BAIR Retreat 3/28/17

Trevor Darrell
UC Berkeley

Overview

Adversarial Domain Adaptation

Learning end-to-end driving models from crowdsourced dashcams

Vision and Language: Learning to reason to answer and explain

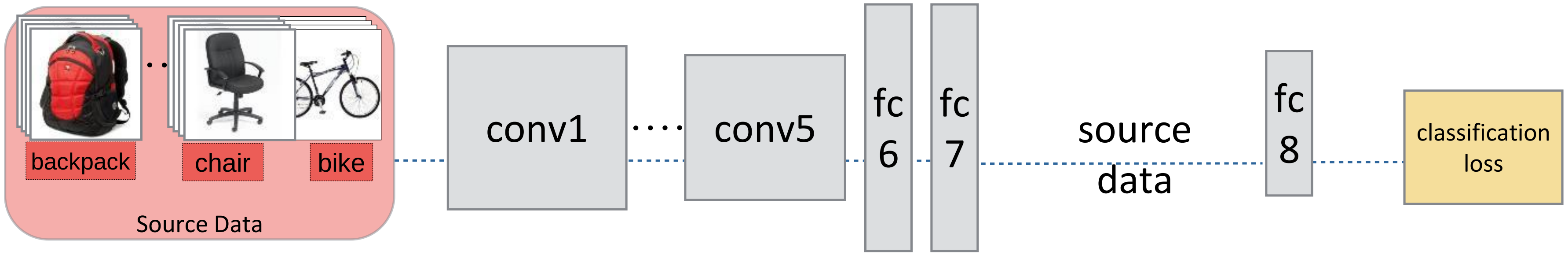
Adversarial Domain Adaptation

Eric Tzeng
UC Berkeley

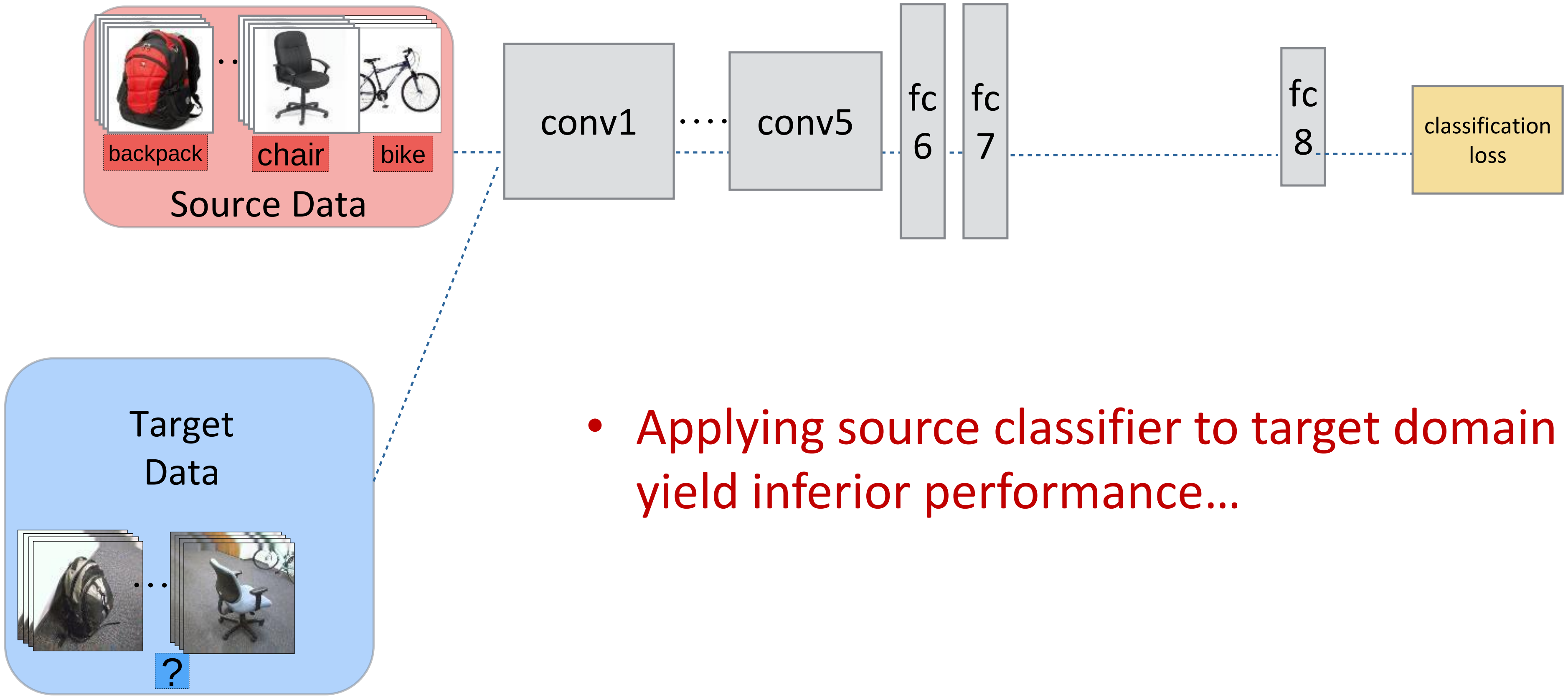
Judy Hoffman
UC Berkeley/ Stanford

Trevor Darrell
UC Berkeley

Adapting across domains ?

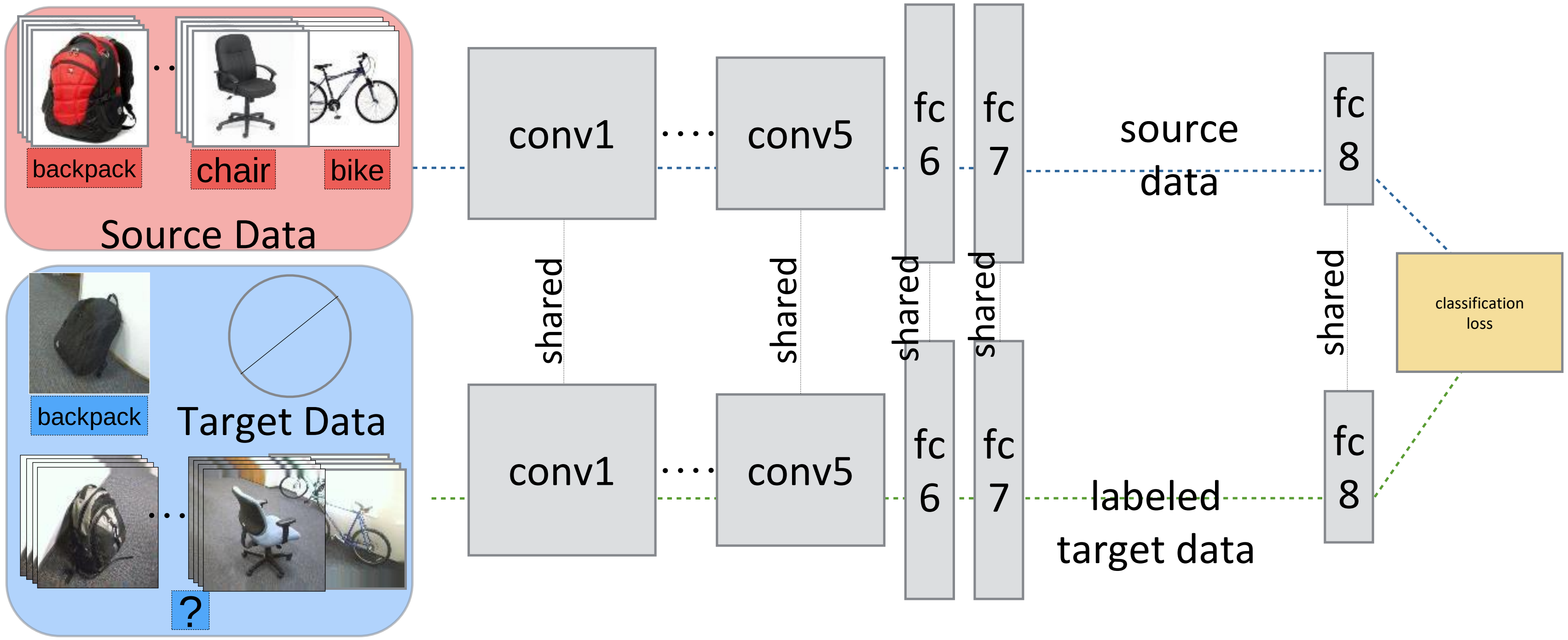


Adapting across domains ?



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

Adapting across domains ?

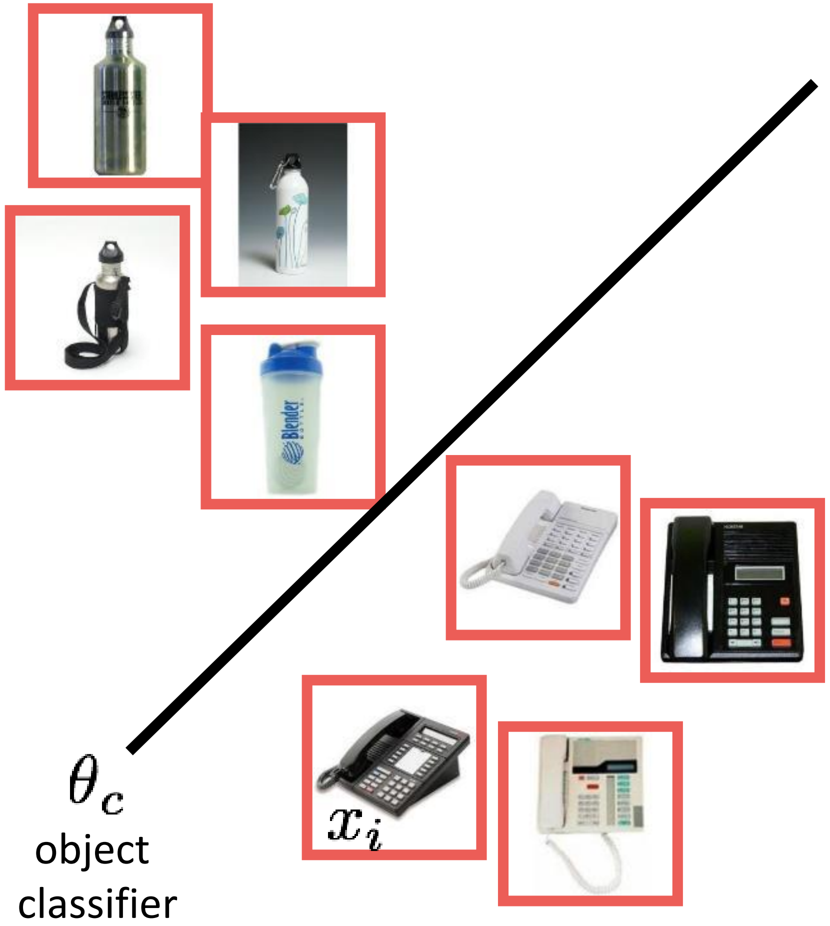


- Fine tune?

  .....Zero or few labels in target domain
- Siamese network?

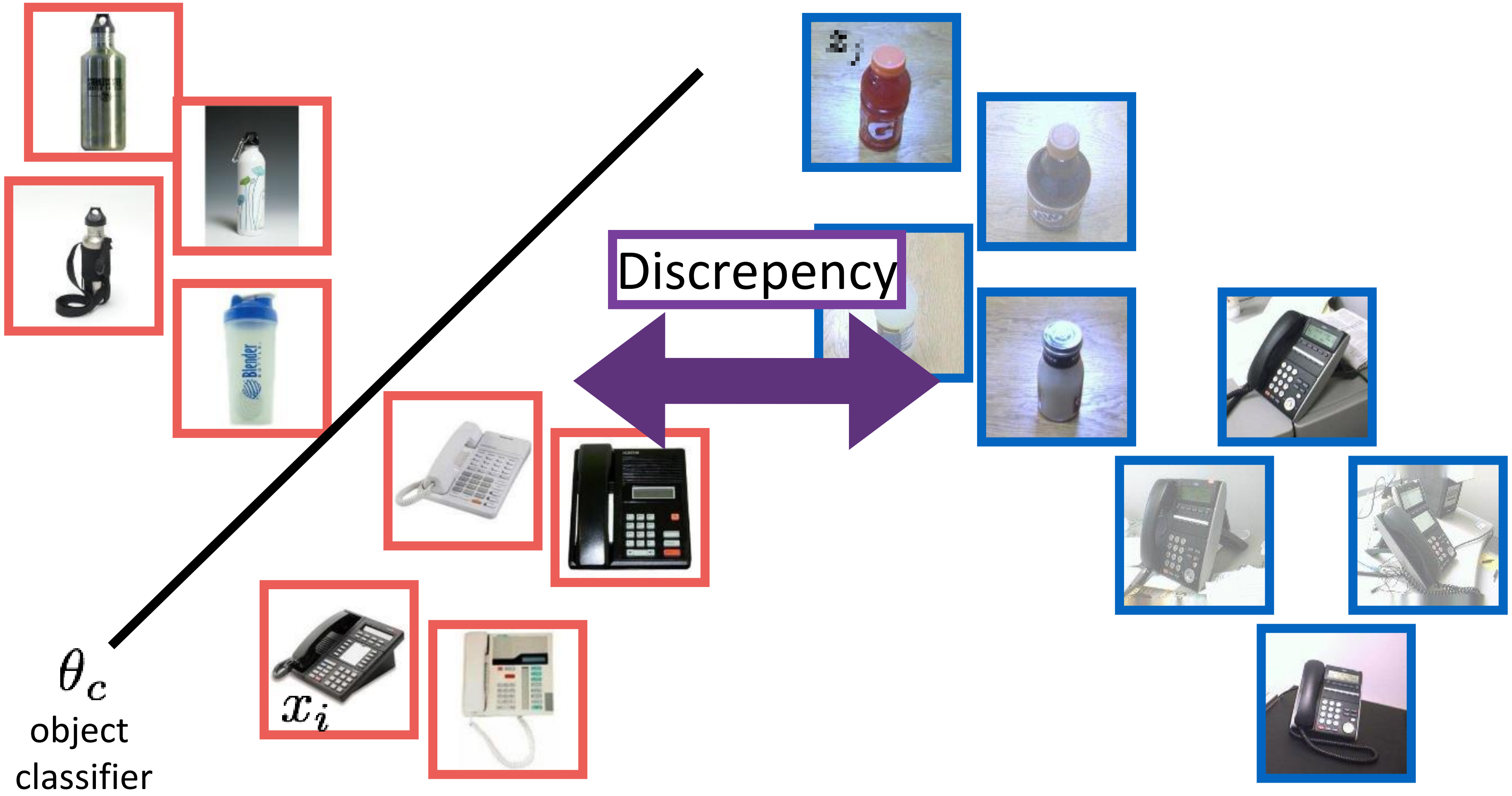
  .....No paired / aligned instance examples!

Adapting across domains: **minimize discrepancy**



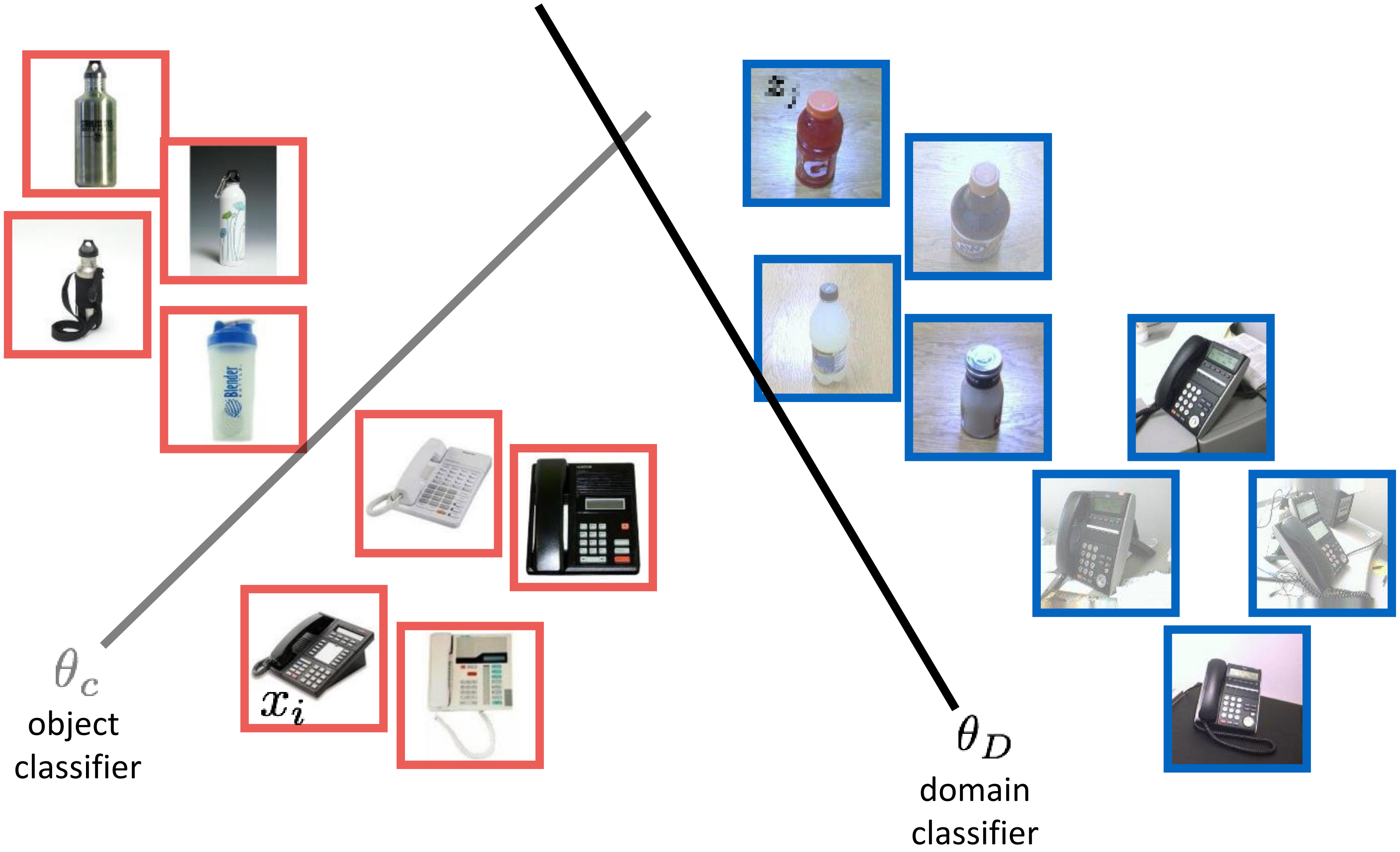
[ICCV 2015]

Adapting across domains: **minimize discrepancy**



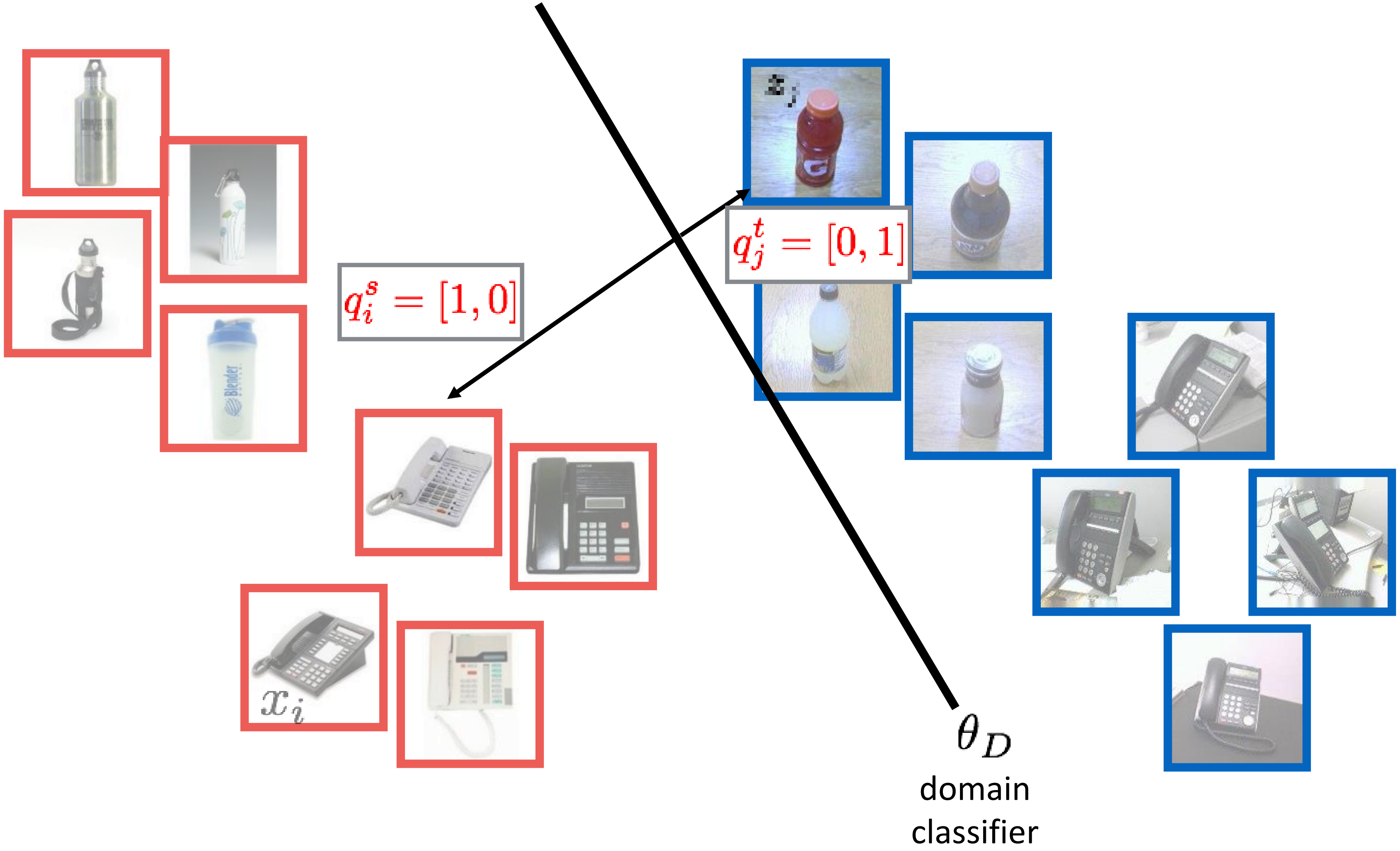
[ICCV 2015]

Adapting across domains: **minimize discrepancy**



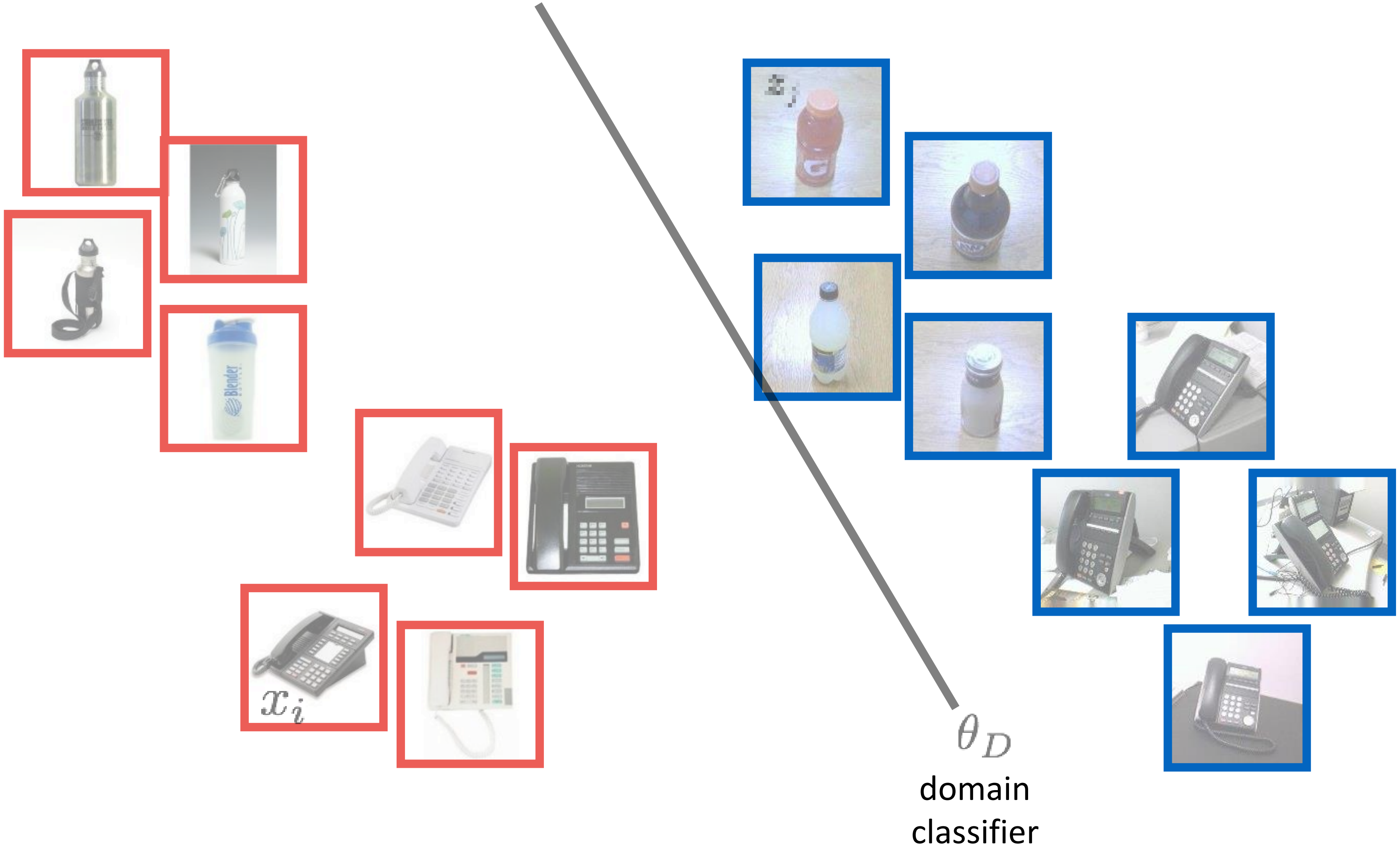
[ICCV 2015]

Adapting across domains: **minimize discrepancy**



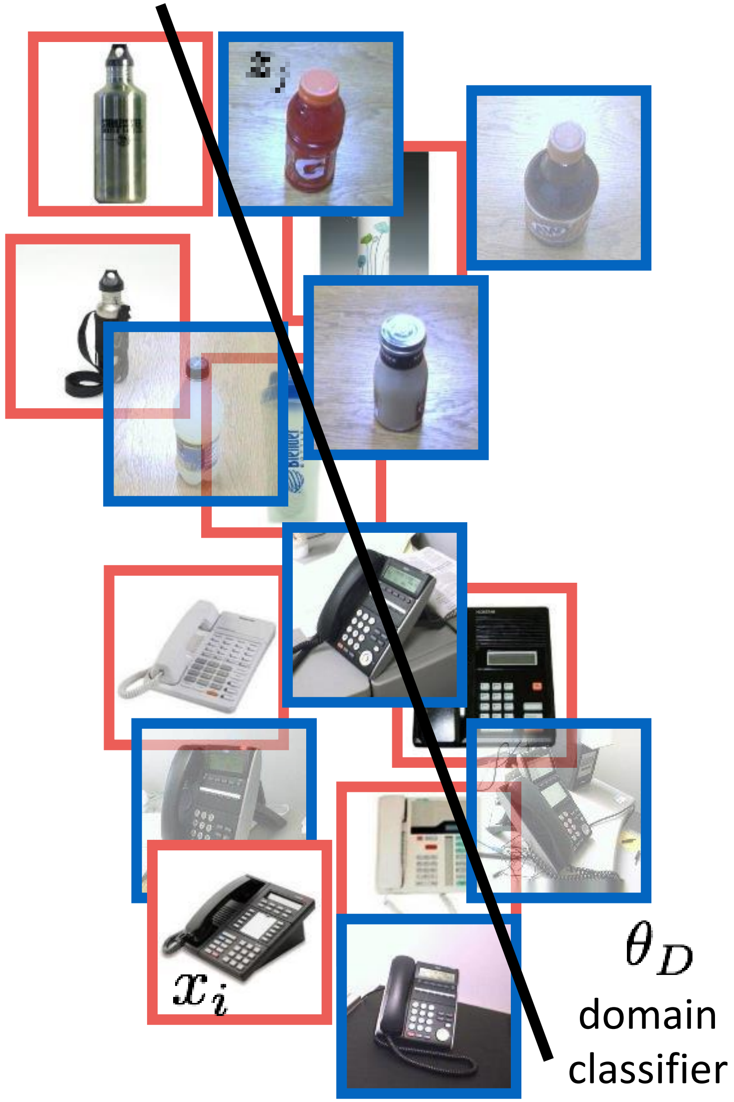
[ICCV 2015]

Adapting across domains: **minimize discrepancy**



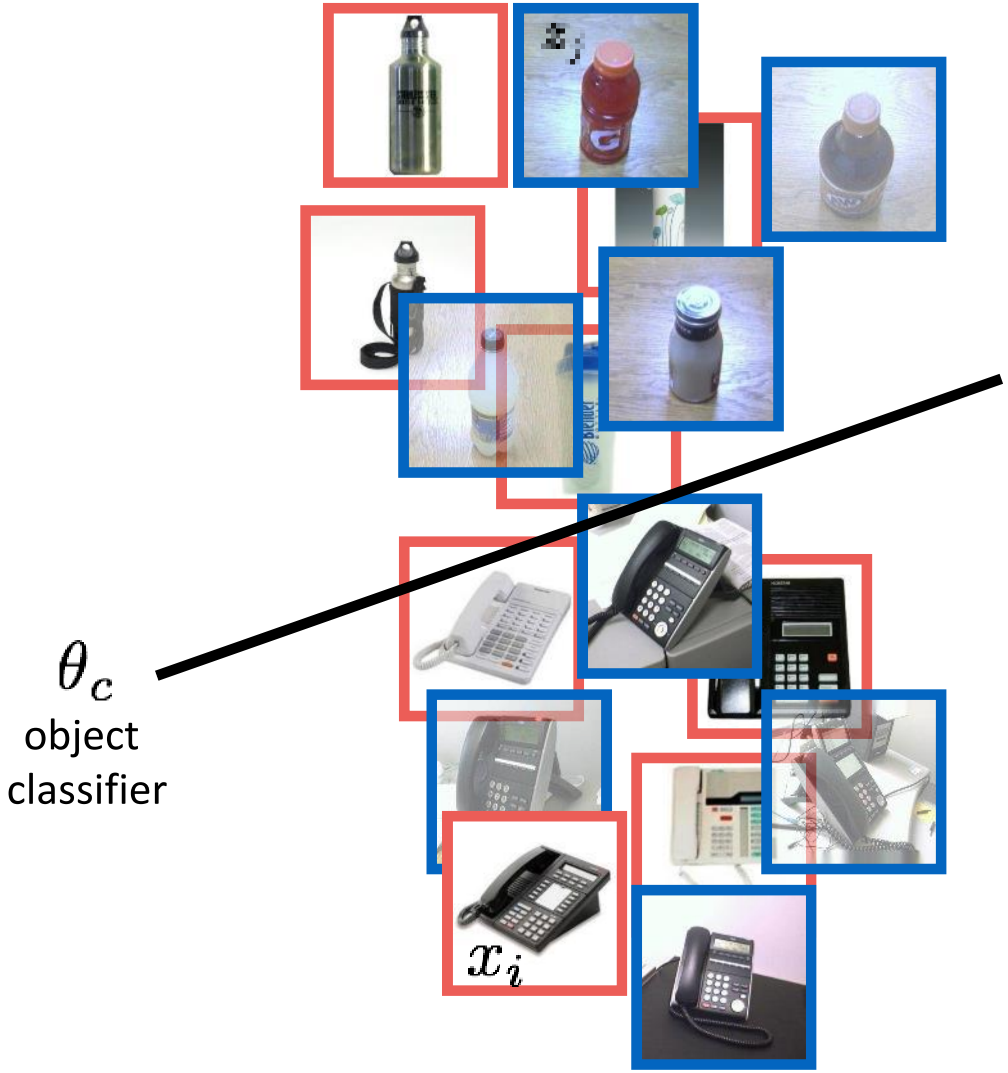
[ICCV 2015]

Adapting across domains: **minimize discrepancy**



[ICCV 2015]

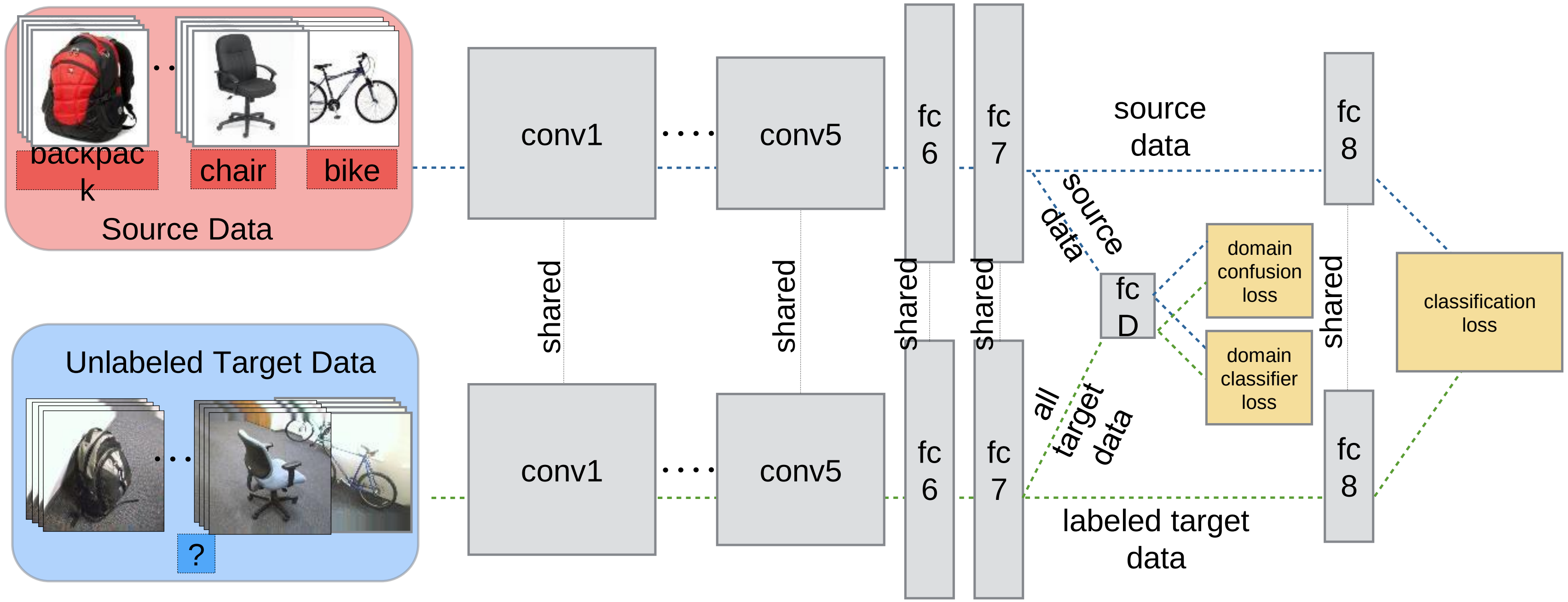
Adapting across domains: **minimize discrepancy**



[ICCV 2015]

Deep domain confusion

[Tzeng ICCV15]



Adversarial Training of domain label predictor and ***domain confusion*** *loss:*

$$\min_{\theta_D} \mathcal{L}_D(x_S, x_T, \theta_{\text{repr}}; \theta_D)$$

$$\min_{\theta_{\text{repr}}} \mathcal{L}_{\text{conf}}(x_S, x_T, \theta_D; \theta_{\text{repr}}).$$

$$\mathcal{L}_D(x_S, x_T, \theta_{\text{repr}}; \theta_D) = -\sum_d \mathbb{1}[y_D = d] \log q_d$$

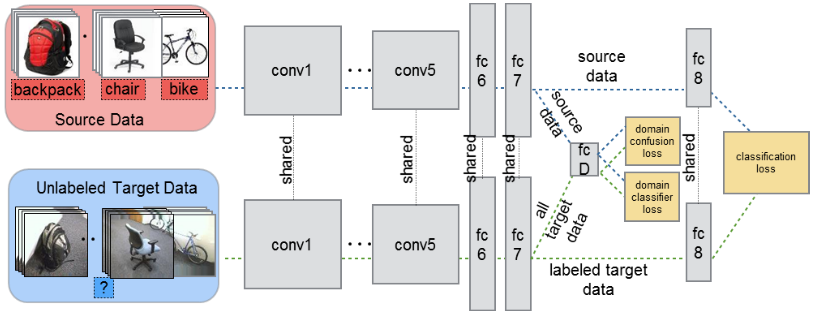
$$\mathcal{L}_{\text{conf}}(x_S, x_T, \theta_D; \theta_{\text{repr}}) = -\sum_d \frac{1}{D} \log q_d.$$

Domain Label Cross-entropy with uniform distribution

Deep domain confusion

[Tzeng ICCV15]

Train a network to minimize classification loss AND confuse two domains



source inputs | target inputs | network parameters (fixed) | domain classifier (learn) | domain classifier loss

$$\mathcal{L}_D(x_S, x_T, \theta_{\text{repr}}; \theta_D) = -\sum_d \mathbb{1}[y_D = d] \log q_d$$

domain classifier prediction

$$q = \text{softmax}(\theta_D^T f(x; \theta_{\text{repr}})) = p(y_D = 1|x)$$

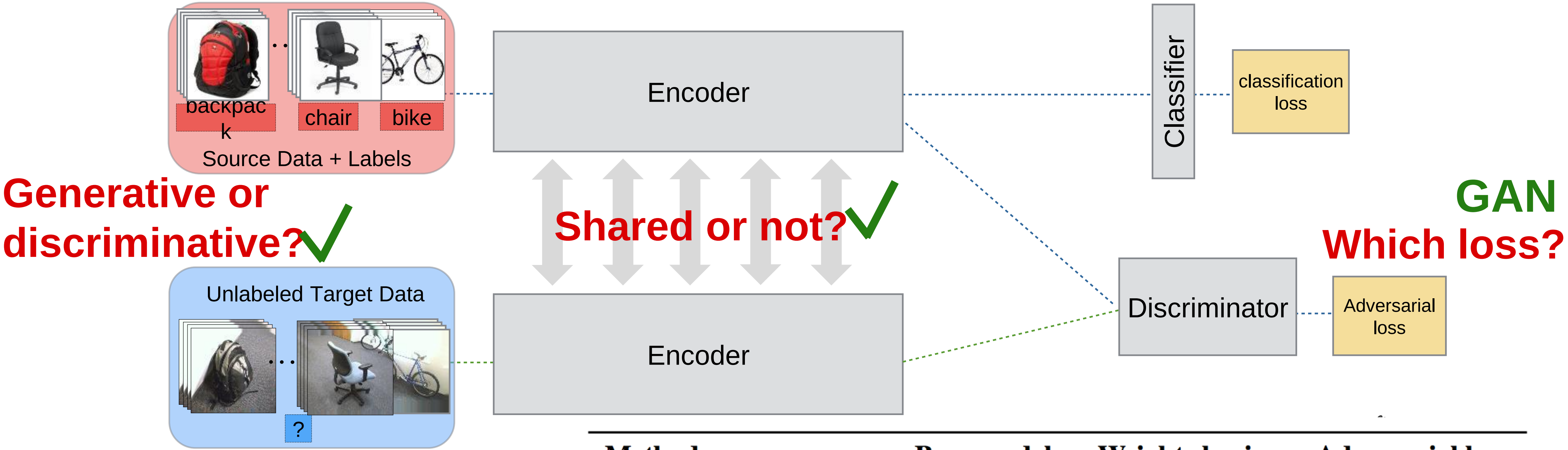
iterate

domain classifier (fixed) | network parameters (learn) | domain confusion loss

$$\mathcal{L}_{\text{conf}}(x_S, x_T, \theta_D; \theta_{\text{repr}}) = -\sum_d \frac{1}{D} \log q_d$$

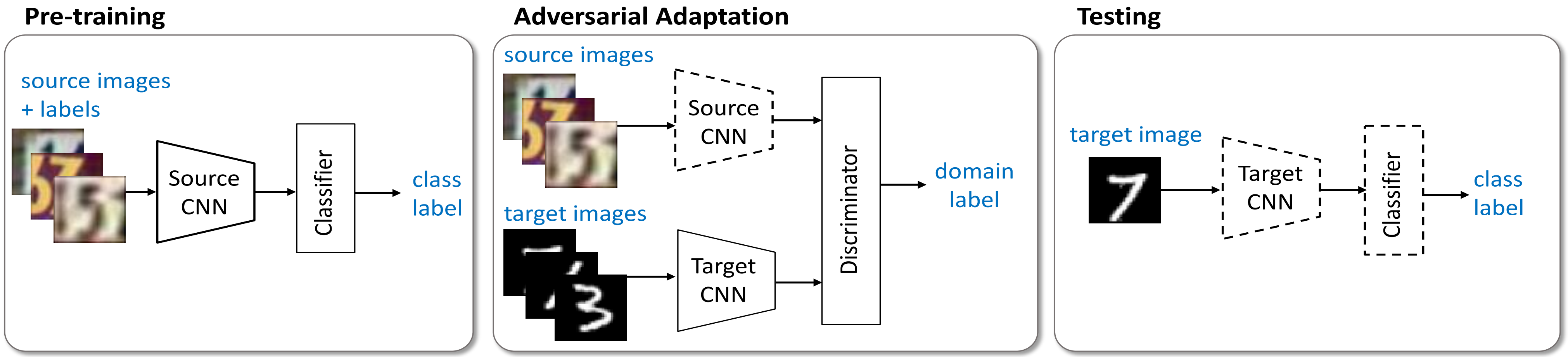
(cross-entropy with uniform distribution)

Adversarial Discriminative Domain Adaptation (ADDA) (in submission)



Method	Base model	Weight sharing	Adversarial loss
Gradient reversal [16]	discriminative	shared	minimax
Domain confusion [12]	discriminative	shared	confusion
CoGAN [13]	generative	unshared	GAN
ADDA (Ours)	discriminative	unshared	GAN

Adversarial Discriminative Domain Adaptation (ADDA) (in submission)



$$\min_{M_s, C} \mathcal{L}_{\text{cls}}(\mathbf{X}_s, Y_s) = -\mathbb{E}_{(\mathbf{x}_s, y_s) \sim (\mathbf{X}_s, Y_s)} \sum_{k=1}^{K} \mathbb{1}_{[k = y_s]} \log C(M_s(\mathbf{x}_s))$$

$$\min_{D} \mathcal{L}_{\text{adv}_D}(\mathbf{X}_s, \mathbf{X}_t, M_s, M_t) = -\mathbb{E}_{\mathbf{x}_s \sim \mathbf{X}_s}[\log D(M_s(\mathbf{x}_s))] - \mathbb{E}_{\mathbf{x}_t \sim \mathbf{X}_t}[\log(1 - D(M_t(\mathbf{x}_t)))]$$

$$\min_{M_s, M_t} \mathcal{L}_{\text{adv}_M}(\mathbf{X}_s, \mathbf{X}_t, D) = -\mathbb{E}_{\mathbf{x}_t \sim \mathbf{X}_t}[\log D(M_t(\mathbf{x}_t))].$$

ADDA: Adaptation on digits

(in submission)

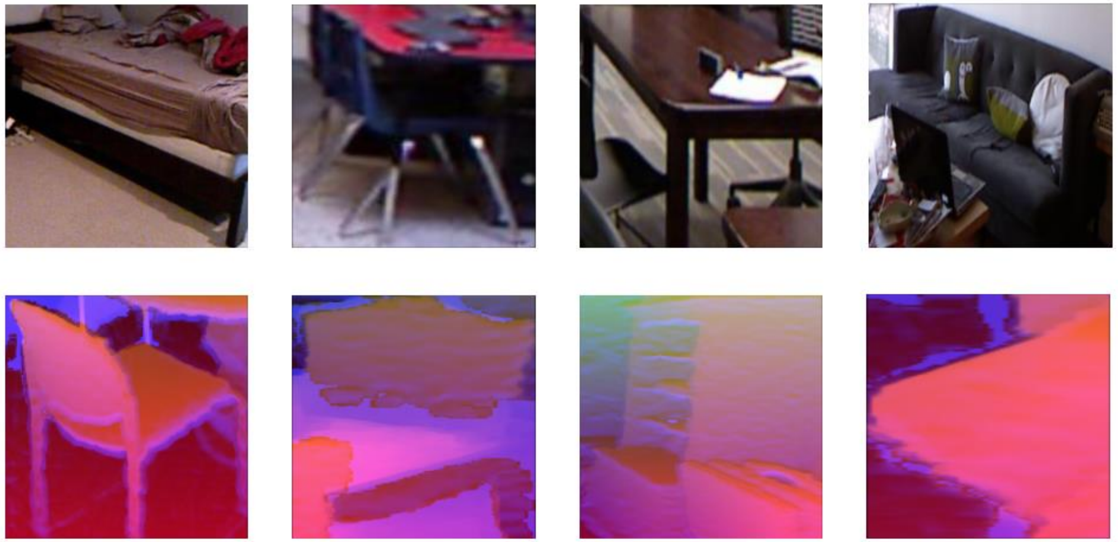
Method	MNIST → USPS	USPS → MNIST	SVHN → MNIST
Source only	0.752 ± 0.016	0.571 ± 0.017	0.601 ± 0.011
Gradient reversal	0.771 ± 0.018	0.730 ± 0.020	0.739 [16]
Domain confusion	0.791 ± 0.005	0.665 ± 0.033	0.681 ± 0.003
CoGAN	0.912 ± 0.008	0.891 ± 0.008	did not converge
ADDA (Ours)	0.894 ± 0.002	0.901 ± 0.008	0.760 ± 0.018

ADDA: Adaptation on RGB-D

(in submission)

Train on RGB

Test on depth



	bathtub	bed	bookshelf	box	chair	counter	desk	door	dresser	garbage bin	lamp	monitor	night stand	pillow	sink	sofa	table	television	toilet	**overall**
# of instances	19	96	87	210	611	103	122	129	25	55	144	37	51	276	47	129	210	33	17	2401
Source only	0.000	0.010	0.011	0.124	0.188	0.029	0.041	0.047	0.000	0.000	0.069	0.000	0.039	0.587	0.000	0.008	0.010	0.000	0.000	0.139
ADDA (Ours)	0.000	0.146	0.046	0.229	0.344	0.447	0.025	0.023	0.000	0.018	0.292	0.081	0.020	0.297	0.021	0.116	0.143	0.091	0.000	0.211
Train on target	0.105	0.531	0.494	0.295	0.619	0.573	0.057	0.636	0.120	0.291	0.576	0.189	0.235	0.630	0.362	0.248	0.357	0.303	0.647	0.468

Autonomous Driving Paradigms

1) Learn affordances to predict state; apply rules or learned classic controllers

2) Abandon engineering principles, learn “end-to-end” policy

Autonomous Driving Paradigms

1) Learn affordances to predict state; apply rules or learned classic controllers

How can visual sensing be robust to new enviroments?

2) Abandon engineering principles, learn “end-to-end” policy

How to learn generic driving policies from diverse data?

Autonomous Driving Paradigms

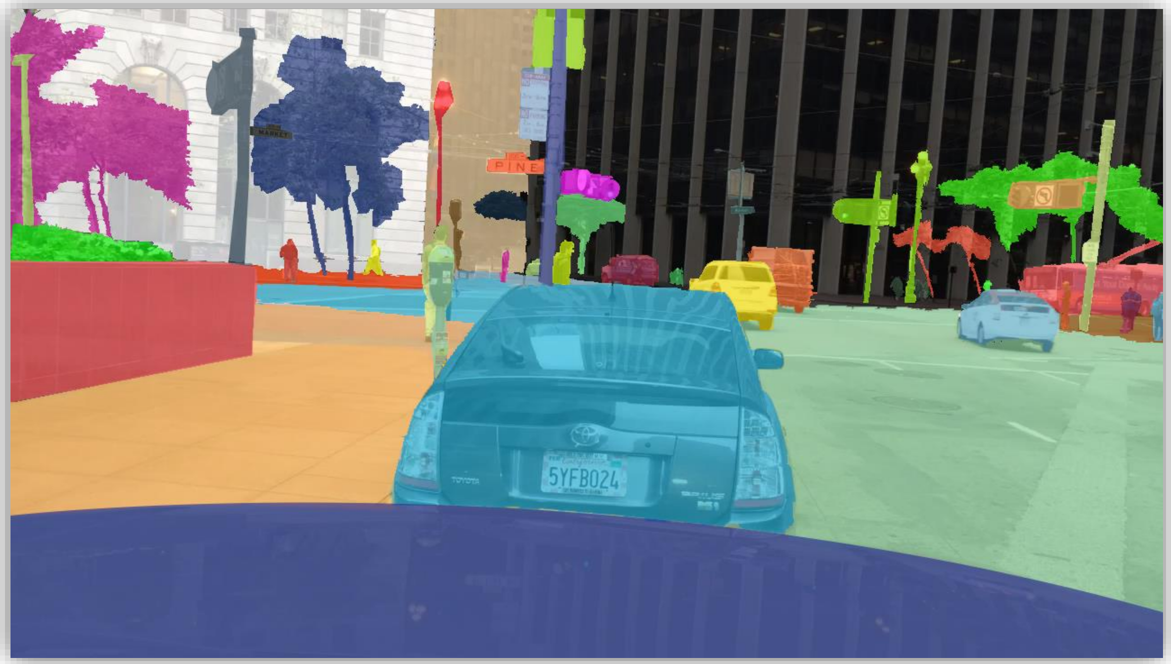
1) Learn affordances to predict state; apply rules or learned classic controllers

How can visual sensing be robust to new enviroments?
…Fully Convolutional Domain Adaptation "in the wild"

2) Abandon engineering principles, learn "end-to-end" policy

How to learn generic driving policies from diverse data?
…Learning end-to-end driving policy/model from crowdsourced videos

BDD Dataset



BDD Video

- 720p 30fps 40s video clips
- ~50K clips
- GPS + IMU

BDD Segmentation

- 720p images
- Fine instance segmentation
- Compatible with Cityscapes

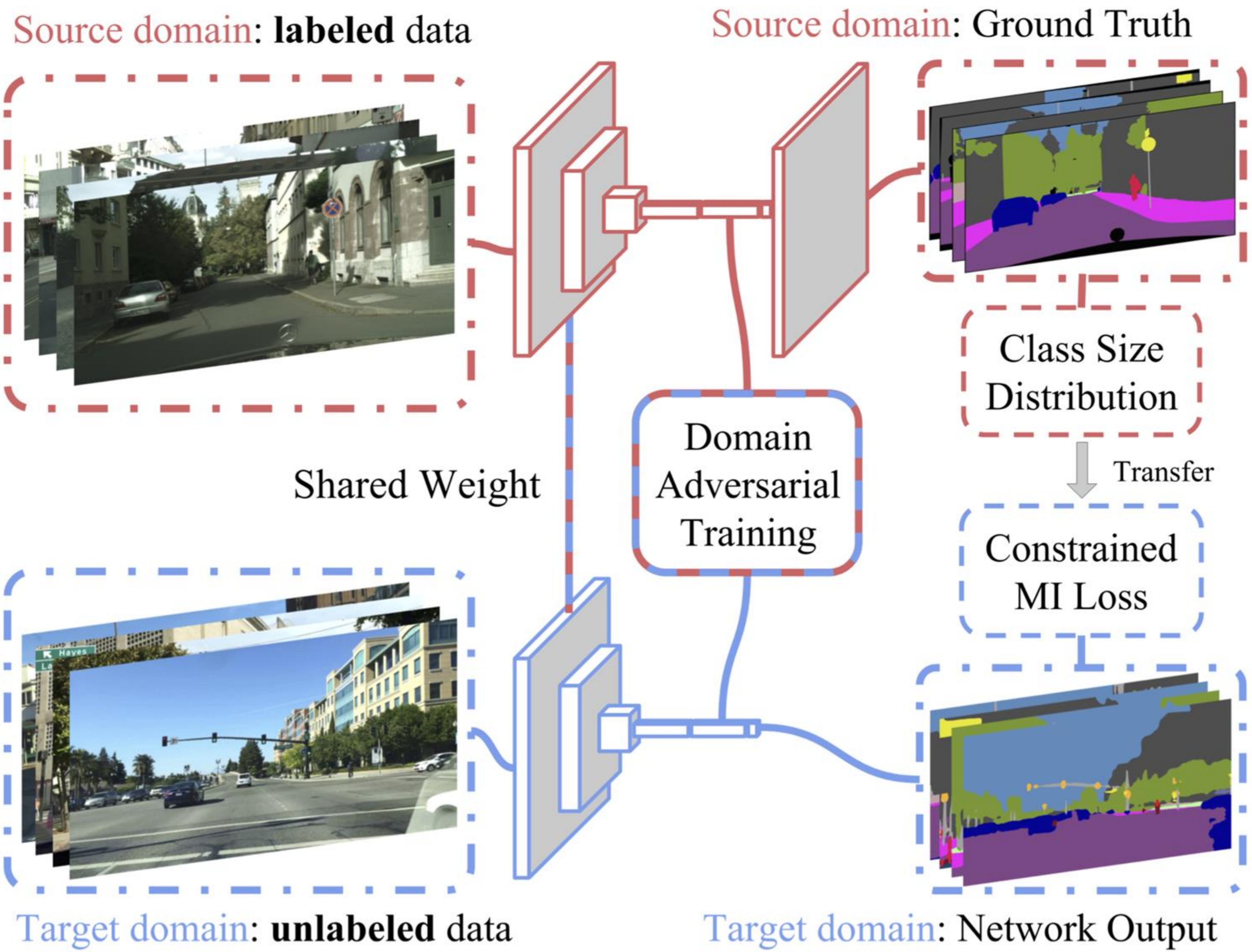
In-domain fully supervised FCN

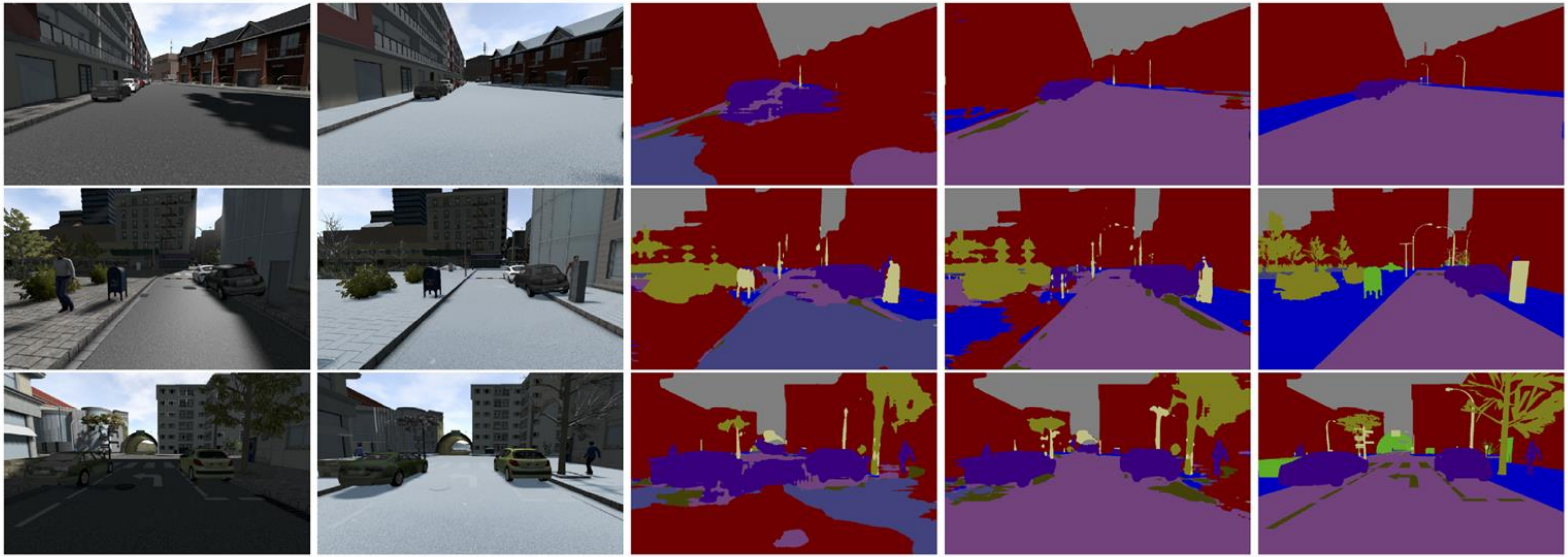
Train on Cityscapes, Test on Cityscapes

Domain shift: Cityscapes to SF

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

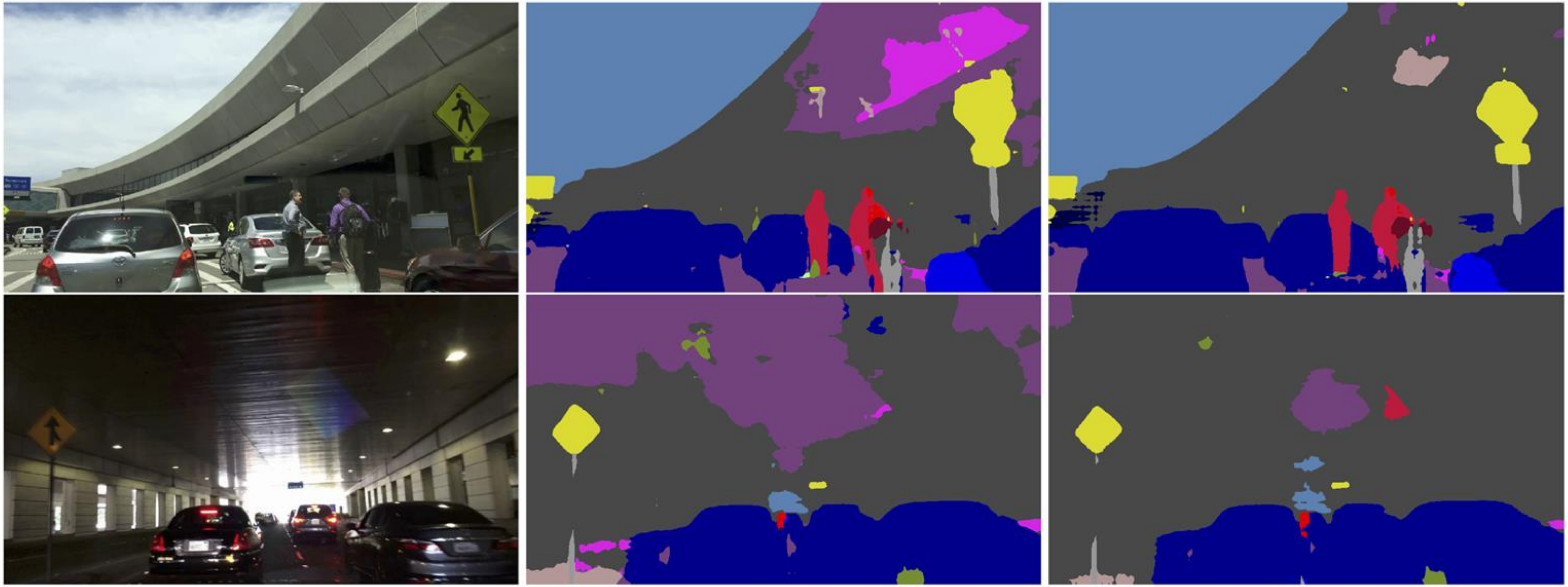
No tunnels in CityScapes?...


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



(a) Fall Image (b) Winter Image (c) Before Adaptation (d) After Adaptation (e) Ground Truth

Medium Shift: Cross Seasons Adaptation



(a) Original Image (b) Before Adaptation (c) After Adaptation

Small Shift: Cross City Adaptation

Effect of domain confusion loss

Before domain confusion

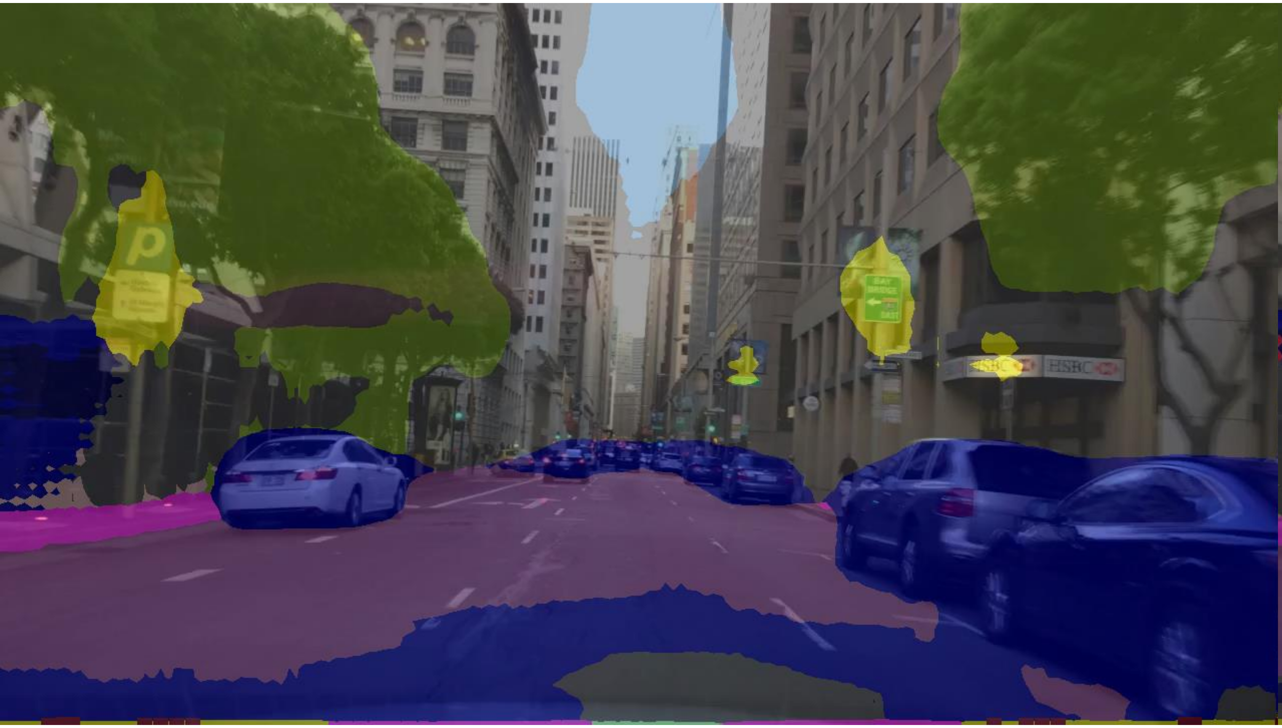
After domain confusion

Effect of domain confusion loss

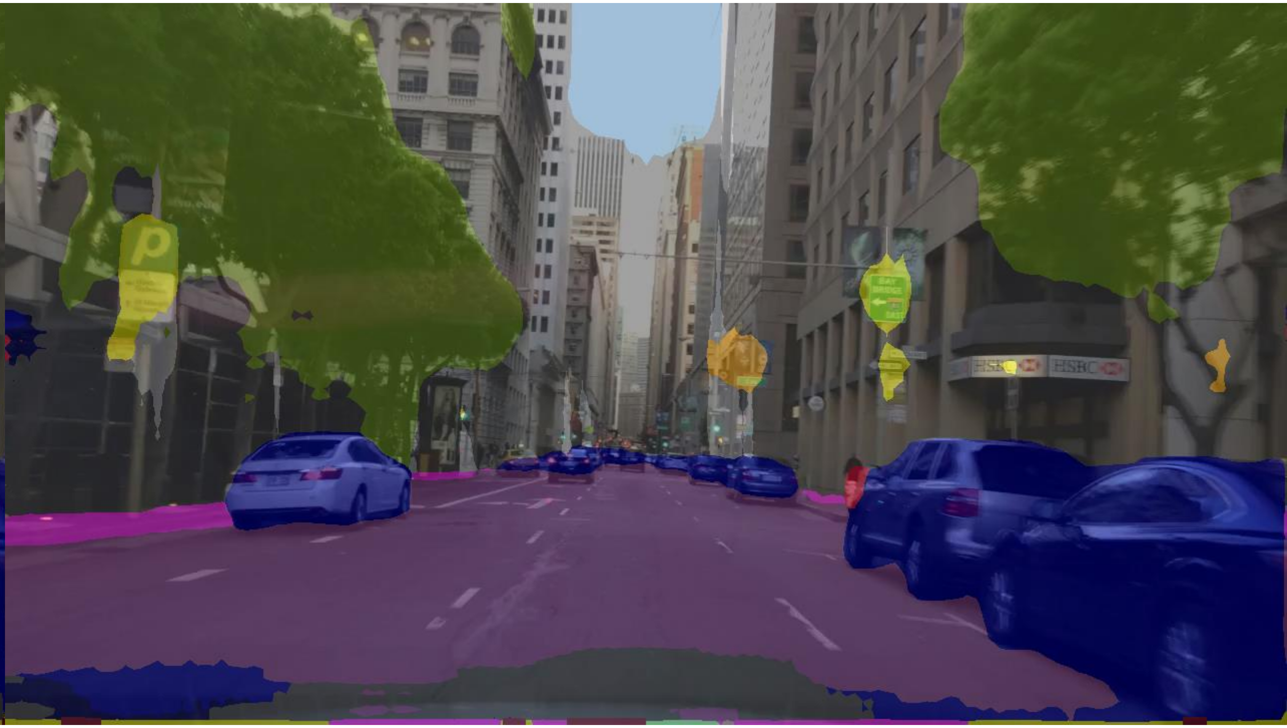
Before domain confusion

After domain confusion

Effect of domain confusion loss

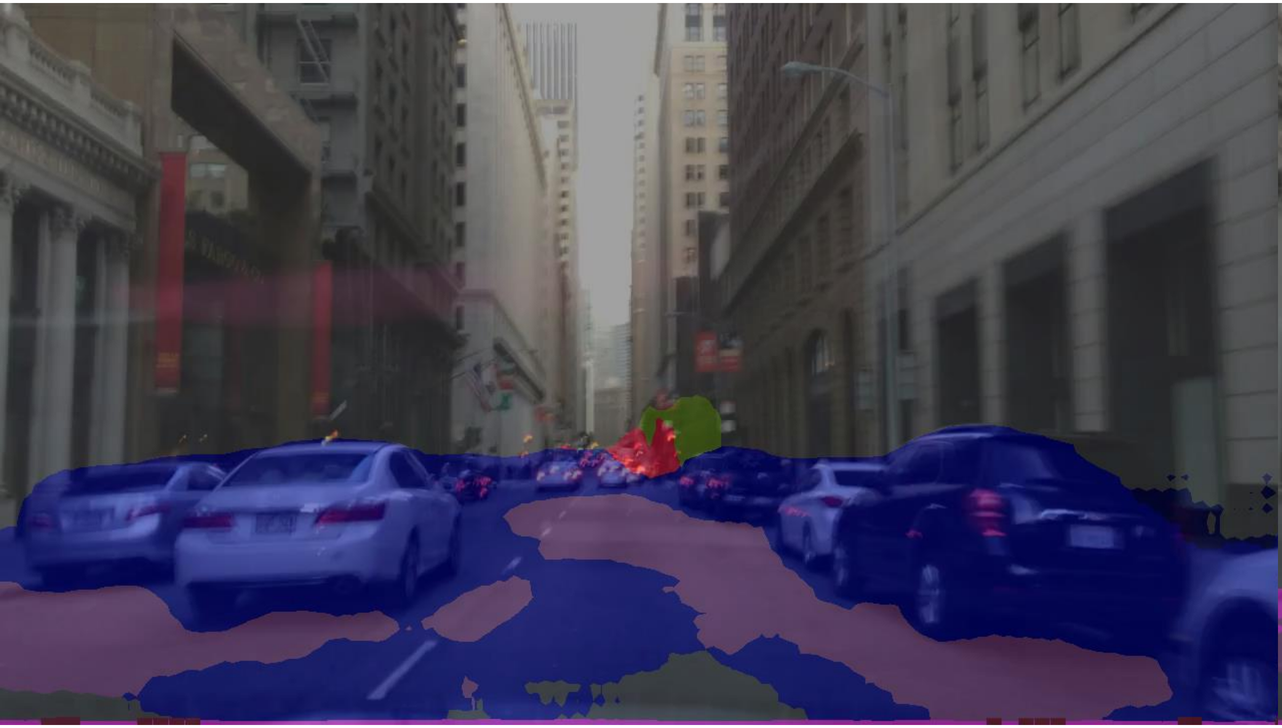


Before domain confusion



After domain confusion

Effect of domain confusion loss



Before domain confusion

After domain confusion

BDD Dataset – static

arXiv.org > cs > arXiv:1612.02649

Search or Article ID inside arXiv | All papers | Broaden yo

(Help | Advanced search)

Computer Science > Computer Vision and Pattern Recognition

FCNs in the Wild: Pixel-level Adversarial and Constraint-based Adaptation

Judy Hoffman, Dequan Wang, Fisher Yu, Trevor Darrell

(Submitted on 8 Dec 2016)

Fully convolutional models for dense prediction have proven successful for a wide range of visual tasks. Such models perform well in a supervised setting, but performance can be surprisingly poor under domain shifts that appear mild to a human observer. For example, training on one city and testing on another in a different geographic region and/or weather condition may result in significantly degraded performance due to pixel-level distribution shift. In this paper, we introduce the first domain adaptive semantic segmentation method, proposing an unsupervised adversarial approach to pixel prediction problems. Our method consists of both global and category specific adaptation techniques. Global domain alignment is performed using a novel semantic segmentation network with fully convolutional domain adversarial learning. This initially adapted space then enables category specific adaptation through a generalization of constrained weak learning, with explicit transfer of the spatial layout from the source to the target domains. Our approach outperforms baselines across different settings on multiple large-scale datasets, including adapting across various real city environments, different synthetic sub-domains, from simulated to real environments, and on a novel large-scale dash-cam dataset.

Overview

Adversarial Domain Adaptation

Learning end-to-end driving models from crowdsourced dashcams

Vision and Language: Learning to reason to answer and explain

Learning and Adapting from Large-Scale Driving Data

- Fully Convolutional Domain Adaptation “in the wild”
- **Learning end-to-end driving policy/model from dashcam videos**

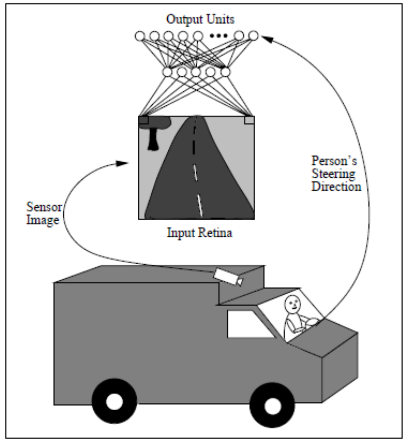
End-to-End Paradigm

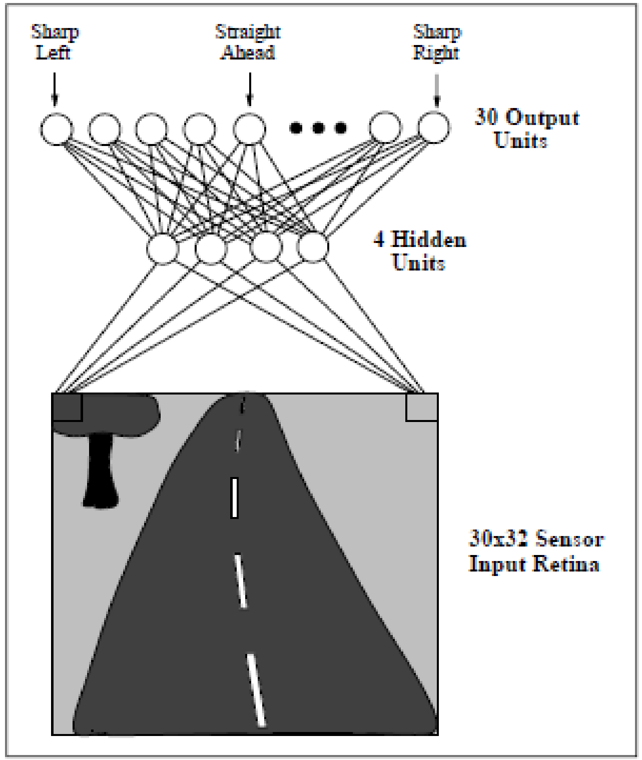
- ALVINN
- DAVE
- NVIDIA
- BDD RC Cars
- BDD WebCam

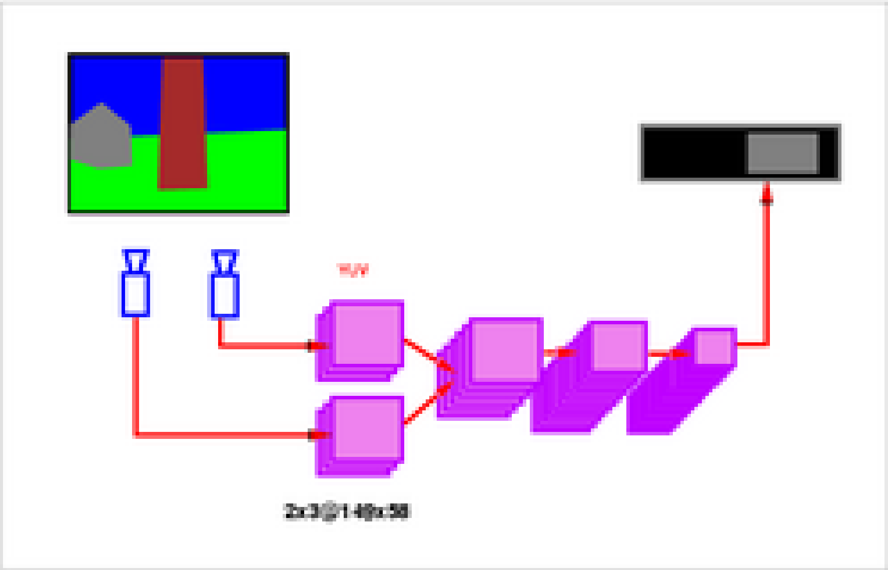
AVLINN (1989)

ALVINN:
An Autonomous Land Vehicle In a Neural Network

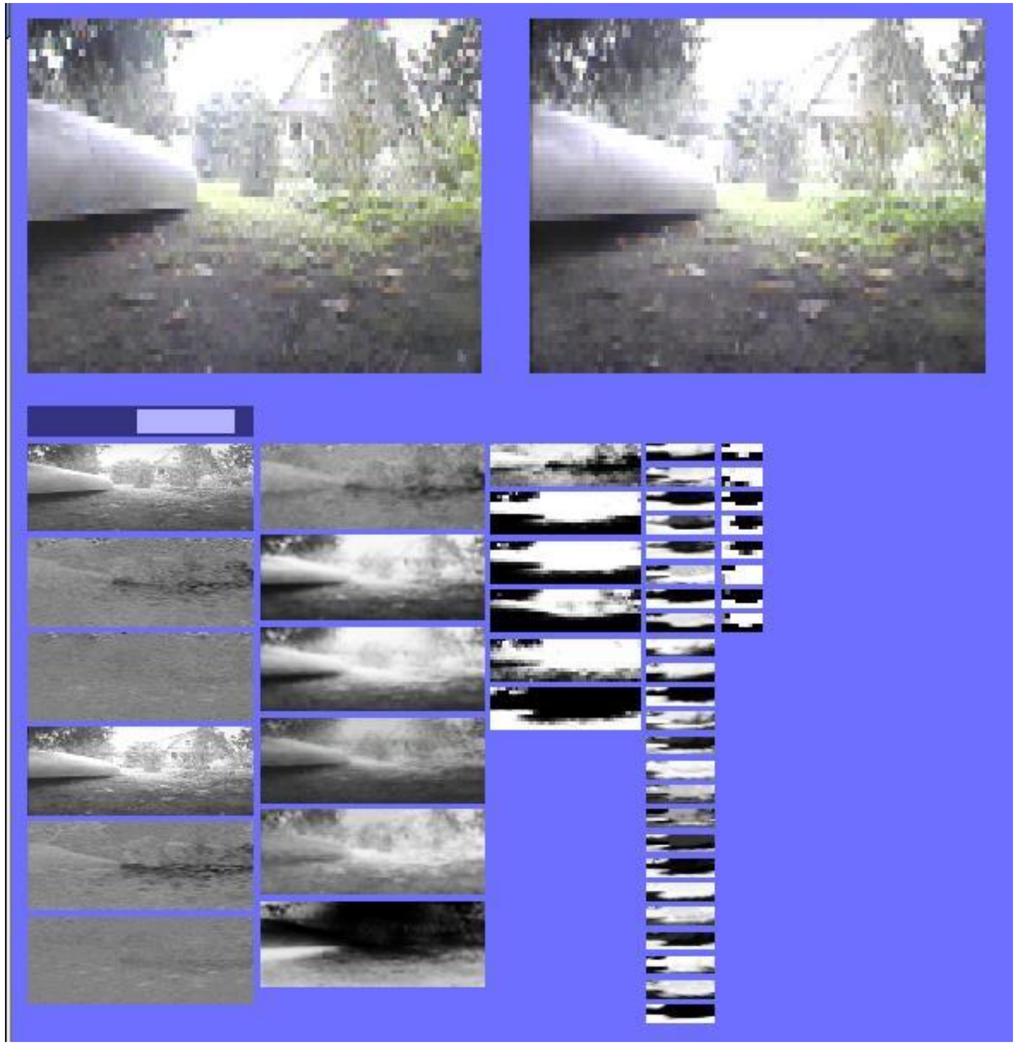
Dean A. Pomerleau
January 1989
CMU-CS-89-107-





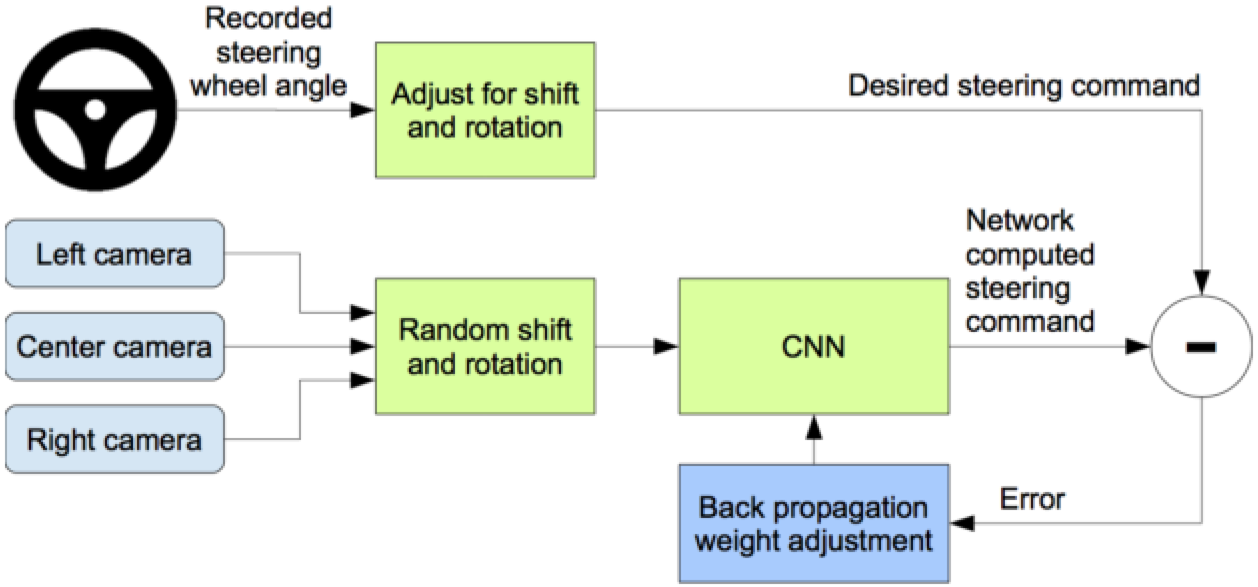


DAVE (2003)

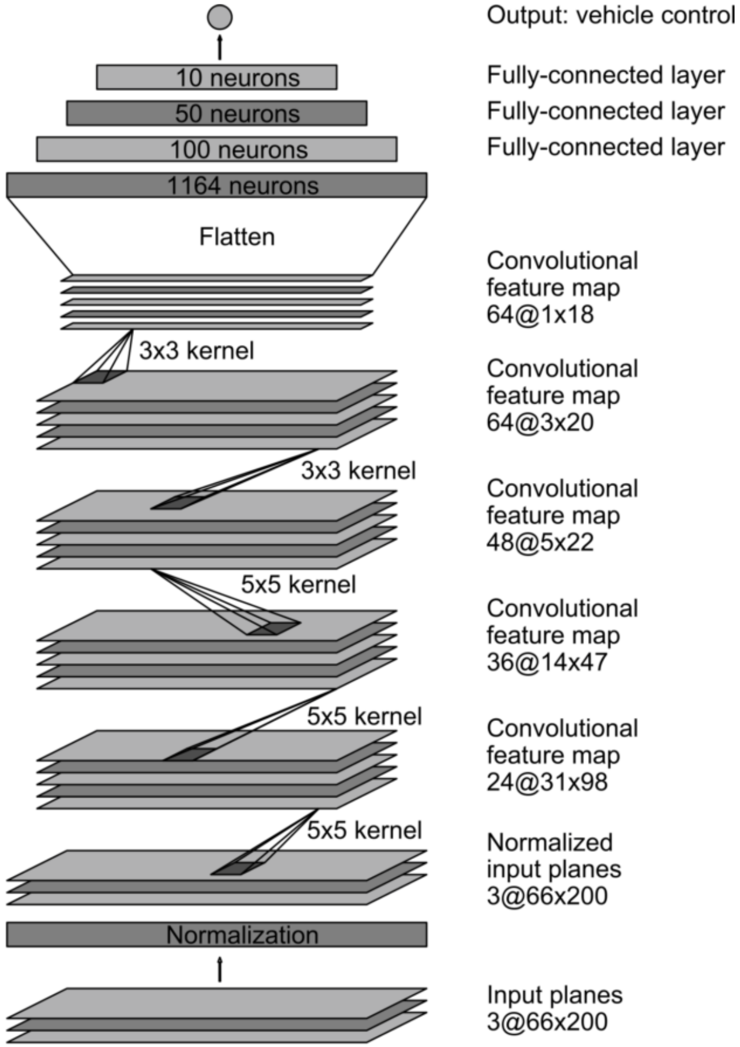


Yann LeCun, Eric Cosatto, Jan Ben, Urs Muller, Beat Flepp: *End-to-End Learning of Vision-Based Obstacle Avoidance for Off-Road Robots*. Delivered at the Learning@Snowbird Workshop, April 2004.

NVIDIA (2016)



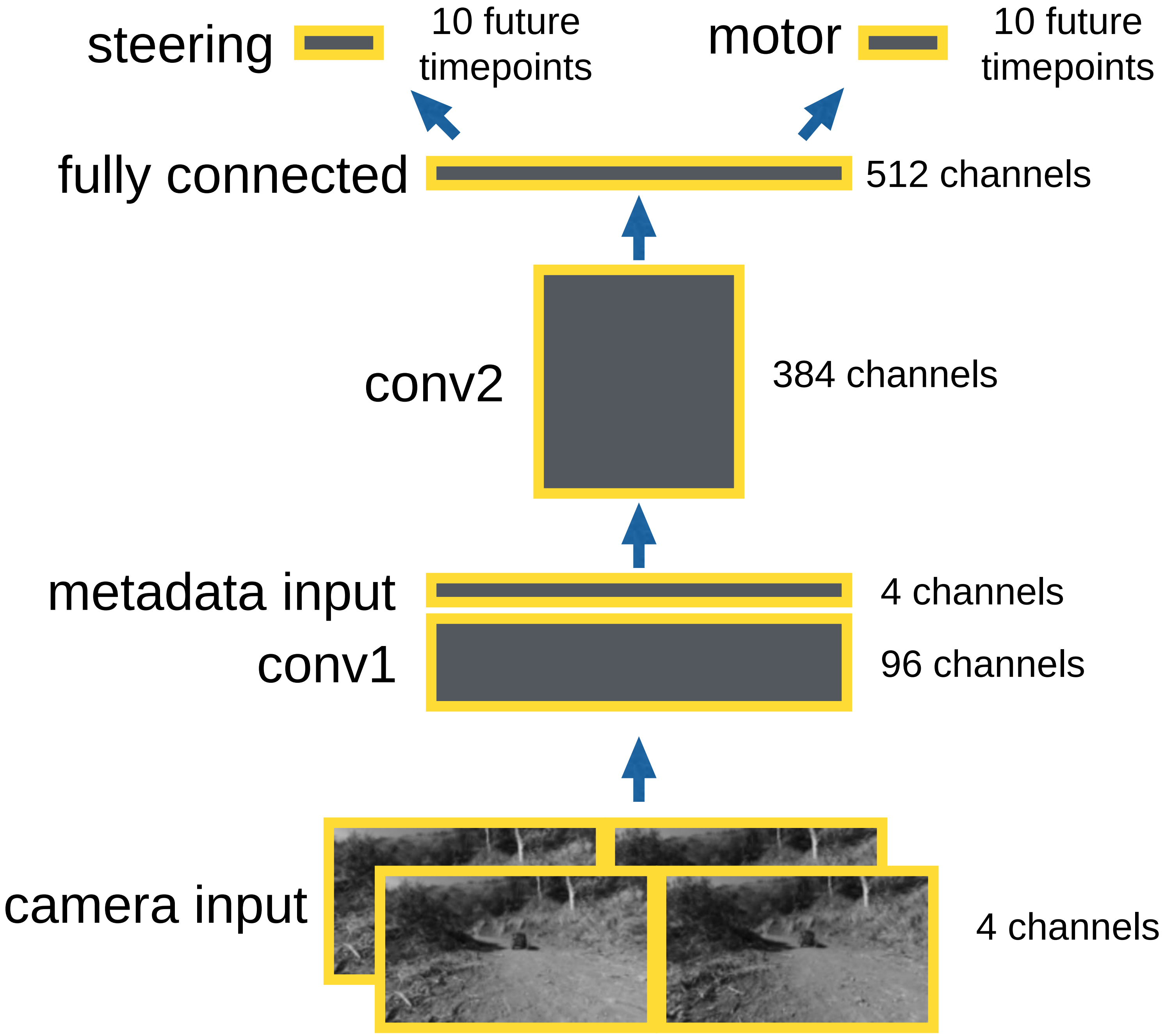
Recorded steering wheel angle
Adjust for shift and rotation
Desired steering command
Left camera
Center camera
Right camera
Random shift and rotation
CNN
Network computed steering command
Back propagation weight adjustment
Error



Output: vehicle control
10 neurons
Fully-connected layer
50 neurons
Fully-connected layer
100 neurons
Fully-connected layer
1164 neurons
Flatten
Convolutional feature map 64@1x18
3x3 kernel
Convolutional feature map 64@3x20
3x3 kernel
Convolutional feature map 48@5x22
5x5 kernel
Convolutional feature map 36@14x47
5x5 kernel
Convolutional feature map 24@31x98
5x5 kernel
Normalized input planes 3@66x200
Normalization
Input planes 3@66x200

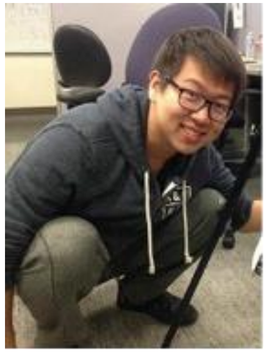
[Karl Zipser]

model driving car, ‘direct’ mode


steering
10 future timepoints
motor
10 future timepoints
fully connected
512 channels
conv2
384 channels
metadata input
4 channels
conv1
96 channels
camera input
4 channels


R8
R8

Driving Policy



Huazhe

Yang

Trevor

Learning a universal driving policy

- Self driving as egomotion prediction
- Learn general driving policy that is applicable to all car models.
- Use a large number of easily accessible dashcam videos as self-supervision.

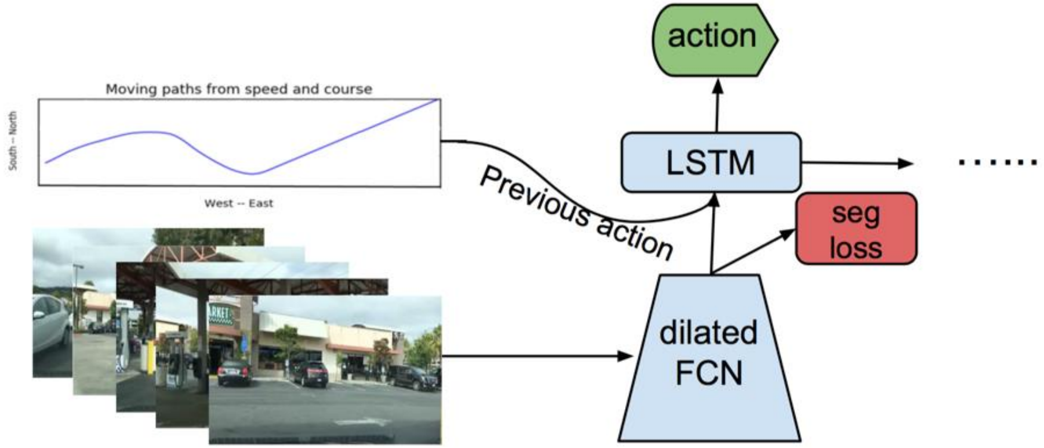
FCN-LSTM

Visual Encoder

- Dilated Fully Convolutional Nets could provide more spatial details than CNN

Temporal Fusion

- Fuse the visual information, vehicle state (speed and angular velocity) from each frame



FCN-LSTM

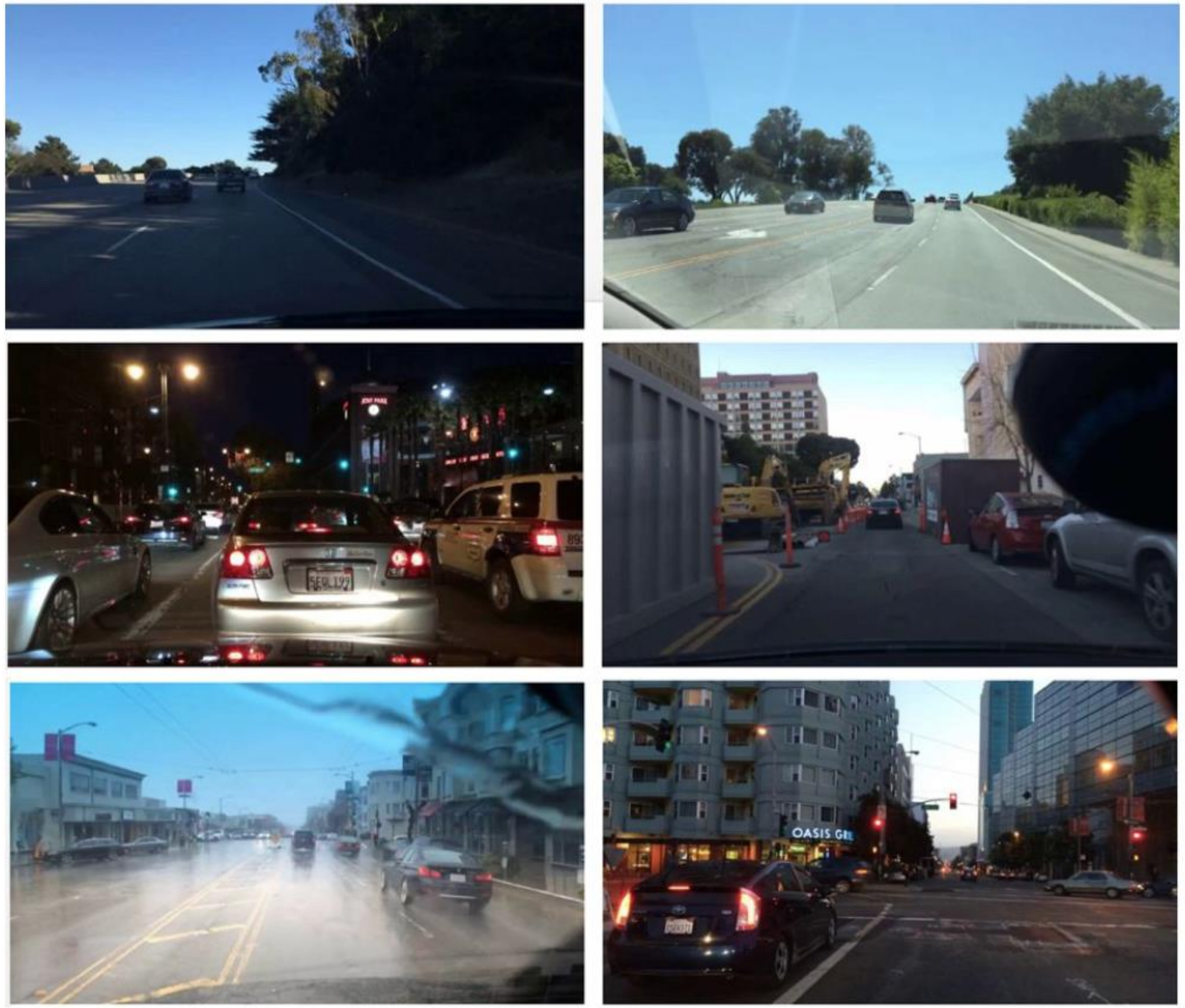
Privileged Learning

- The model should implicitly know what objects are in the scene
- We use the semantic segmentation mask from Cityscapes as extra source of supervision
- It ultimately improves the learnt representation of the dilated FCN

Vapnik V, Vashist A. A new learning paradigm: Learning using privileged information[J]. Neural Networks the Official Journal of the International Neural Network Society, 2009, 22(5-6):544-57.

Dataset

- Real first person driving videos
- Diverse
  - City
  - Highway
  - Rainy days
  - Nights and evenings
  - Construction zones



Sample frames from the dataset

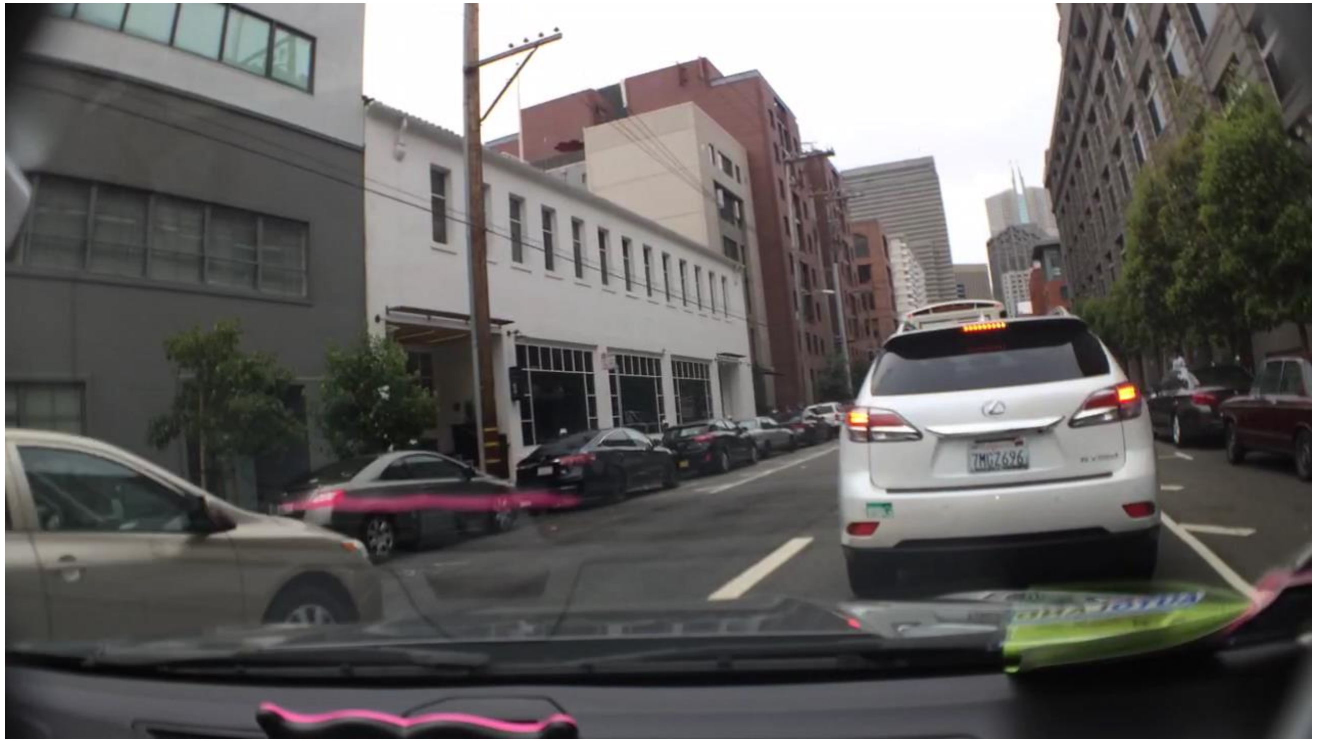
Scene and Trajectory Reconstruction of Crowd-sourced Driving Videos using Semantic Filtered SfM

Yang Gao*, Huazhe Xu*, Christian Hane, Fisher Yu, Trevor Darrell

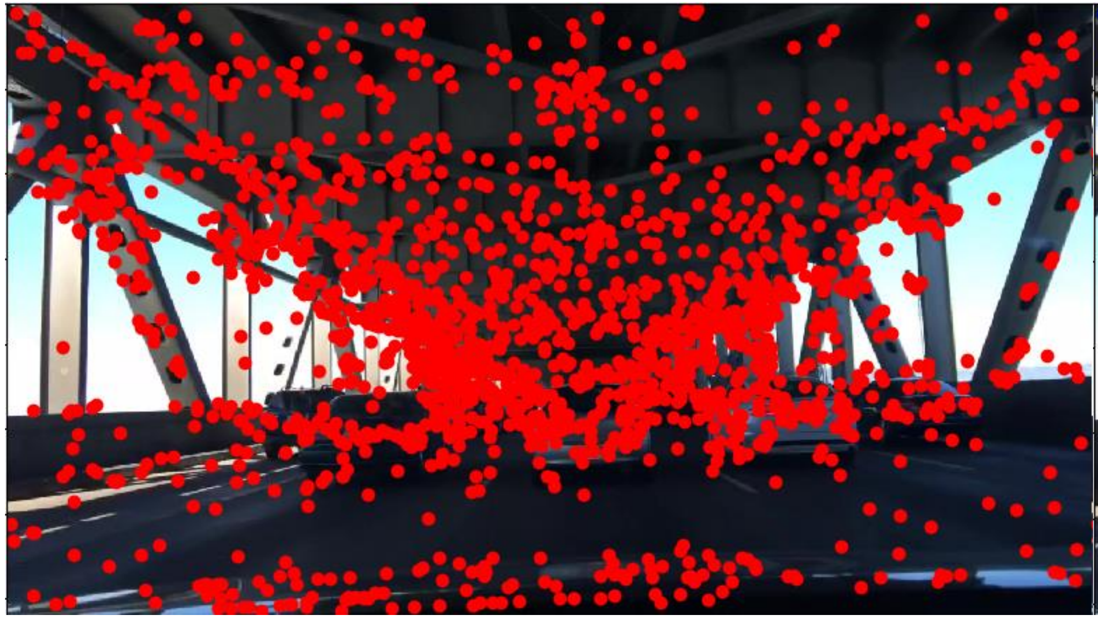
Challenging Driving Videos in the Wild

Challenges

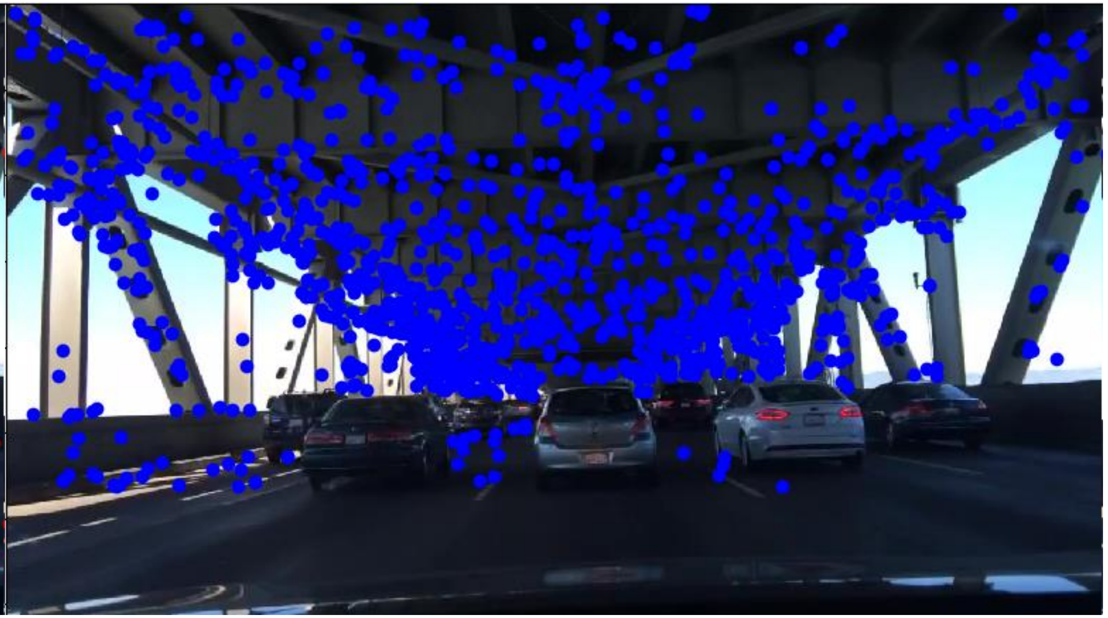
- Moving Objects
- Subtle behaviors
  - Lane changing
  - Slight Steering
- Unknown Camera Calibration
- Rolling Shutters



Existing Motion-Based Method Failed to Reject Moving Object from the Scene



Keypoints from motion-based keypoints rejection methods



Keypoints from our Semantic Filtered SfM pipeline. Most moving keypoints have been filtered out.

Semantically Filtered SfM: $(Sf)^2M$

Classical keypoints matching as points pair preference ranking

$$M(i_1, i_2) = \frac{1}{\left|\left|d(I_1, i_1), d(I_2, i_2)\right|\right|_2}$$

M is the preference score over point pair (i_1, i_2), defined by distance between two low level descriptors $d(\cdot,\cdot)$.

Classical matchings could be formulated as ranking based on $M(\cdot,\cdot)$

Semantics should be incorporated in SfM to be robust to moving objects

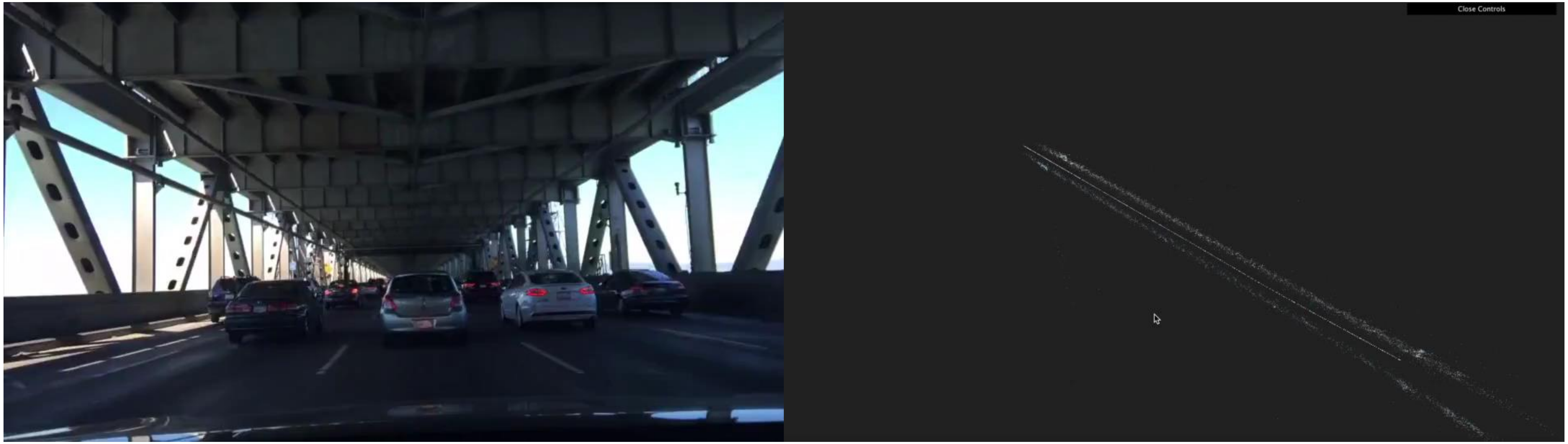
$$M(i_1, i_2) = \frac{Semantic(I_1, I_2)[i_1, i_2]}{\left|\left|d(I_1, i_1), d(I_2, i_2)\right|\right|_2}$$

Use the FCN as a semantic term

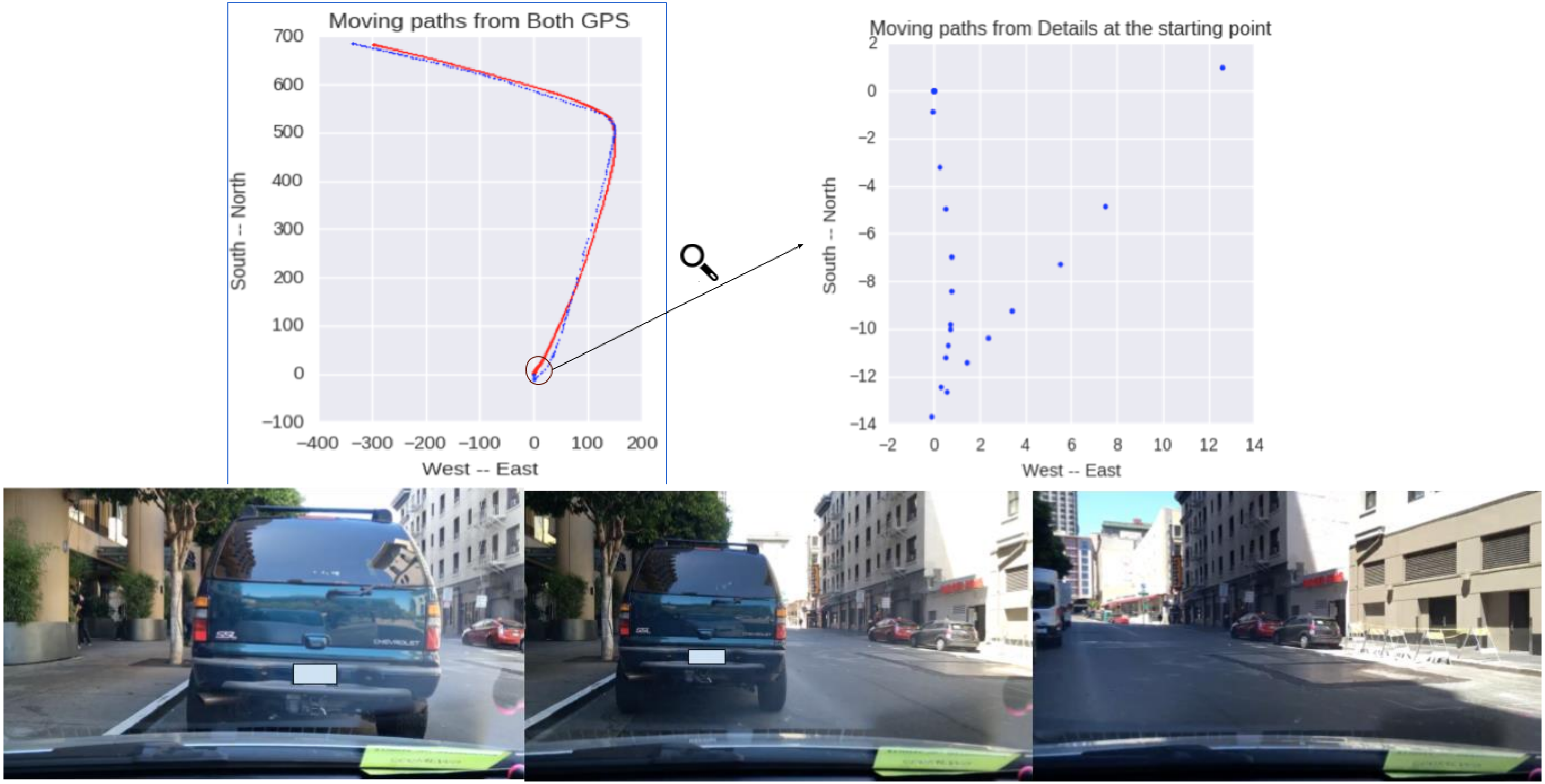
$$Semantic(I_1, I_2)[i_1, i_2] = FCN(I_1)[i_1] \cdot FCN(I_2)[i_2]$$

City Turning Example

Lots of Moving Vehicles Example



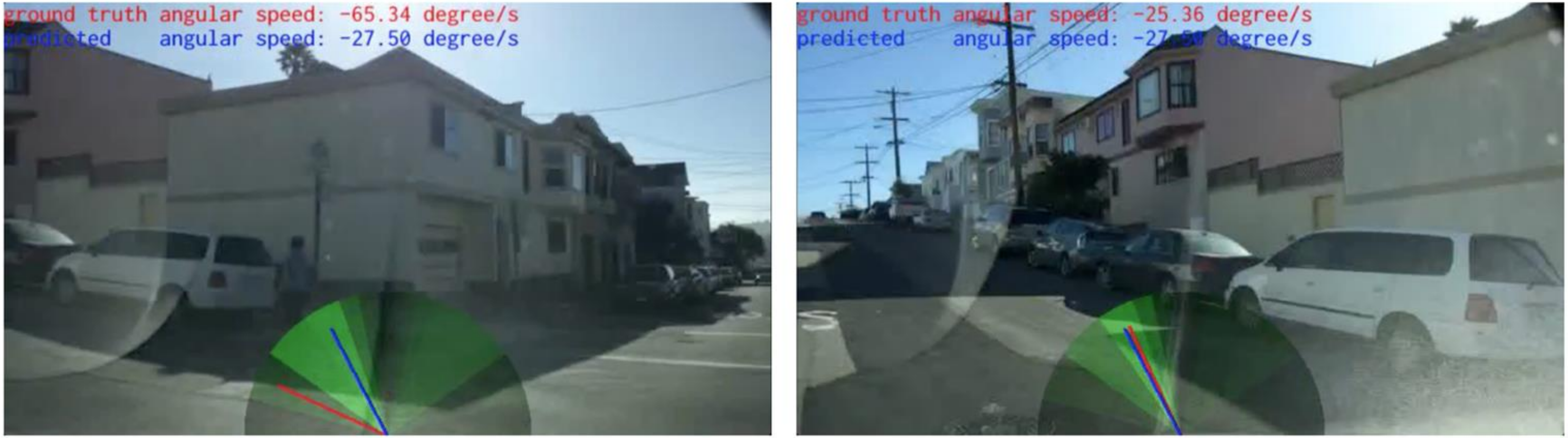
Recover the subtle car backing behavior



Experiments – Continuous Actions

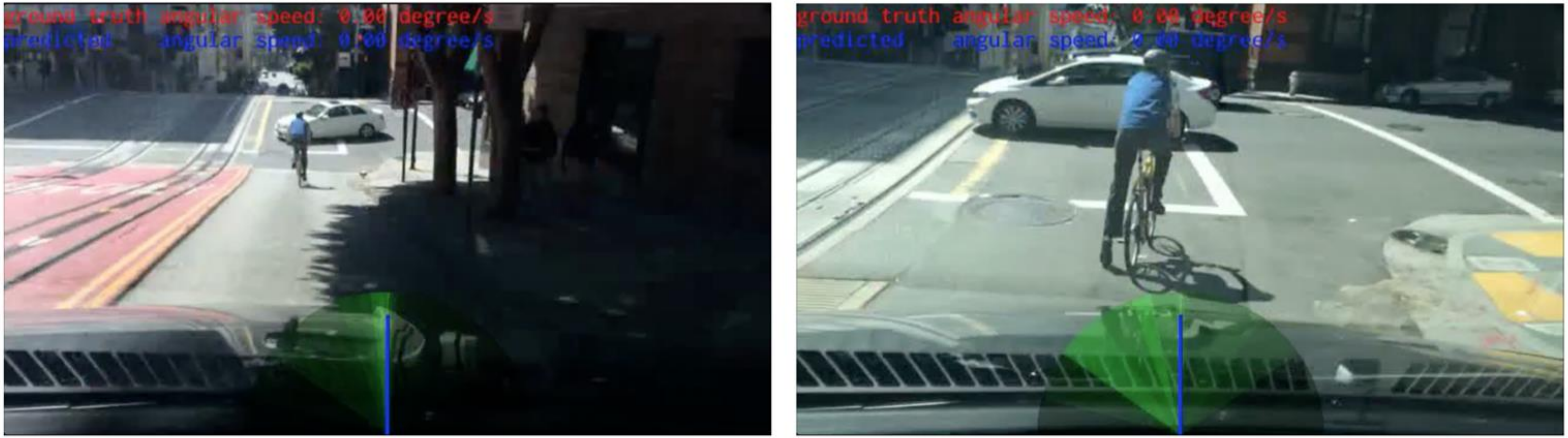
Lane following: left and right

Experiments – Continuous Actions



Intersection

Experiments – Continuous Actions



Side Walk

BDD Dataset – video

arXiv.org > cs > arXiv:1612.01079

Search or Article ID inside arXiv | All papers | Broaden y

(Help | Advanced search)

Computer Science > Computer Vision and Pattern Recognition

End-to-end Learning of Driving Models from Large-scale Video Datasets

Huazhe Xu, Yang Gao, Fisher Yu, Trevor Darrell

(Submitted on 4 Dec 2016)

Robust perception-action models should be learned from training data with diverse visual appearances and realistic behaviors, yet current approaches to deep visuomotor policy learning have been generally limited to in-situ models learned from a single vehicle or a simulation environment. We advocate learning a generic vehicle motion model from large scale crowd-sourced video data, and develop an end-to-end trainable architecture for learning to predict a distribution over future vehicle egomotion from instantaneous monocular camera observations and previous vehicle state. Our model incorporates a novel FCN-LSTM architecture, which can be learned from large-scale crowd-sourced vehicle action data, and leverages available scene segmentation side tasks to improve performance under a privileged learning paradigm.

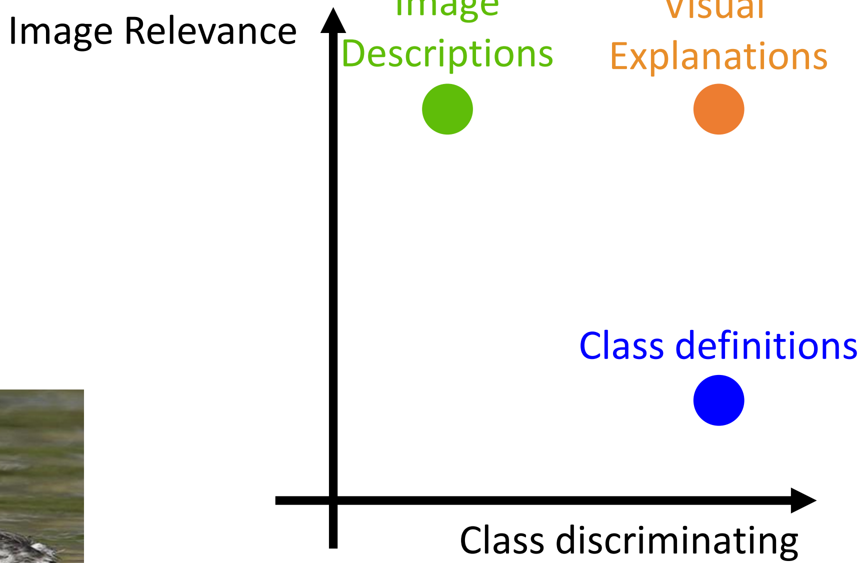
Overview

Adversarial Domain Adaptation

Learning end-to-end driving models from crowdsourced dashcams

Vision and Language: Learning to reason to answer and explain

Explainable AI (XAI): Visual Explanations



This bird has black and white feathers, with a white neck and a yellow beak.

Western Grebe

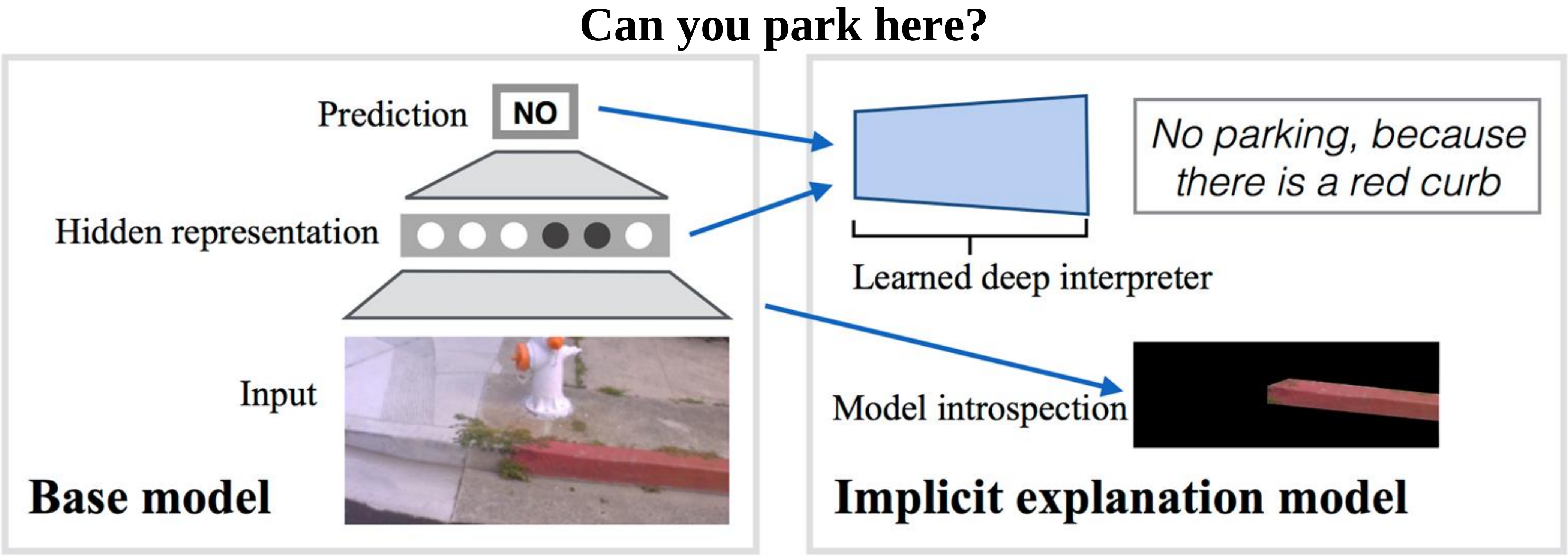
This is a *Western Grebe* because this bird has a long white neck, pointy yellow beak and red eye.

Laysan Albatros

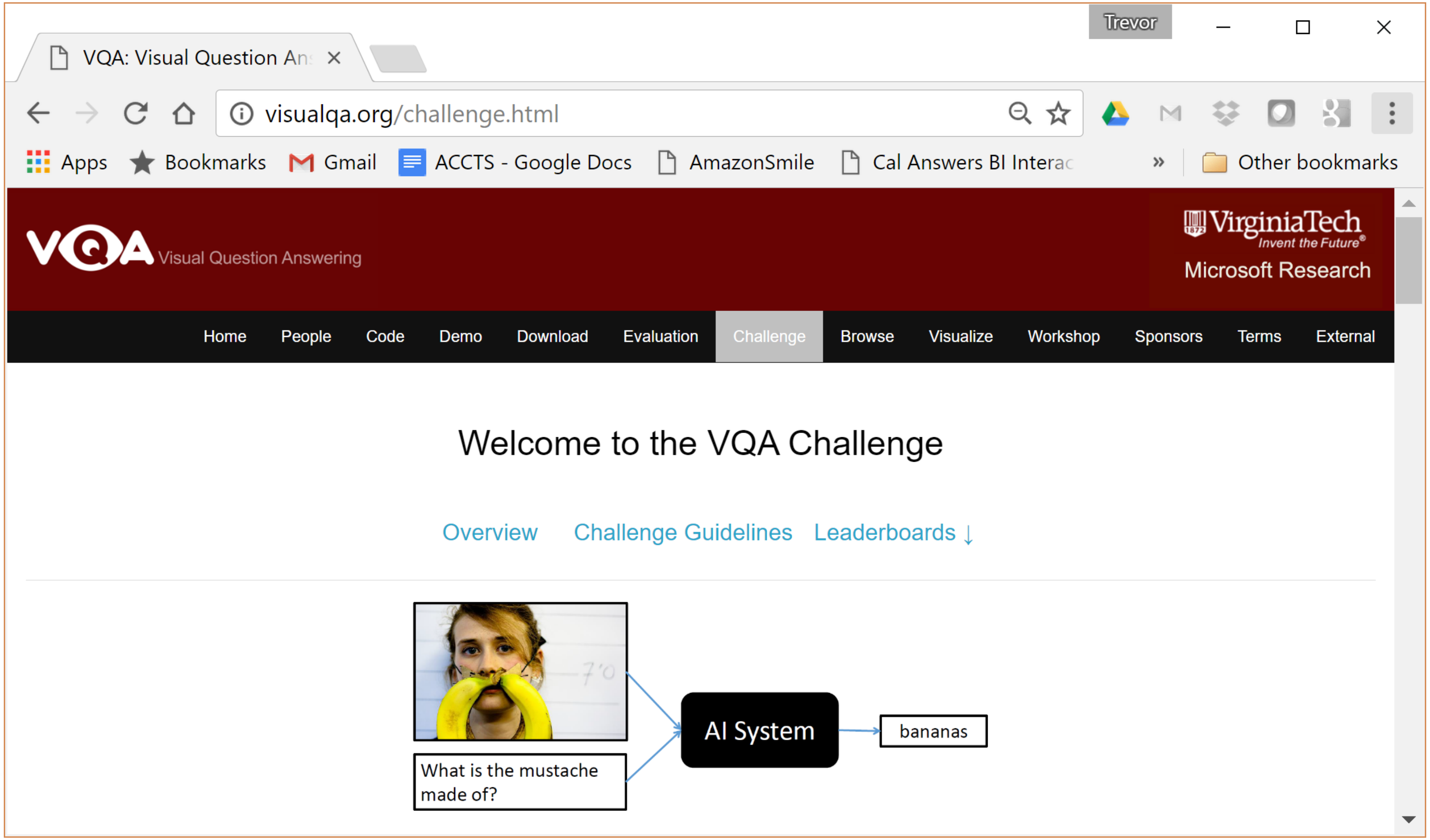
This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck and black back.

Explainable Models with Implicit Capabilities

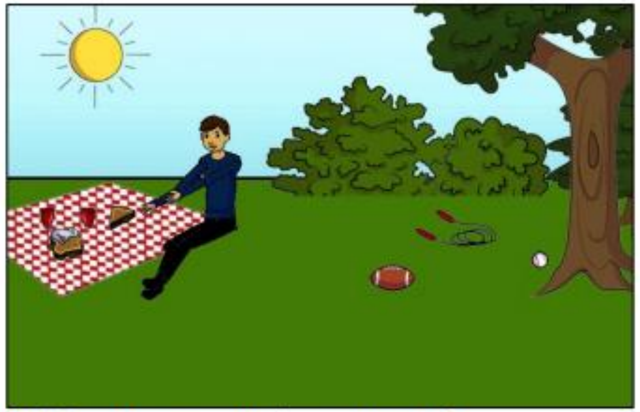
- Translate DNN hidden state into
  - human-interpretable language
  - visualizations and exemplars



Visual Question Answering



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Who is wearing glasses?

man

woman

Where is the child sitting?

fridge

arms

Is the umbrella upside down?

yes

no

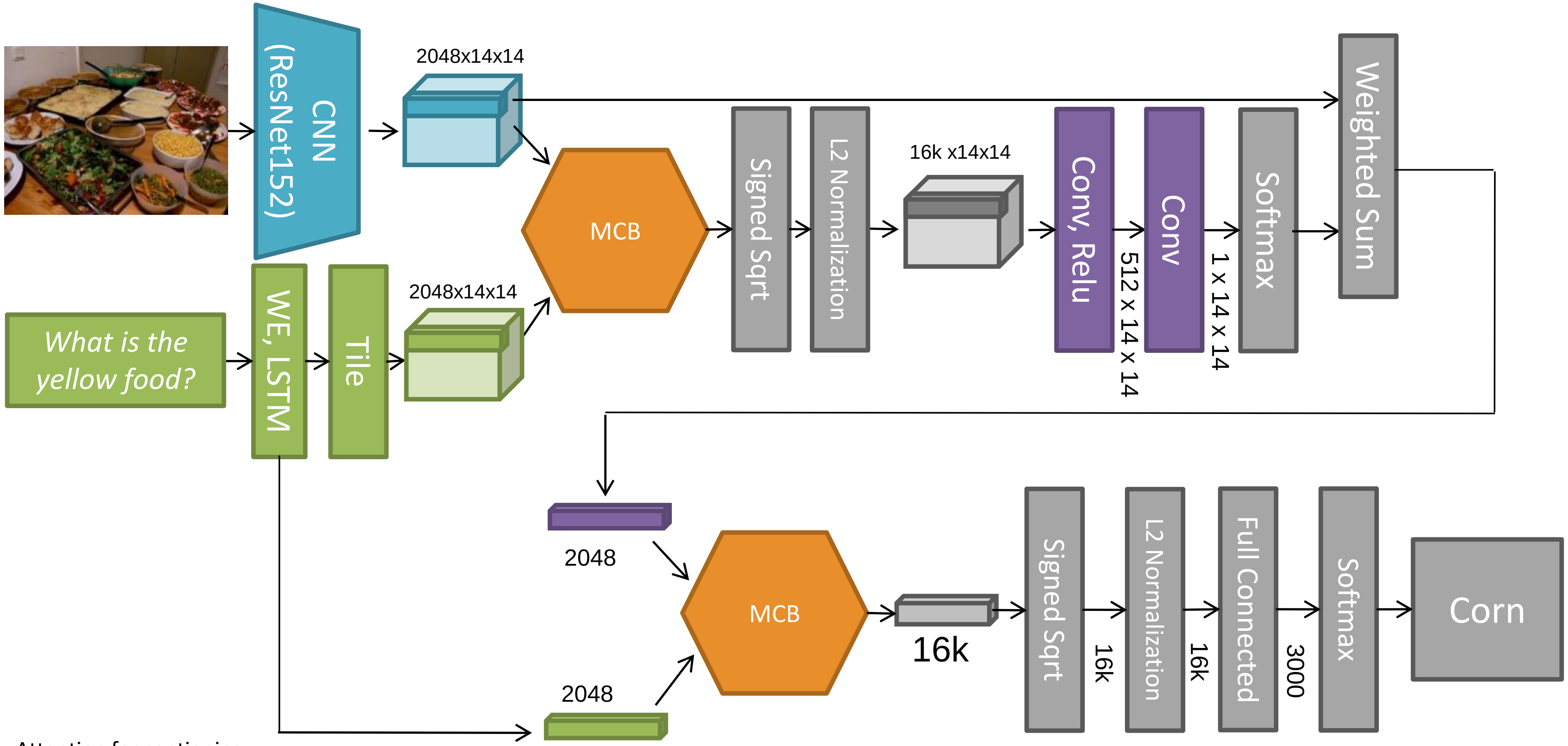
How many children are in the bed?

2

1

NAACL 2016: MCB with Attention

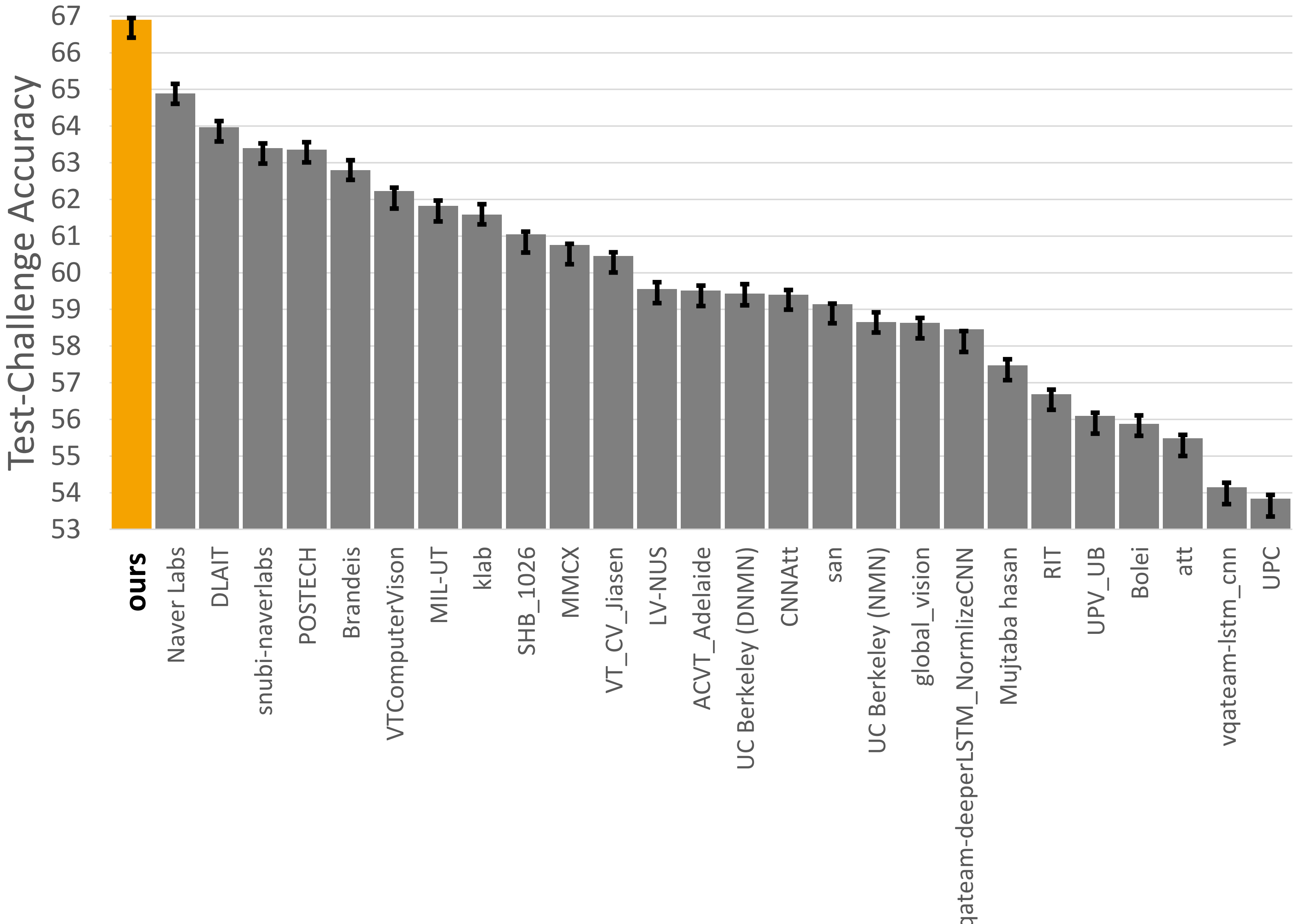
- Predict spatial attentions with MCB



Attention for captioning :
- K. Xu, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

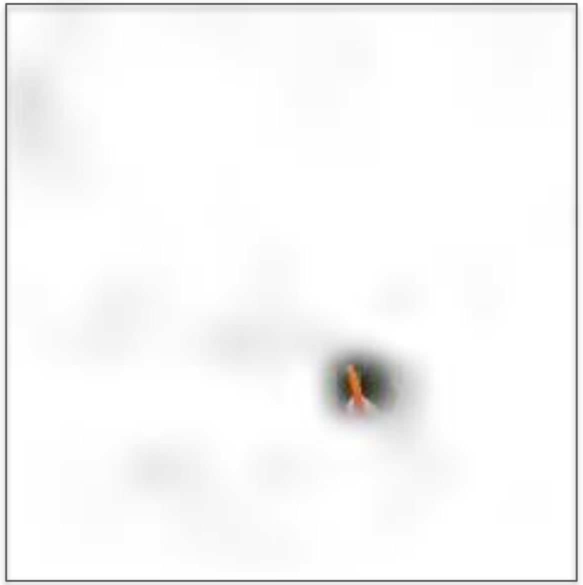
Attention for VQA :
- H. Xu, K. Saenko Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering
- J.Lu Hierarchal Question-Image Co-Attention for Visual Question Answering

Winner VQA Challenge 2016 (real open ended)



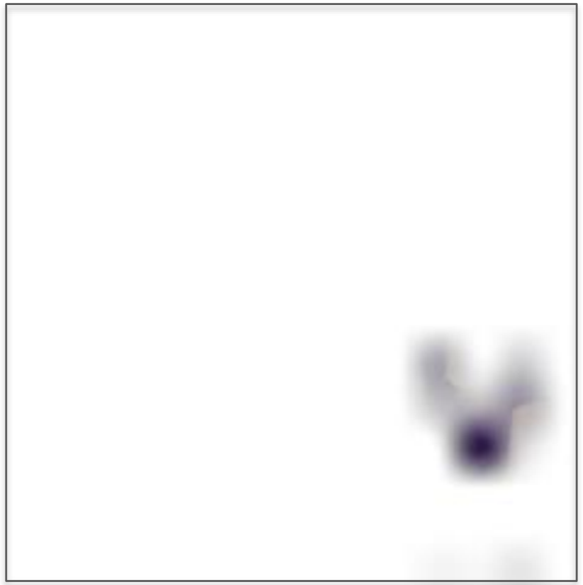
Attention Visualizations

What is the woman feeding the giraffe?
Carrot
[Groundtruth: Carrot]



Attention Visualizations

What color is her shirt?
Purple
[Groundtruth: Purple]



Attention Visualizations

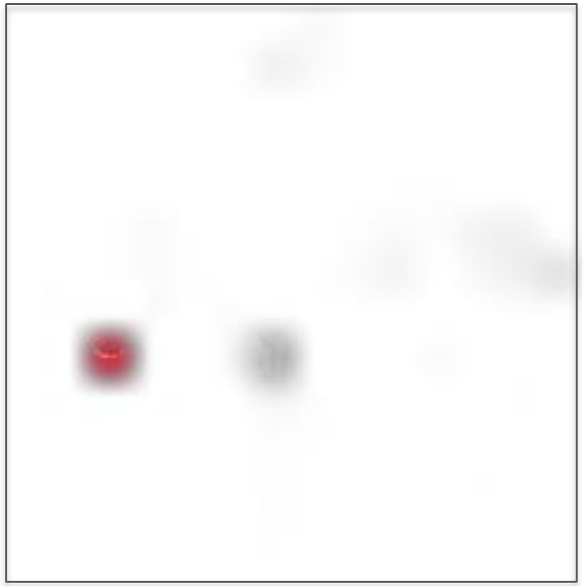
What is her hairstyle for the picture?
Ponytail
[Groundtruth: Ponytail]

Attention Visualizations

What color is the chain on the red dress?

Pink

[Groundtruth: Gold]



- Correct Attention, Incorrect Fine-grained Recognition

Attention Visualizations

Is the man going to fall down?

No

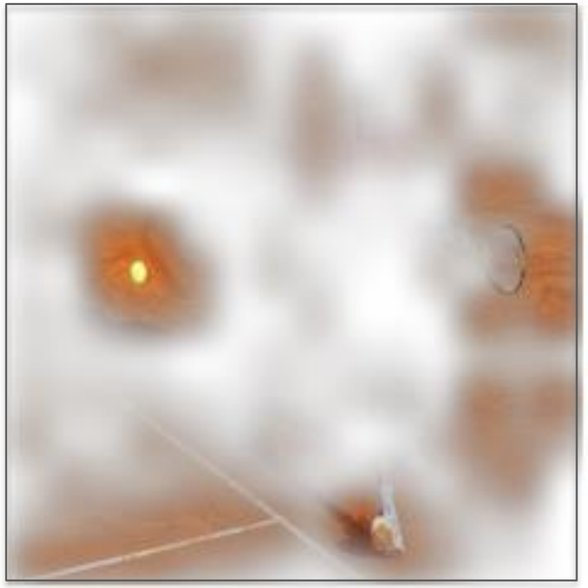
[Groundtruth: No]

Attention Visualizations

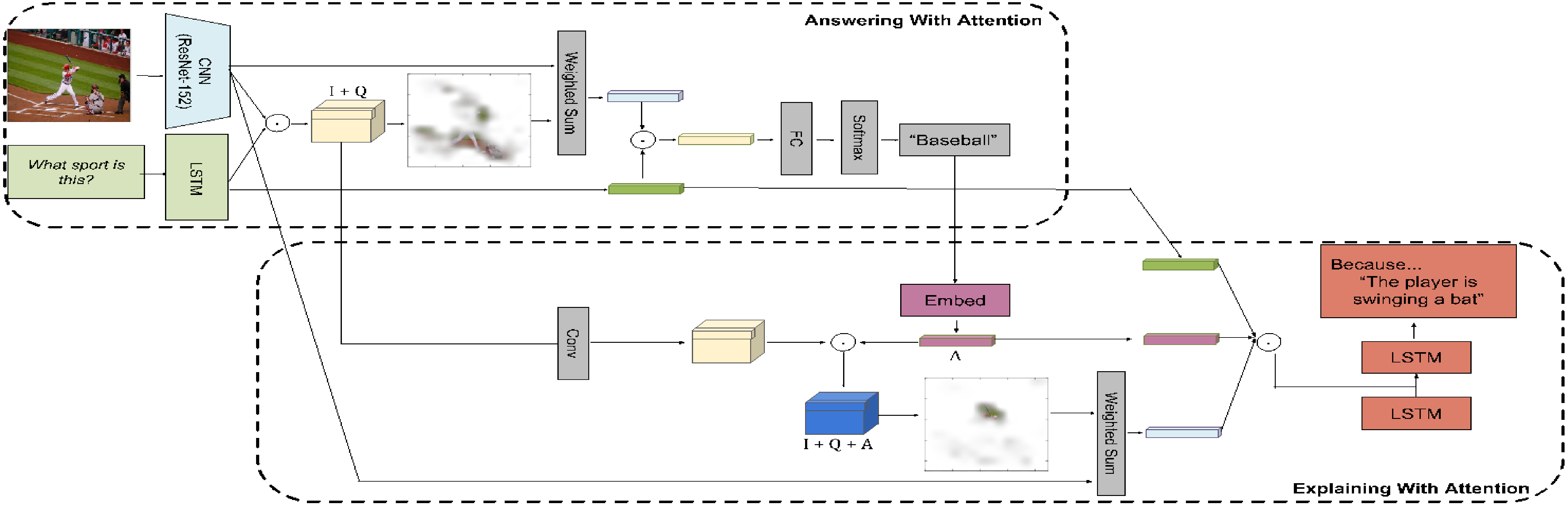
What is the surface of the court made of?
Clay
[Groundtruth: Clay]

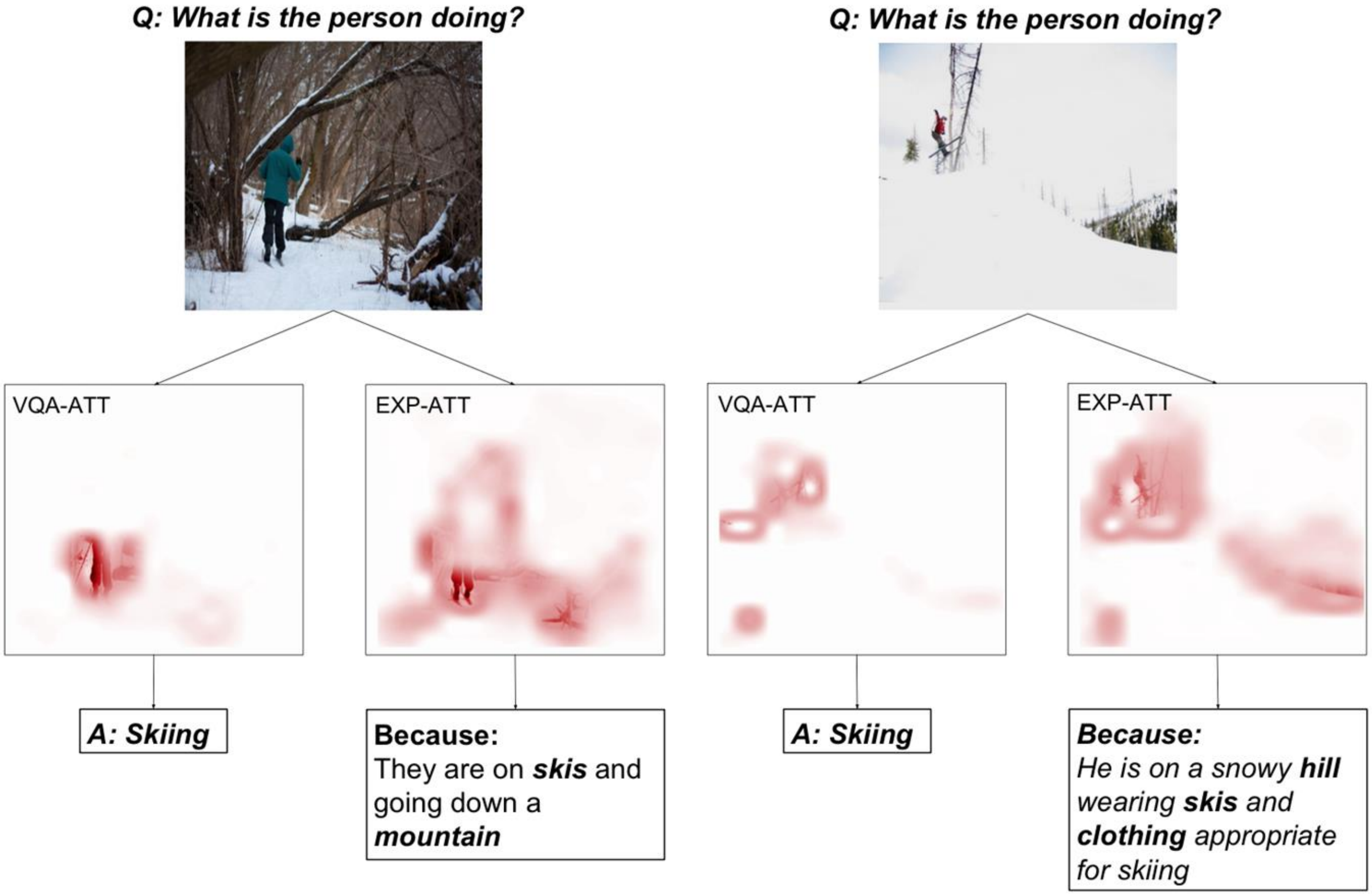
Attention Visualizations

What sport is being played?
Tennis
[Groundtruth: Tennis]



Attentive Explanations:
Justifying Decisions and Pointing to the Evidence




Q: What is the person doing?
VQA-ATT
EXP-ATT
A: Skiing
Because:
They are on skis and going down a mountain
Q: What is the person doing?
VQA-ATT
EXP-ATT
A: Skiing
Because:
He is on a snowy hill wearing skis and clothing appropriate for skiing

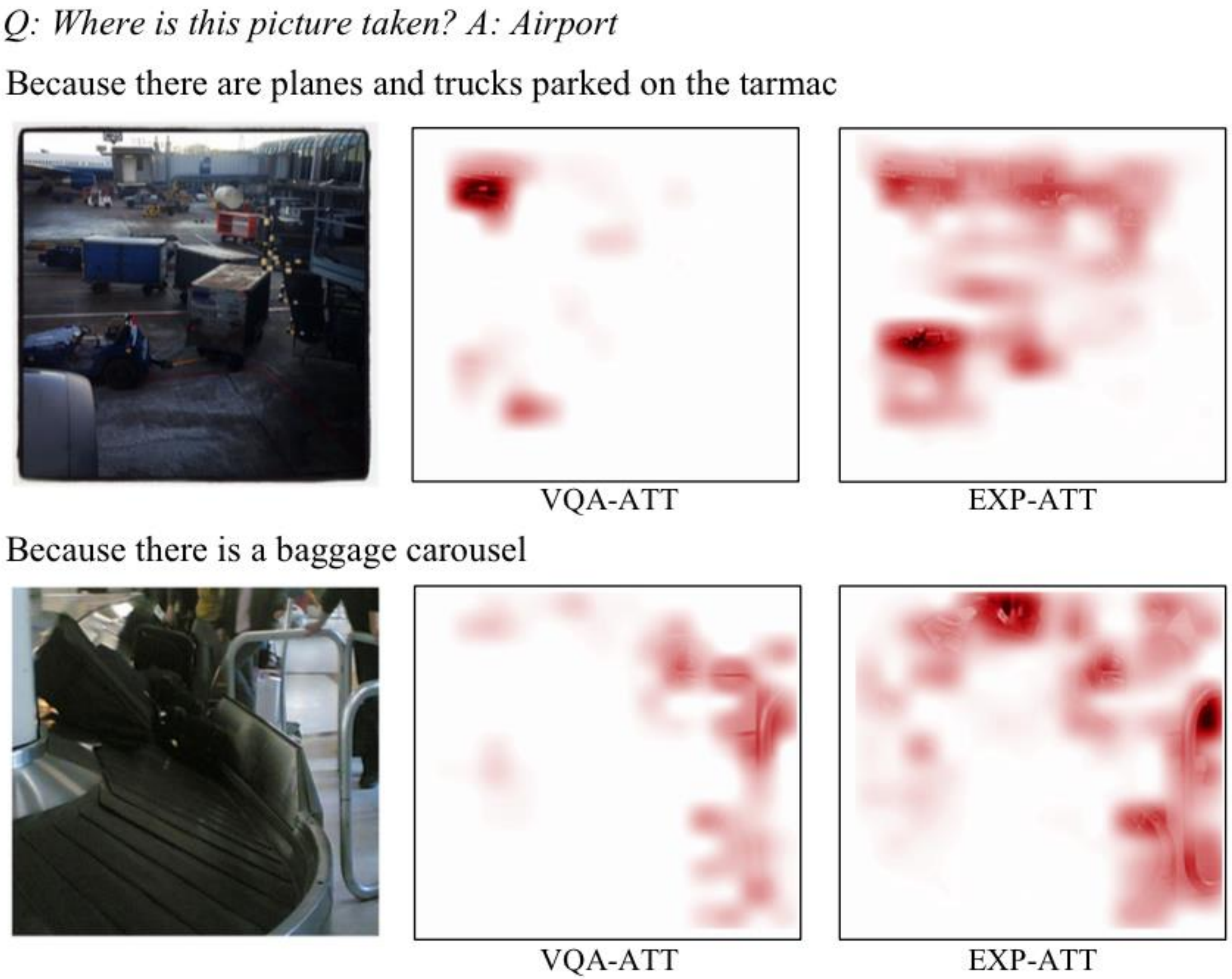
Human ground truth for the textual justification task.

	Description	Explanation
	A man on a snowboard is on a ramp.	Q: What is the person doing? A: Snowboarding Because... they are on a snowboard in snowboarding outfit.
	A gang of biker police riding their bikes in formation down a street.	Q: Can these people arrest someone? A: Yes Because... they are Vancouver police.

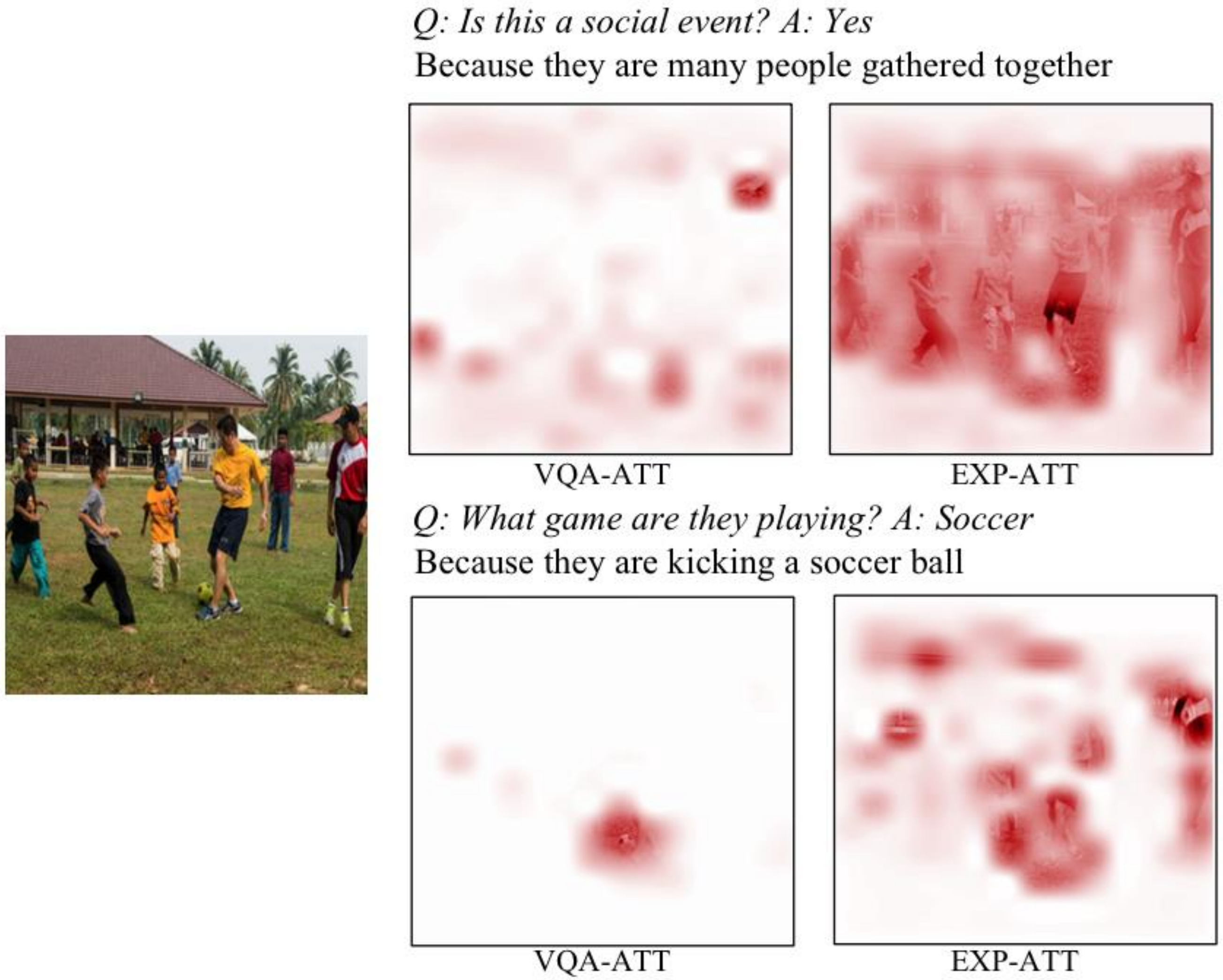
	Description	Explanation
		I can tell the person is juggling
	A man in a black shirt and blue jeans is holding a glowing ball.	Because... he holds two balls in one hand, while another ball is aloft just above the other hand.
	A man standing wearing a pink shirt and grey pants near a ball.	Because... he has two balls in his hands while two are in the air.

Human ground truth for the pointing task.

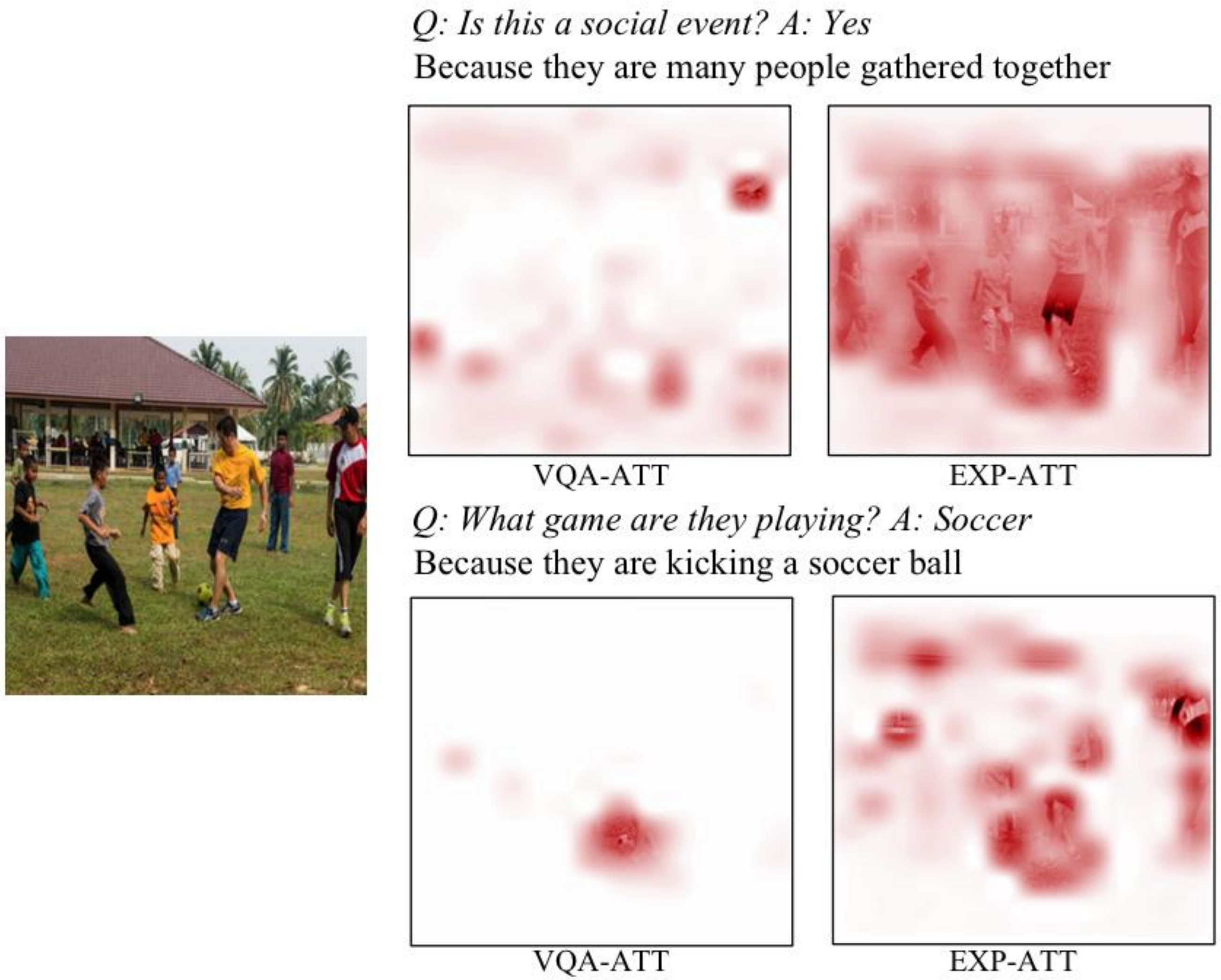
Discussing different evidence for different images.



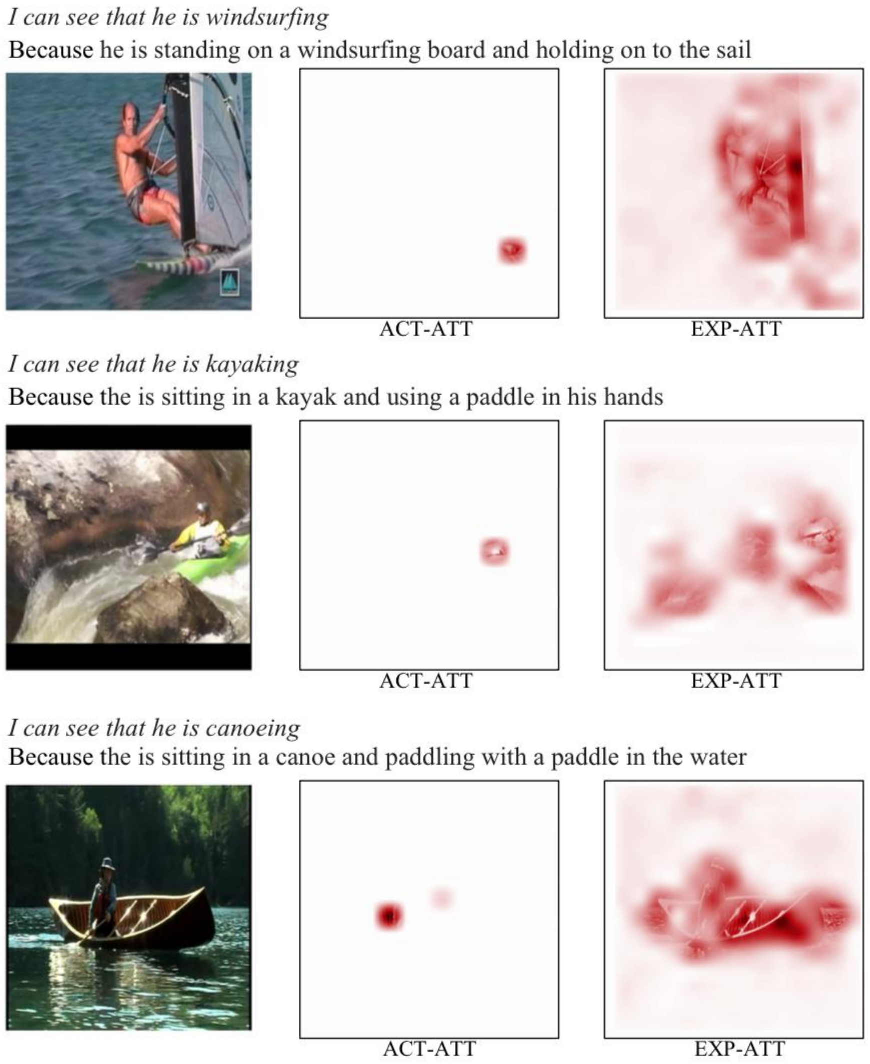
Discussing different evidence for different questions.



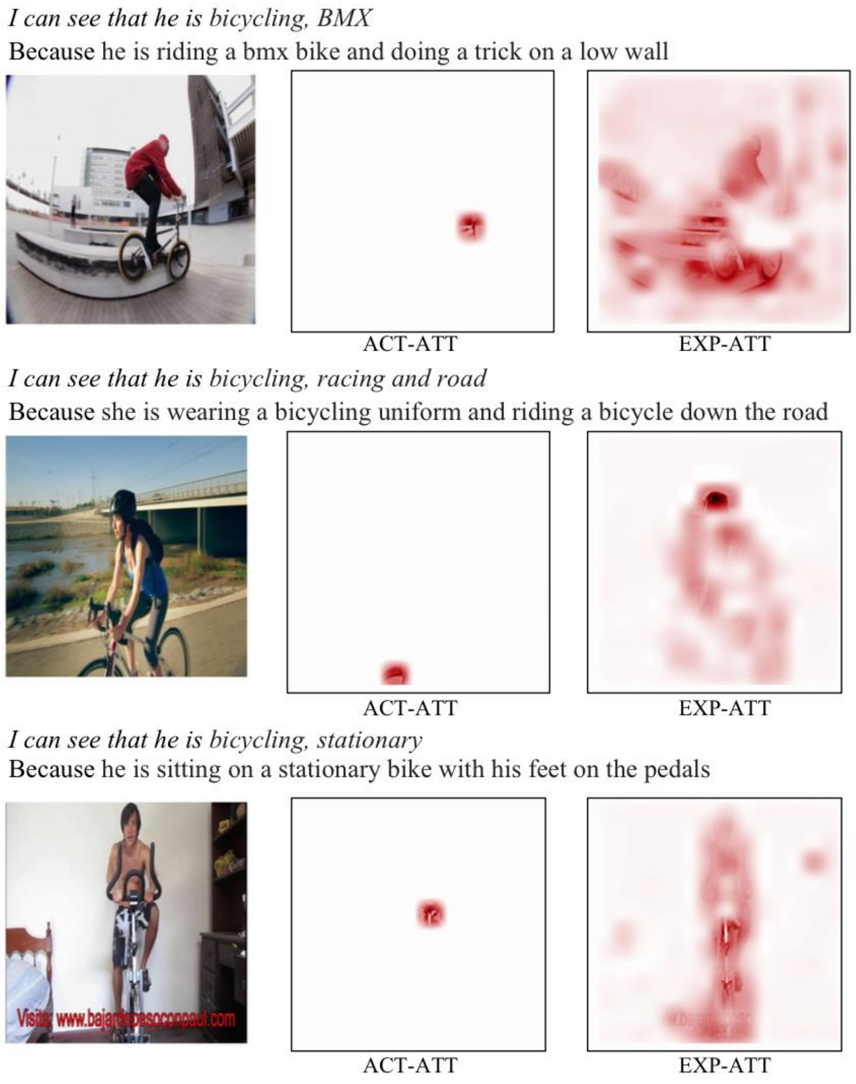
Discussing different evidence for different questions.



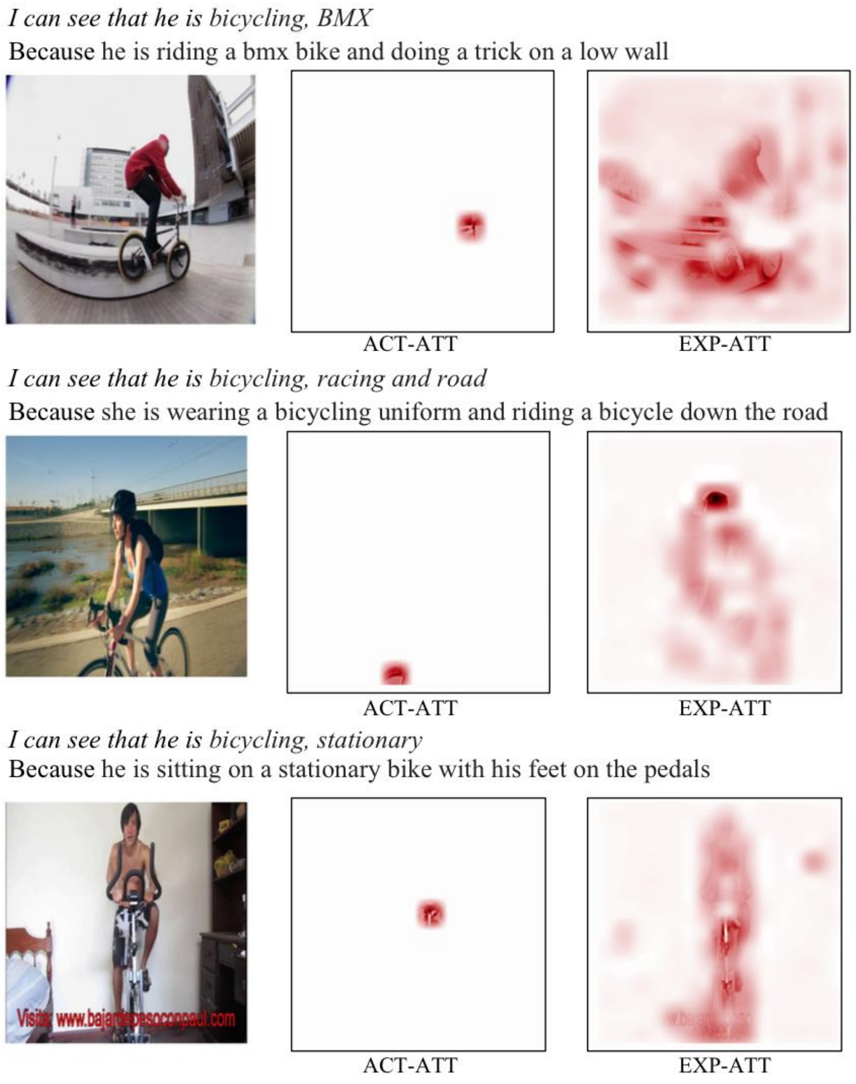
Differentiating between some activities requires understanding special equipment.



Differentiating between some activities requires recognizing specific context.



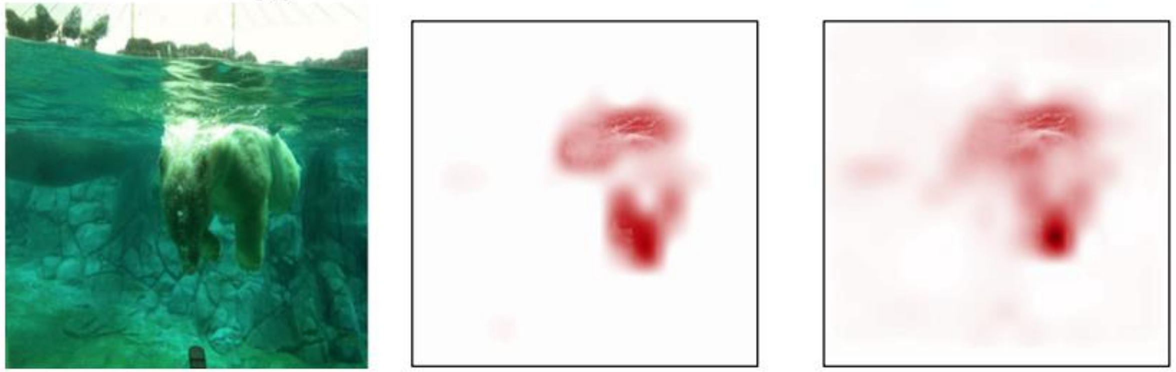
Differentiating between some activities requires recognizing specific context.



Explanations when the model predicts the wrong answer.

Q: What is the bear doing? GT = Swimming, P = Eating

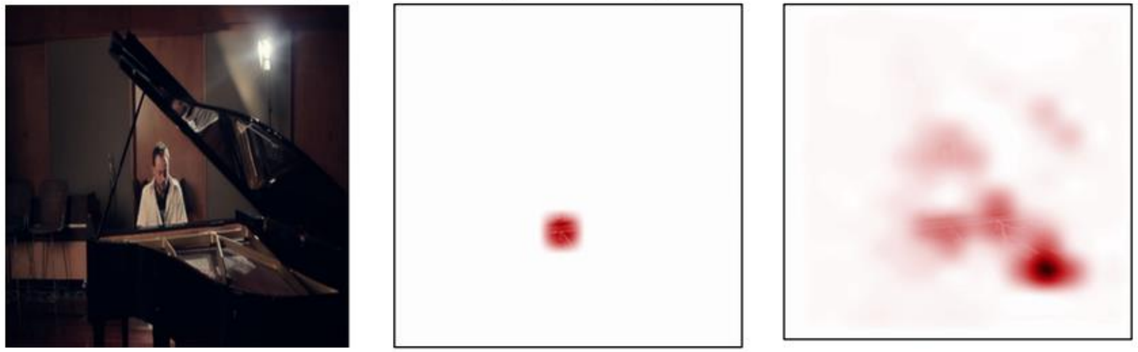
Because it is hungry and likes food



VQA-ATT EXP-ATT

GT = Piano, Sitting, P = Carpentry, General

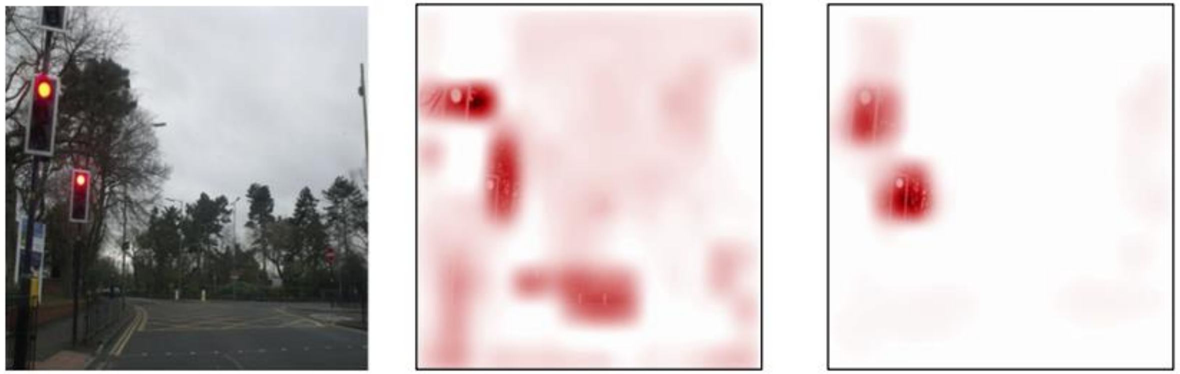
Because he is standing in a workshop with many tools on the table



ACT-ATT EXP-ATT

Q: Should we stop? GT = Yes, P = No

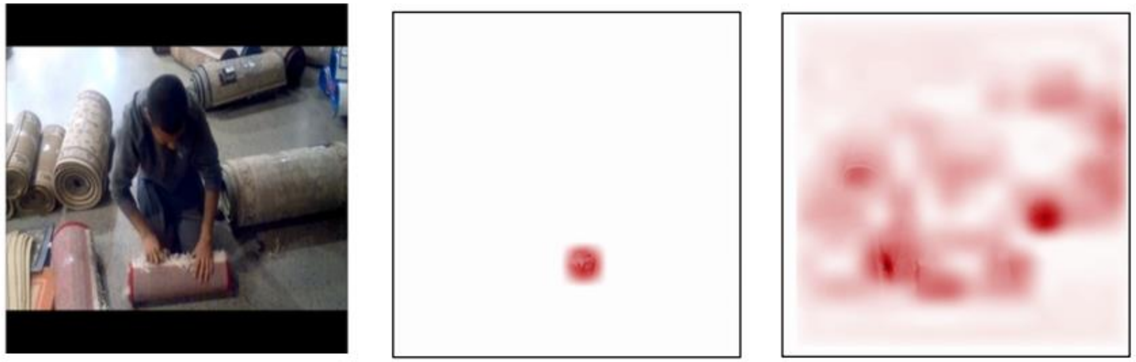
Because the light is green



VQA-ATT EXP-ATT

GT = Manual or Unskilled Labor, P = Yoga, Power

Because he is sitting on a yoga mat and holding a yoga pose

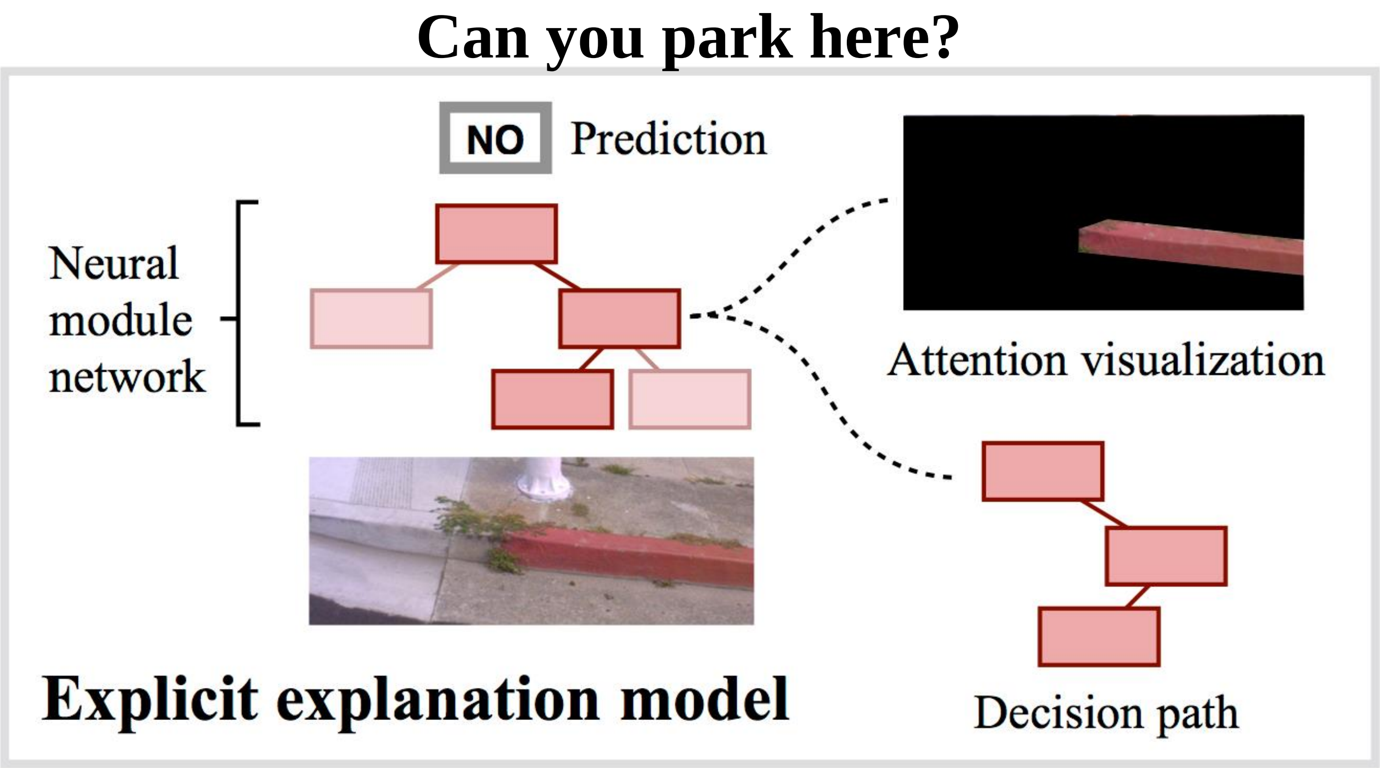


ACT-ATT EXP-ATT

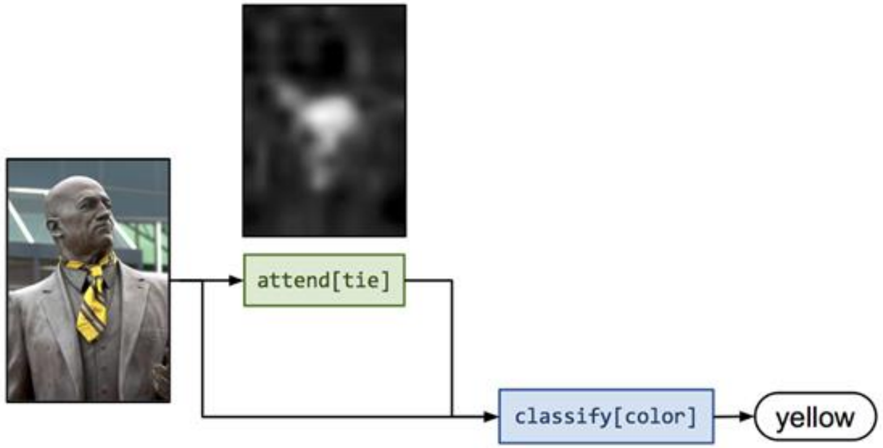
Explainable Models with Explicit Capabilities

Explain higher-level reasoning in DNNs

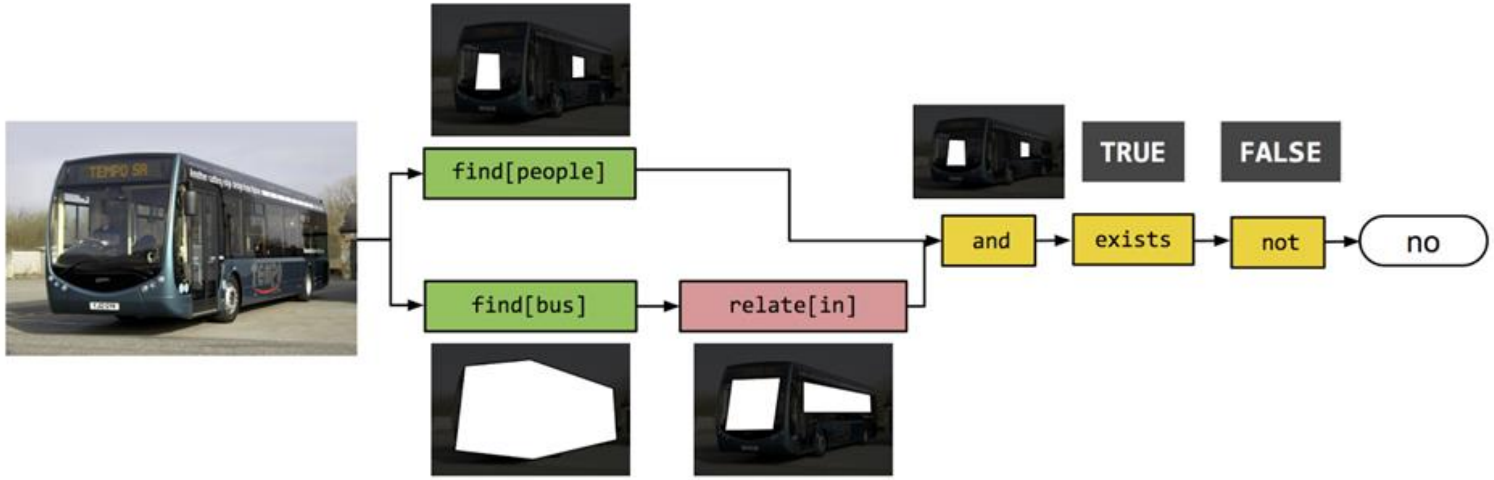
- Explainable decision path for multi-task, control and planning
- Provide structure and intermediate state



Explainable Models with
Explicit **and** Implicit Capabilities



(a) NMN for the question *What color is his tie?*

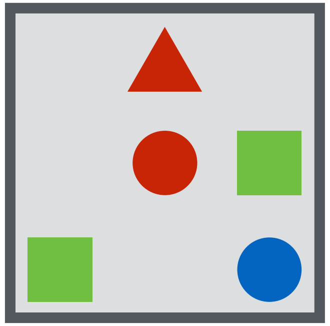


(b) NMN for the question *Is the bus empty?*

No, because there is a person in the bus.

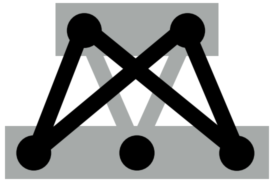
Grounded question answering

Is there a red shape above a circle?

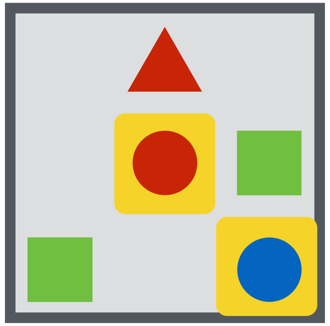


yes

Neural nets learn lexical groundings



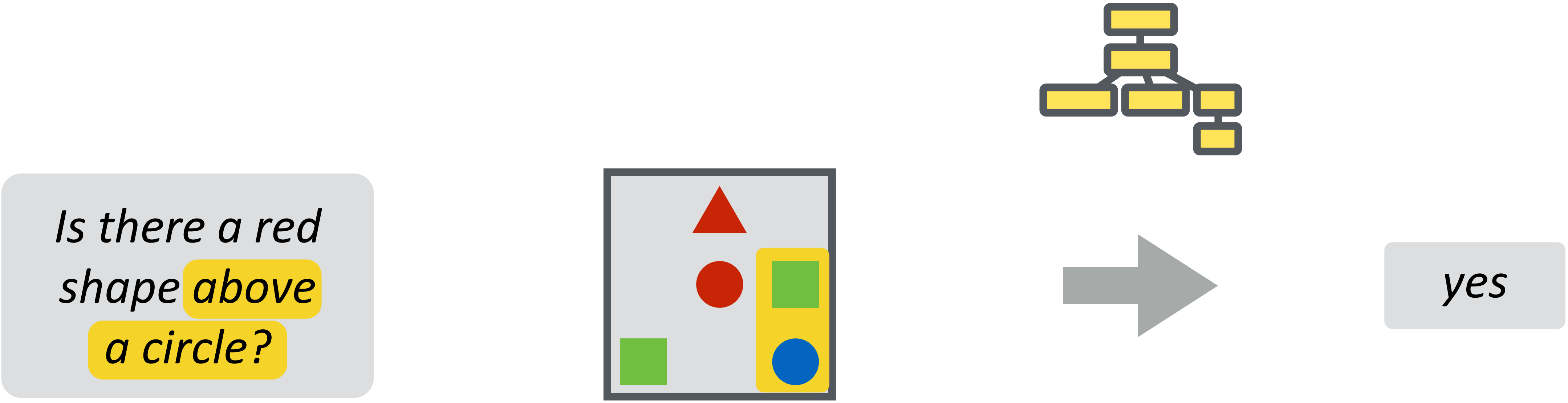
Is there a red shape above a circle?



yes

[Iyyer et al. 2014, Bordes et al. 2014,
Yang et al. 2015, Malinowski et al., 2015]

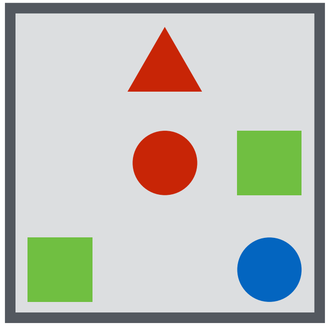
Semantic parsers learn composition

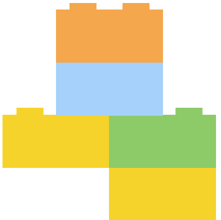


[Wong & Mooney 2007, Kwiatkowski et al. 2010, Liang et al. 2011, A et al. 2013]

Neural module networks learn both!

Is there a red shape above a circle?

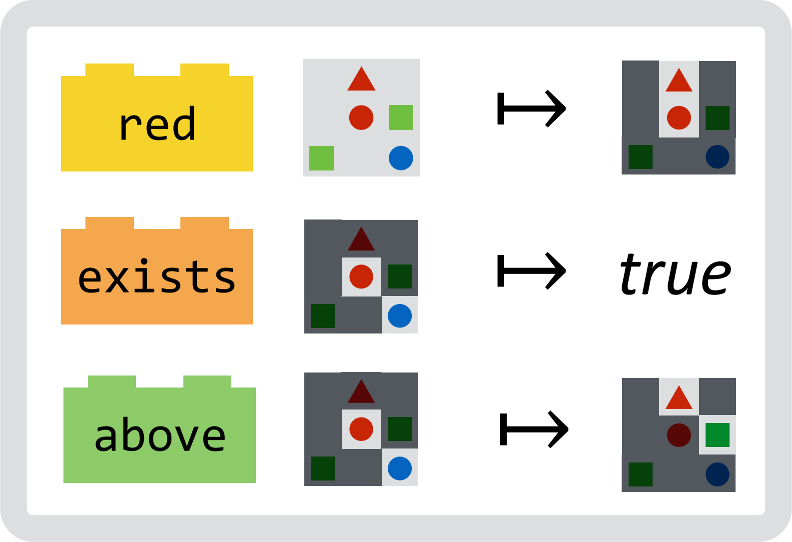




yes

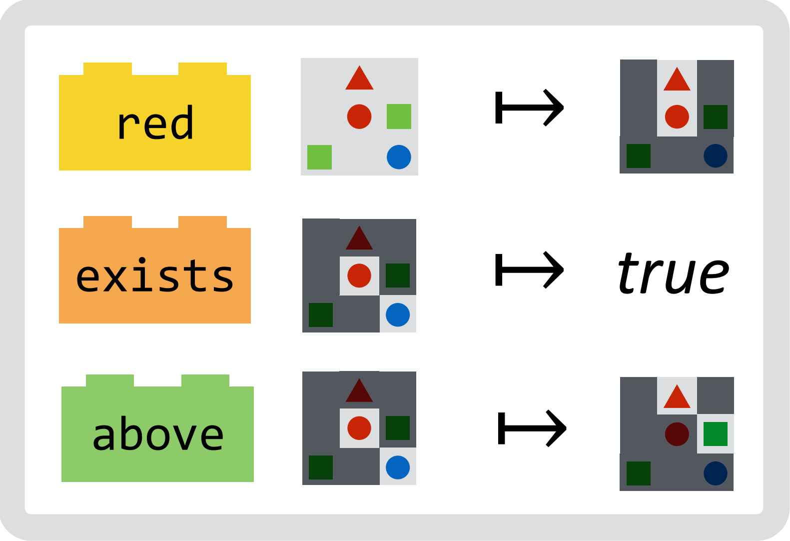
Neural module networks

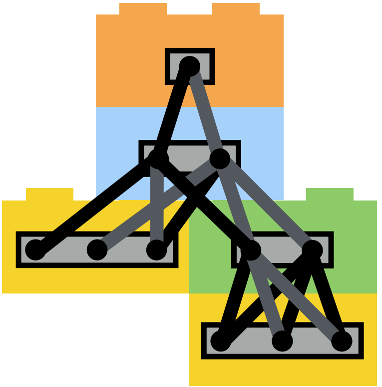
Is there a red shape above a circle?



Neural module networks

Is there a red shape above a circle?

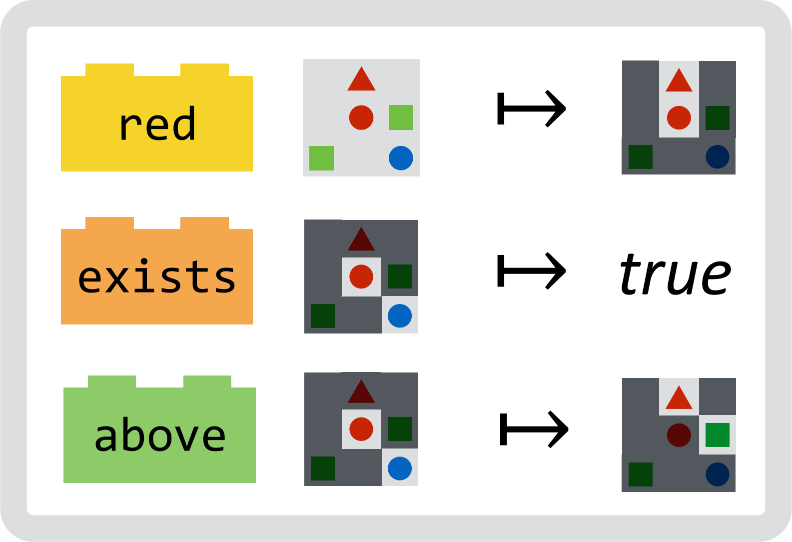


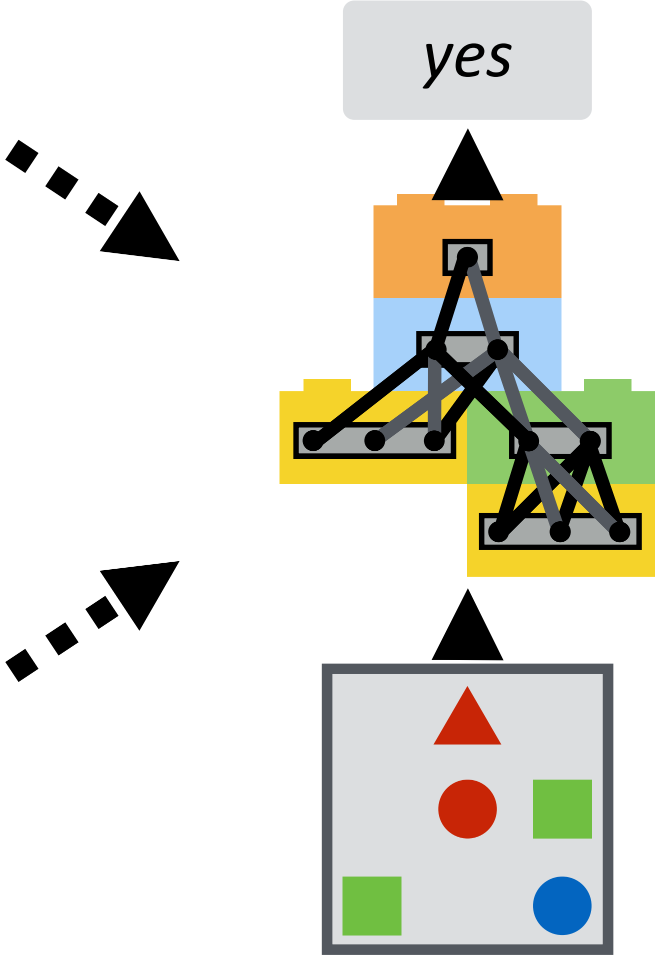


Neural module networks

Is there a red shape above a circle?

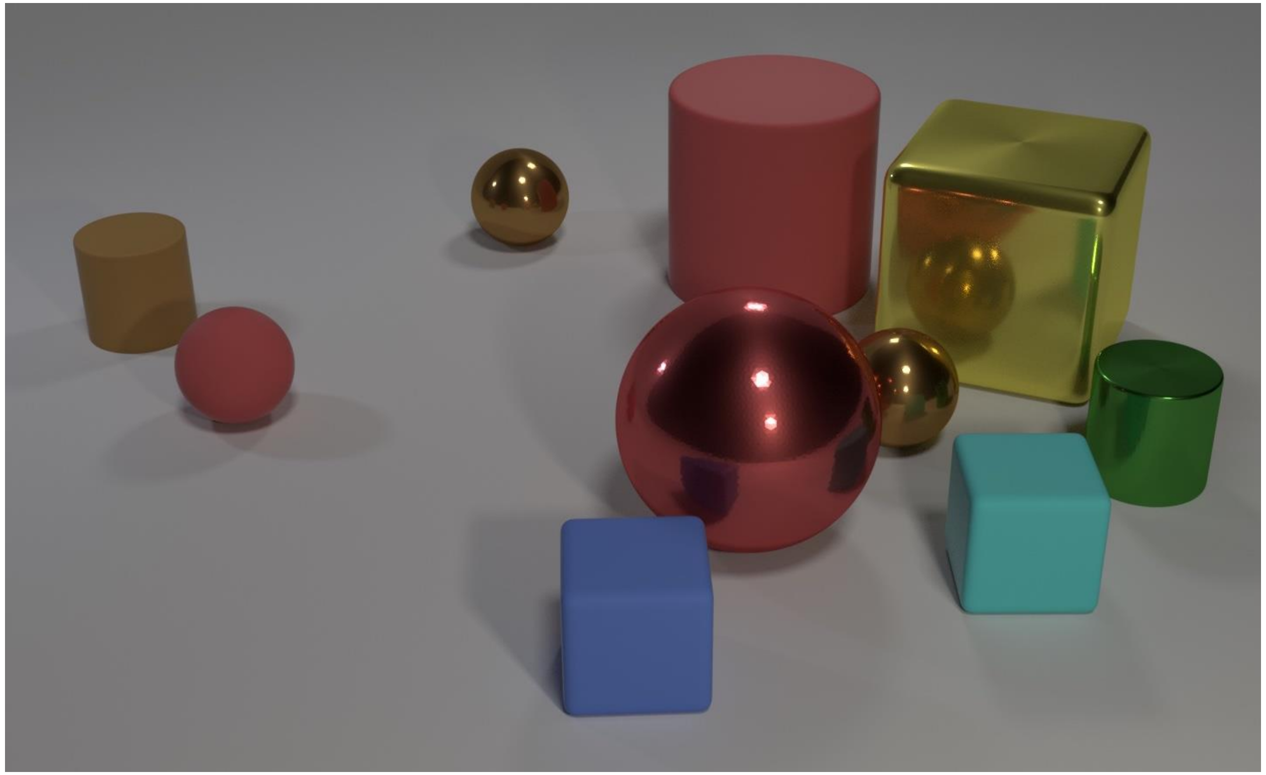
101





CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning

Justin Johnson[1,2*] Bharath Hariharan[2] Laurens van der Maaten[2]
Li Fei-Fei[1] C. Lawrence Zitnick[2] Ross Girshick[2]
[1]Stanford University [2]Facebook AI Research



Q: Are there an **equal number** of **large things** and **metal spheres**?
Q: **What size** is the **cylinder** **that is left of** the **brown metal** thing**that is left of** the **big sphere**?
Q: There is a **sphere** with the **same size as** the **metal cube**; is it **made of the same material as** the **small red sphere**?
Q: **How many** objects are **either** **small cylinders** or **red** things?

Questions in CLEVR test various aspects of visual reasoning including **attribute identification**,**counting**, **comparison**, **spatial relationships**, and**logical operations**.

Learning to Reason: End-to-End Module Networks for Visual Question Answering

R. Hu, J. Andreas, M. Rohrbach, T. Darrell, K. Saenko

Background

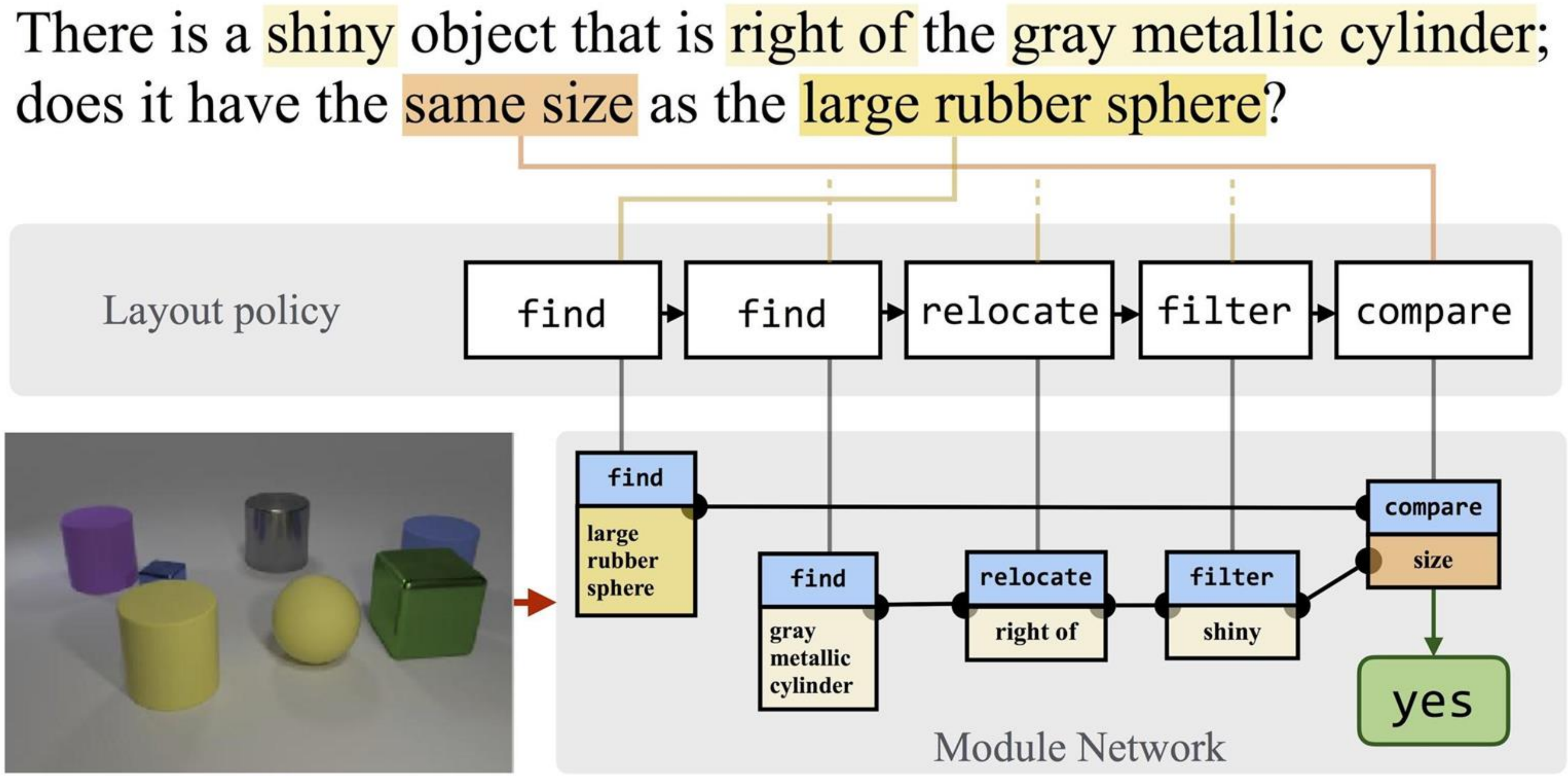
Natural language is **compositional**: the meaning of a sentence comes from the meanings of its components.

Different questions may require significantly different reasoning procedure.

- *What kind of vehicle is the one on the left of the brown car that is next to the building?*
- *Why is the person running away?*

Background

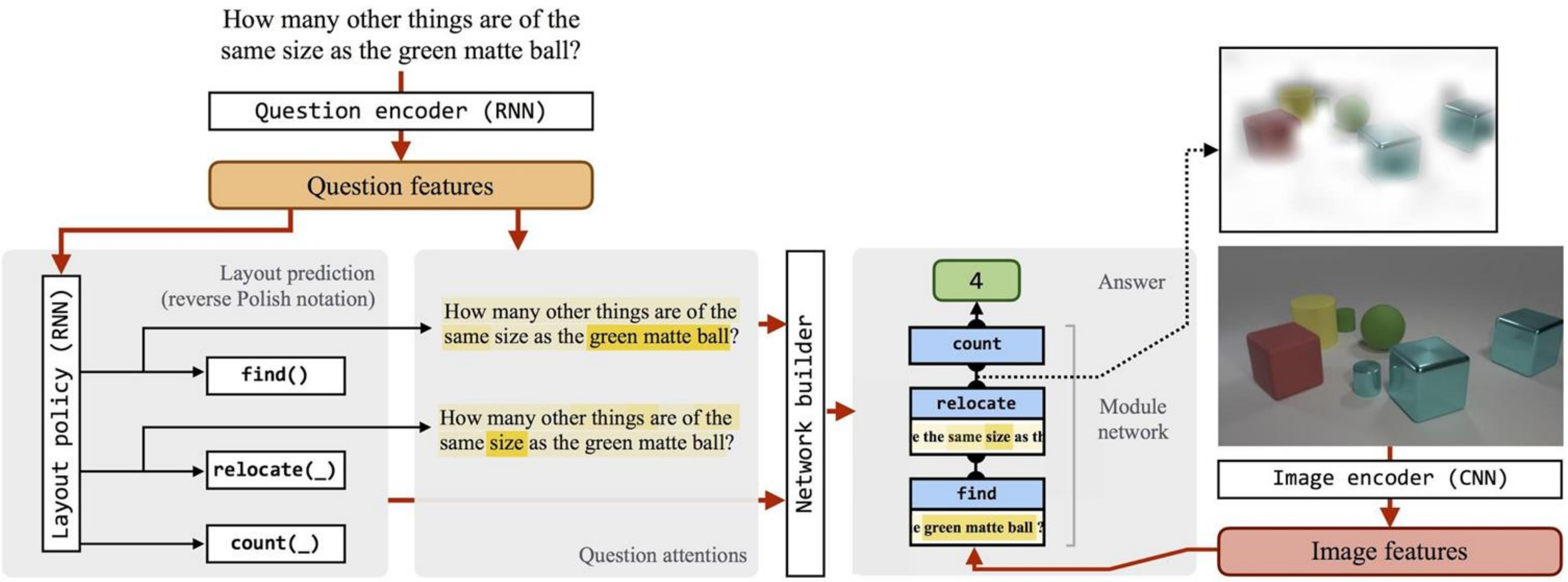
- Neural Module Networks: **dynamic** inference structure for each question
- Previous work: structure from NLP parser or parser re-ranking
- This work: **learned** layout policy to dynamically build a network



End-to-End Module Networks (N2NMN)

Components

- Layout policy *p(l | q)* with sequence-to-sequence RNN
- Neural modules with co-attention, dynamically assembled into a network



End-to-end Training

Loss $L(\theta) = E_{l \sim p(l|q;\theta)}[\tilde{L}(\theta, l; q, I)]$

where $\tilde{L}(\theta, l; q, I)$ is the softmax loss of the answer

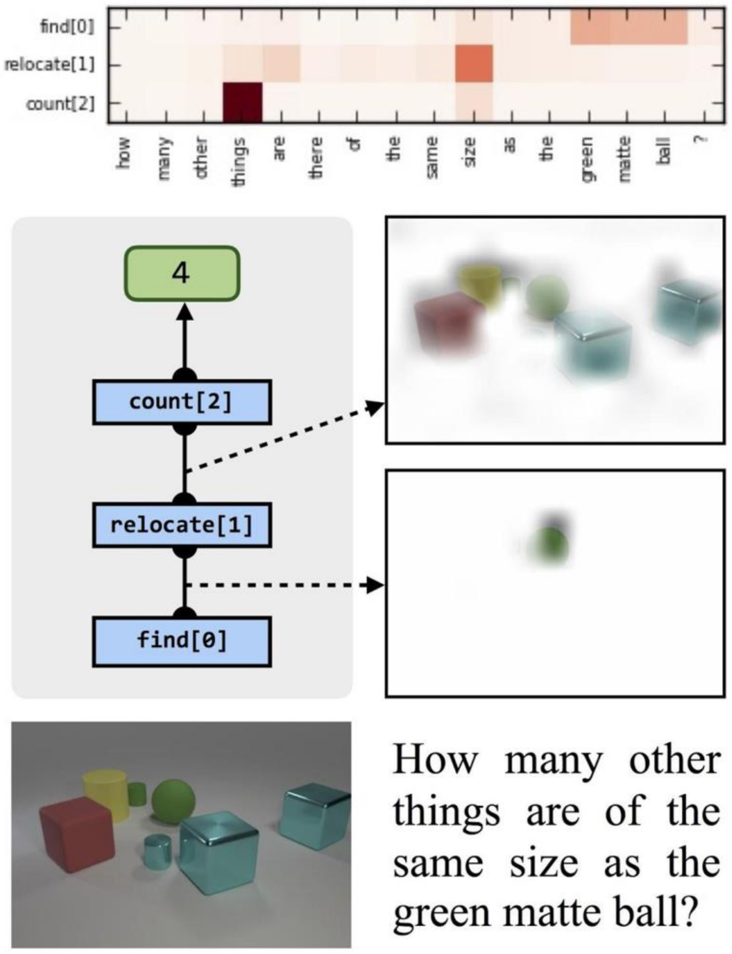
Optimization: policy gradient method

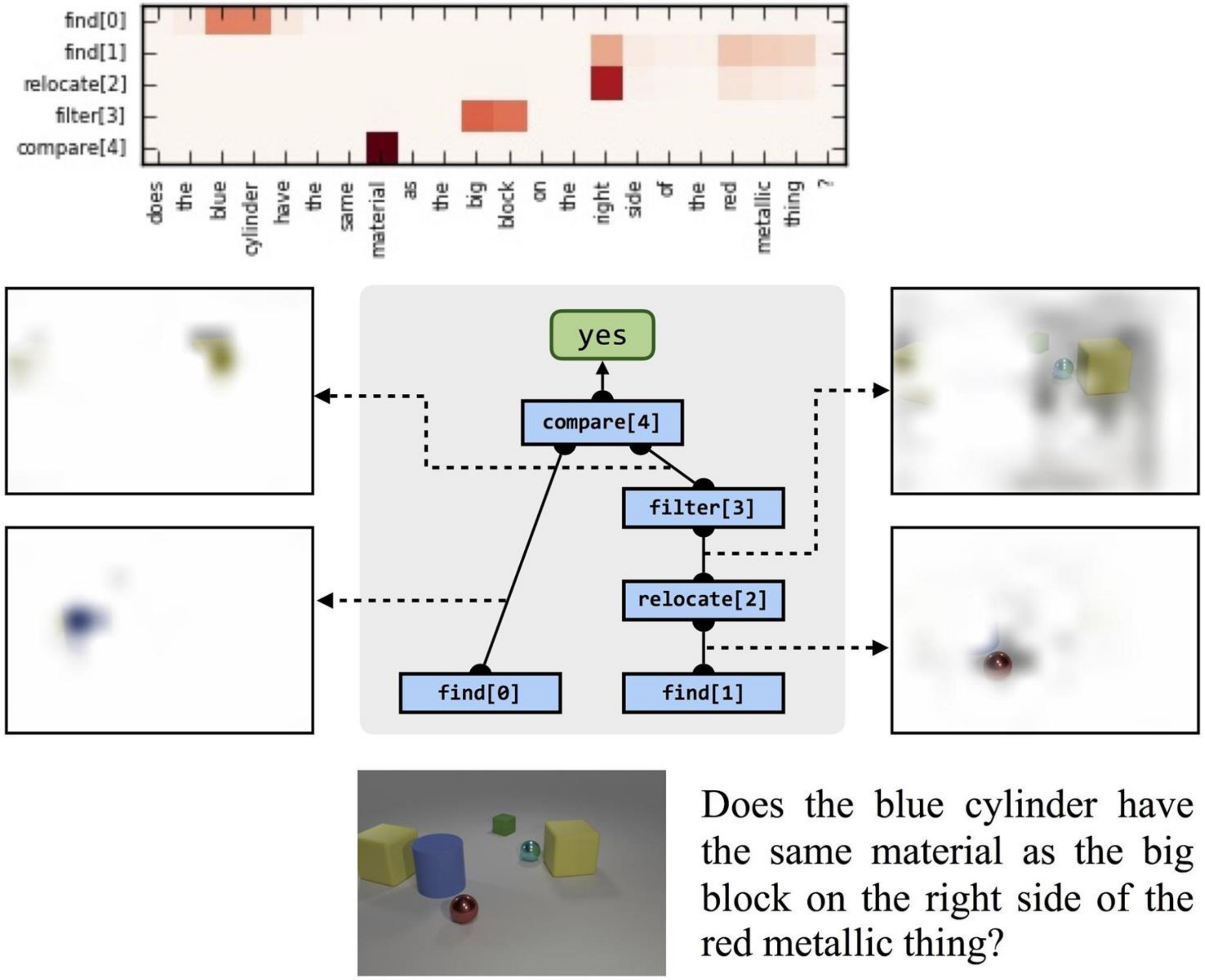
$$\nabla_\theta L = E_{l \sim p(l|q;\theta)} \left[\tilde{L}(\theta, l) \nabla_\theta \log p(l|q;\theta) + \nabla_\theta \tilde{L}(\theta, l) \right]$$

$$\approx \frac{1}{M} \sum_{m=1}^{M} \left(\tilde{L}(\theta, l_m) \nabla_\theta \log p(l_m|q;\theta) + \nabla_\theta \tilde{L}(\theta, l_m) \right)$$

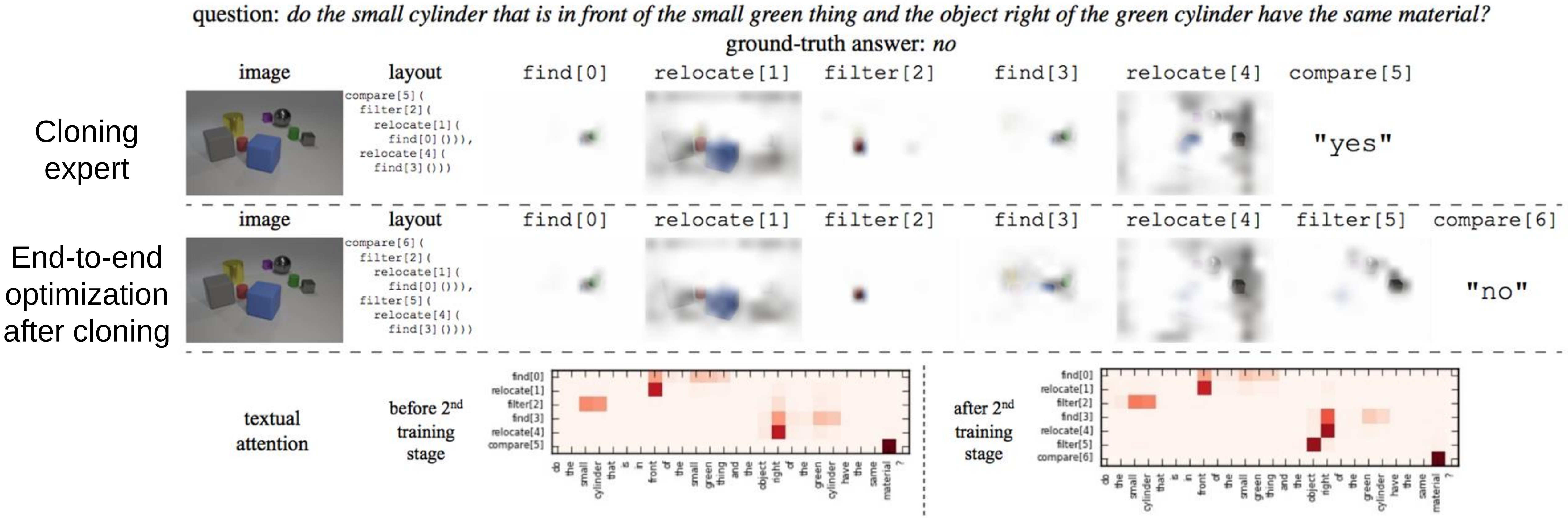
Easier: behavior cloning from expert layout policy

Experiments on the CLEVR dataset





Method	**Overall**	Exist	Count	Compare Integer			Query Attribute				Compare Attribute			
				equal	less	more	size	color	material	shape	size	color	material	shape
CNN+BoW [25]	48.4	59.5	38.9	50	54	49	56	32	58	47	52	52	51	52
CNN+LSTM [4]	52.3	65.2	43.7	57	72	69	59	32	58	48	54	54	51	53
CNN+LSTM+MCB [9]	51.4	63.4	42.1	57	71	68	59	32	57	48	51	52	50	51
CNN+LSTM+SA [24]	68.5	71.1	52.2	60	82	74	87	81	88	85	52	55	51	51
ours - cloning expert	78.9	83.3	63.3	68.2	87.2	85.4	90.5	80.2	88.9	88.3	89.4	52.5	85.4	86.7
ours - policy search after cloning	**83.7**	**85.7**	**68.5**	**73.8**	**89.7**	**87.7**	**93.1**	**84.8**	**91.5**	**90.6**	**92.6**	**82.8**	**89.6**	**90.0**



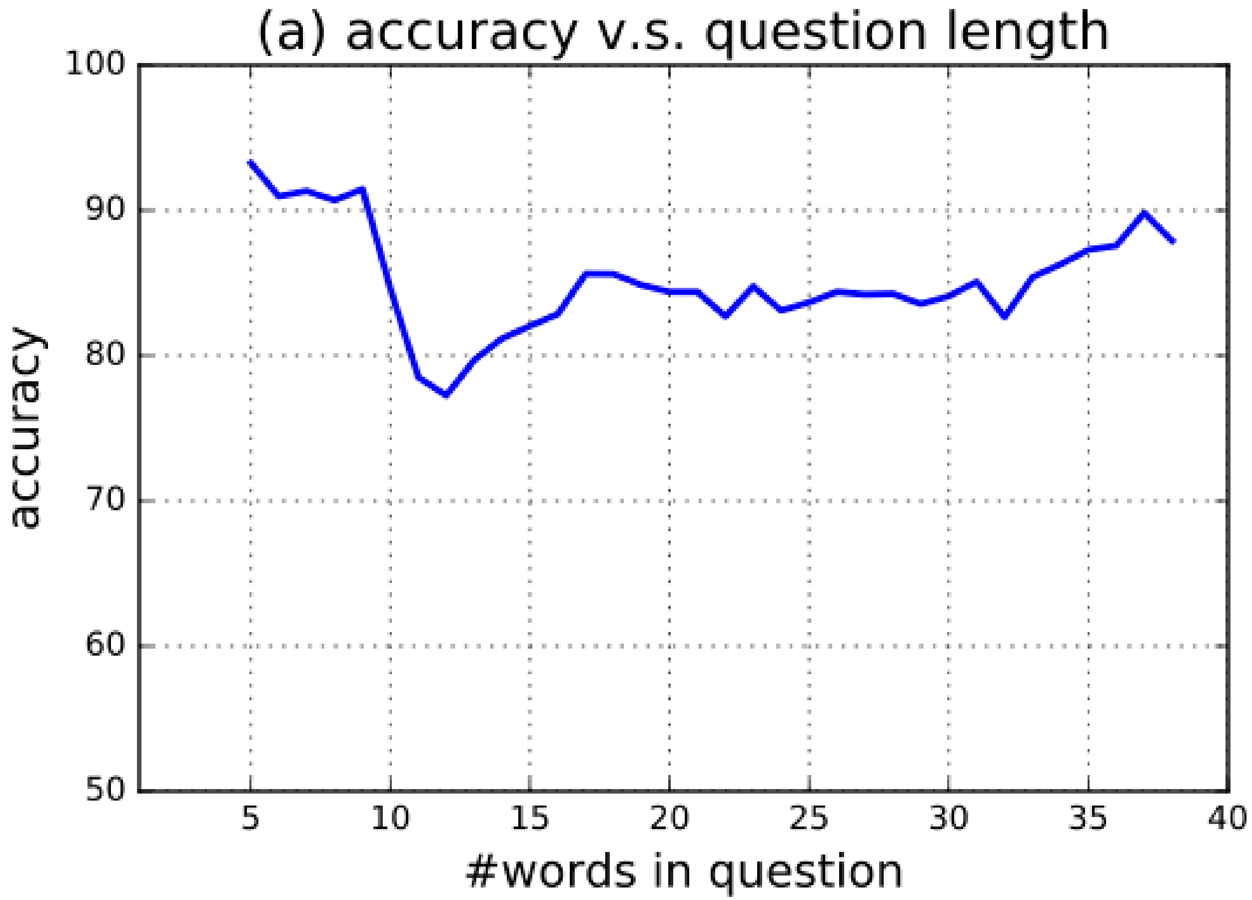
Policy Search from scratch (no experts used)

Even without resorting to an expert policy during training, our method still achieves state-of-the-art performance with reinforcement learning from scratch.

Method	Overall accuracy
CNN+BoW [4]	48.4
CNN+LSTM [1]	52.3
CNN+LSTM+MCB [2]	51.4
CNN+LSTM+SA [3]	68.5
ours - policy search from scratch	68.5
ours - cloning expert	78.9
ours - policy search after cloning	**83.7**

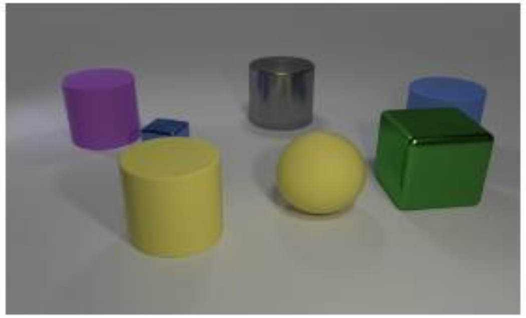
Accuracy v.s. Question length

Even on long questions with 30+ words, our method still achieves relatively high accuracy (Figure a).



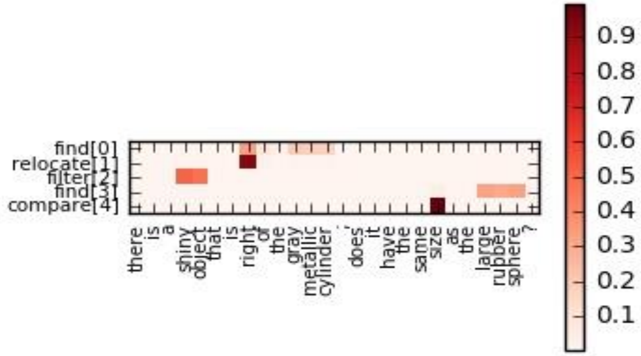
question: there is a shiny object that is right of the gray metallic cylinder ; does it have the same size as the large rubber sphere ?

ground-truth answer: "yes" predicted answer: "yes"



predicted layout:

```
compare[4](
   filter[2](
      relocate[1](
         find[0]())),
   find[3]())
```



output from find[0]

output from relocate[1]

output from filter[2]

output from find[3]

output from compare[4]

"yes"

Overview

Adversarial Domain Adaptation

Learning end-to-end driving models from crowdsourced dashcams

Vision and Language: Learning to reason to answer and explain