly and breast**, **brown wings**, and a short **black bill**.

Score: 12.9394



This is an *Anna Hummingbird* because this is a **small brown** and **green bird** with a **pink throat** and a **long black beak**.

This is a *Brewer Blackbird* because … this is a **black bird** with a **white eye** and **long pointy black beak**.



Best 3 Captions:	Score
… this is a **black bird** with a **white eye** and **long pointy black beak**.	1.46
… this is a **black bird** with a **white eye** and a **pointy black beak**.	1.45
… this **all black bird** has a **short black beak** and a **lighter colored eye - coloring**.	1.32
Worst 3 Captions:	
… this is a **shiny all blue bird** with a **yellow eye** and **black beak**.	-2.88
… this bird is **completely green** with a **short blunt bill** and a **white wingbar**.	-3.08
… this bird is **all black with a yellow eye** and **large thick to gray feathers on its crown.**	-3.40

Grounds constituent visual attributes + generates an explanation

Explanation ranker: scoring explanation based on image consistency

Higher ranked explanation → more image relevant

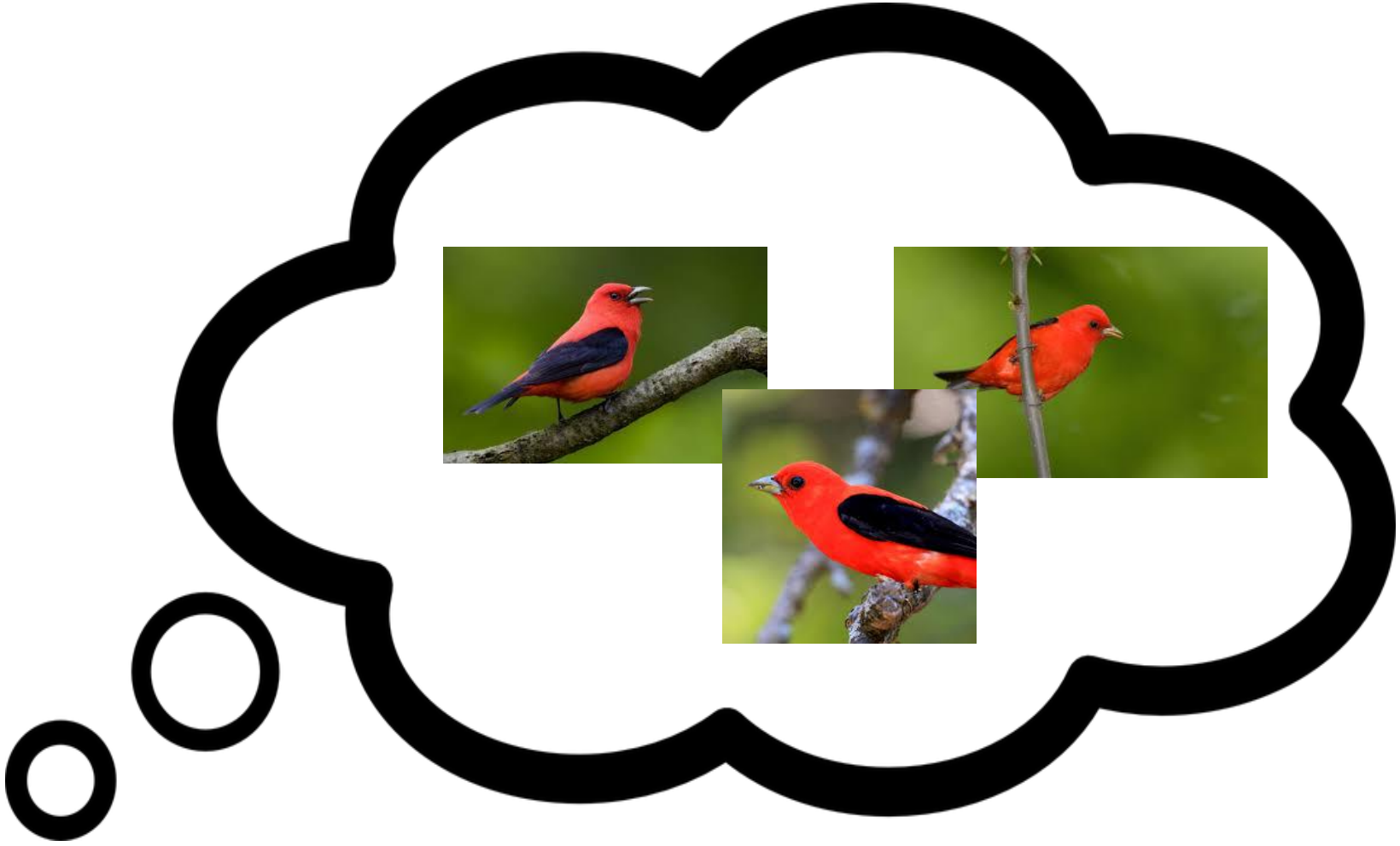
Generating Counterfactual Explanations with Natural Language

Lisa Anne Hendricks[1], Ronghang Hu[1], Trevor Darrell[1], Zeynep Akata[2]

[1]UC Berkeley, [2]University of Amsterdam

ICML 2018, Workshop on Human Interpretability in Machine Learning

[1] Hendricks et al. *Generating Visual Explanations*. ECCV 2016.
[2] Park et al. *Multimodal Explanations: Justifying Decisions and Pointing to the Evidence*. CVPR 2018.
[3] Hendricks et al. *Grounding Visual Explanations*. ECCV 2018 (to appear).



This bird is not a ***Vermillion Flycatcher*** because it does not have black wings.

Textual Counterfactual Explanations

Challenges

Challenges

- Difficult to selectively change attributes of an image

Challenges

- Difficult to selectively change attributes of an image
  - Instead, reason about visual evidence in natural language

Challenges

- Difficult to selectively change attributes of an image
  - Instead, reason about visual evidence in natural language
- Do not have training data with textual counterfactual explanations

Challenges

- Difficult to selectively change attributes of an image
  - Instead, reason about visual evidence in natural language
- Do not have training data with textual counterfactual explanations
  - Mine ground truth by reasoning about which evidence is present in one class, but not another

Pipeline

1. Predict counterfactual evidence
2. Check if there is counterfactual evidence in original image
3. Construct explanation

Why is this a cardinal, but not a scarlet tanager?

Step 1: Predict candidate counterfactual evidence.

Explanation Generator

This is a ***Scarlet Tanager*** because it is a *red bird* with a *pointy beak* and *black eyes*.
…
This is a ***Scarlet Tanager*** because it is a *red bird* with *black wings* and a *pointy beak*.

Hendricks et al. *Generating Visual Explanations*. ECCV 2016.

Why is this a cardinal, but not a scarlet tanager?

Step 1: Predict candidate counterfactual evidence.

Explanation Generator

This is a ***Scarlet Tanager*** because it is a ***red bird*** with a ***pointy beak*** and ***black eyes***.
…
This is a ***Scarlet Tanager*** because it is a ***red bird*** with ***black wings*** and a ***pointy beak***.

Why is this a cardinal, but not a scarlet tanager?

Step 2: Evidence Checker

Scarlet Tanager features:

red bird, pointy beak, black eyes, black wings

Cardinal features:

red bird, black face, red beak

?

Why is this a cardinal, but not a scarlet tanager?

Step 2: Evidence Checker

Scarlet Tanager features:

red bird, pointy beak, black eyes, black wings

Cardinal features:

red bird, black face, red beak

?

Counterfactual evidence: Black wings.

Evidence checker: Random choice (baseline)

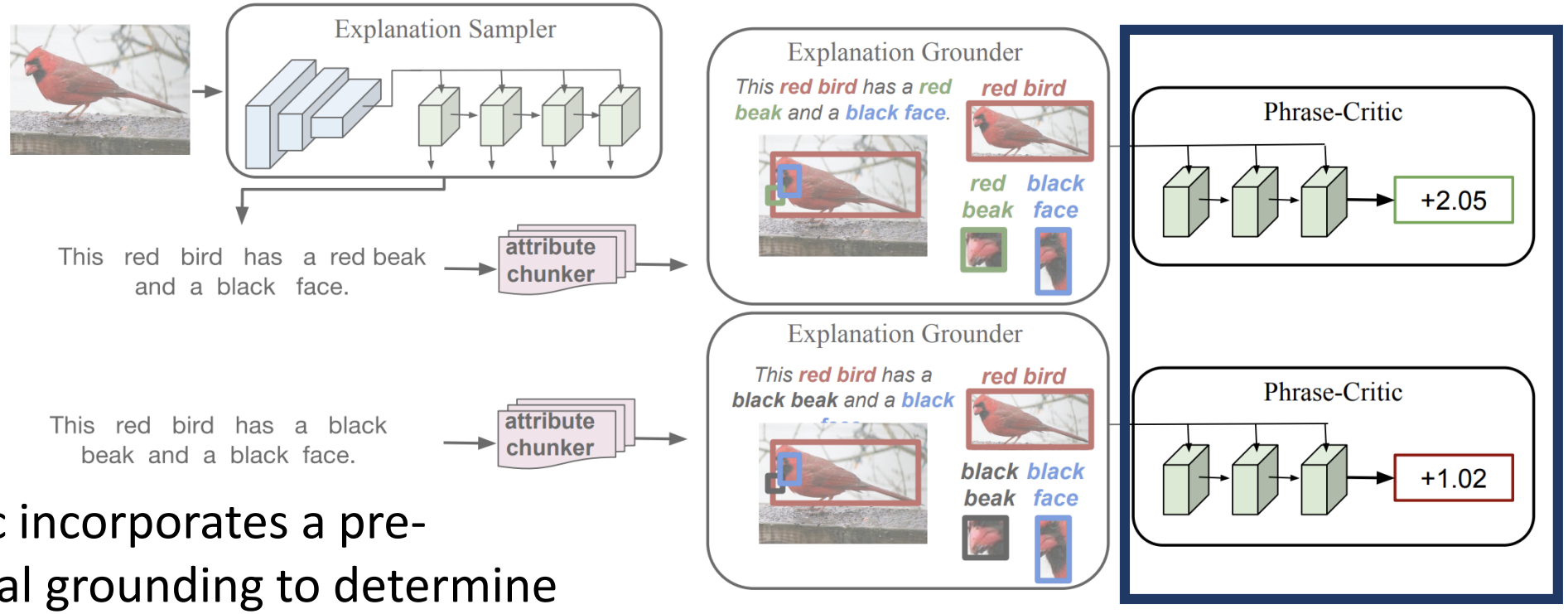
Scarlet Tanager features:

red bird, pointy beak, black eyes, black wings, pointy beak

Counterfactual evidence: Pointy beak

Evidence checker: Train from scratch (CF: Classifier)

Evidence checker: Phrase Critic (CF: Phrase Critic)



Phrase critic incorporates a pre-trained visual grounding to determine if evidence is in an image.

Hu et al. CVPR 2017.

[1] Hendricks et al. *Grounding Visual Explanations*. NIPS Interpretable ML Symposium 2017.

[2] Hendricks et al. *Grounding Visual Explanations*. ECCV 2018 (to appear).

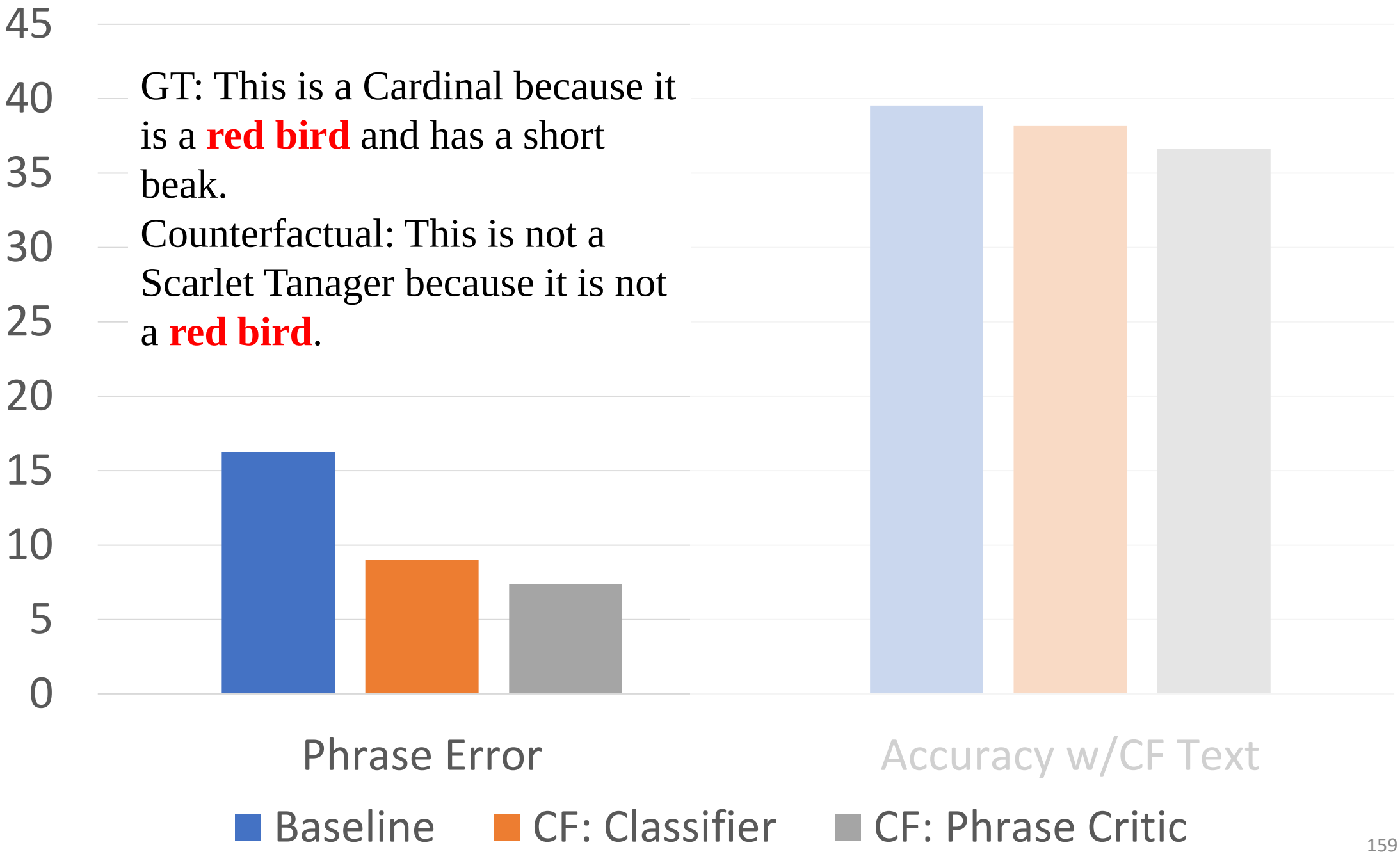
Why is this a cardinal, but not a scarlet tanager?

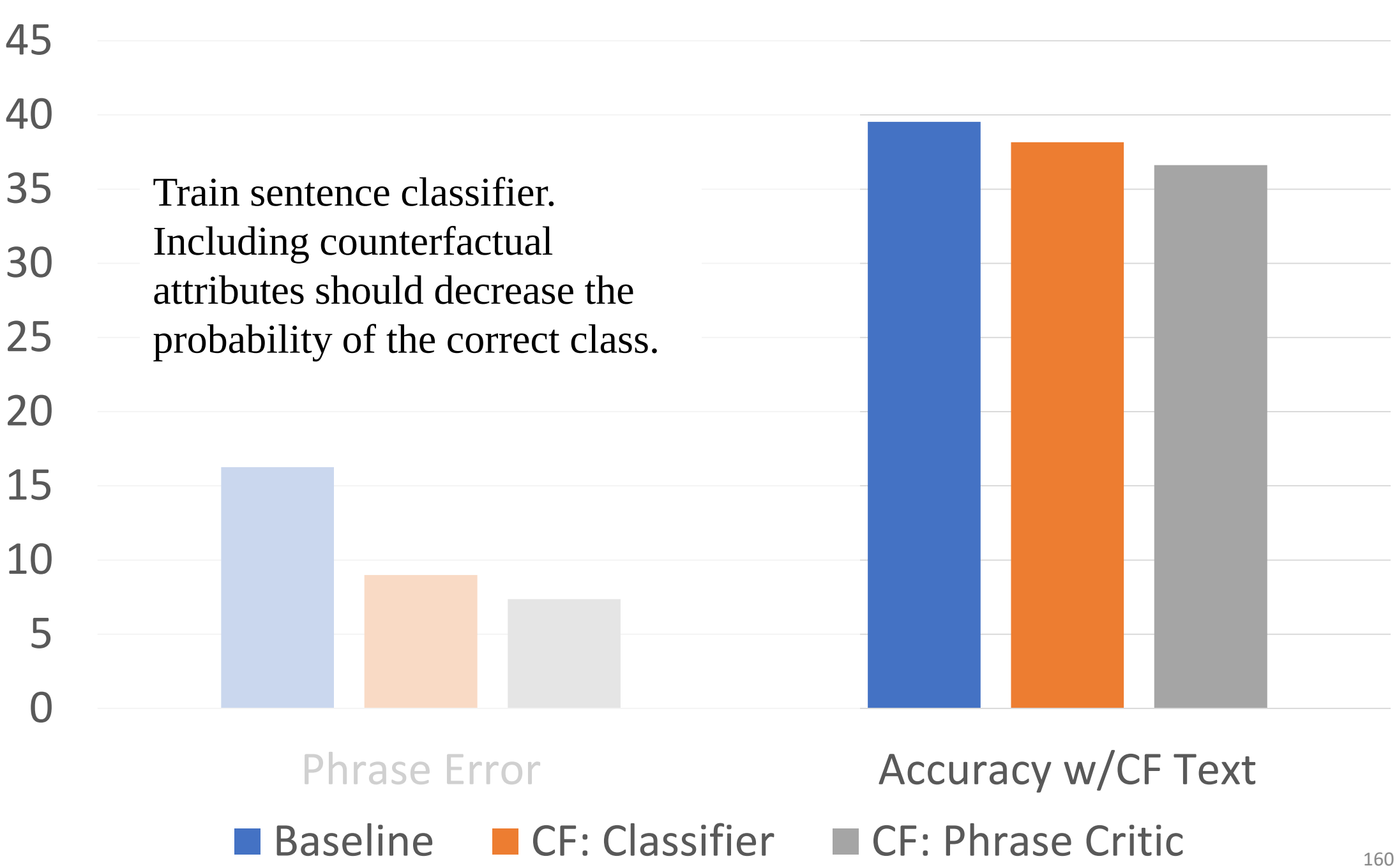
Step 3: Explanation Generator

Counterfactual class: Scarlet Tanager
Counterfactual evidence: Black wings.

*This is not a **Scarlet Tanager** because it does not have black wings.*

RESULTS


45
40
35
30
25
20
15
10
5
0
GT: This is a Cardinal because it is a red bird and has a short beak.
Counterfactual: This is not a Scarlet Tanager because it is not a red bird.
Phrase Error
Accuracy w/CF Text
Baseline
CF: Classifier
CF: Phrase Critic


45
40
35
30
25
20
15
10
5
0
Train sentence classifier.
Including counterfactual
attributes should decrease the
probability of the correct class.
Phrase Error
Accuracy w/CF Text
Baseline
CF: Classifier
CF: Phrase Critic

Qualitative Examples

Explanation [1]: This is a ***White Necked Raven*** because this is a black bird with a white nape and a large beak.

Counterfactual: This is not a ***Bobolink*** because it does not have a yellow nape.

Class:
White Necked Raven

Counter-Class:
Bobolink

[1] Hendricks et al. *Generating Visual Explanations*. ECCV 2016.

Qualitative Examples

Explanation [1]: This is a ***Blue Winged Warbler*** because this is a yellow bird with a black wing and a black pointy beak.

Counterfactual: This is not a ***Common Yellowthroat*** because it does not have a black face.

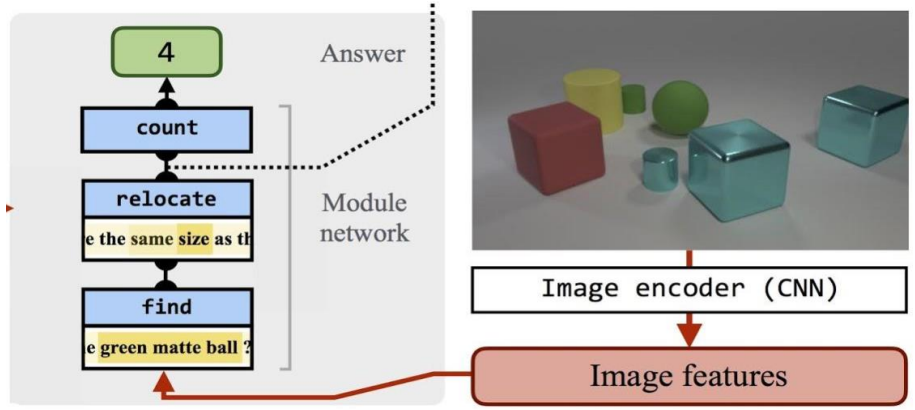
Class:
Blue-Winged Warbler

Counter-Class:
Common Yellowthroat

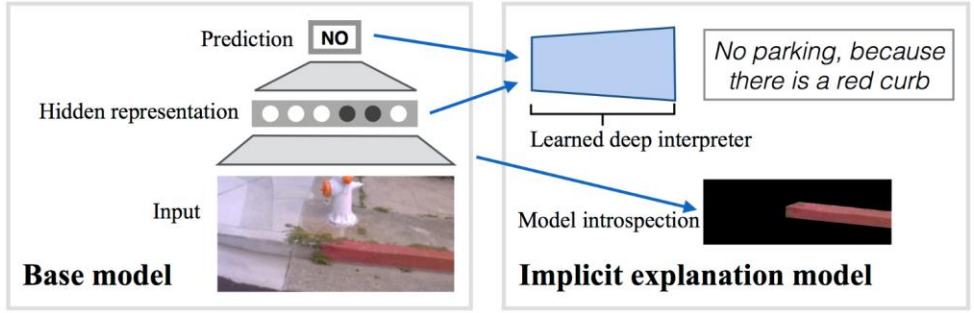
[1] Hendricks et al. *Generating Visual Explanations*. ECCV 2016.

Deep Explanation Models

Explicit / Introspective models: interpretable internal visualization



Implicit / Justification models: post-hoc rationalization



RISE:
Randomized Importance Sampling for Explanation of Black-box Models

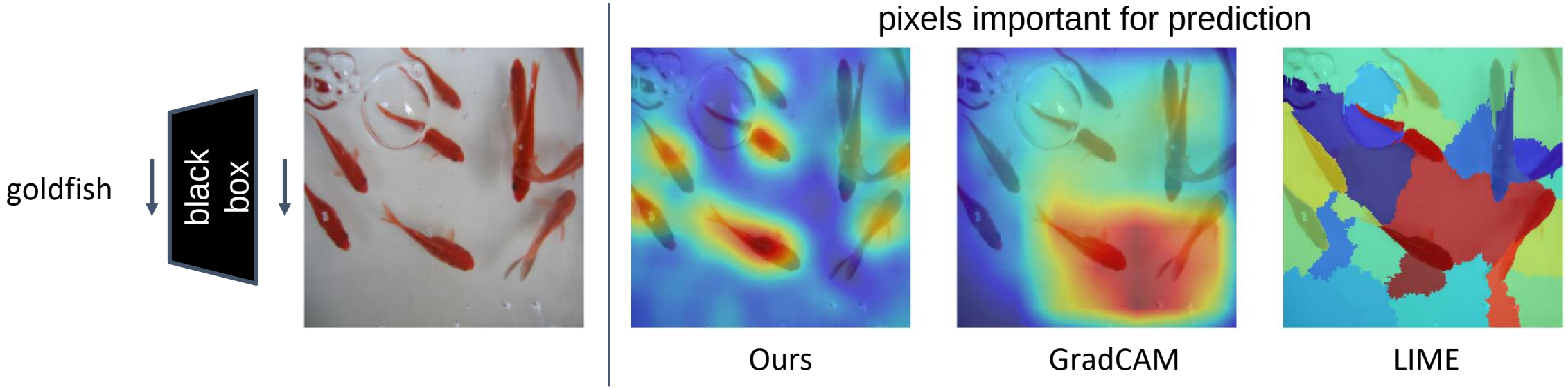
Vitali Petsiuk, Abir Das, Kate Saenko

Boston University



“High-fidelity” salience for black-box networks

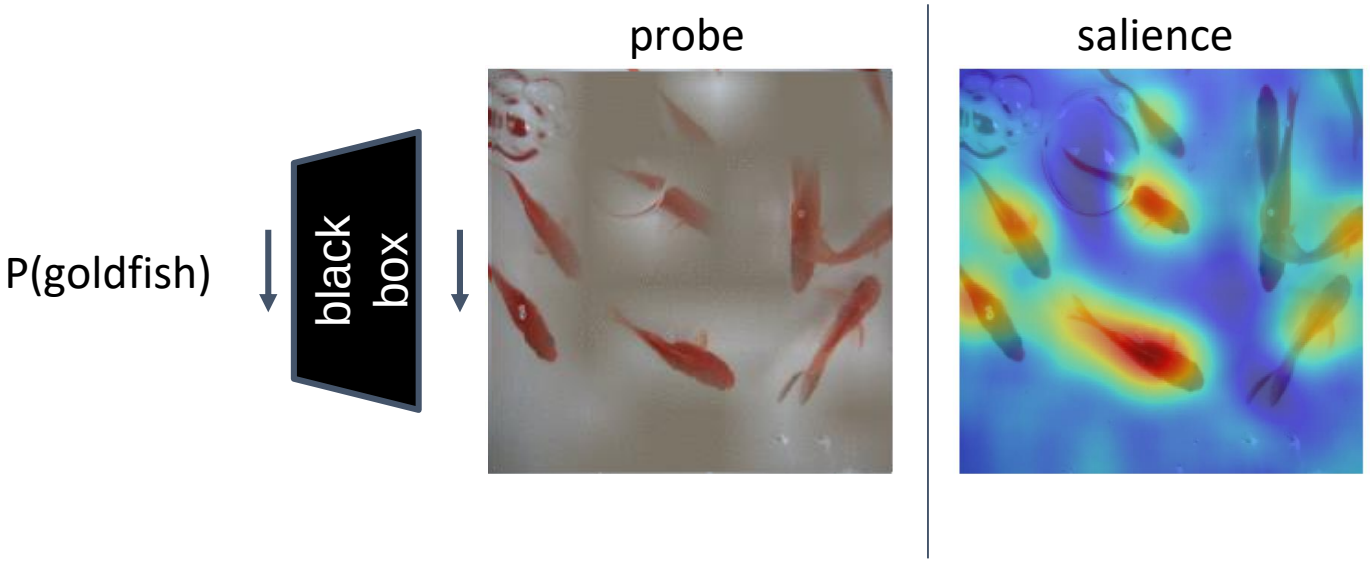
- Estimate important pixels without access to parameters (black box)
- What the network actually sees, not what a human sees: “high-fidelity” explanation



RISE: Randomized Input Sampling for Explanation of Black-box Models, Vitali Petsiuk, Abir Das, Kate Saenko, BMVC 2018

“High-fidelity” salience for black-box networks

- How it works:



Salience is estimated by probing black box with randomly masked images

RISE: Randomized Input Sampling for Explanation of Black-box Models, Vitali Petsiuk, Abir Das, Kate Saenko, BMVC 2018

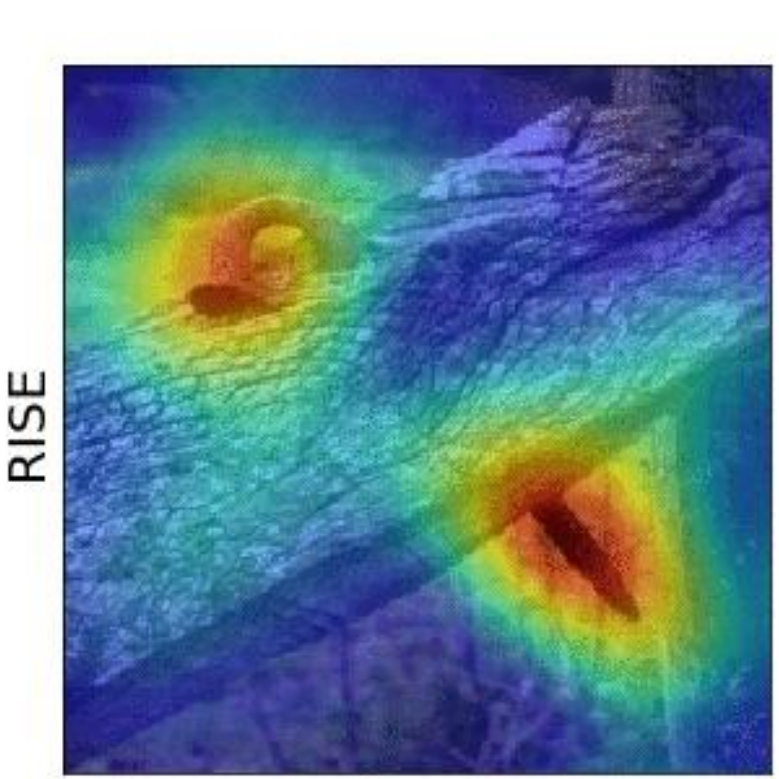
“High-fidelity” salience for black-box networks

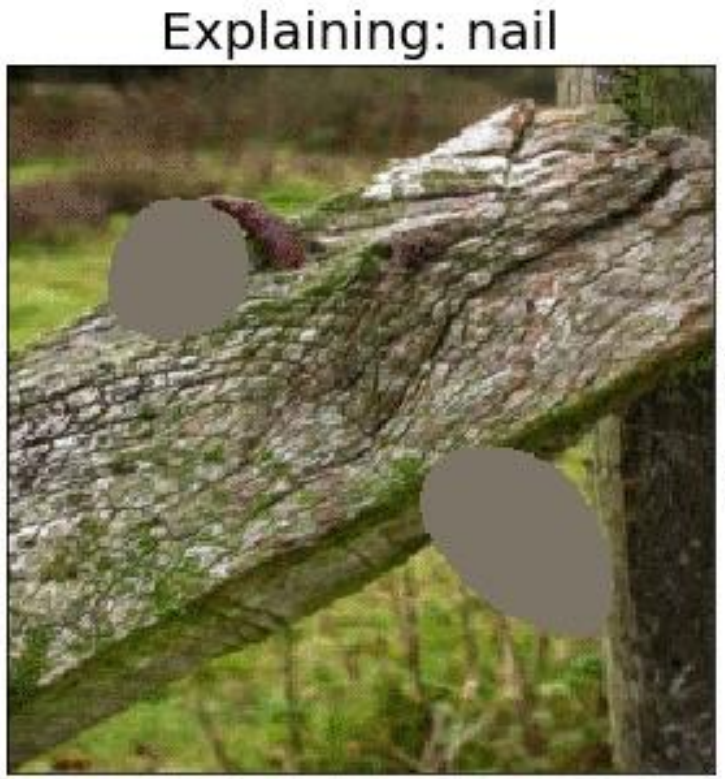
- Black-box salience illustration

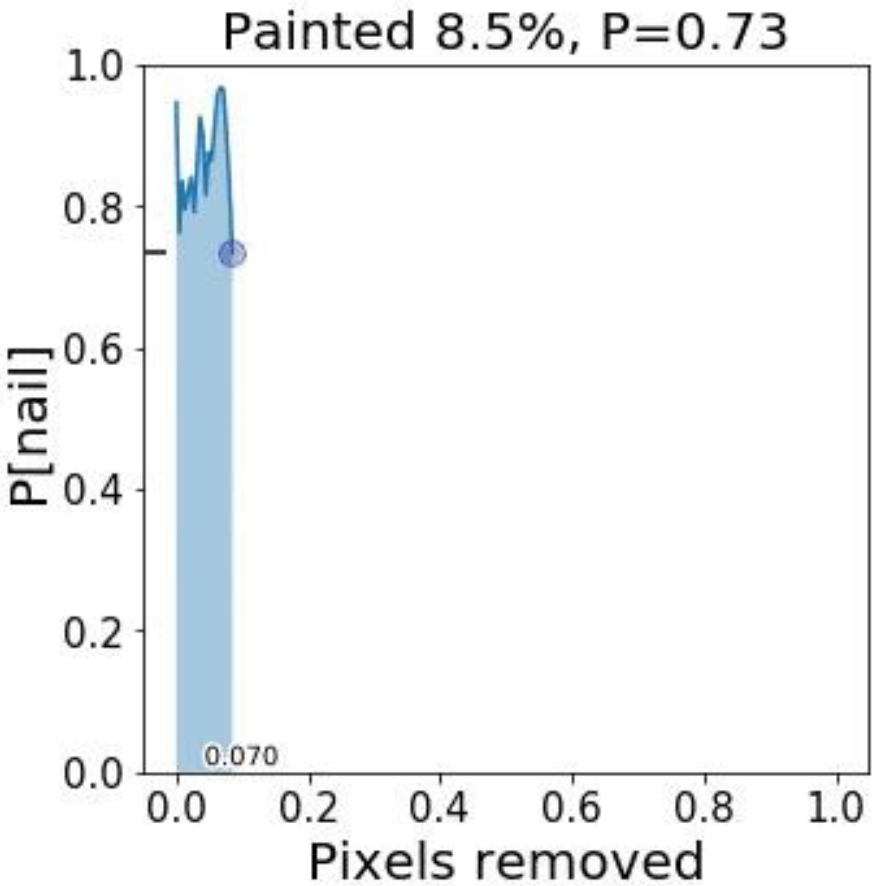
RISE: Randomized Input Sampling for Explanation of Black-box Models, Vitali Petsiuk, Abir Das, Kate Saenko, BMVC 2018

“High-fidelity” salience for black-box networks

- Black-box salience illustration



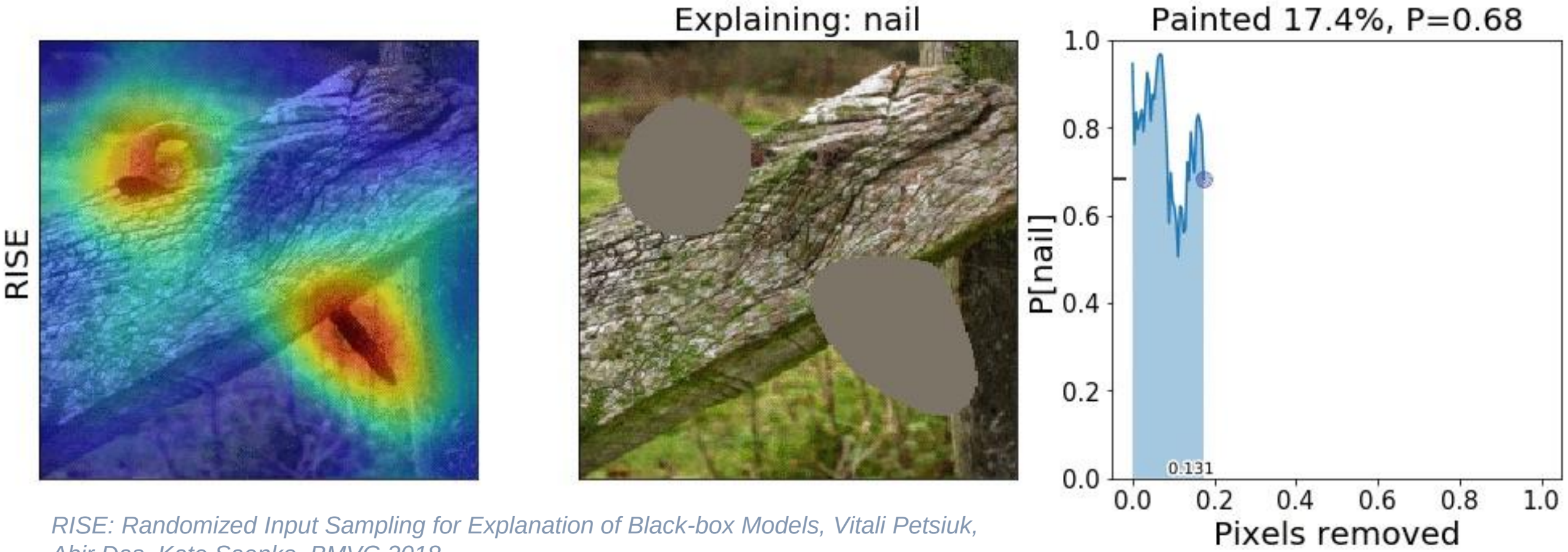




RISE: Randomized Input Sampling for Explanation of Black-box Models, Vitali Petsiuk, Abir Das, Kate Saenko, BMVC 2018

“High-fidelity” salience for black-box networks

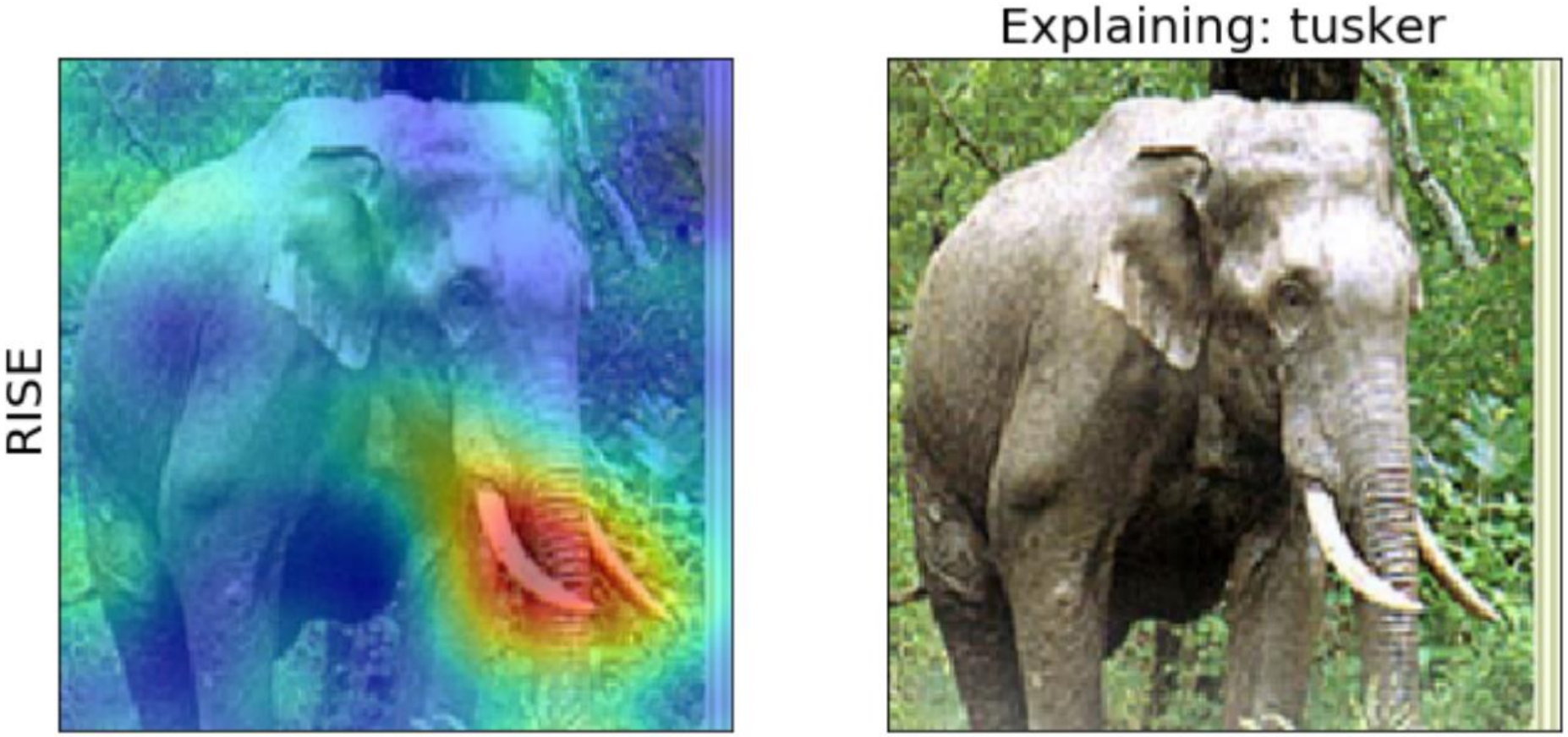
- Black-box salience illustration



RISE: Randomized Input Sampling for Explanation of Black-box Models, Vitali Petsiuk, Abir Das, Kate Saenko, BMVC 2018

Using XAI to discover salient regions

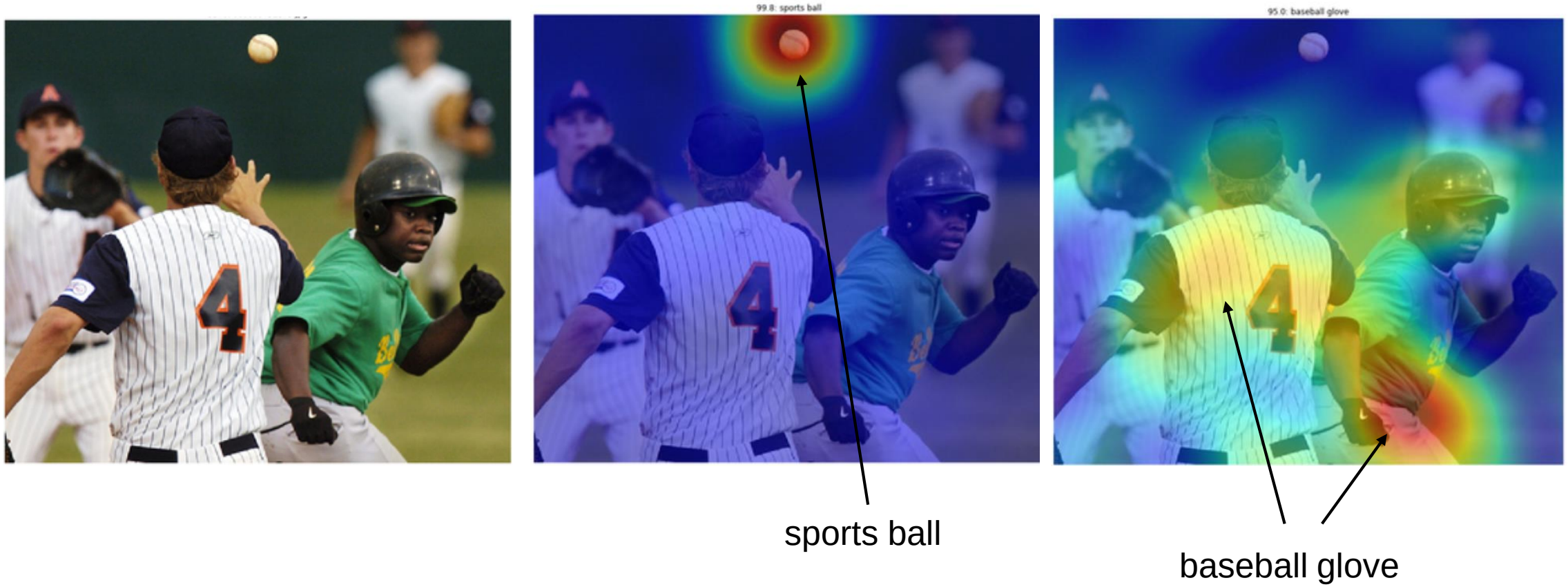
Why did the classifier predict "tusker"?



RISE: Randomized Input Sampling for Explanation of Black-box Models, Vitali Petsiuk, Abir Das, Kate Saenko, BMVC 2018

Using XAI to discover salient regions

Why did the classifier predict "sports ball" and "baseball glove"?

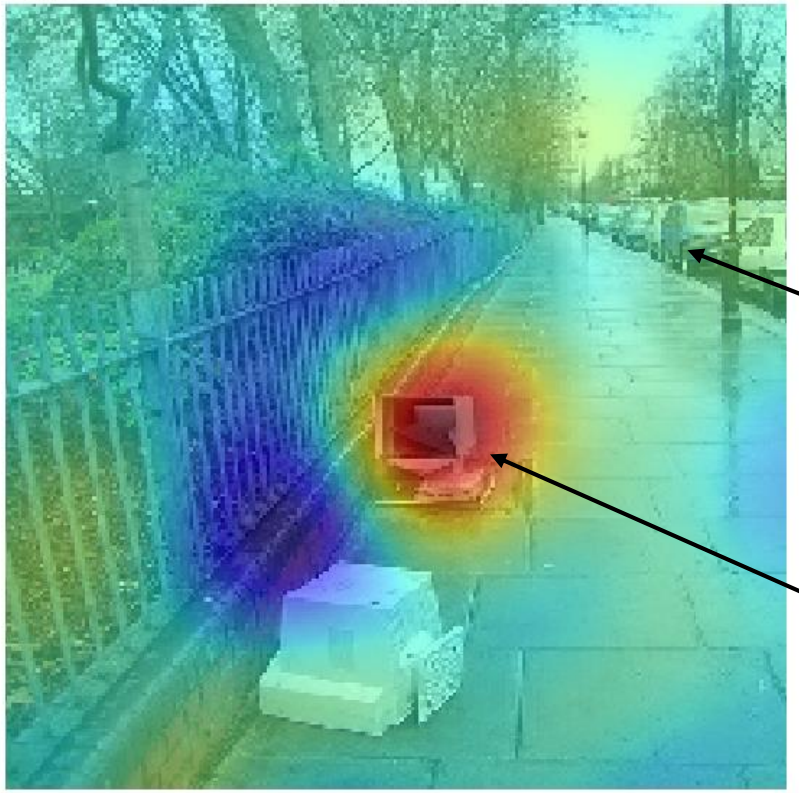


RISE: Randomized Input Sampling for Explanation of Black-box Models, Vitali Petsiuk, Abir Das, Kate Saenko, BMVC 2018

Using XAI to discover salient regions

Why did the classifier predict "car"? –correct but for wrong reasons!

(e) car – 64%



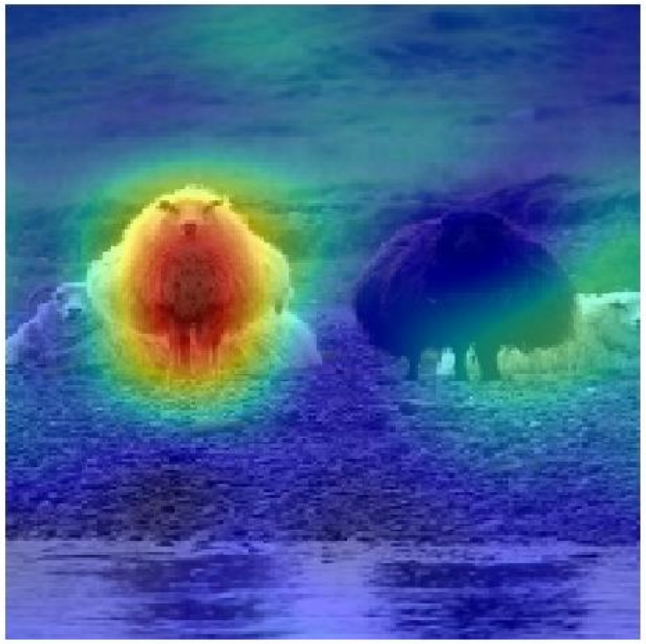
(f) '*car*' importance

RISE: Randomized Input Sampling for Explanation of Black-box Models, Vitali Petsiuk, Abir Das, Kate Saenko, BMVC 2018

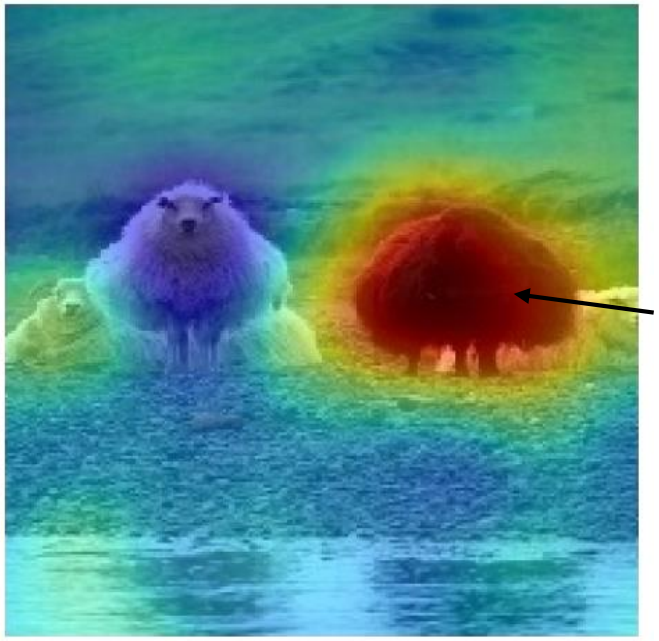
Using XAI to discover salient regions

Why did the classifier predict "sheep" (correct) and "cow" (incorrect)?

(a) Sheep - 26%, Cow - 17%



(b) Importance map of '*sheep*'



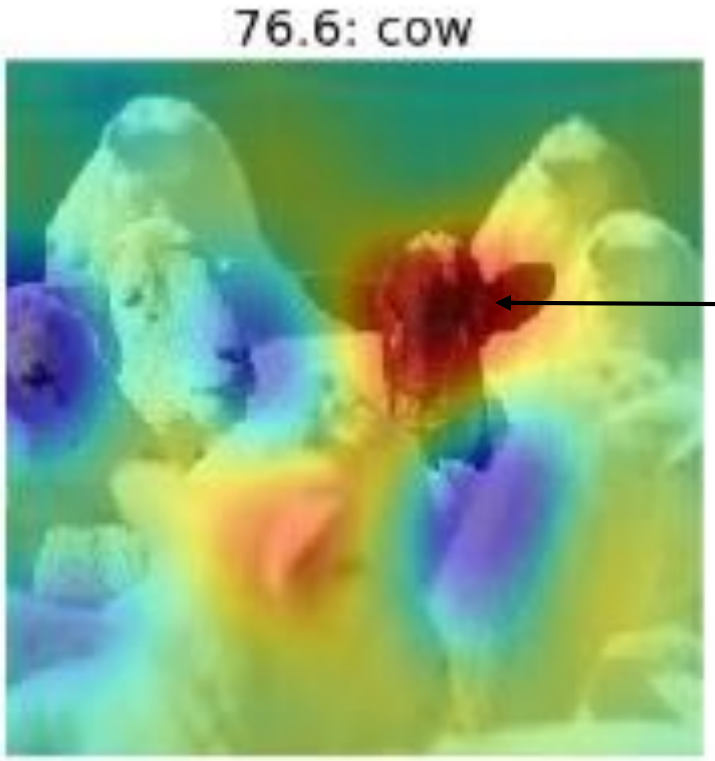
(c) Importance map of '*cow*'

Most sheep are white, so model mistakes black sheep for cows

RISE: Randomized Input Sampling for Explanation of Black-box Models, Vitali Petsiuk, Abir Das, Kate Saenko, BMVC 2018

Using XAI to discover salient regions

Why did the classifier predict "cow" (incorrect)?



Most sheep are white, so model mistakes black sheep for cows

RISE: Randomized Input Sampling for Explanation of Black-box Models, Vitali Petsiuk, Abir Das, Kate Saenko, BMVC 2018

Women Also Snowboard: Overcoming Bias in Captioning Models

Lisa Anne Hendricks*, Kaylee Burns*, Kate Saenko, Trevor Darrell, Anna Rohrbach

Motivation

- Bias can effect predictions in different ways…

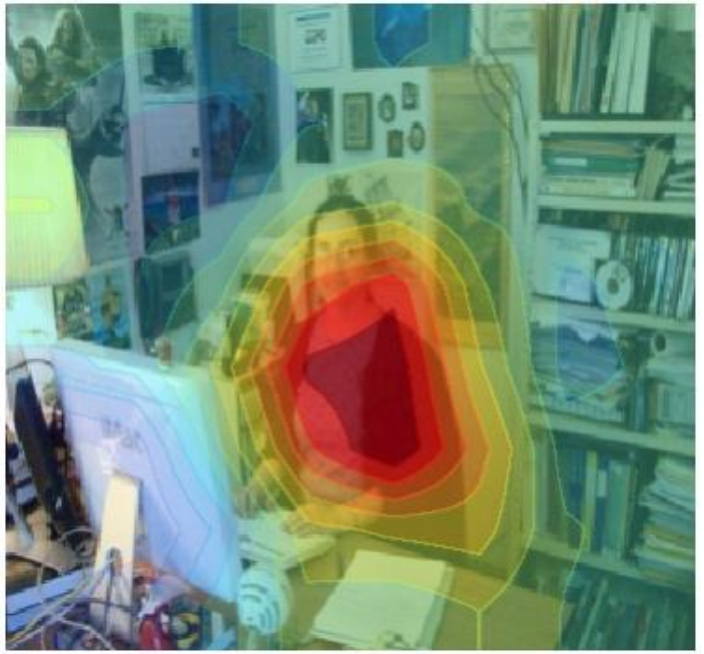
Motivation

- We may make errors:

Wrong

Baseline: *A **man** sitting at a desk with a laptop computer.*

Right for the Right Reasons



Our Model: *A **woman** sitting in front of a laptop computer.*

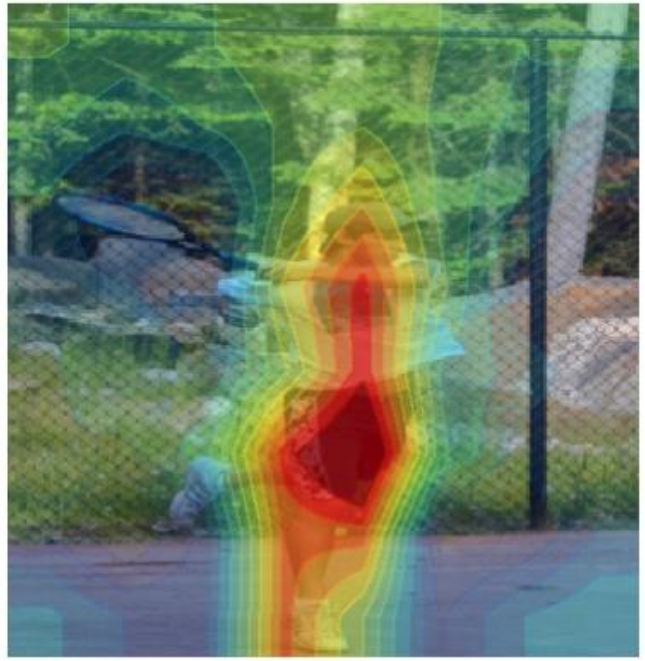
Motivation

- We may base our prediction on wrong evidence:

Right for the Wrong Reasons

Baseline: *A **man** holding a tennis racquet on a tennis court.*

Right for the Right Reasons



Our Model: *A **man** holding a tennis racquet on a tennis court.*

This work

- Overcome bias in image captioning
- Focus on gender bias, inspired by [1]
- Make gender prediction use appropriate cues, e.g. person's appearance

[1] Zhao et al. "Men also like shopping: Reducing gender bias amplification using corpus-level constraints." EMNLP 2017

Our approach: Equalizer

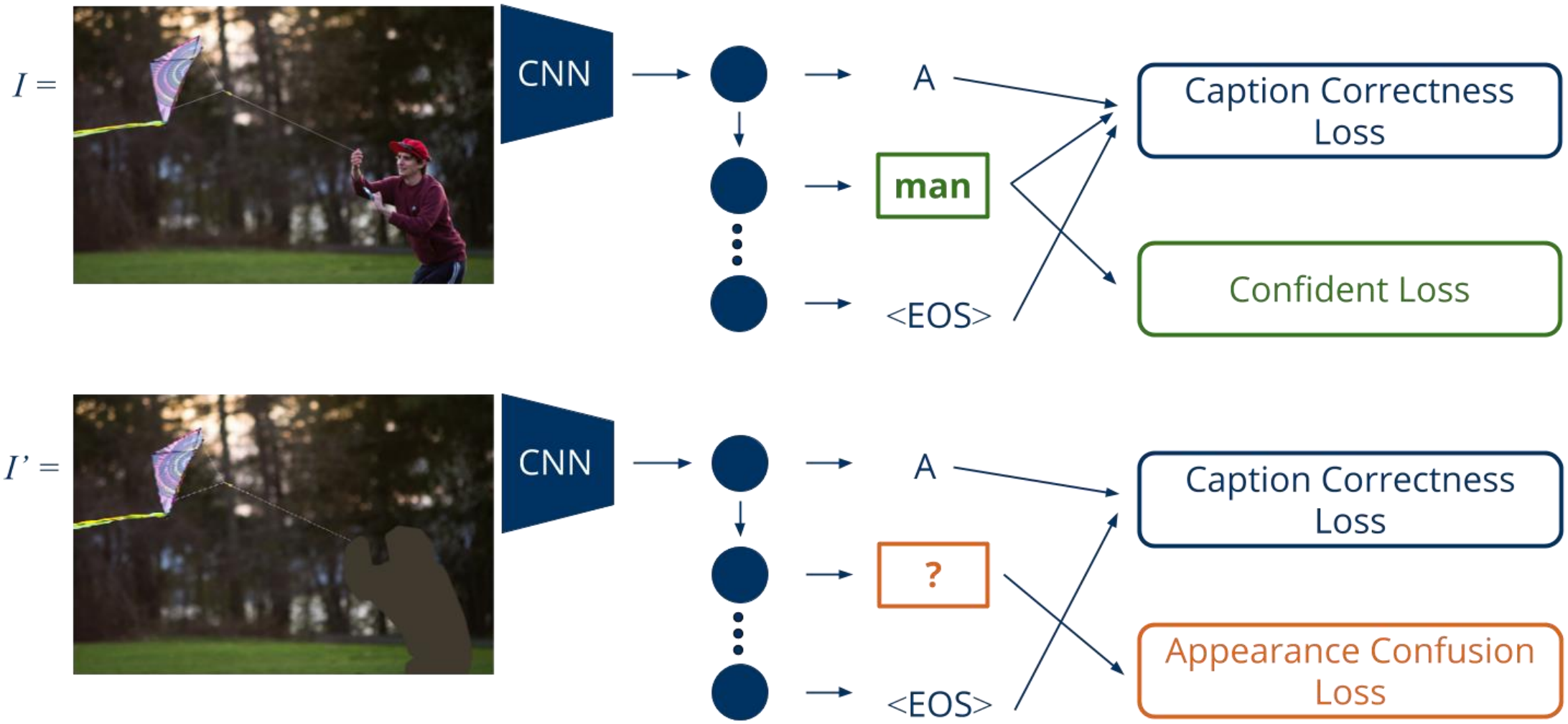
$I =$ 

A man …

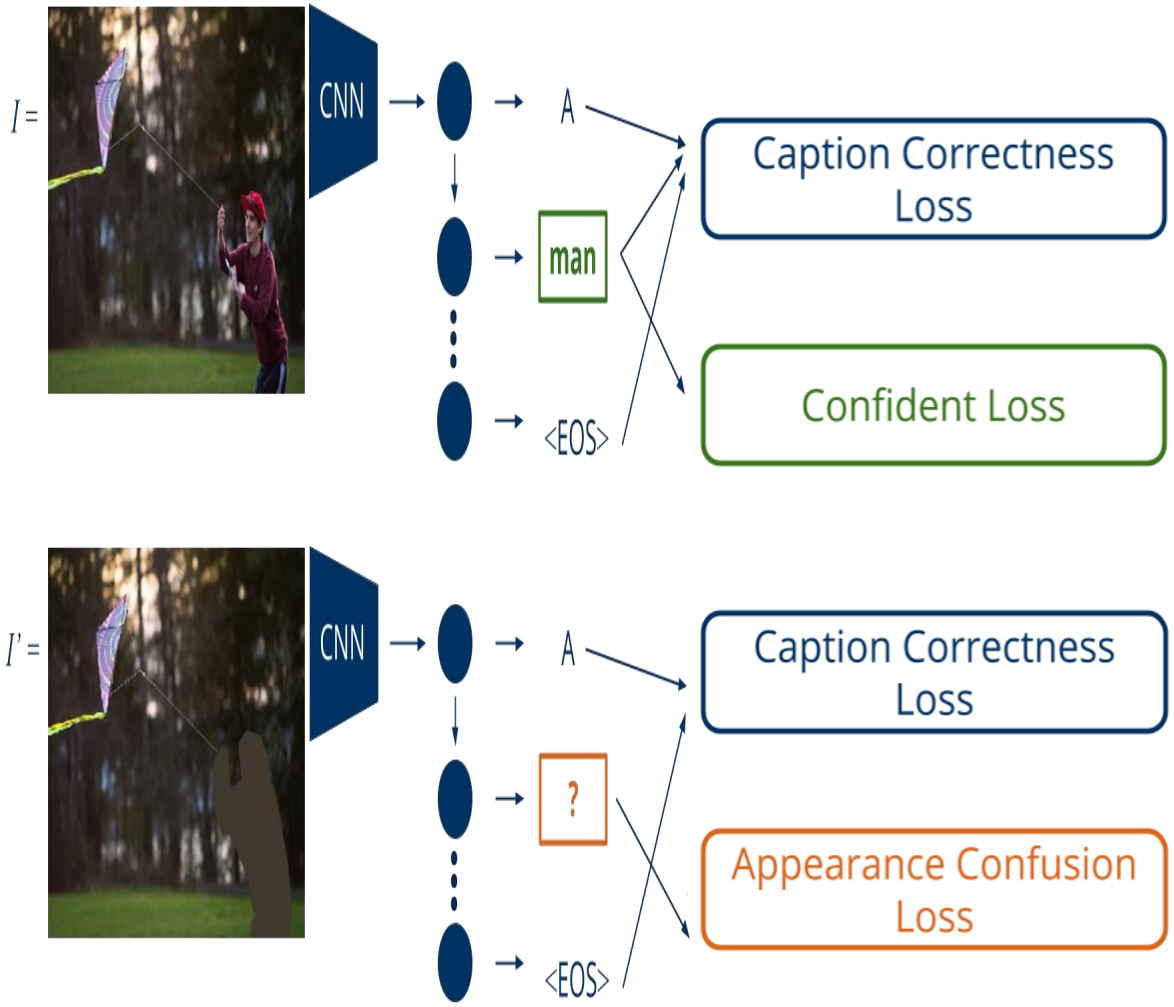
$I' =$ 

A ? …

Our approach: Equalizer



Our approach: Equalizer



A **Confident Loss** on images with men or women

$$\mathcal{L}^{Con} = \frac{1}{N}\sum_{n=0}^{N}\sum_{t=0}^{T}(\mathbb{1}(w_t \in \mathcal{G}_w)\mathcal{F}^W(\tilde{w}_t, I) + \mathbb{1}(w_t \in \mathcal{G}_m)\mathcal{F}^M(\tilde{w}_t, I))$$

$$\mathcal{F}^W(\tilde{w}_t, I) = \sum_{g_w \in \mathcal{G}_w}\sum_{g_m \in \mathcal{G}_m} p(\tilde{w}_t = g_m | w_{0:t-1}, I)/(p(\tilde{w}_t = g_w | w_{0:t-1}, I) + \epsilon)$$

An **Appearance Confusion Loss** on images where men and women are blocked out

$$\mathcal{L}^{AC} = \frac{1}{N}\sum_{n=0}^{N}\sum_{t=0}^{T}\mathbb{1}(w_t \in \mathcal{G}_w \cup \mathcal{G}_m)\mathcal{C}(\tilde{w}_t, I')$$

$$\mathcal{C}(\tilde{w}_t, I') = |\sum_{g_w \in \mathcal{G}_w} p(\tilde{w}_t = g_w | w_{0:t-1}, I') - \sum_{g_m \in \mathcal{G}_m} p(\tilde{w}_t = g_m | w_{0:t-1}, I')|$$

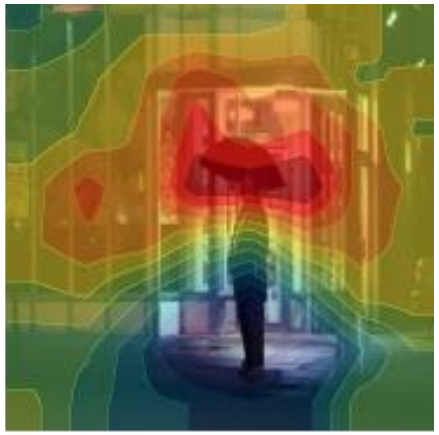
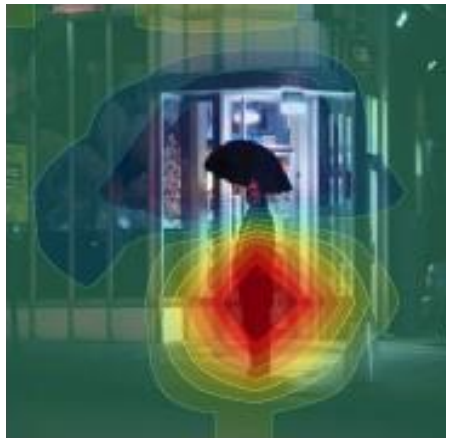
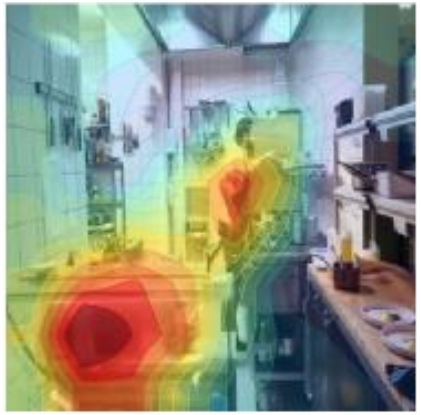
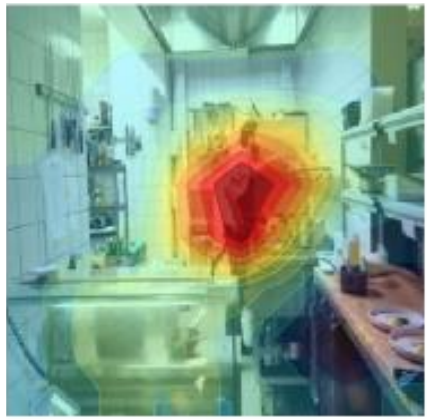
Results

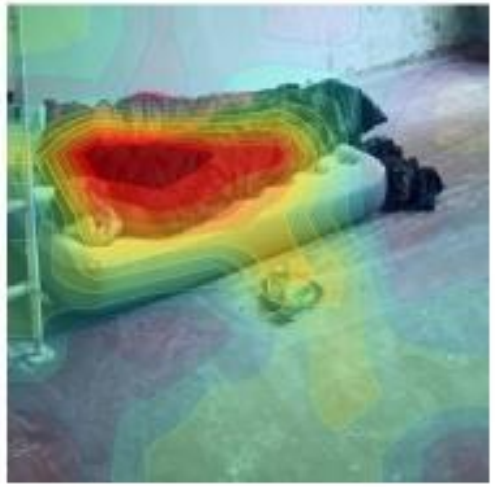
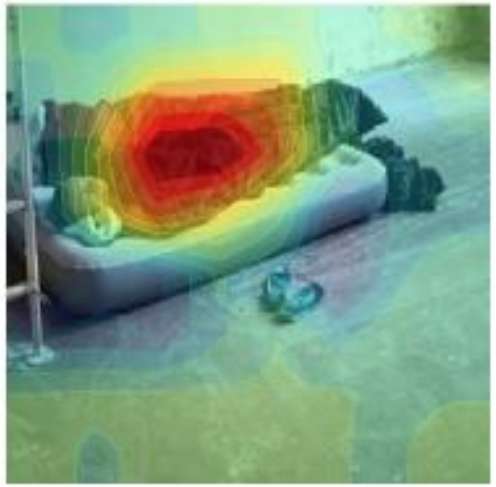
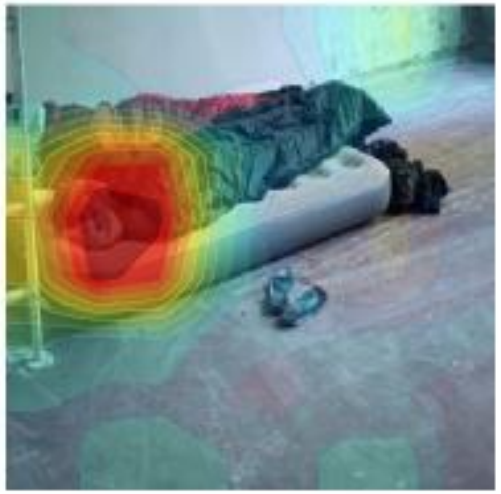
- **Error rate** : misclassification of women as men and vice versa
- **Gender ratio** : women / men in predicted captions
- **Right for the right reasons** : pointing game with visual explanations

Results: Different Distributions

Model	MSCOCO-Biased		MSCOCO-Confident		MSCOCO-Balanced	
	Error	Ratio	Error	Ratio	Error	Ratio
GT	-	0.466	-	0.548	-	1.000
Baseline-FT	0.129	0.265	0.143	0.384	0.203	0.597
Balanced	0.129	0.270	0.142	0.393	0.204	0.610
UpWeight	0.134	0.315	0.116	0.472	0.157	0.712
Equalizer w/o ACL	0.079	0.369	0.081	0.499	0.106	0.777
Equalizer w/o Conf	0.098	0.318	0.116	0.425	0.165	0.673
Equalizer	**0.070**	**0.437**	**0.071**	**0.563**	**0.081**	**0.973**

- Equalizer has the lowest error rate.
- The gender ratio of Equalizer more closely follows the gender ratio of the ground truth captions.

Baseline-FT	UpWeight	Equalizer
		
*A **woman** walking down a street holding an umbrella.*	*A **woman** walking down a street holding an umbrella.*	*A **man** walking down a street holding an umbrella.*
		
*A **man** standing in a kitchen preparing food.*	*A **man** standing in a kitchen preparing food.*	*A **man** standing in a kitchen preparing food.*

Baseline-FT	UpWeight	Equalizer
		
*A **woman** is feeding a giraffe in a zoo.*	*A **woman** is feeding a giraffe at a zoo.*	*A **person** feeding a giraffe through a fence.*
		
*A **person** laying on a bed in a room.*	*A **man** laying on a bed in a room.*	*A **person** laying on a bed in a room.*

Summary

- Equalizer is a novel approach to overcome bias in image captioning
- We obtain correct ratio of women/men for different gender distributions
- We show that our model is right for the right reason, i.e. is looking at people
- When uncertain Equalizer "backs off" to a gender-neural prediction (e.g. person)

XAI Perspectives and Goals

- **All-in for Deep Models**: “The solution to Interpretable Deep Learning is more Deep Learning”

data → ? → prediction

Post-hoc rationalization
“story-telling” models

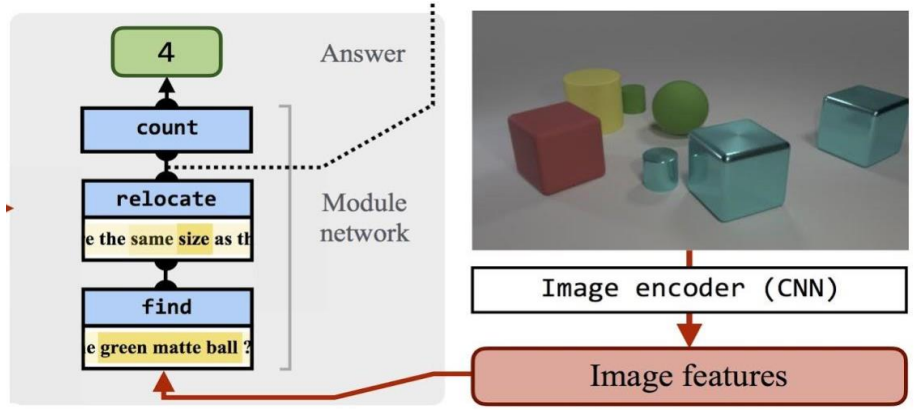
MODULARITY

Introspective /
causal models

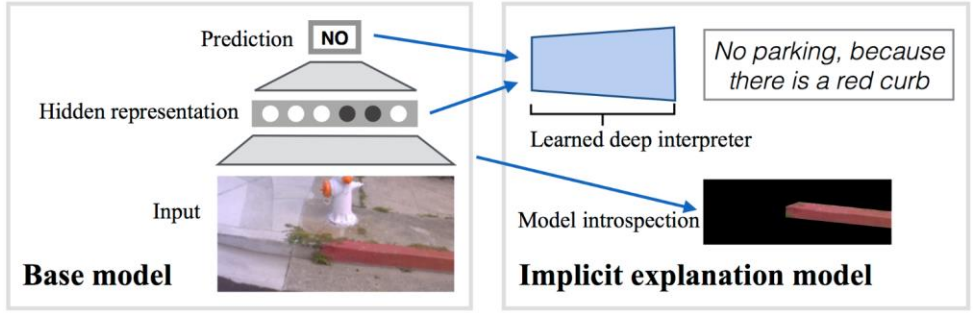
JUSTIFICATION

Deep Explanation Models Summary

Explicit / Introspective models: interpretable internal visualization

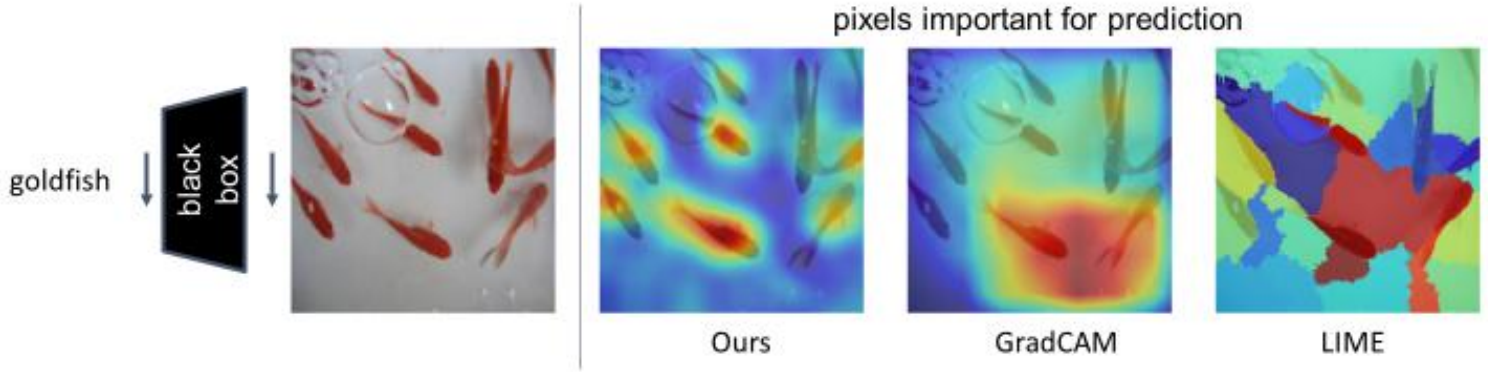


Implicit / Justification models: post-hoc rationalization

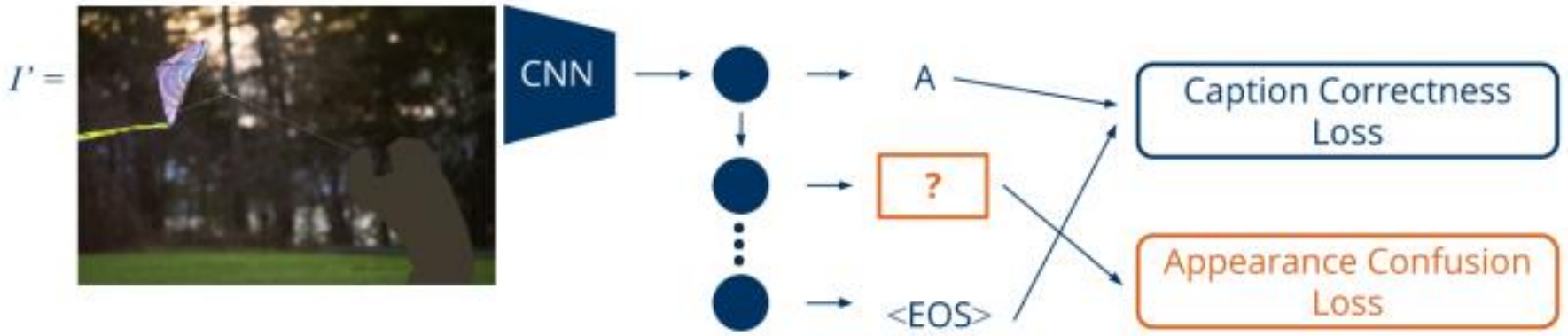


Deep Explanation Models Summary

Saliency / pixel-level causal attribution



Fairness in captioning and Explanation (“right for the right reasons”):



Thanks!