Learning and Reasoning with Visual Knowledge

Xinlei Chen
Carnegie Mellon University

Thesis Committee:
Abhinav Gupta (Chair)
Martial Hebert
Tom Mitchell
Fei-Fei Li, Stanford University
Andrew Zisserman, University of Oxford

AI is Beating Best Humans!

96/97: Deep Blue

AI is Beating Best Humans!

96/97: Deep Blue

16/17: AlphaGo

AI for Understanding Images?


Garry Kasparov
VS
DEEP BLUE

Russian chess player Garry Kasparov playing against computer program Deep Blue.

Clarifai (Zeiler et al. 2014)

people
administration
election
adult
leader
one
indoors
war
competition
man
...

COCO caption (Fang et al. 2015)

I am not really confident, but I think it is a man holding a cake.

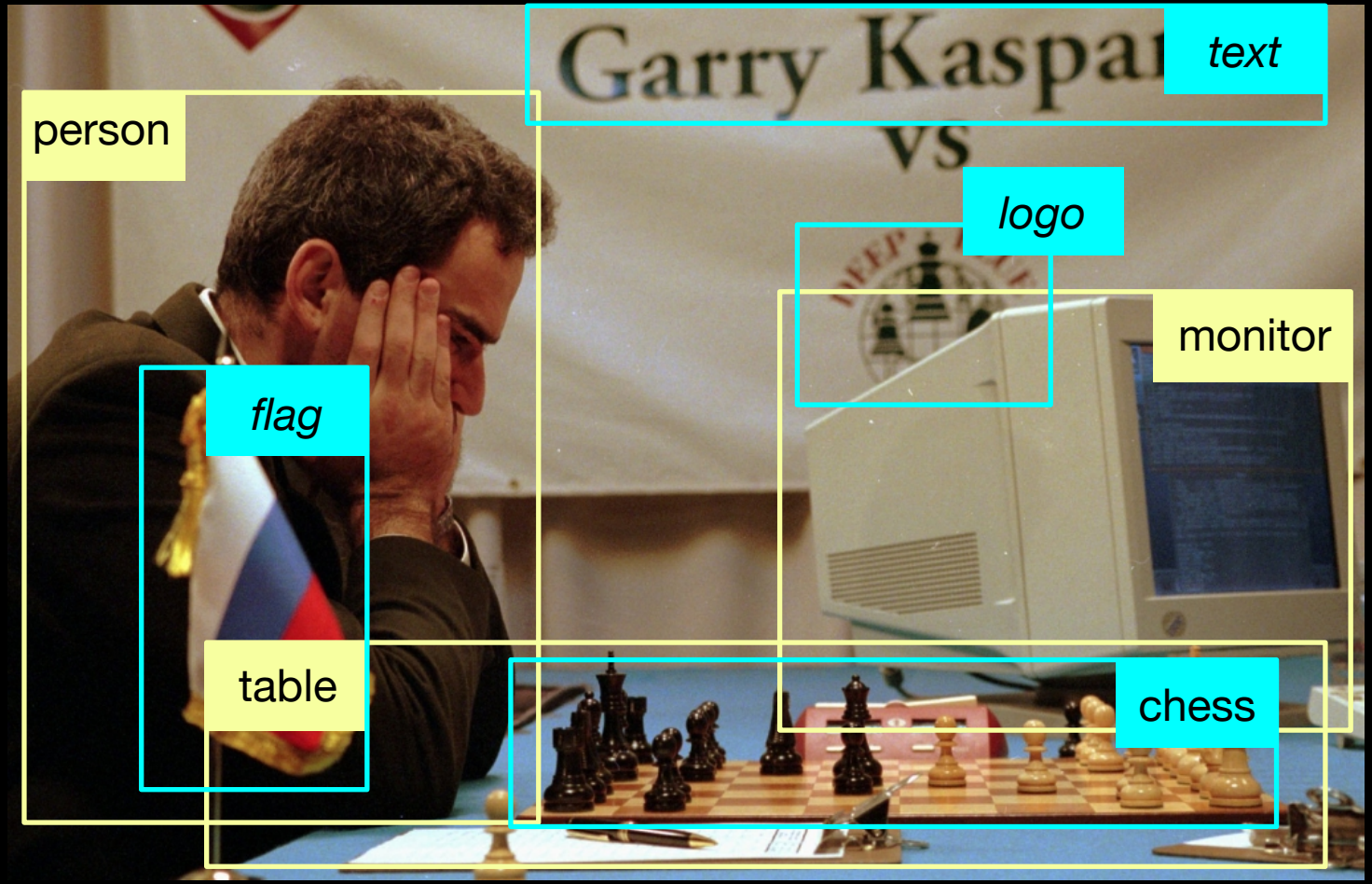
I am not really confident, but I think it is a man holding a cake.

I am not really confident, but I think it is a man holding a cake.

Russian chess player Garry Kasparov playing against computer program Deep Blue.

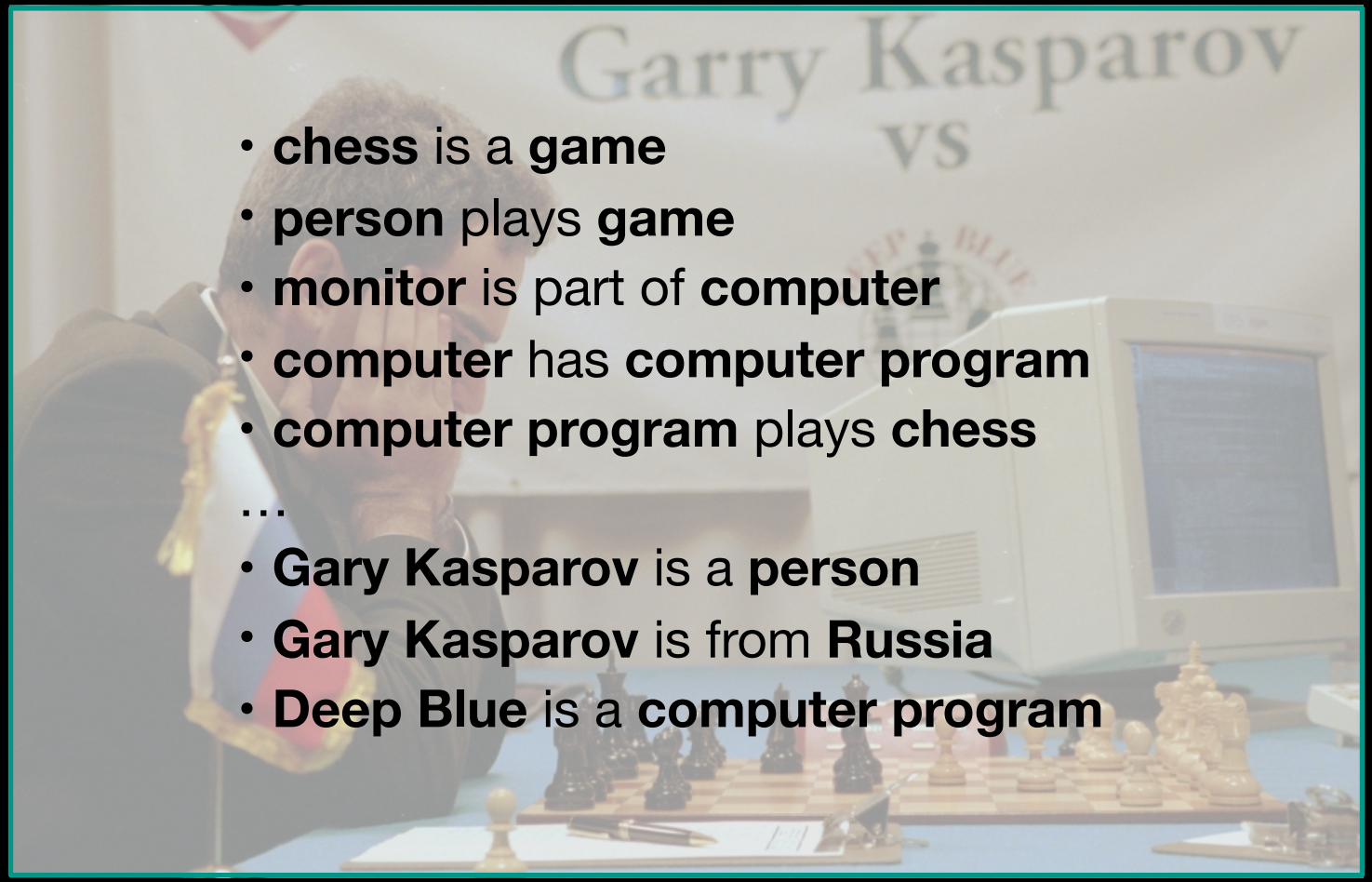
(I)

(I)



(II)

(II)



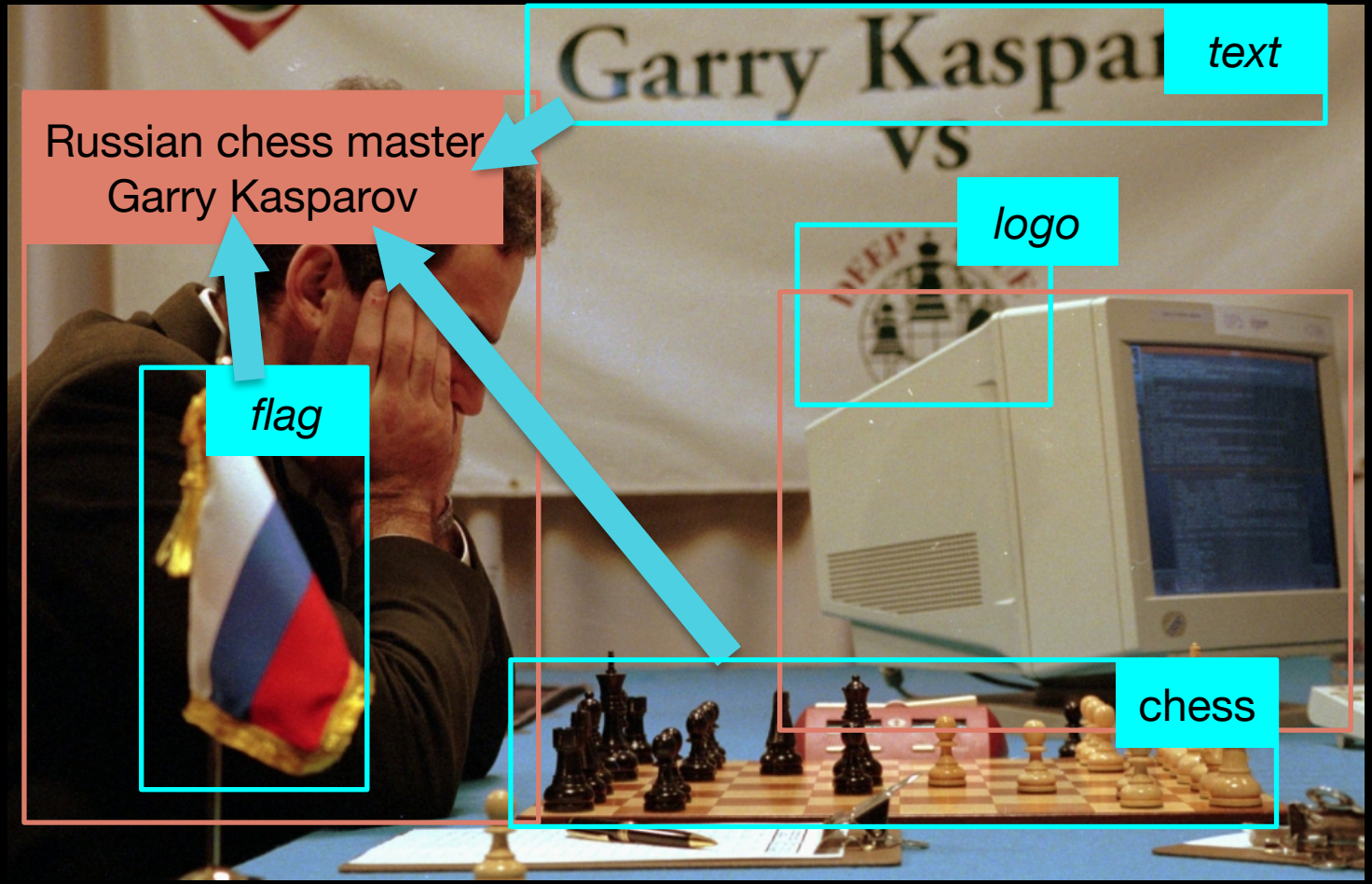
- **chess** is a **game**
- **person** plays **game**
- **monitor** is part of **computer**
- **computer** has **computer program**
- **computer program** plays **chess**

…

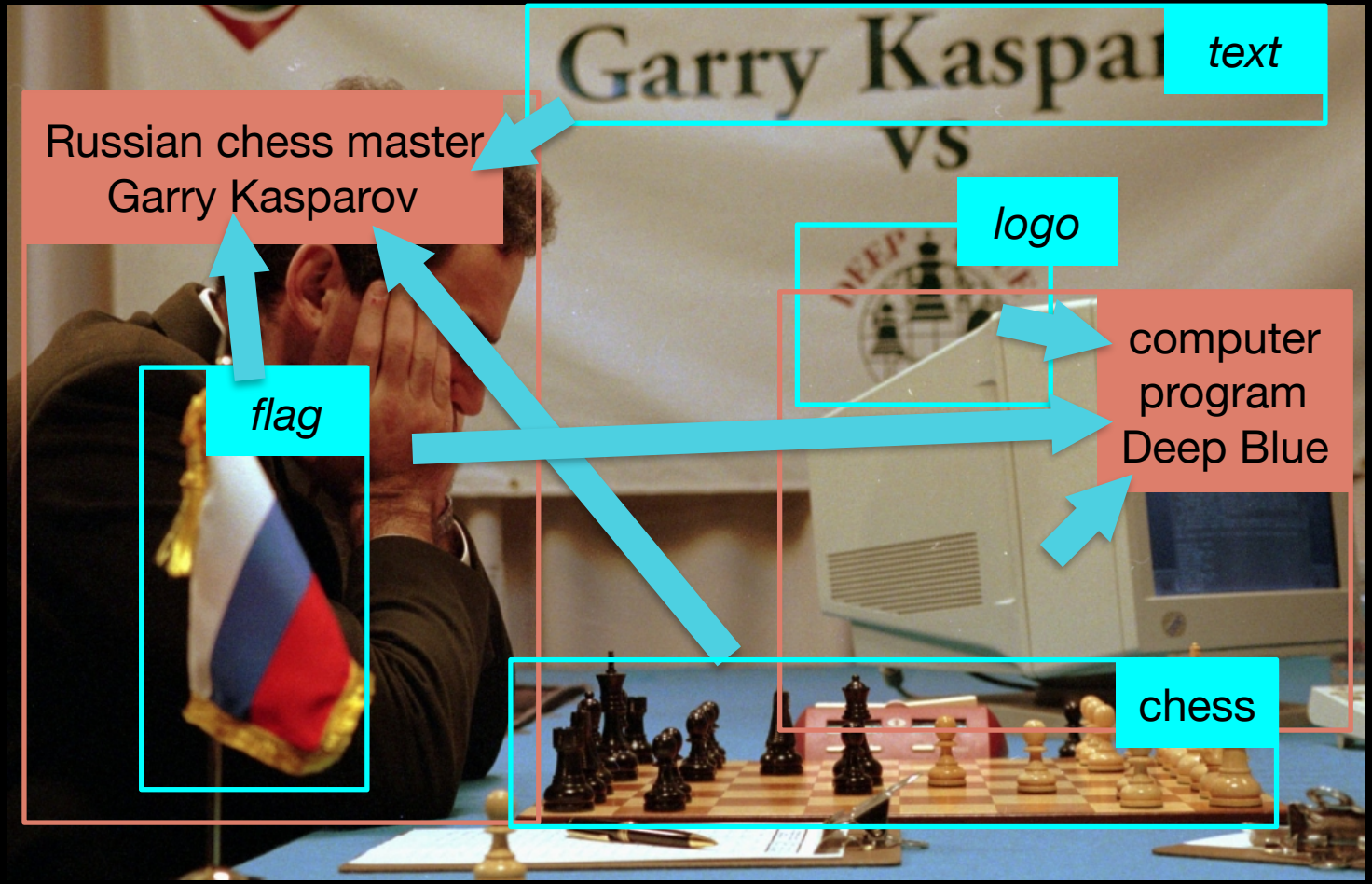
- **Gary Kasparov** is a **person**
- **Gary Kasparov** is from **Russia**
- **Deep Blue** is a **computer program**

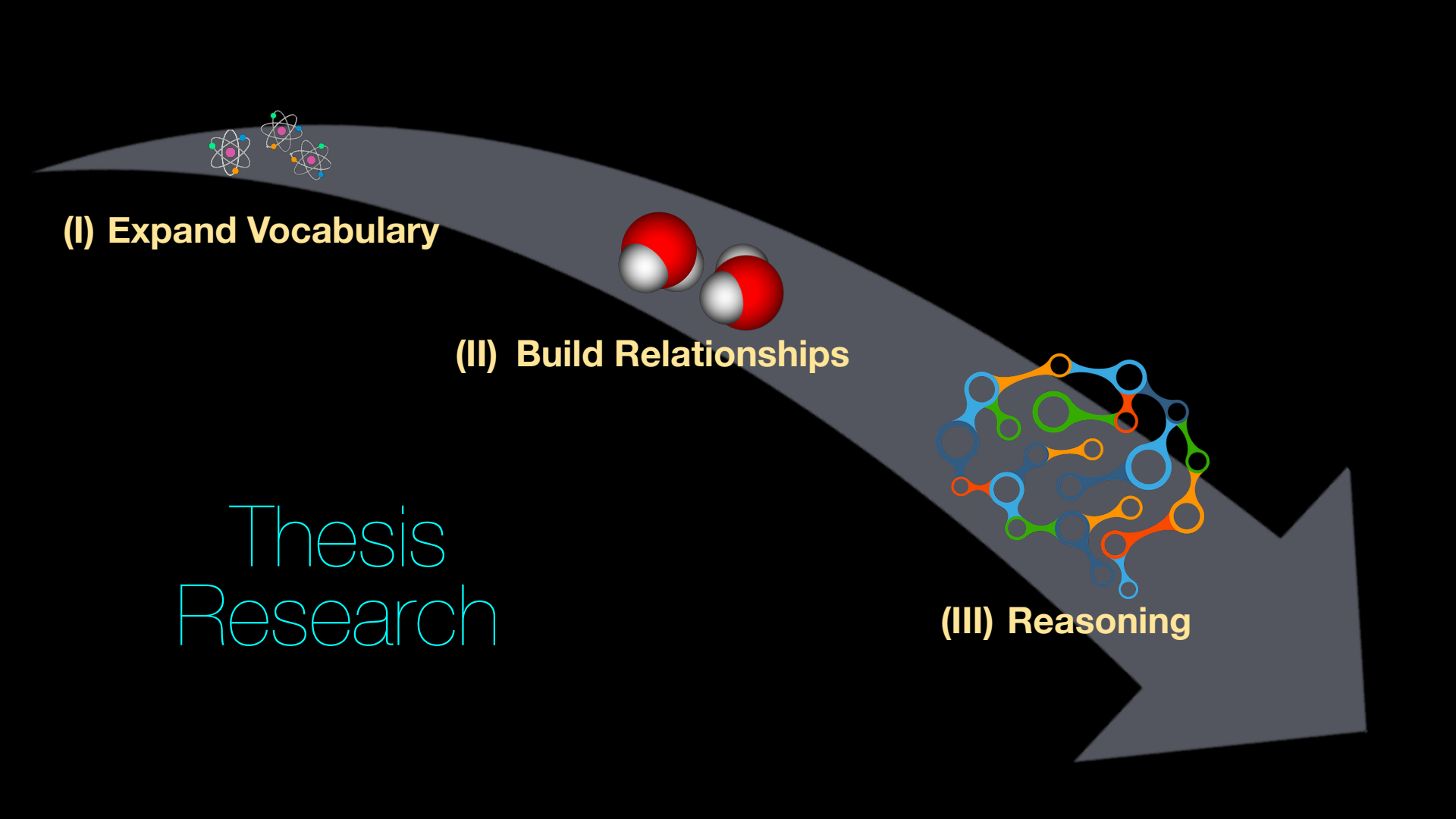
(III)

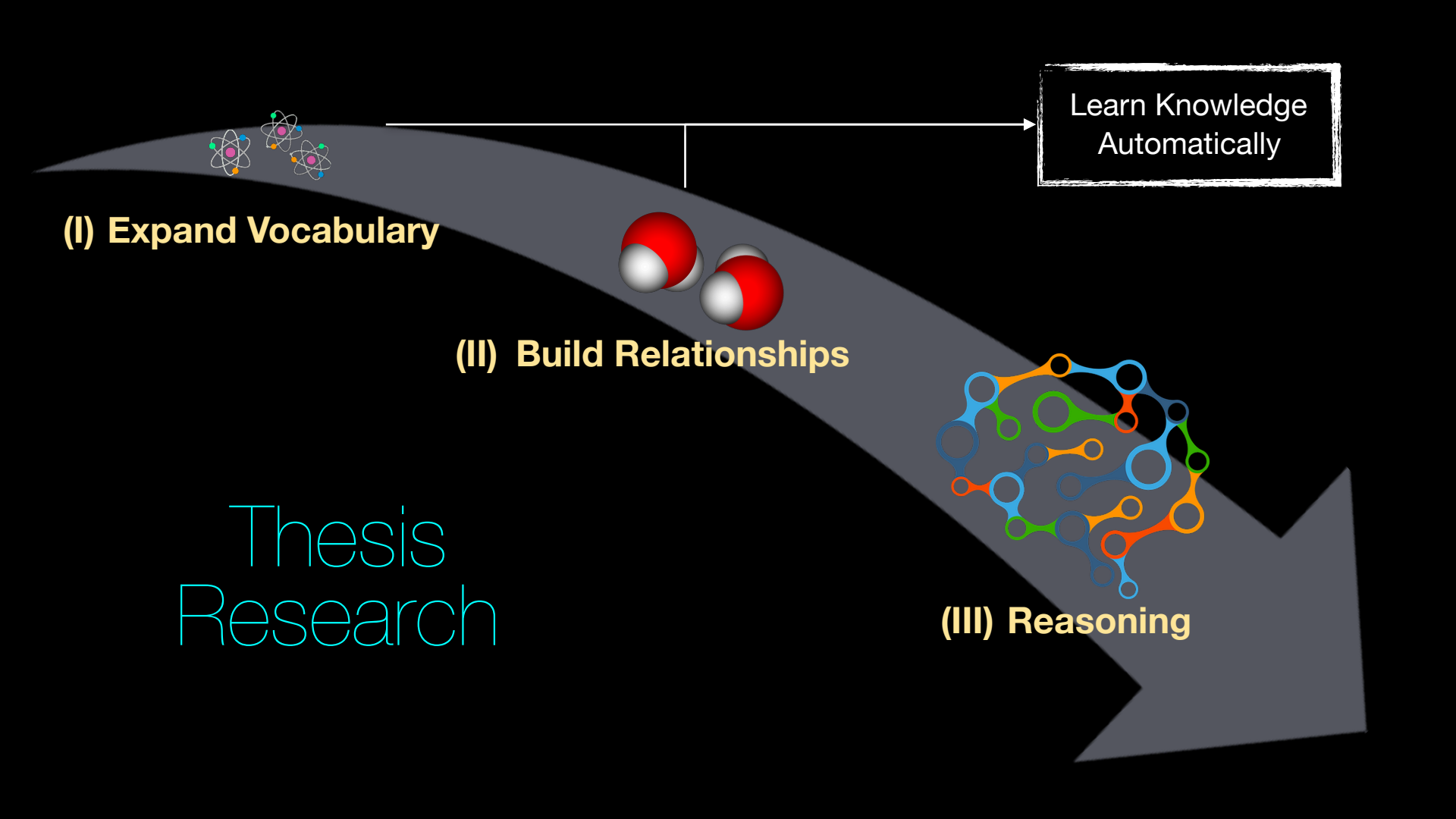
(III)

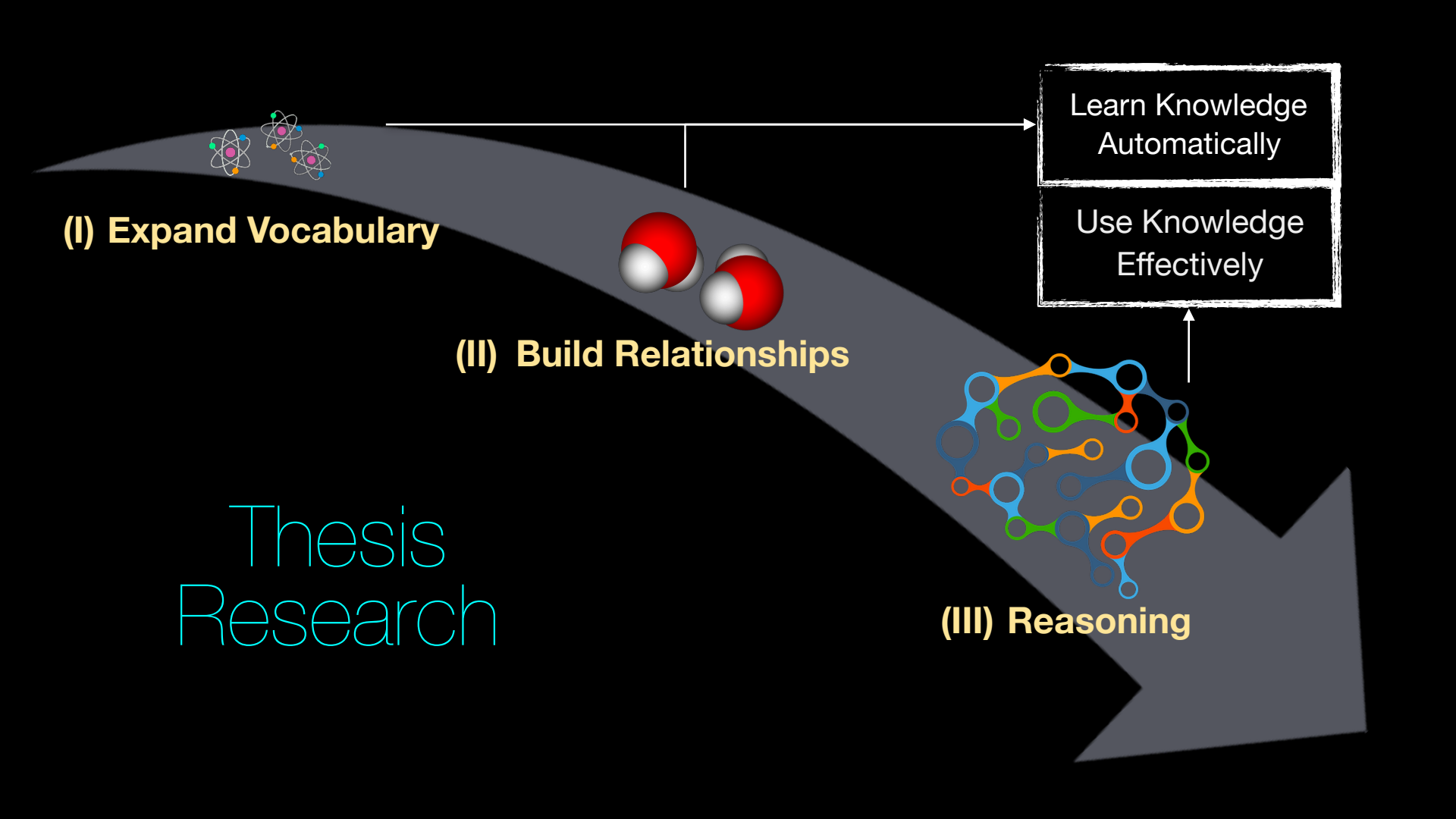

Garry Kaspar
VS
text
Russian chess master
Garry Kasparov
logo
flag
chess

(III)


Garry Kaspar
VS
text
Russian chess master
Garry Kasparov
logo
flag
computer
program
Deep Blue
chess


(I) Expand Vocabulary
(II) Build Relationships
(III) Reasoning
Thesis
Research


Learn Knowledge Automatically
(I) Expand Vocabulary
(II) Build Relationships
(III) Reasoning
Thesis Research


Learn Knowledge Automatically
Use Knowledge Effectively
(I) Expand Vocabulary
(II) Build Relationships
(III) Reasoning
Thesis Research

(I) Expand Vocabulary

- Detectors from the Web [ICCV13/15]

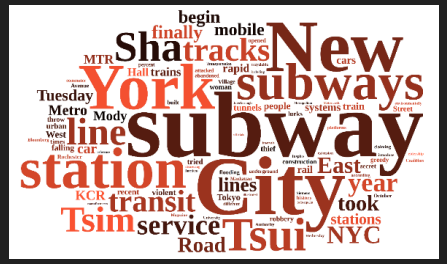
(I) Expand Vocabulary

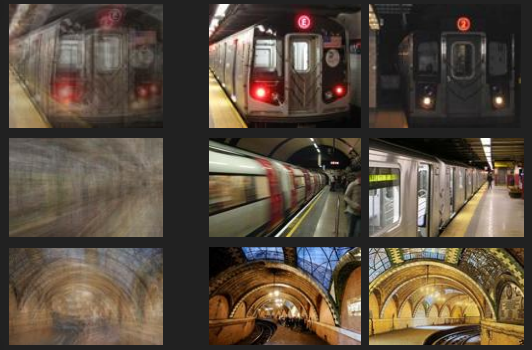
- Detectors from the Web [ICCV13/15]
- Pixel-Level Labeling [CVPR 2014]

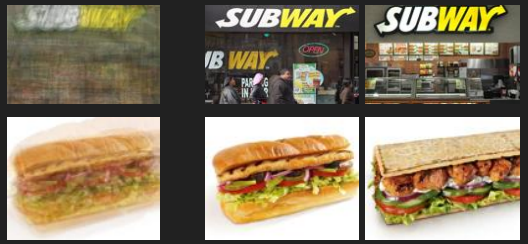
(I) Expand Vocabulary

- Detectors from the Web [ICCV13/15]
- Pixel-Level Labeling [CVPR 2014]
- Sense Discovery [CVPR 2015]

subway





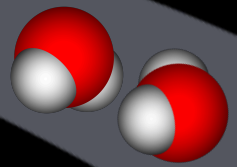


Semantic Senses

Visual Senses

Instances

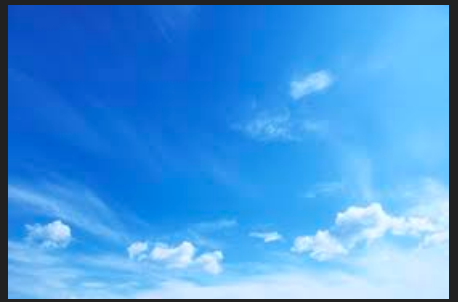
(II) Build Relationships

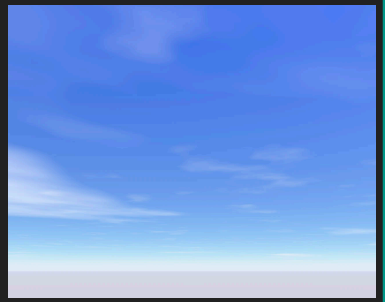


- Never Ending Image Learner [ICCV 2013]

Explicit, Structured Relationships

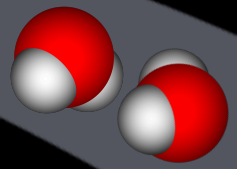
airplane is found in **runway**





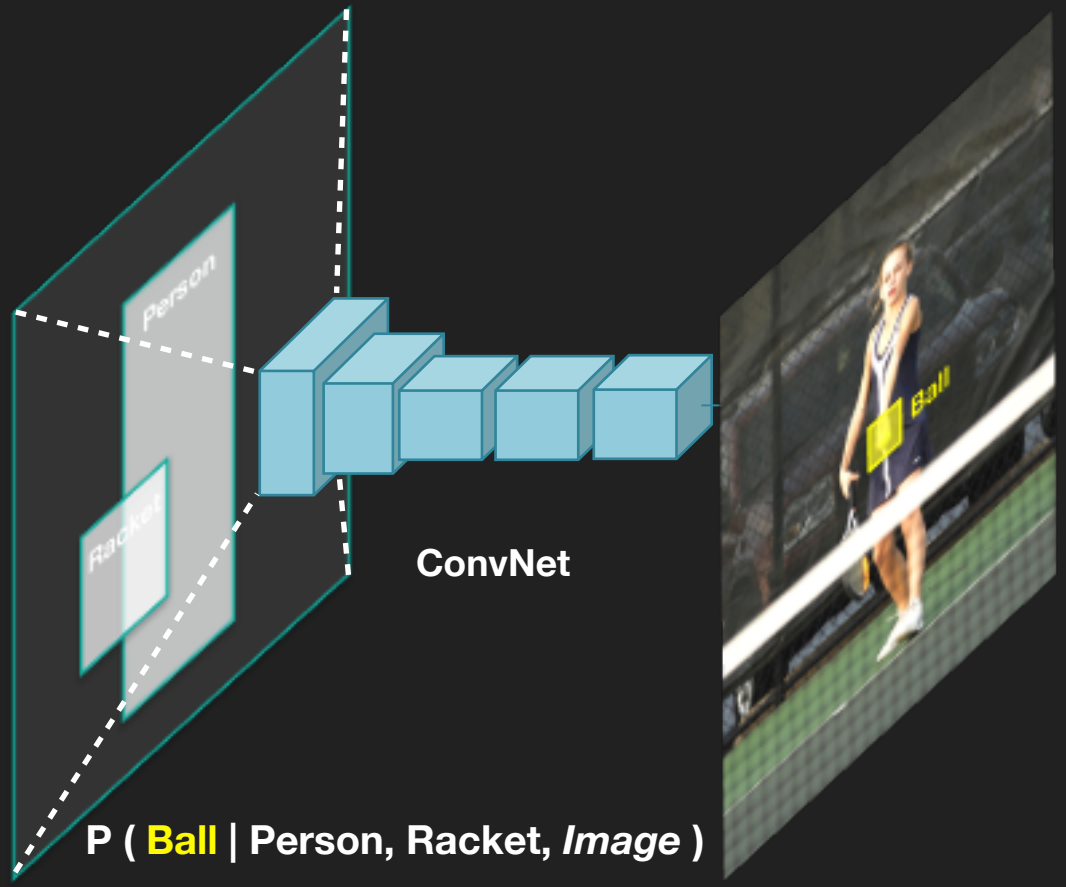
sky is **blue**

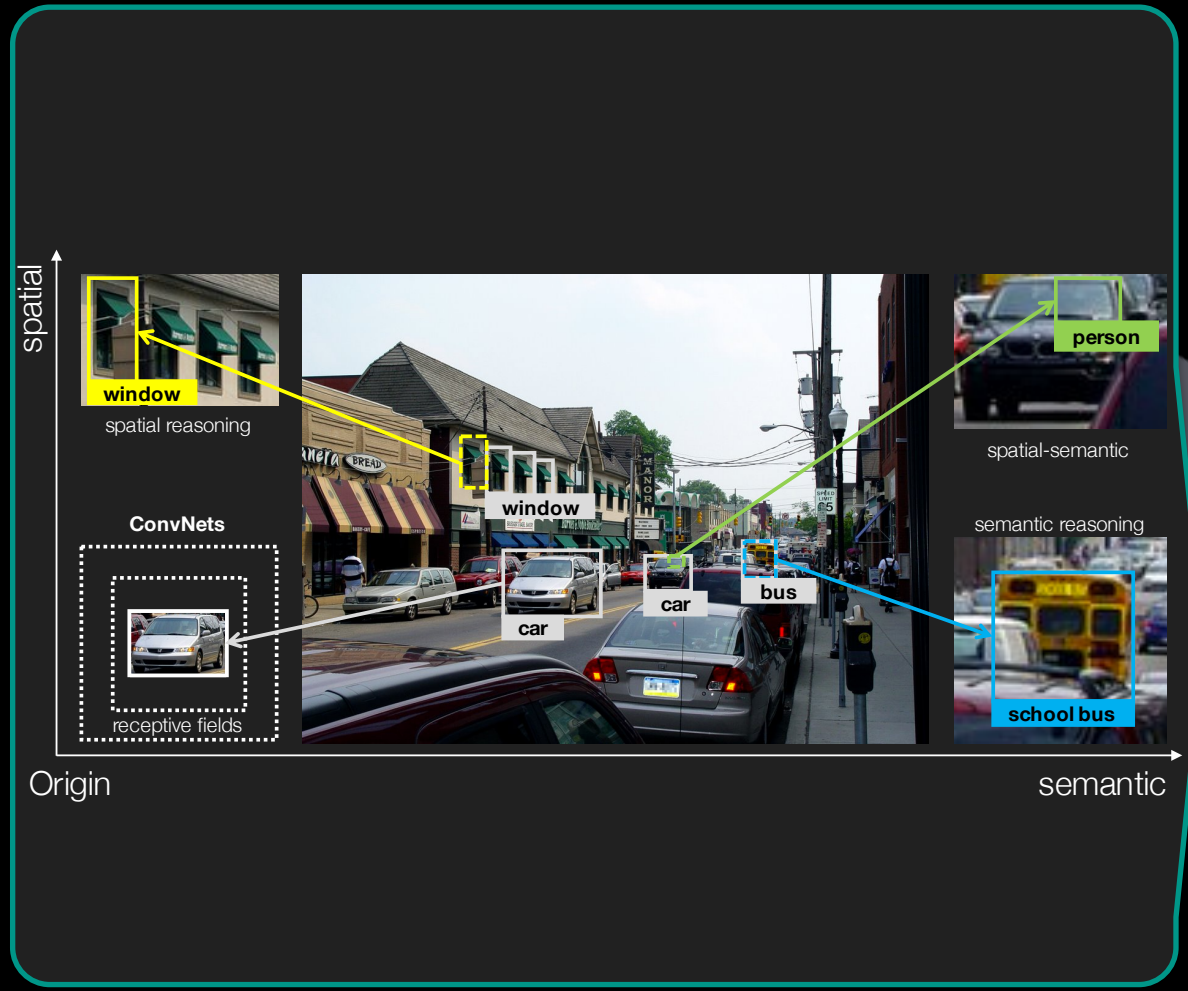
(II) Build Relationships



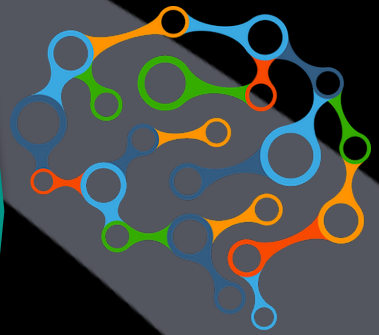
- Never Ending Image Learner [ICCV 2013]
- Spatial Memory Network [ICCV 2017]

Implicit, Contextual Relationships

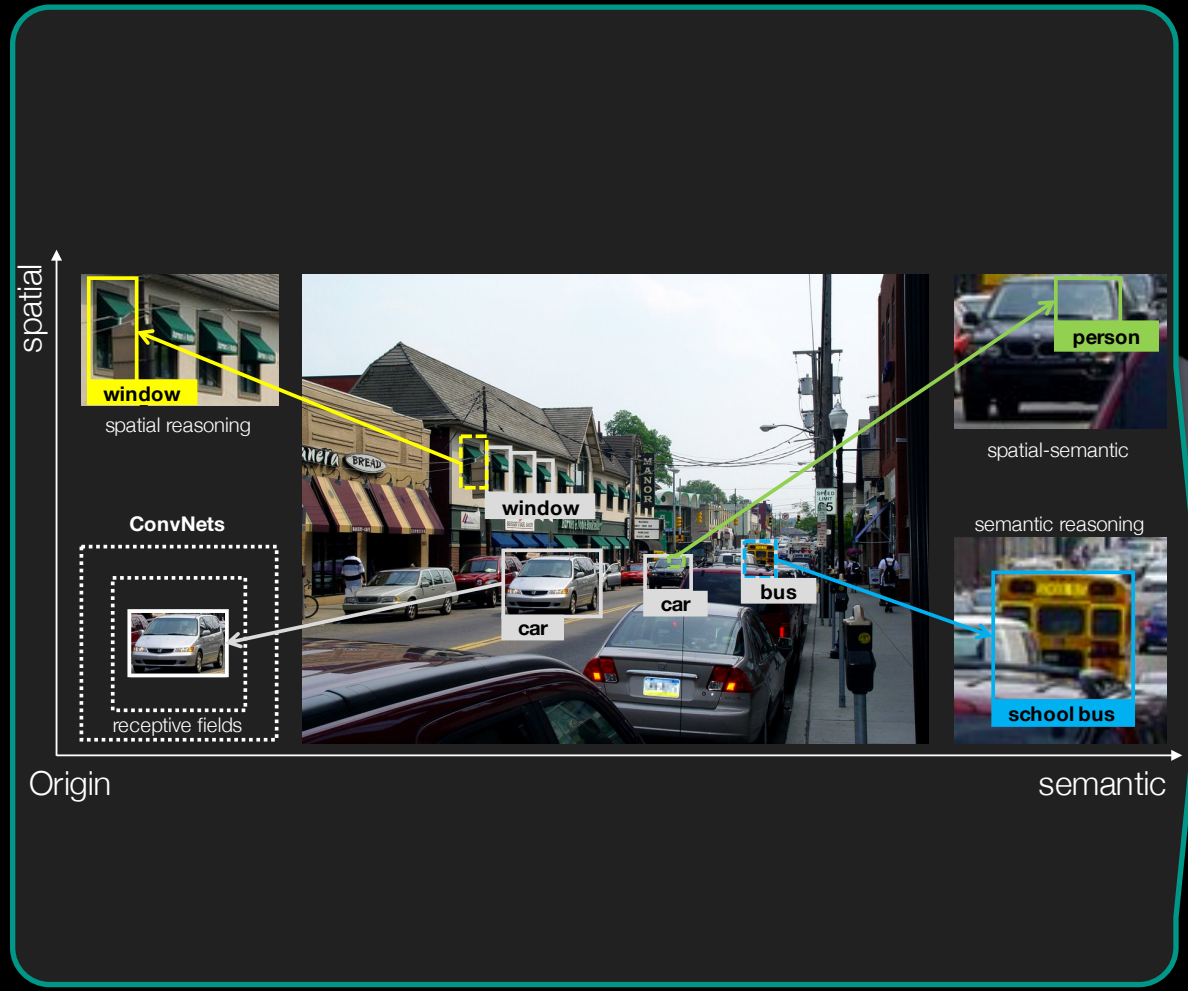




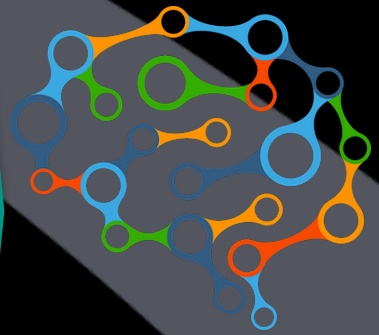
(III) Reasoning



- **Iterative Reasoning [submitted]**

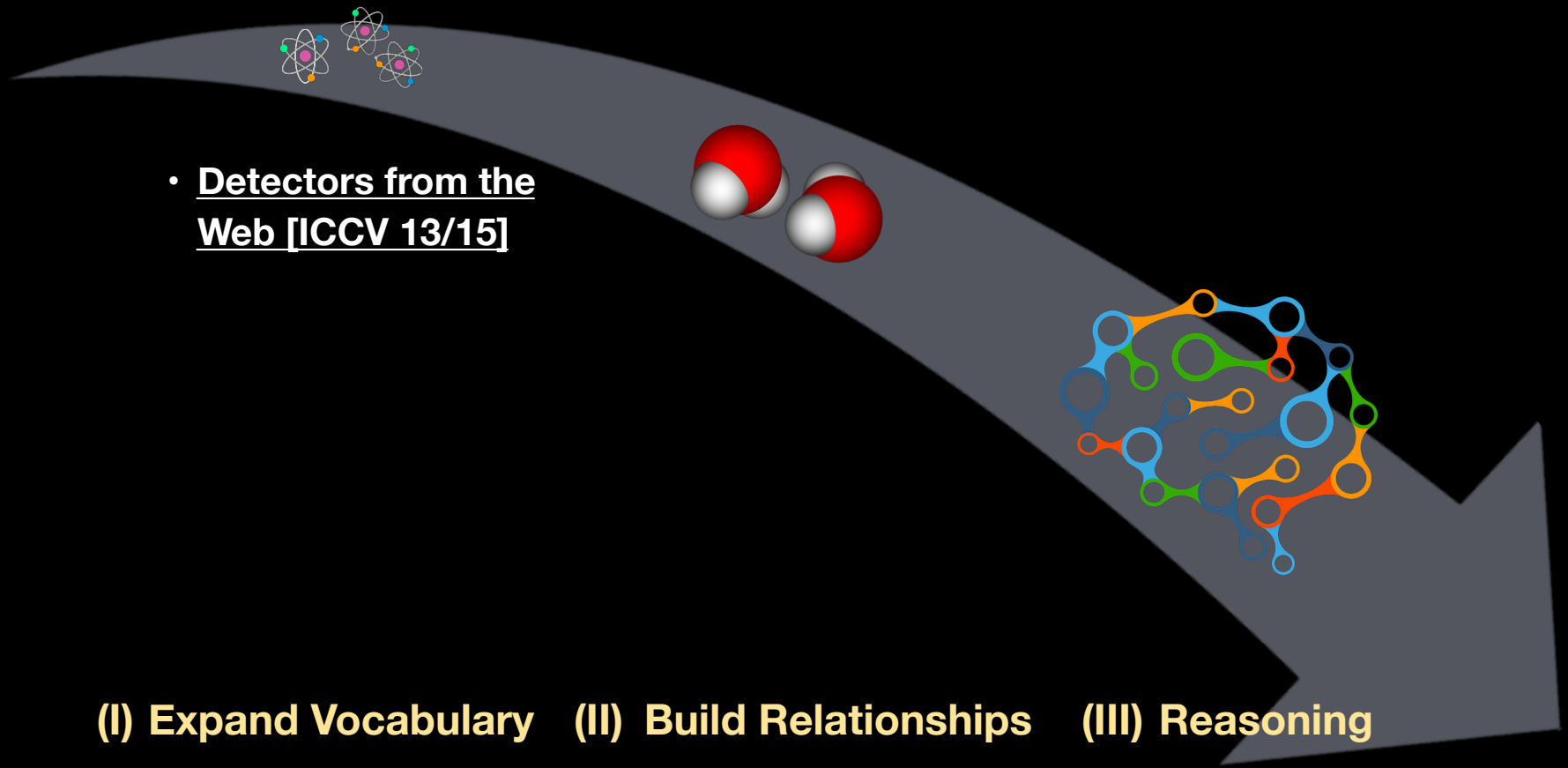


(III) Reasoning



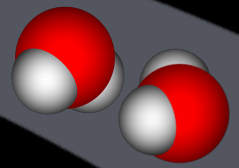
- **Iterative Reasoning [submitted]**

- **Detectors from the Web [ICCV 13/15]**

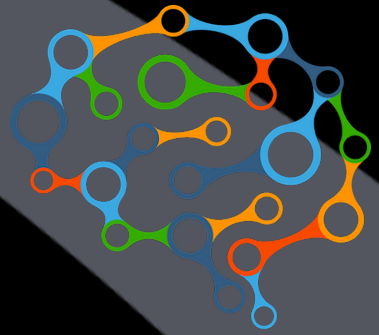


(I) Expand Vocabulary (II) Build Relationships (III) Reasoning

- **Detectors from the Web [ICCV 13/15]**
- Pixel-Level Labeling [CVPR 2014]
- Sense Discovery [CVPR 2015]



- **Never Ending Image Learner [ICCV 2013]**
- Spatial Memory Network [ICCV 2017]



- **Iterative Reasoning [submitted]**

(I) Expand Vocabulary **(II) Build Relationships** **(III) Reasoning**

- **Detectors from the Web [ICCV 2013/2015]**
- Pixel-Level Labeling [CVPR 2014]
- Sense Discovery [CVPR 2015]

(I) Expand Vocabulary

Harness Human Intelligence

amazon mechanical turk beta

LabelMe

(Russell et al., 2007) (Everingham et al., 2010) (Lin et al., 2014)

Harness Human Intelligence

amazon mechanical turk beta

LabelMe

(Russell et al., 2007) (Everingham et al., 2010) (Lin et al., 2014)

Scalable?

amazon mechanical turk beta

LabelMe

(Deng et al., 2009) (Russakovsky et al., 2015) (Kalkowski et al., 2015)

Scalable?

~1M boxes 5 years
~3K classes

(Deng et al., 2009) (Russakovsky et al., 2015) (Kalkowski et al., 2015)

Scalable?

amazon mechanical turk beta

LabelMe

~1M boxes 5 years
~3K classes

~800M images everyday
up to 8M tags

IMAGENET

(Deng et al., 2009) (Russakovsky et al., 2015) (Kalkowski et al., 2015)

(Li & Fei-Fei, 2010) (Chen et al., 2013) (Divvala et al., 2014)

Learning Detectors from the Web

chess

Learning Detectors from the Web

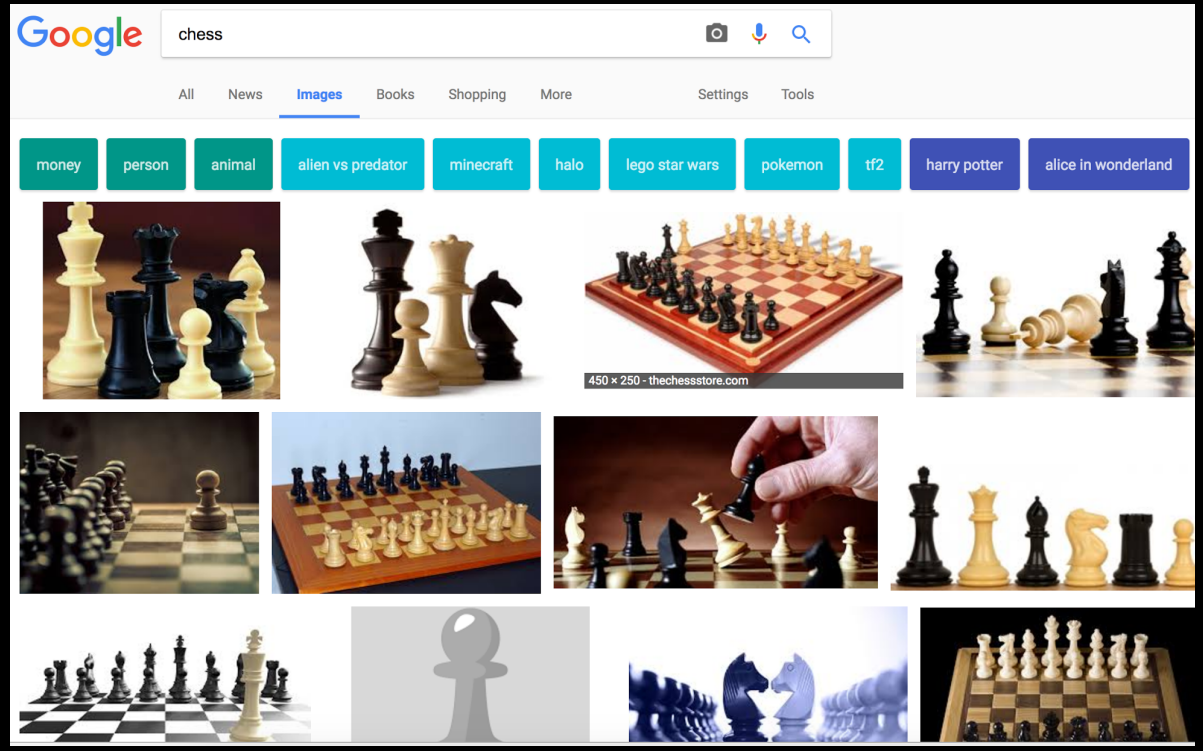
chess

Google Image Search

flickr

Learning Detectors from the Web

chess

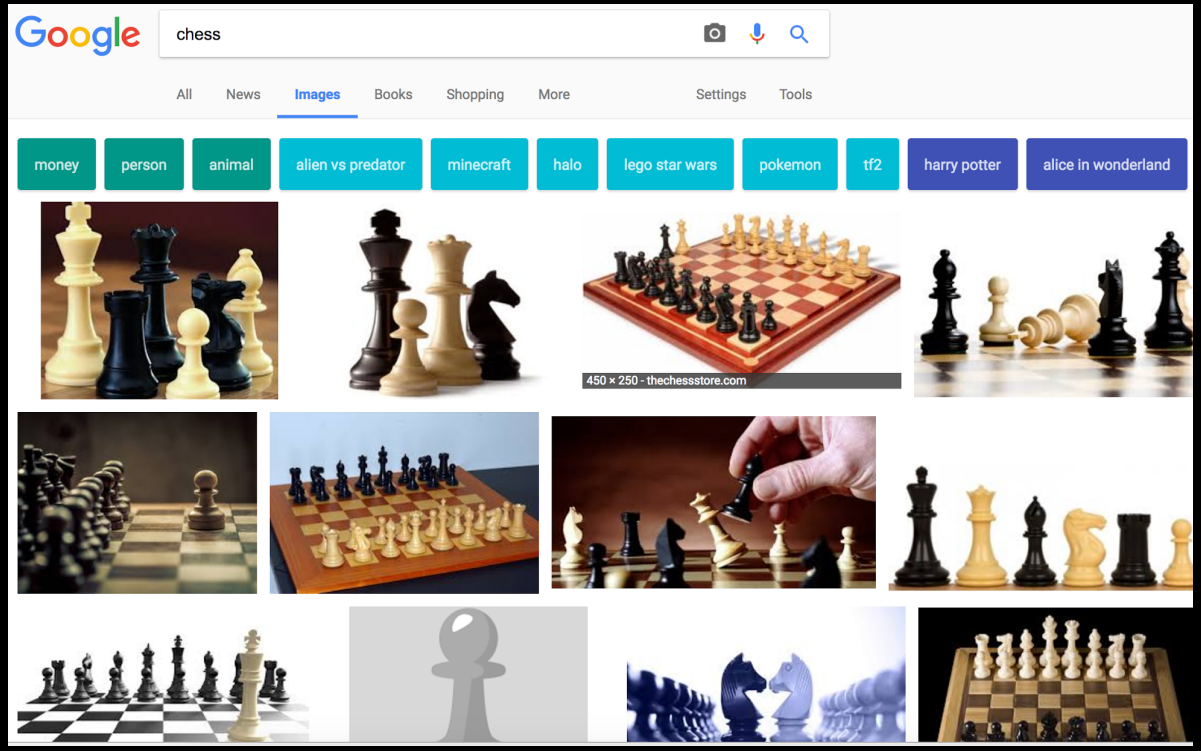


Google Image Search

flickr

Learning Detectors from the Web

chess



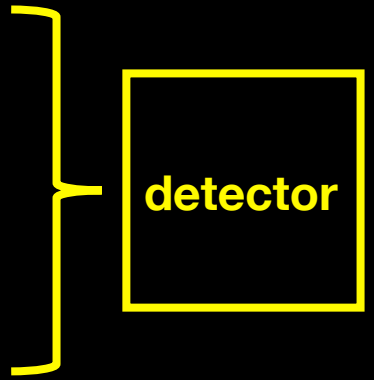
Google Image Search

flickr

chess detector

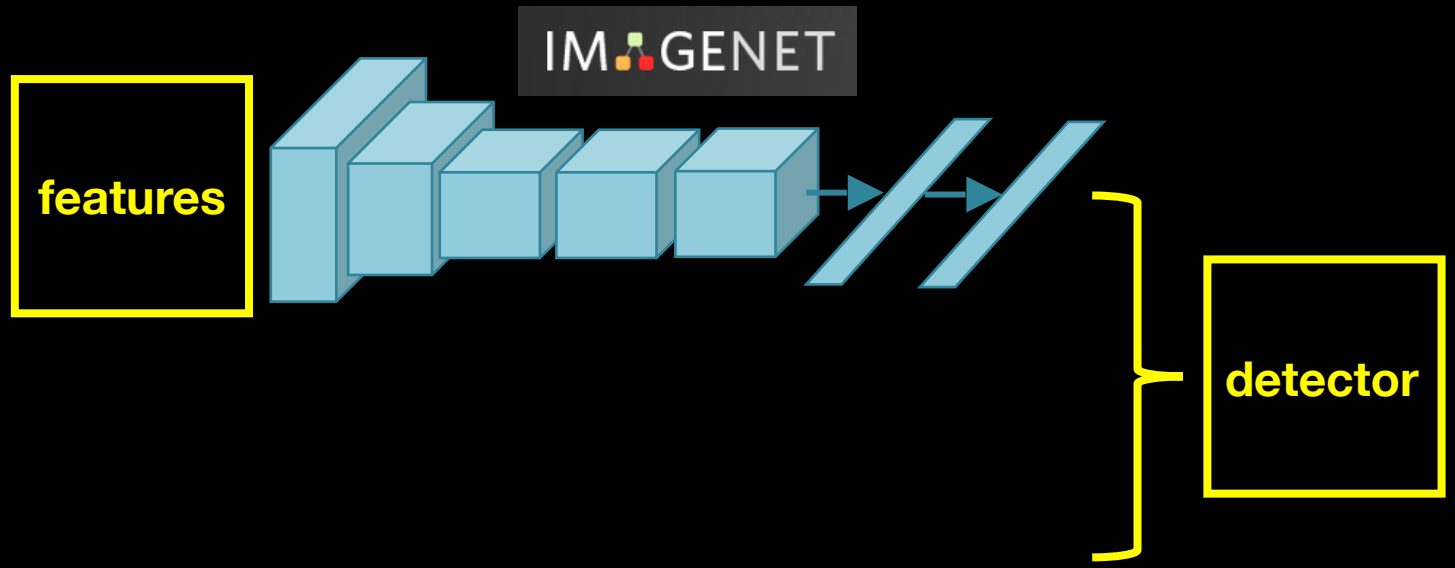

Garry Kasparov
VS
DEEP BLUE
chess

Dissecting Current Detectors



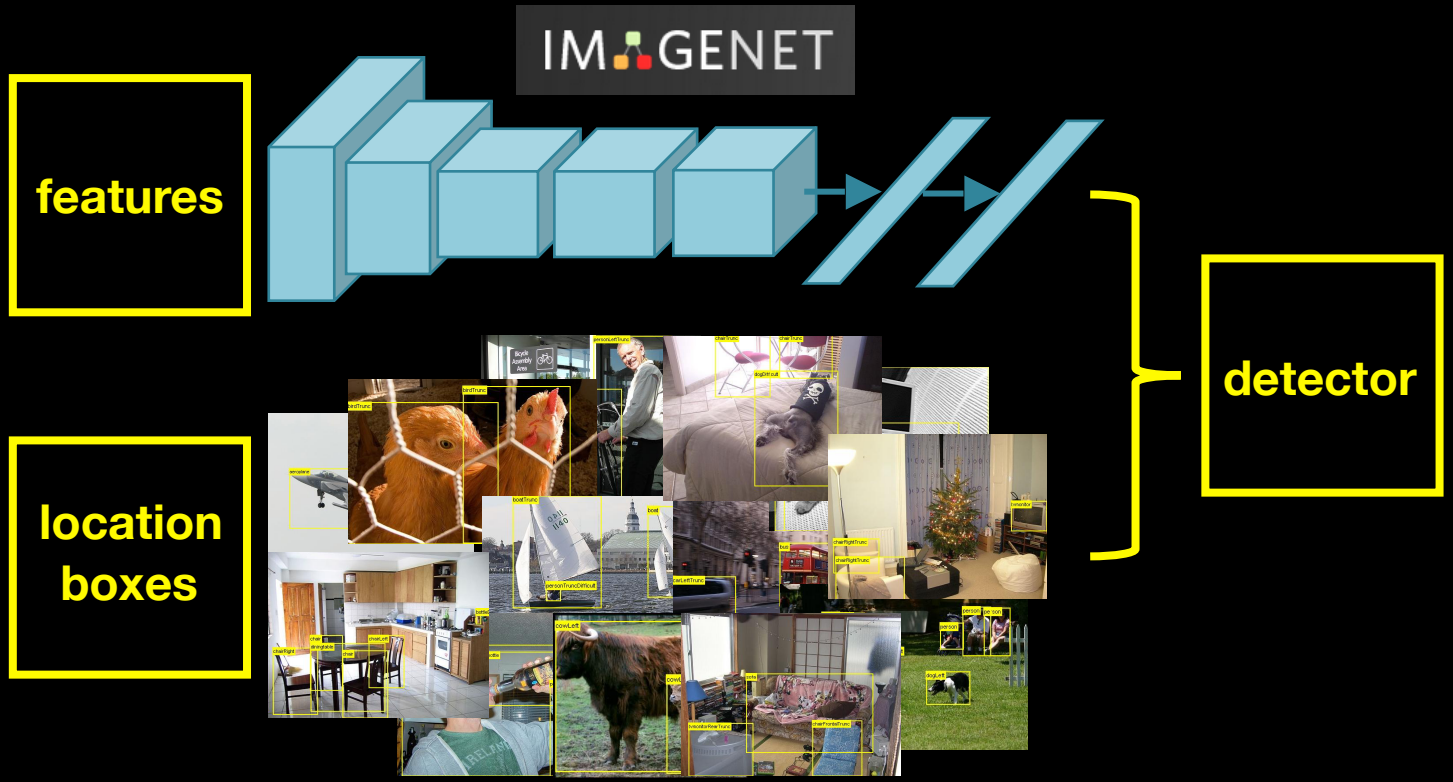
(Girshick et al., 2015) (Ren et al., 2015)
(Liu et al., 2016) (Redmon et al., 2016)

Dissecting Current Detectors



(Girshick et al., 2015) (Ren et al., 2015)
(Liu et al., 2016) (Redmon et al., 2016)

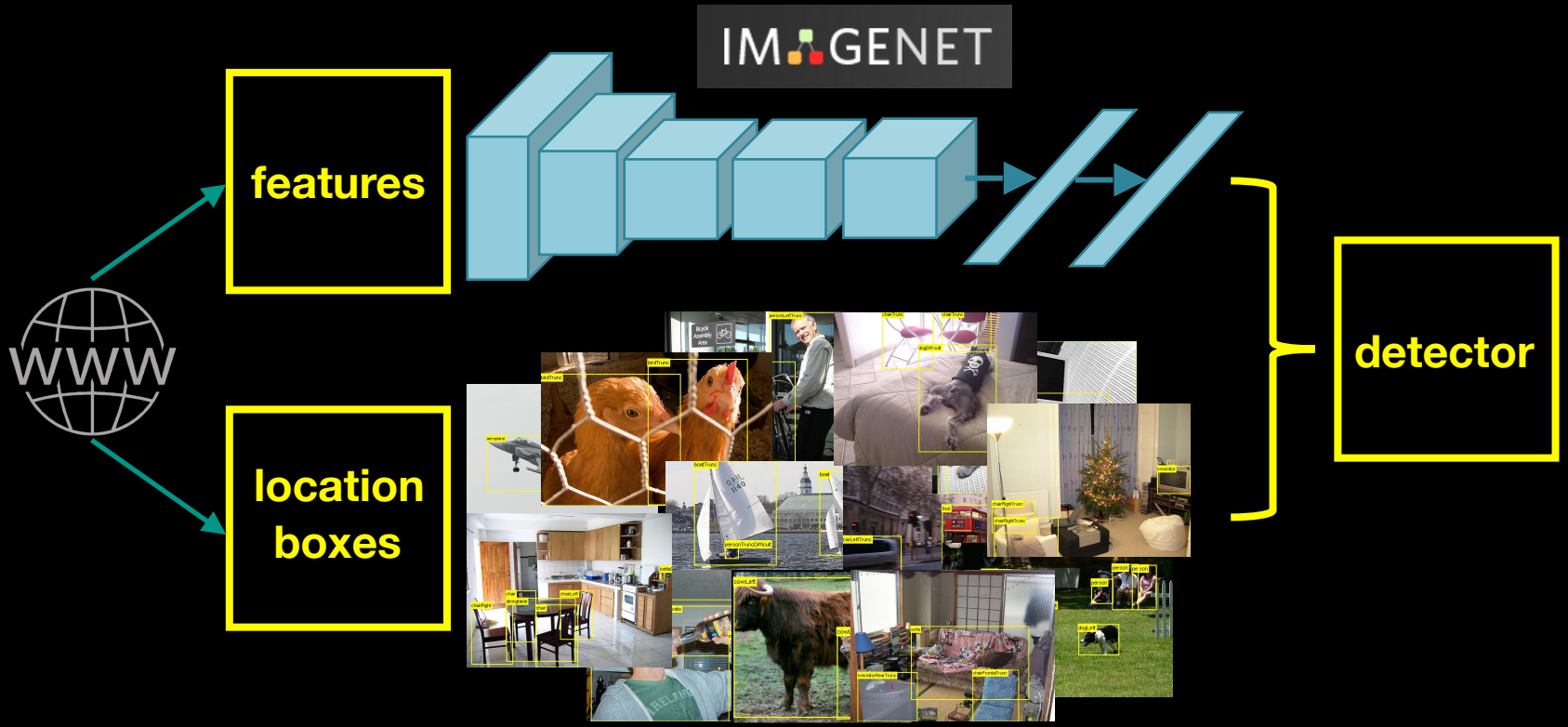
Dissecting Current Detectors



(Girshick et al., 2015) (Ren et al., 2015)
(Liu et al., 2016) (Redmon et al., 2016)

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

Dissecting Current Detectors



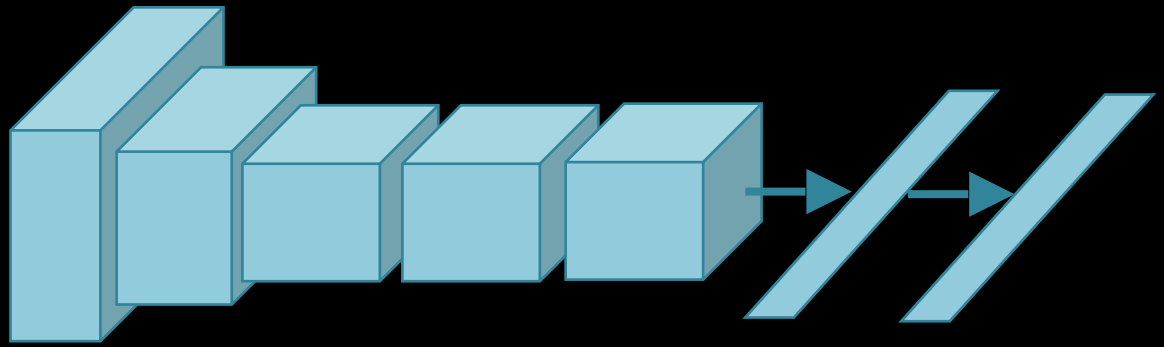
(Girshick et al., 2015) (Ren et al., 2015)
(Liu et al., 2016) (Redmon et al., 2016)

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

Visual Features from the Web

(Chen & Gupta, 2015)

Visual Features from the Web



(Chen & Gupta, 2015)

Basic Setup

@2015

(Deng et al., 2009) (Chen et al., 2013)
(Krizhevesky et al., 2012) (Girshick et al., 2014)

Basic Setup

@2015

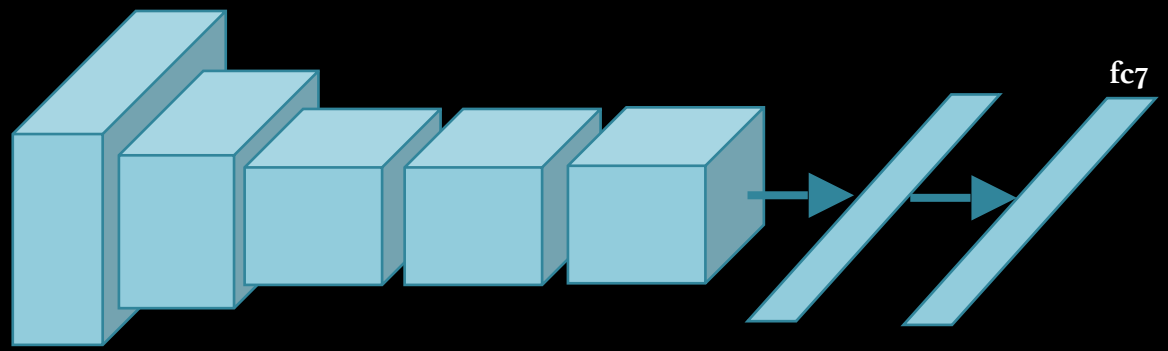
- List of categories**:** ImageNet (1K) + (Object/Attribute) 1.3K

(Deng et al., 2009) (Chen et al., 2013)
(Krizhevesky et al., 2012) (Girshick et al., 2014)

Basic Setup

@2015

- List of categories**:** ImageNet (1K) + (Object/Attribute) 1.3K
- Use category names as **queries**

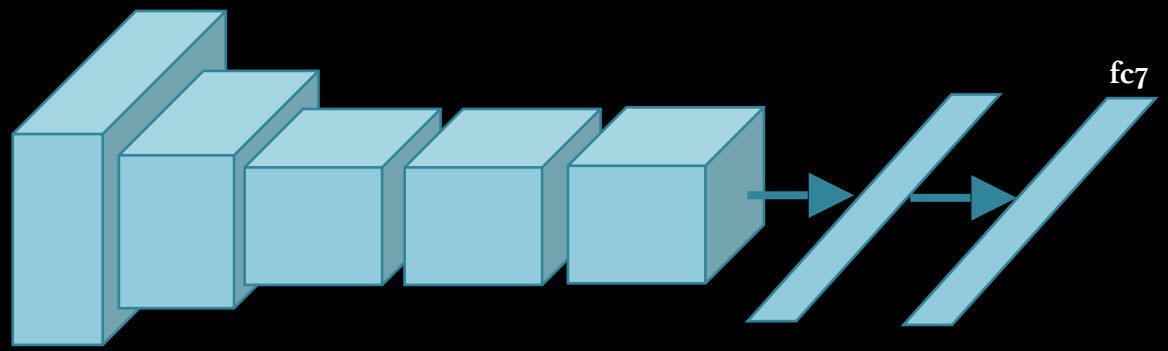


(Deng et al., 2009) (Chen et al., 2013)
(Krizhevesky et al., 2012) (Girshick et al., 2014)

Basic Setup

@2015

- List of categories**:** ImageNet (1K) + (Object/Attribute) 1.3K
- Use category names as **queries**
- Train **AlexNet** to predict category names, fc7 features

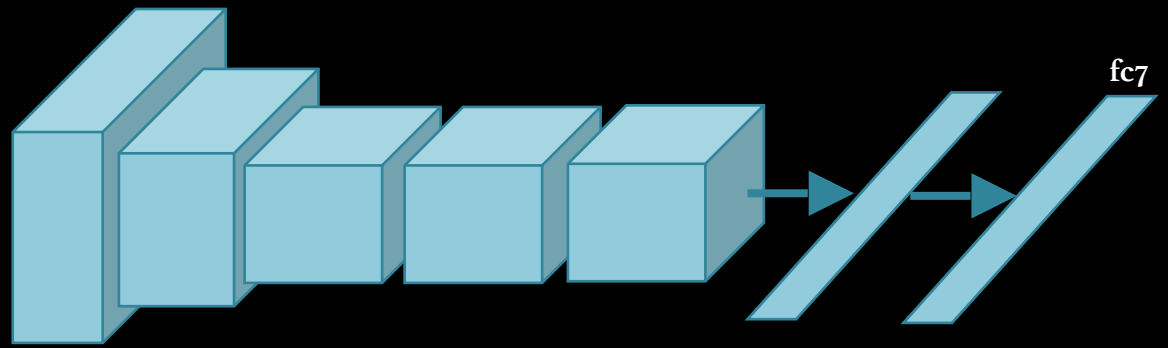


(Deng et al., 2009) (Chen et al., 2013)
(Krizhevesky et al., 2012) (Girshick et al., 2014)

Basic Setup

@2015

- List of categories**:** ImageNet (1K) + (Object/Attribute) 1.3K
- Use category names as **queries**
- Train **AlexNet** to predict category names, fc7 features
- **R-CNN,** VOC boxes, mAP

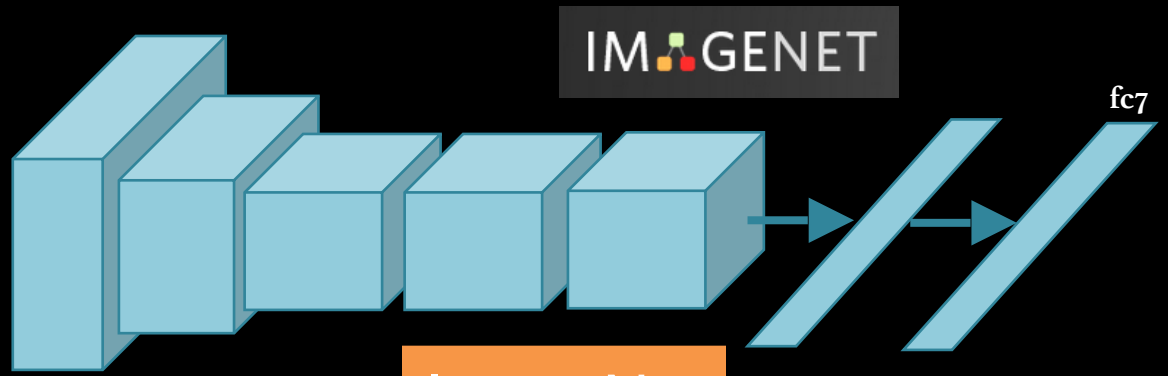


(Deng et al., 2009) (Chen et al., 2013)
(Krizhevesky et al., 2012) (Girshick et al., 2014)

Basic Setup

@2015

- List of categories**:** ImageNet (1K) + (Object/Attribute) 1.3K
- Use category names as **queries**
- Train **AlexNet** to predict category names, fc_7 features
- **R-CNN,** VOC boxes, mAP

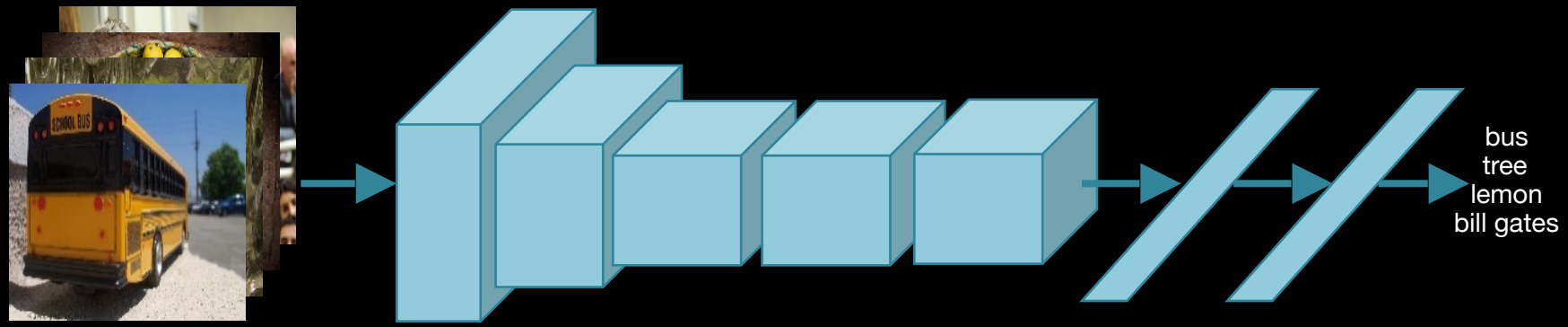


ImageNet
44.7

w/o Fine-Tune!

(Deng et al., 2009) (Chen et al., 2013)
(Krizhevesky et al., 2012) (Girshick et al., 2014)

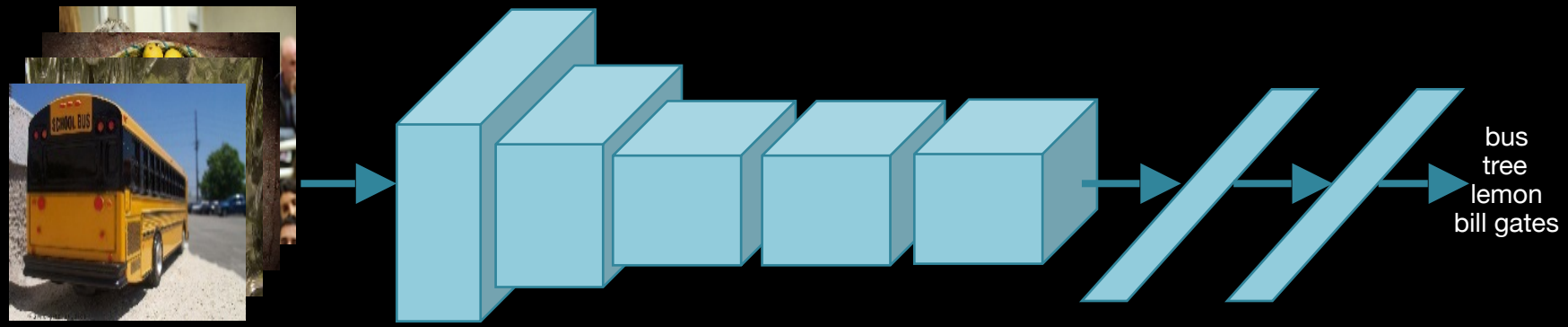
Trial 1: Train from Flickr



flickr

1.2M

Trial 1: Train from Flickr

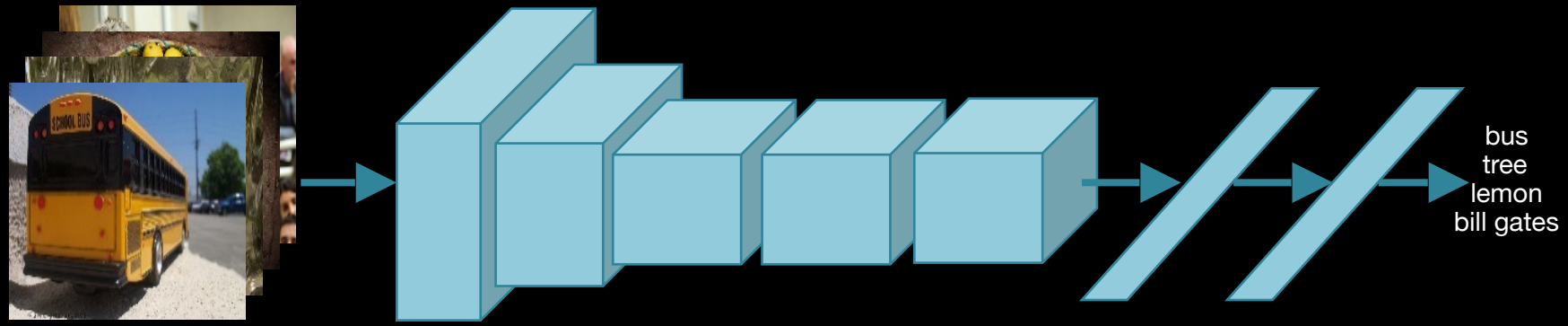


flickr

1.2M

ImageNet	Flickr
44.7	38.1

Trial 1: Train from Flickr



flickr

1.2M

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

ImageNet	Flickr
44.7	38.1

Scratch
40.7

Two Types of Web Data

Two Types of Web Data

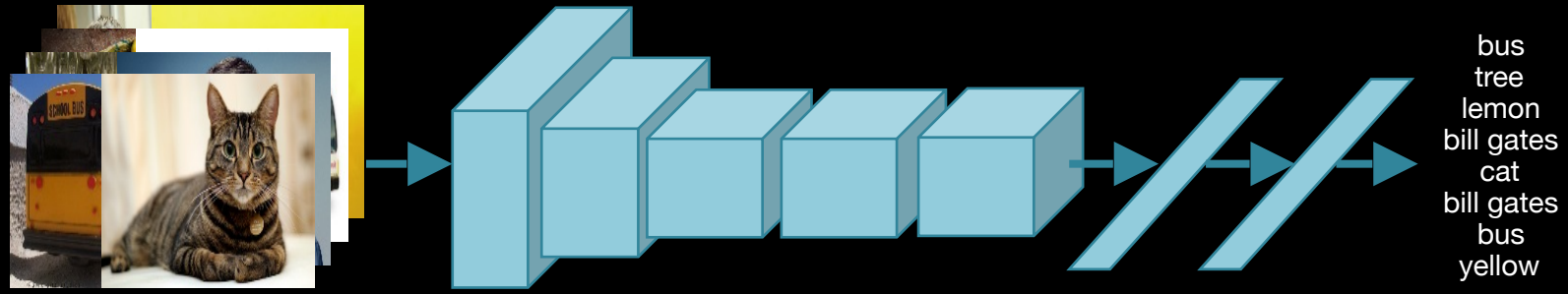
flickr

Two Types of Web Data

flickr

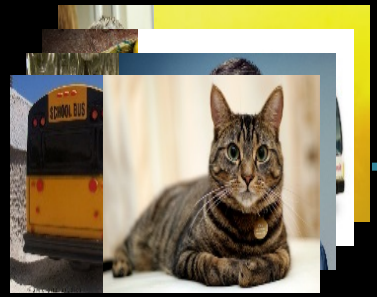
Google

Trial 2: More Data



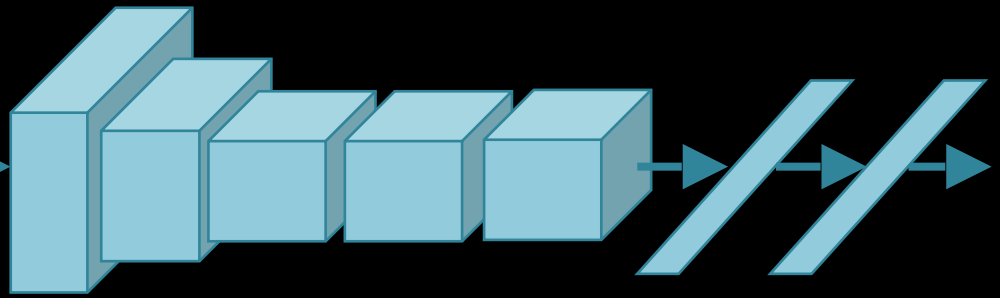
flickr

Trial 2: More Data



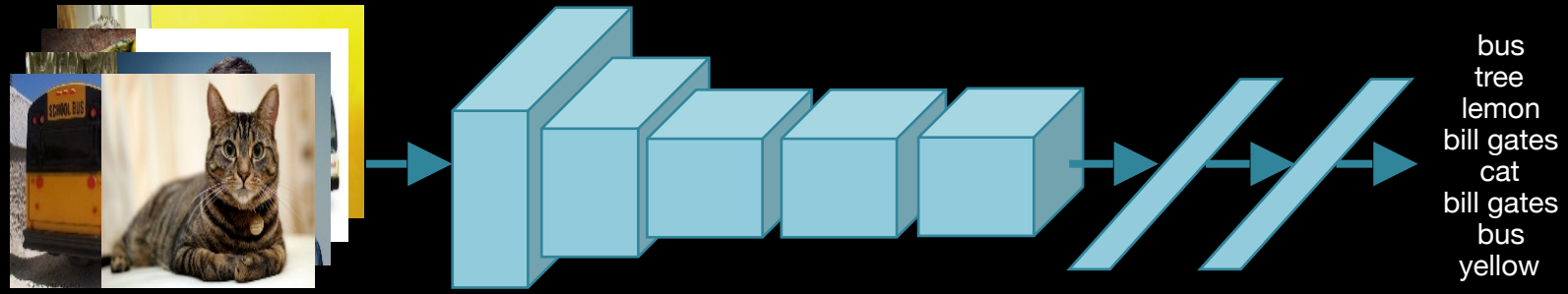
flickr + Google

1.5M



bus
tree
lemon
bill gates
cat
bill gates
bus
yellow

Trial 2: More Data

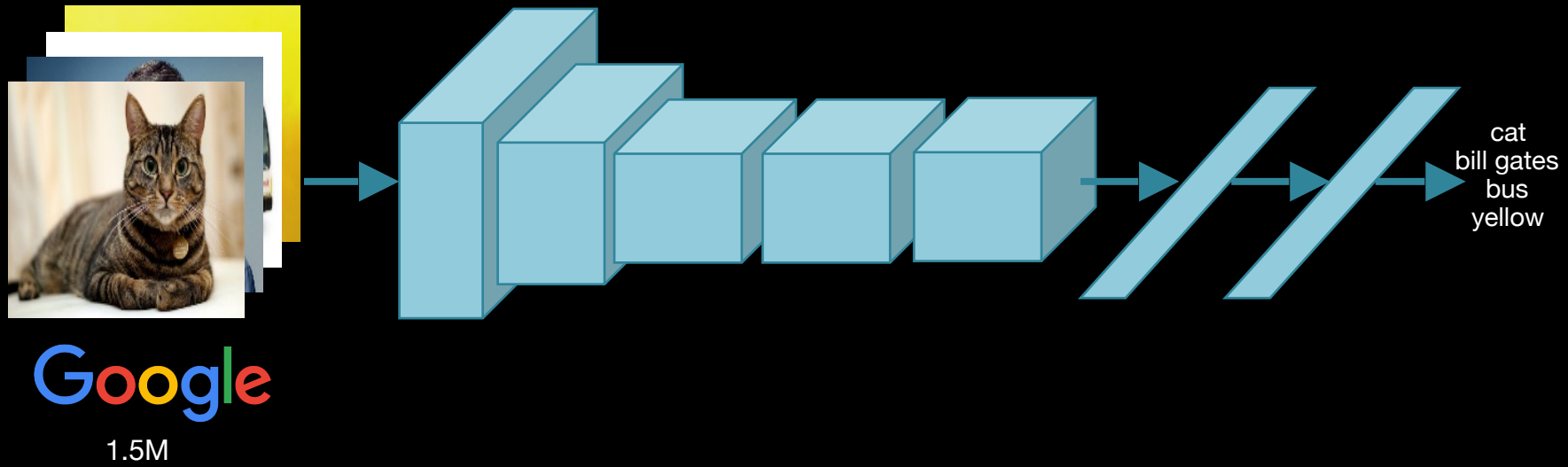


flickr + Google

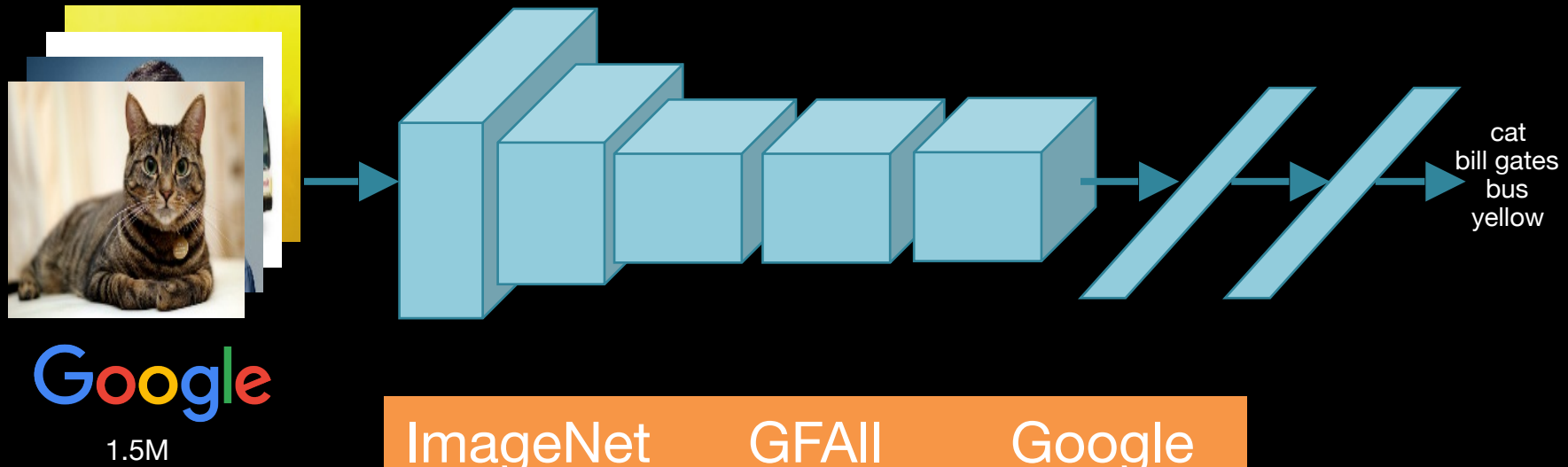
1.5M

ImageNet	FlickrS	GFAll
44.7	38.1	40.5

Trial 3: Train from Google



Trial 3: Train from Google

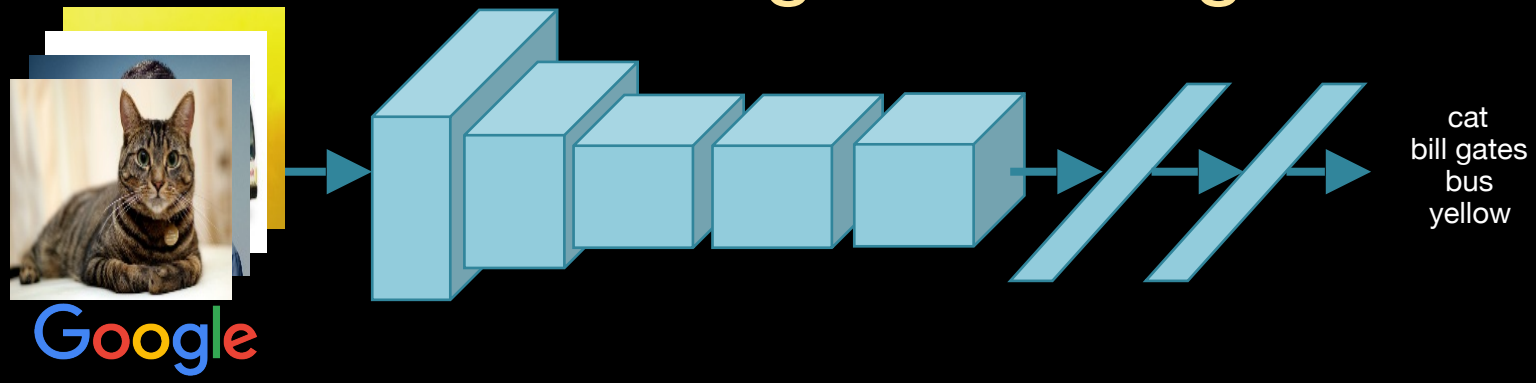


Google

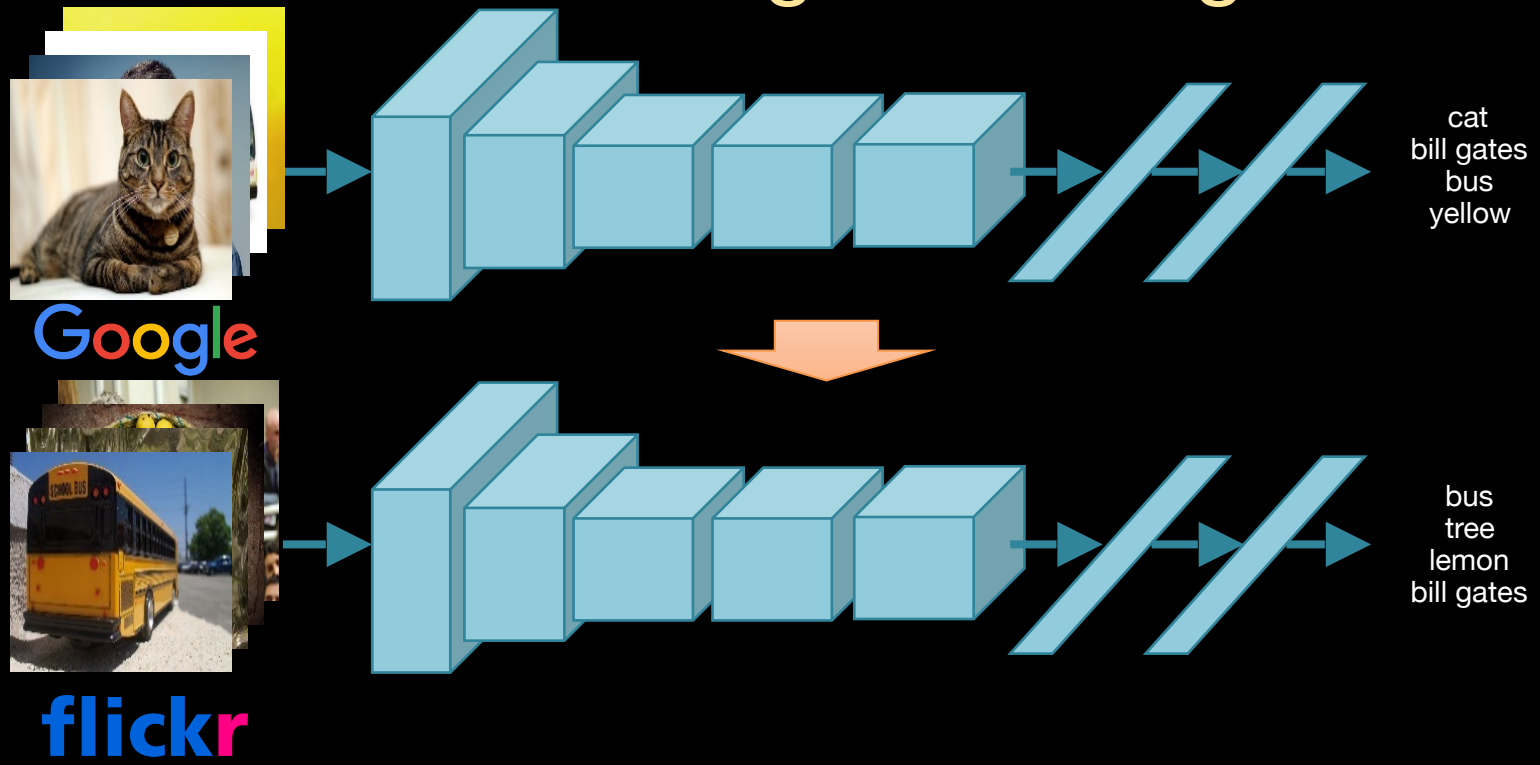
1.5M

ImageNet	GFAll	Google
44.7	40.5	42.7

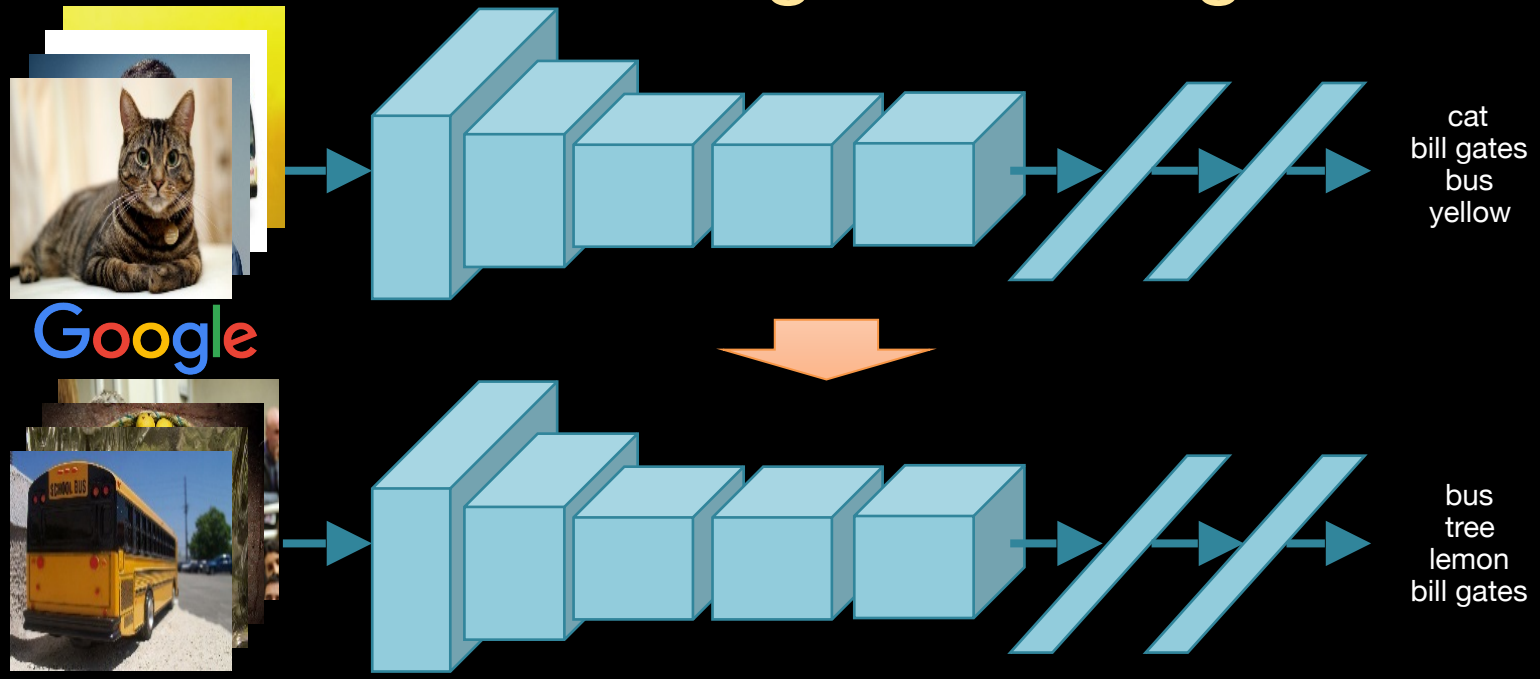
Trial 4: Staged Training



Trial 4: Staged Training

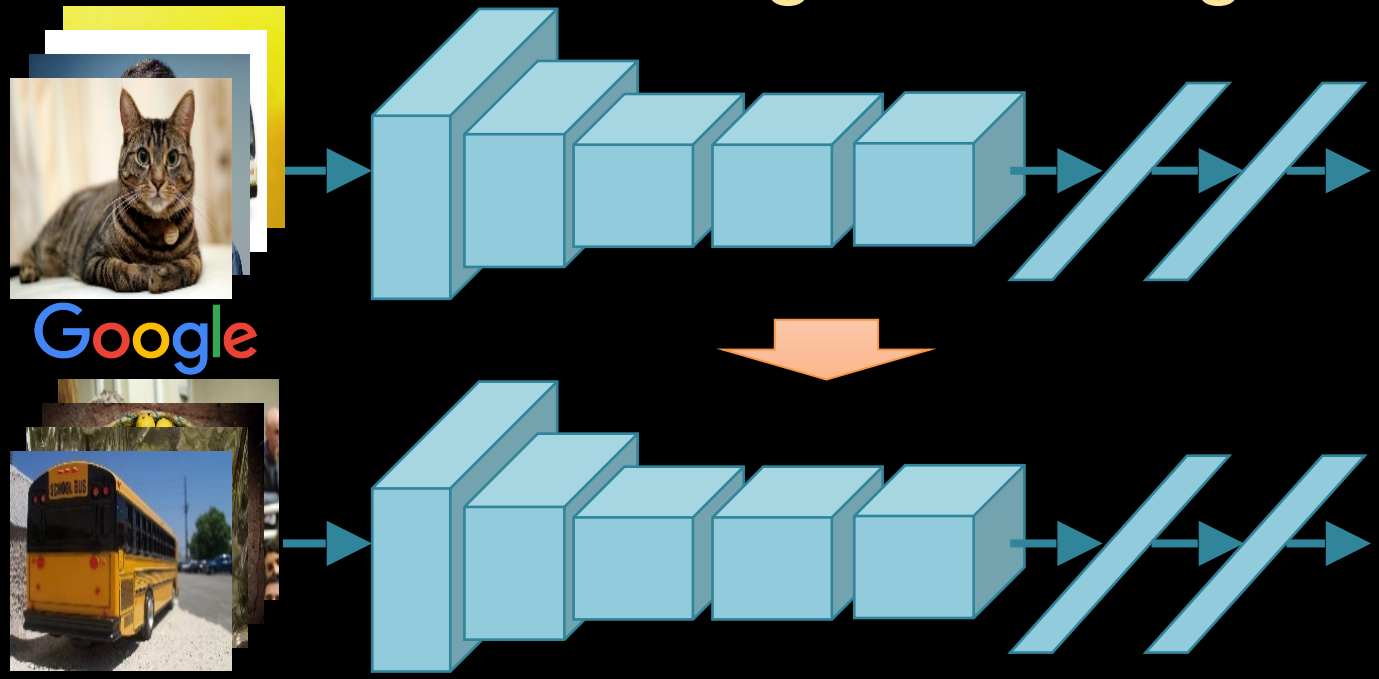


Trial 4: Staged Training



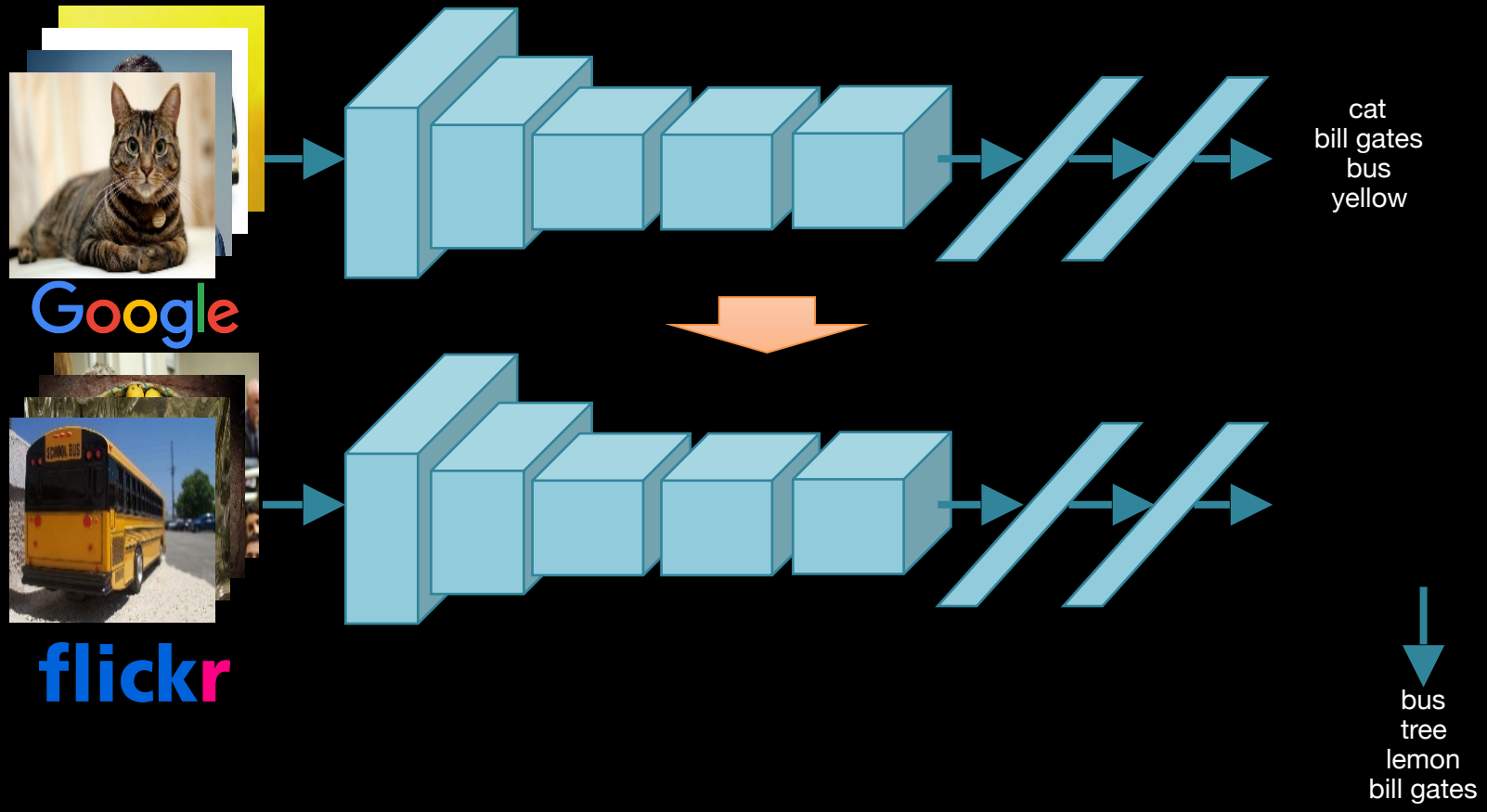
ImageNet	Google	FineTune
44.7	42.7	43.4

Trial 4: Staged Training

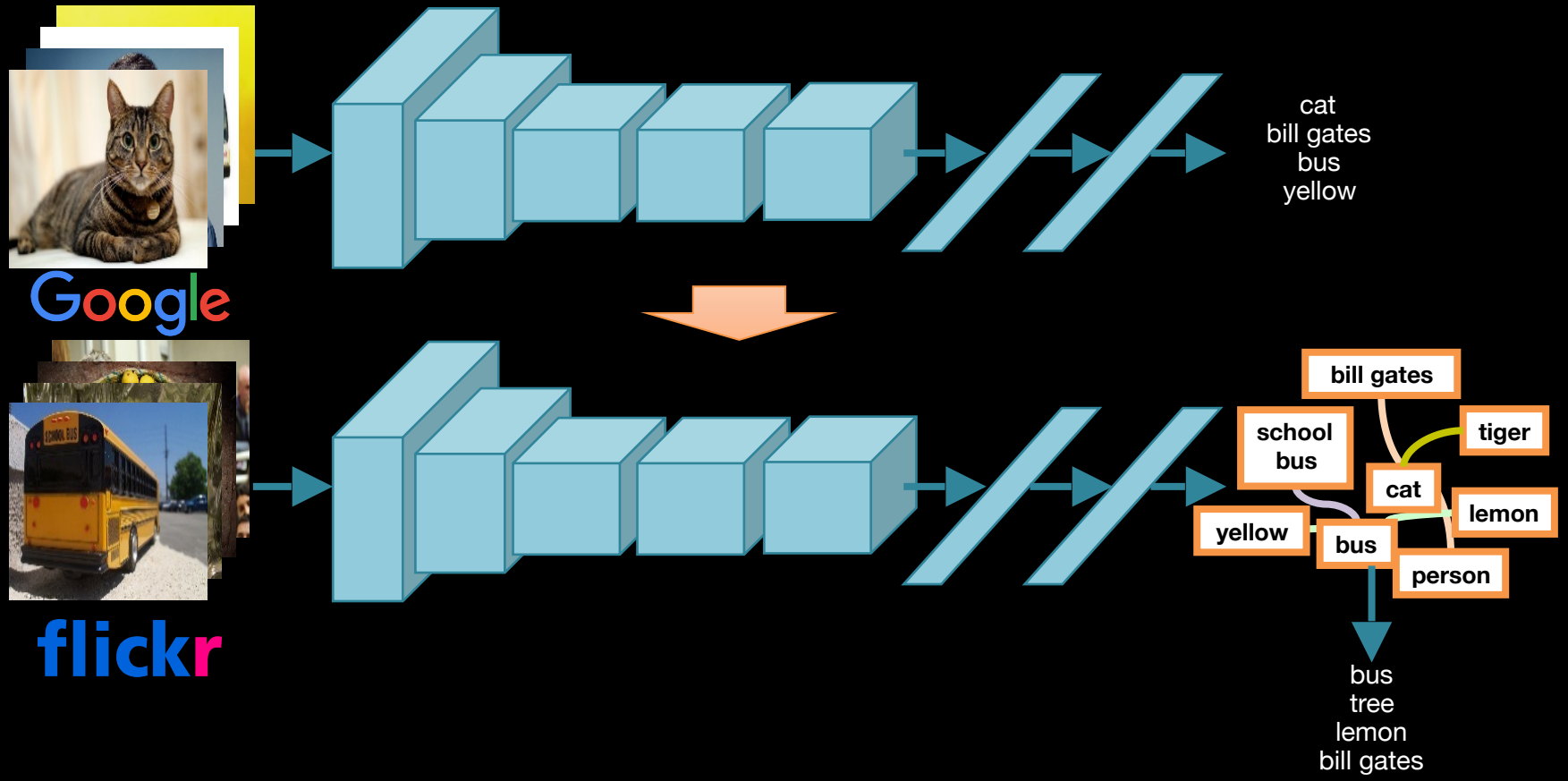


ImageNet	Google	FineTune
44.7	42.7	43.4

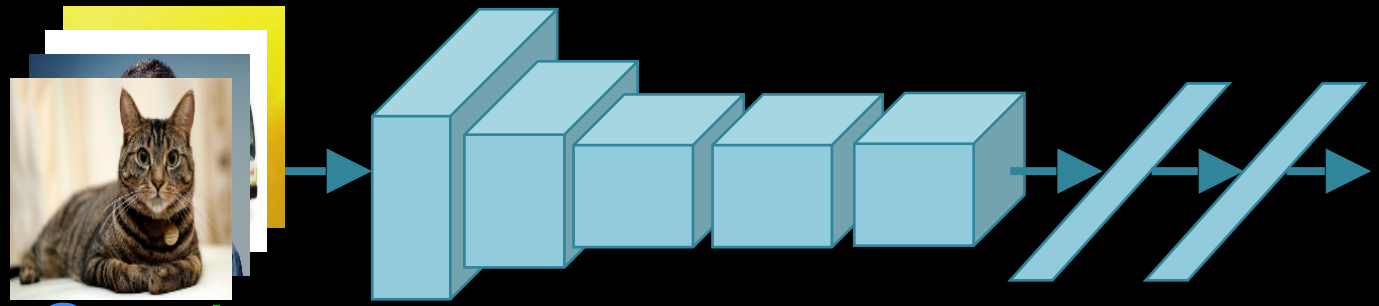
Final Approach: Staged + Graph

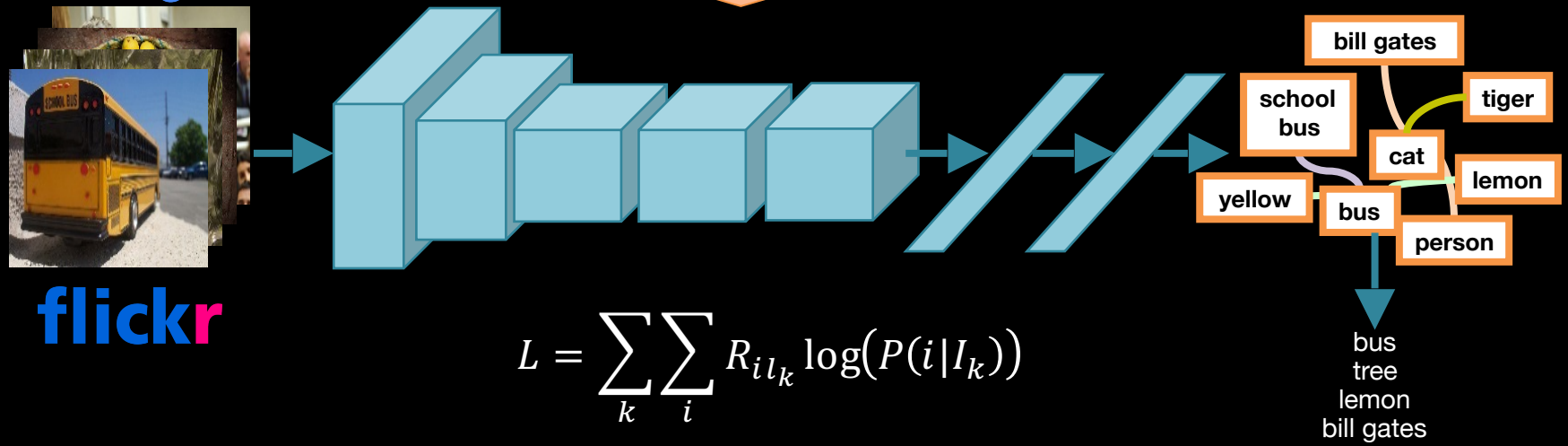


Final Approach: Staged + Graph



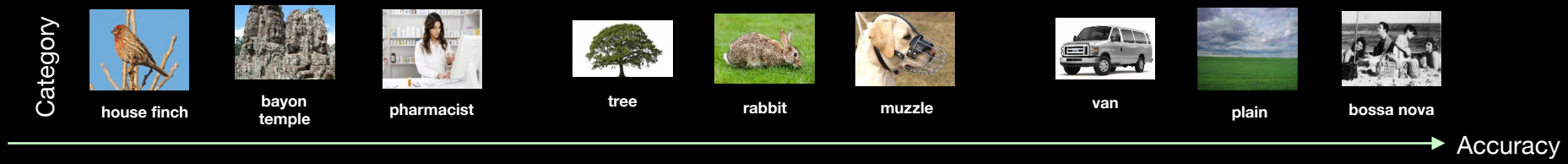
Final Approach: Staged + Graph





$$L = \sum_k \sum_i R_{il_k} \log\left(P(i|I_k)\right)$$

Graph: Confusion Matrix



Graph: Confusion Matrix

Category


house finch


bayon temple


pharmacist


tree


rabbit


muzzle

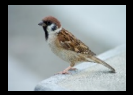

van


plain


bossa nova

Accuracy

Similar/Confusing Categories


sparrow


angkor


lab coat


banyan


hare


malinois


camionnette


open area


guitar


indigo bunting


obelisk


doctor

●●●


buckeye


wood rabbit


german shepherd

●●●


club wagon


rapeseed

downbeat


baya weaver


stupa


tobacco shop


natural


angora


bull mastiff


minibus


valley


ukulele


goldfinch


megalith


stethoscope


tree stump


wallaby

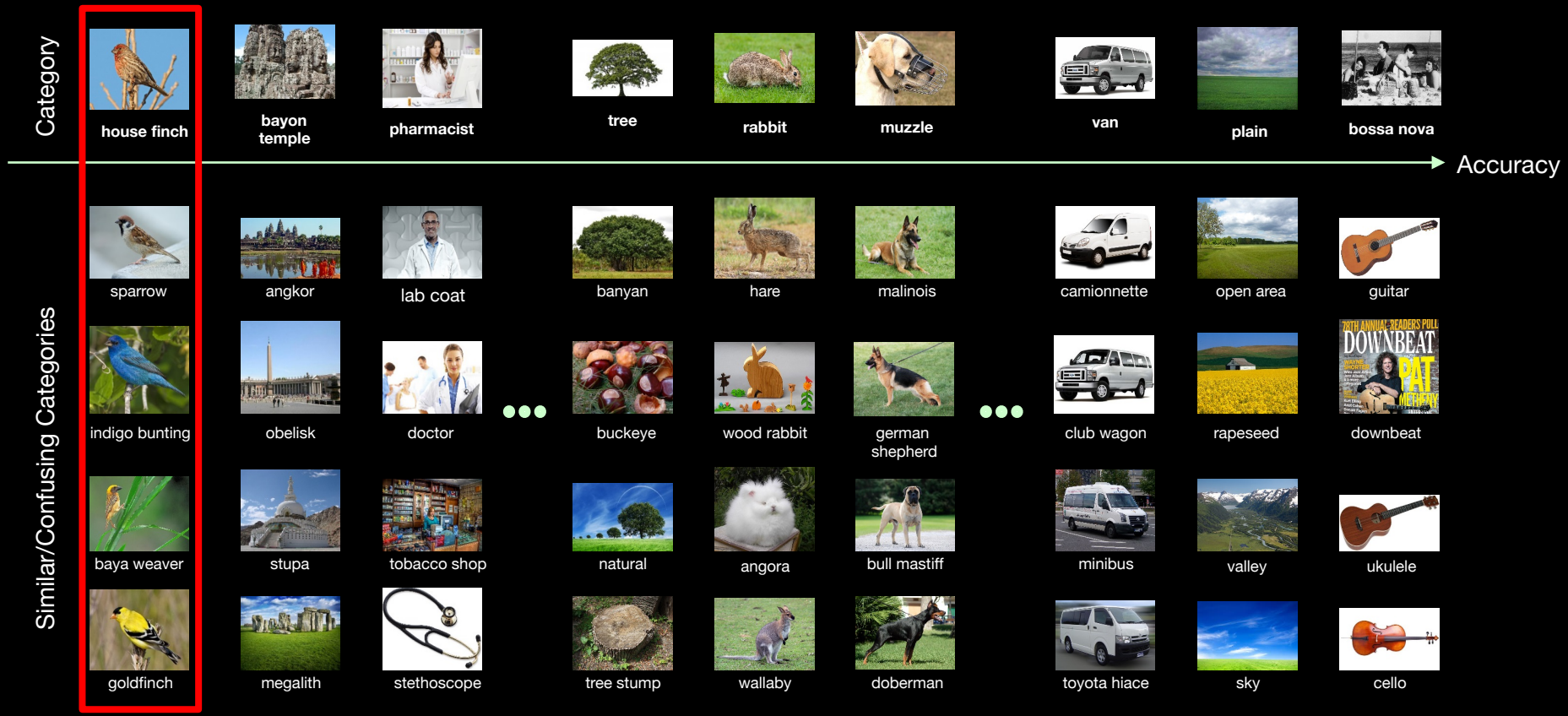

doberman


toyota hiace

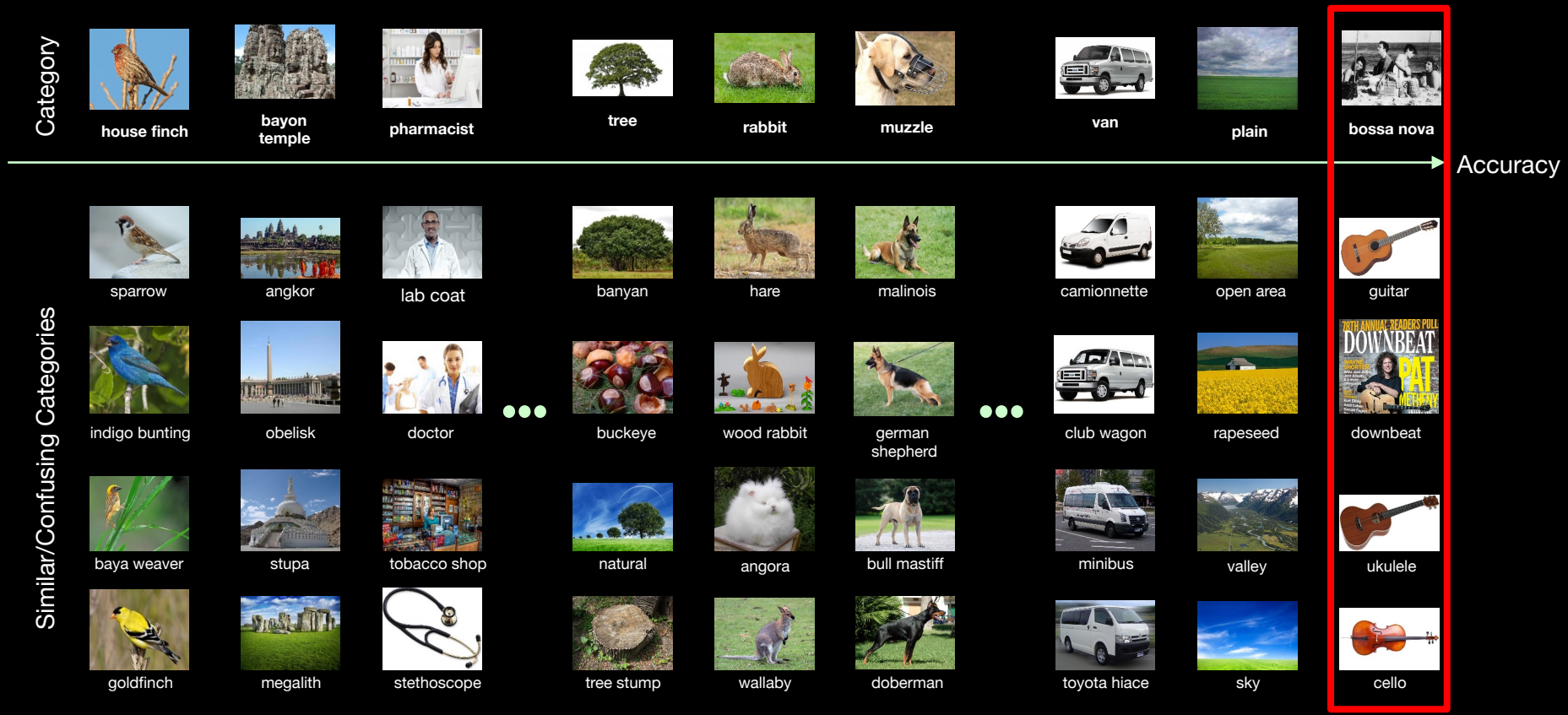

sky


cello

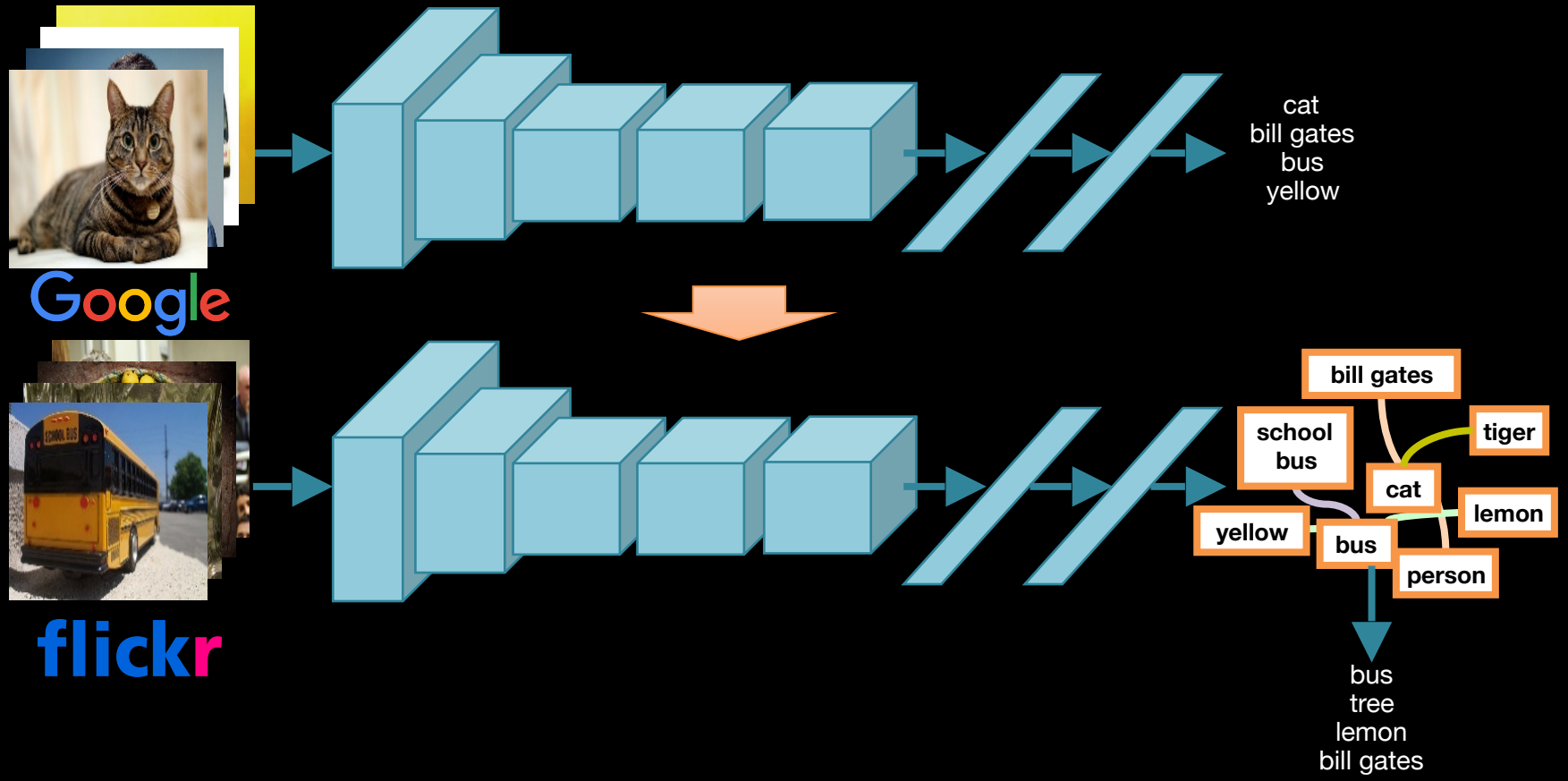
Graph: Confusion Matrix



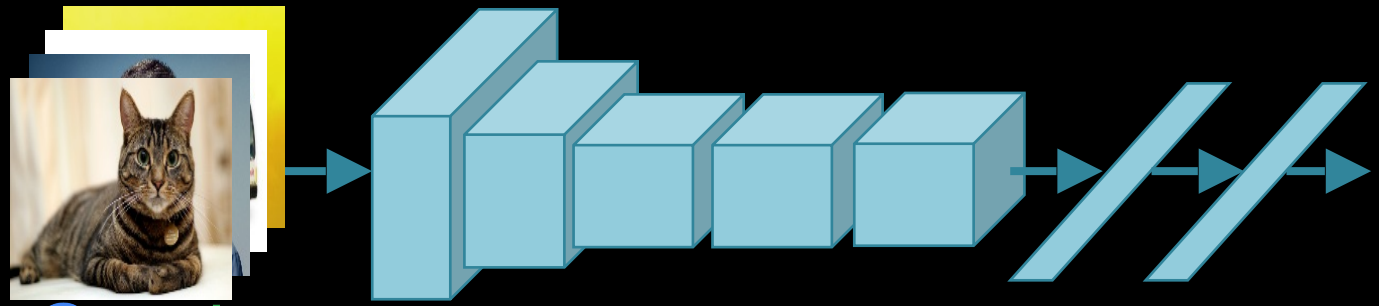
Graph: Confusion Matrix

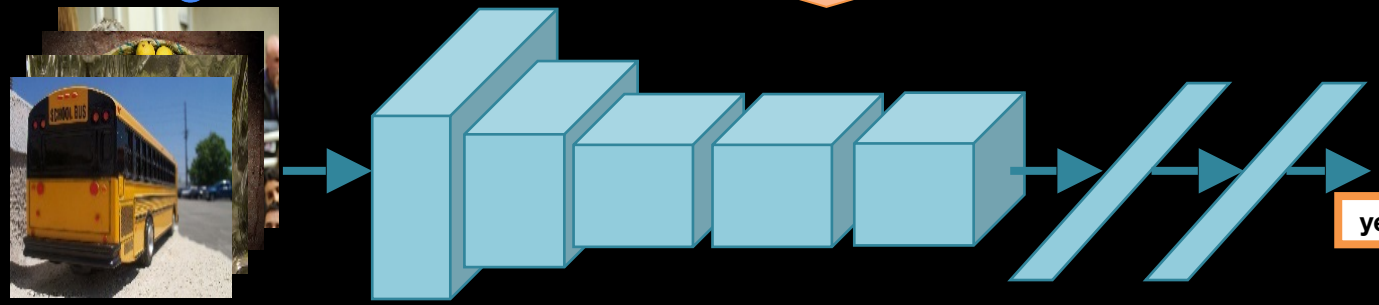


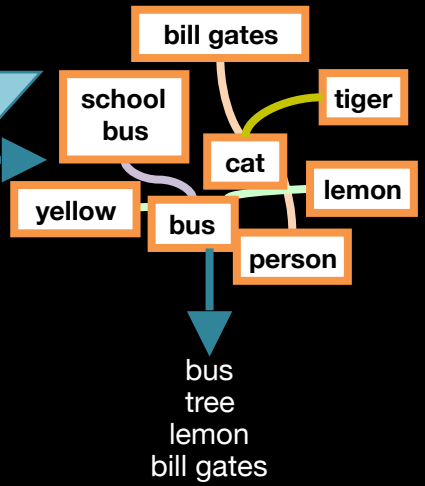
Final Approach: Staged + Graph



Final Approach: Staged + Graph





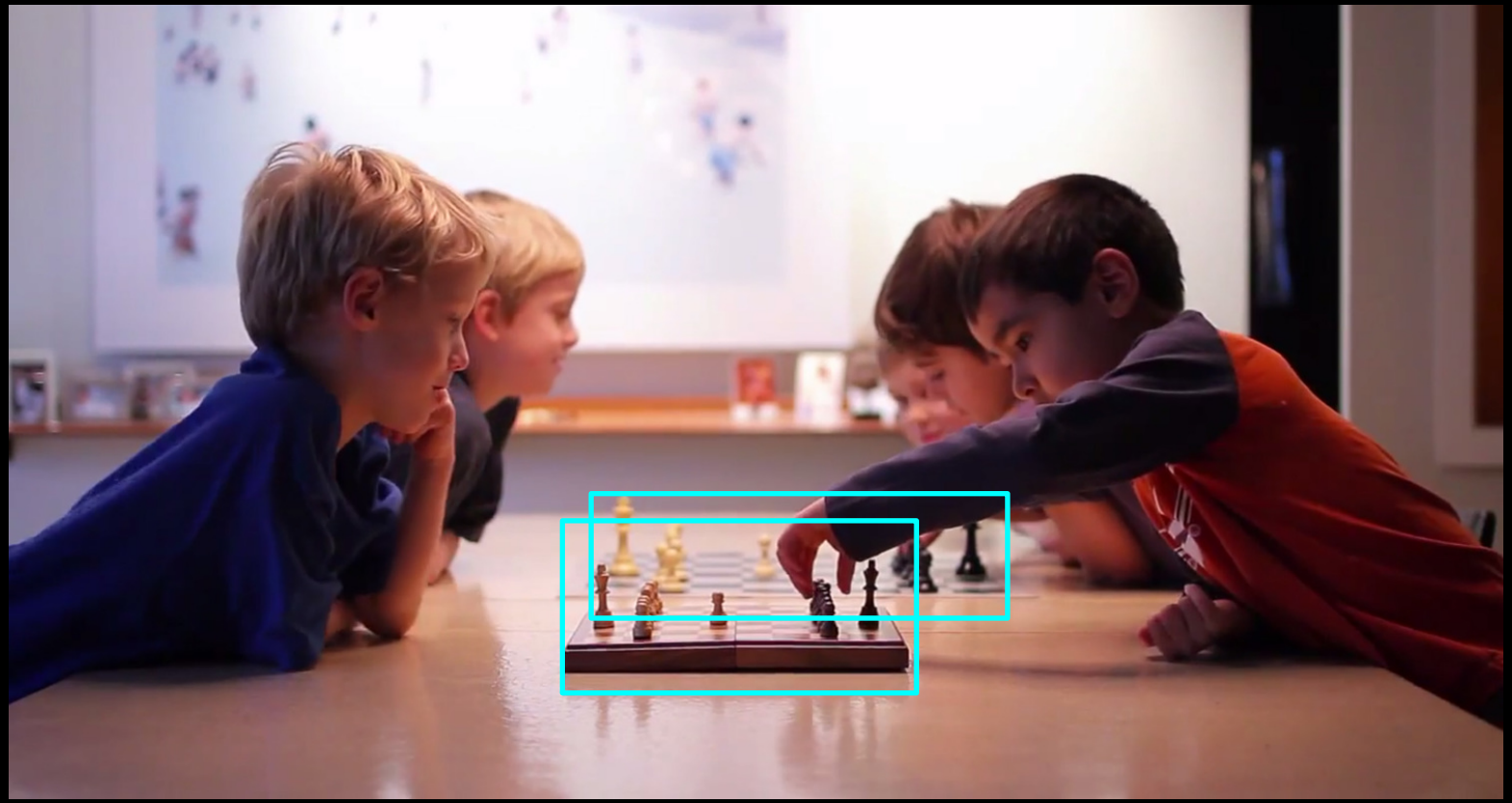


ImageNet	FineTune	Graph
44.7	43.4	44.7

Location Boxes from the Web

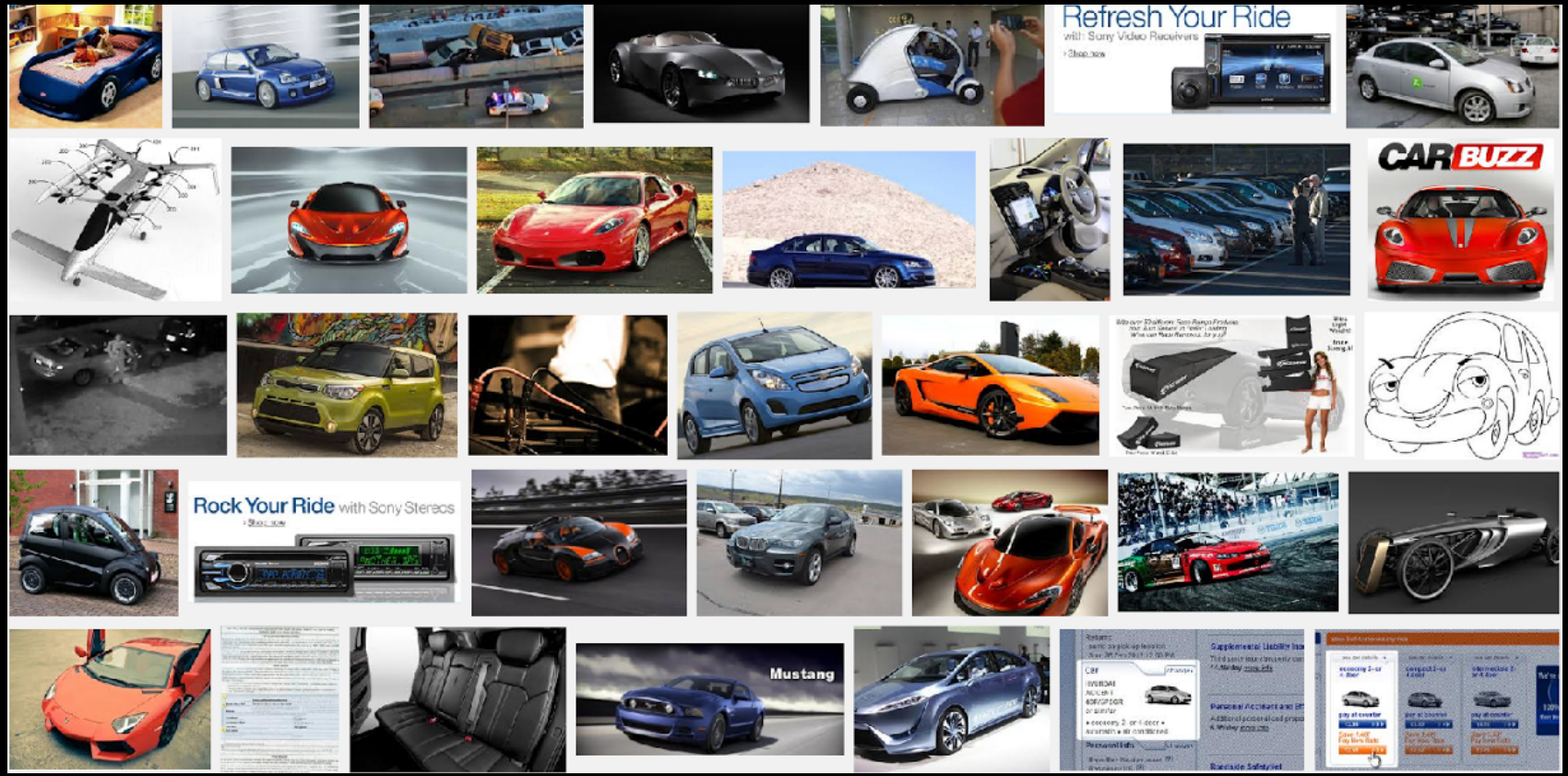
PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

Web Images: No Location



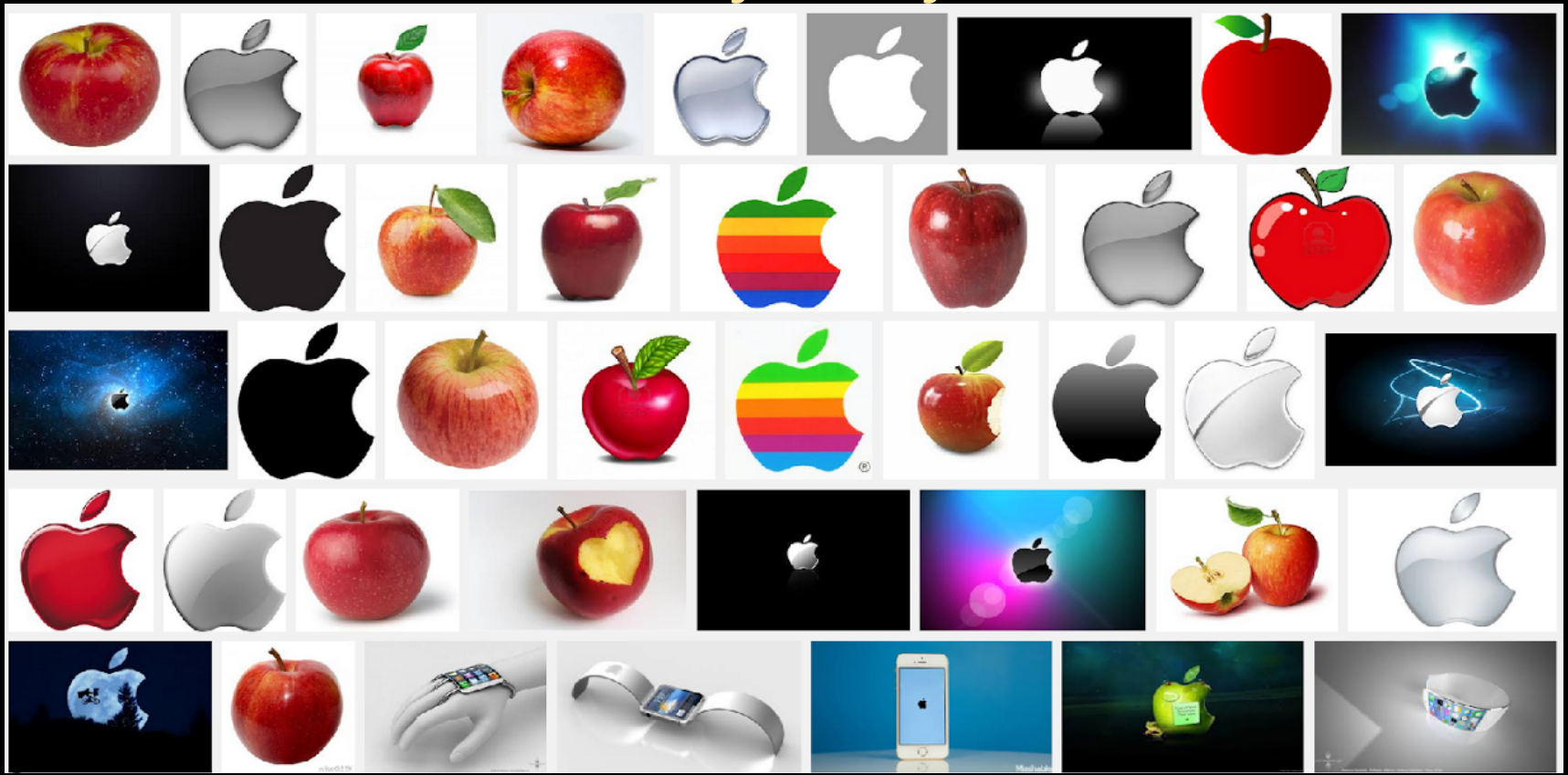
car

Noise

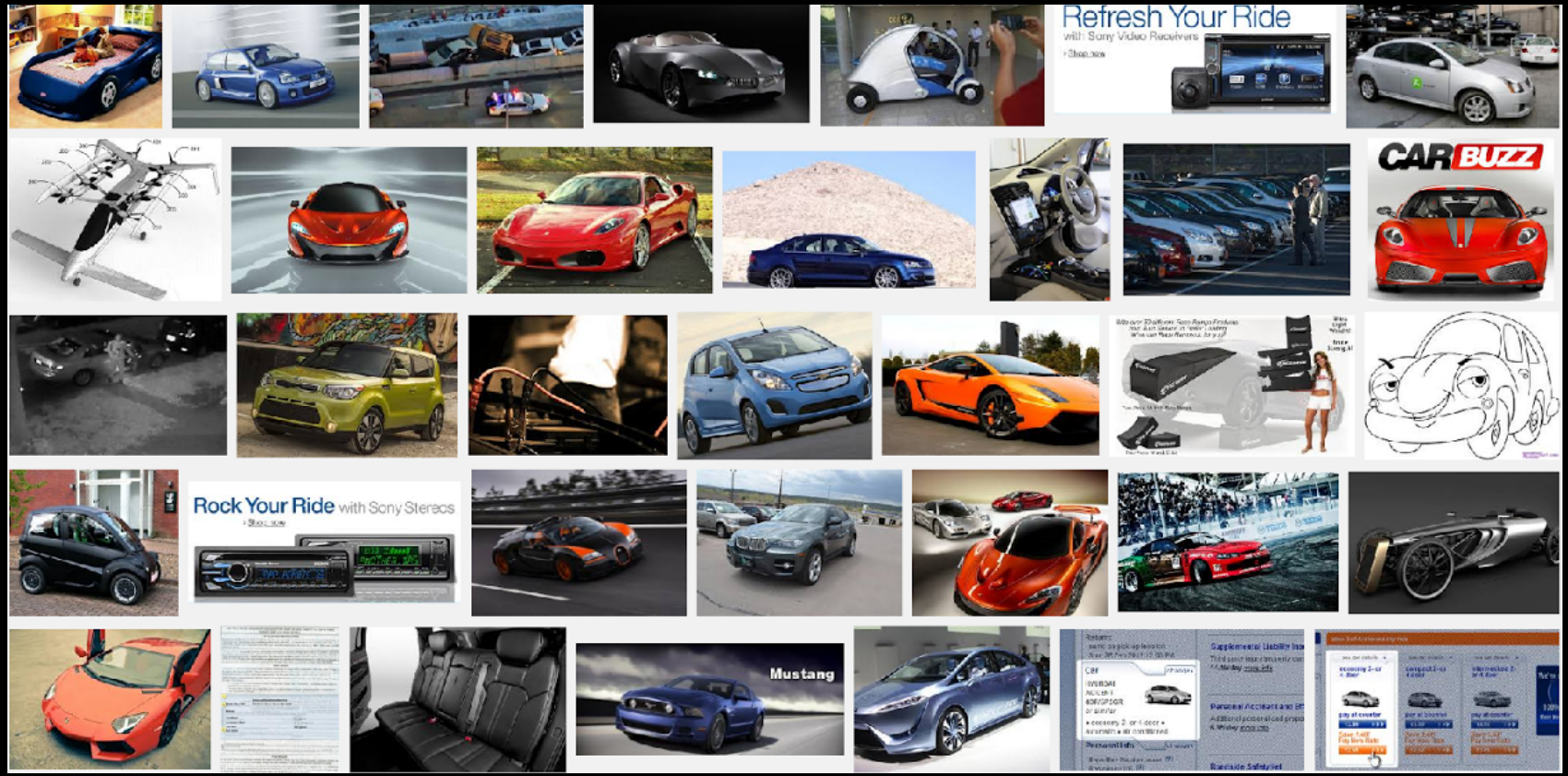

Refresh Your Ride
with Sony Video Receivers
CAR BUZZ
Rock Your Ride with Sony Stereos
Mustang

apple

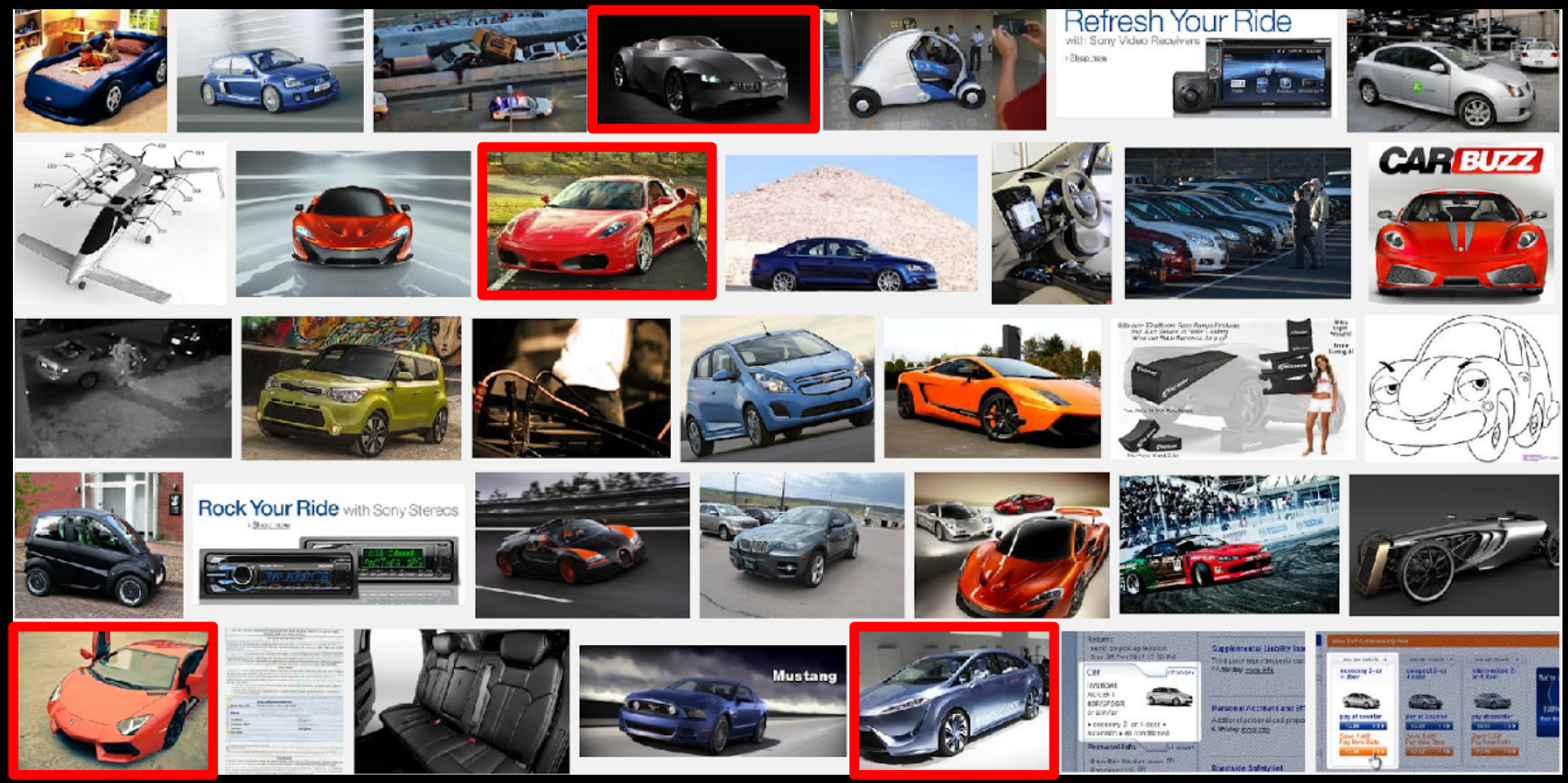
Polysemy



Subcategory Discovery



Subcategory Discovery



Exemplar Detectors

car

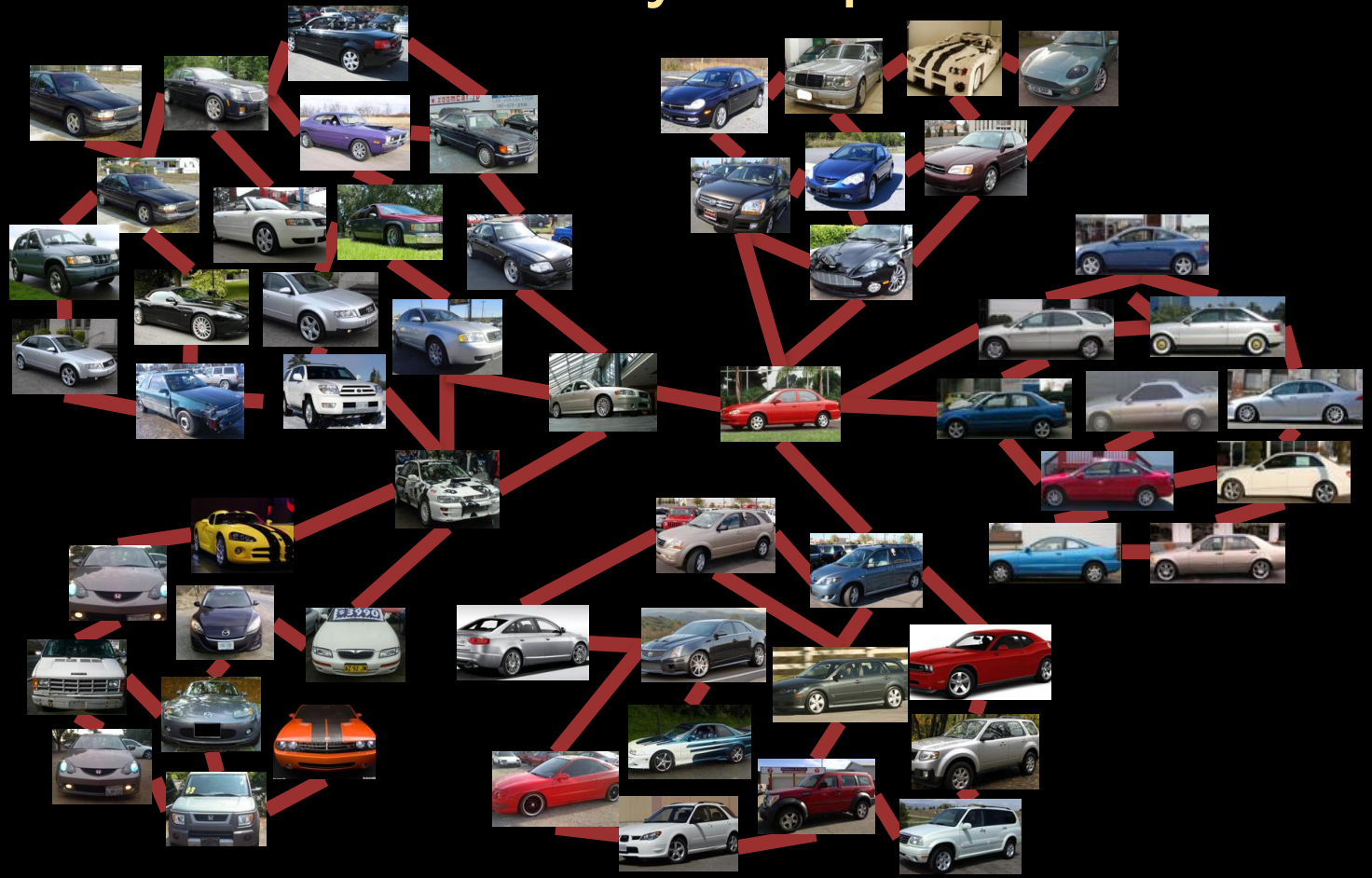
Exemplar Detectors

car

Exemplar Detectors

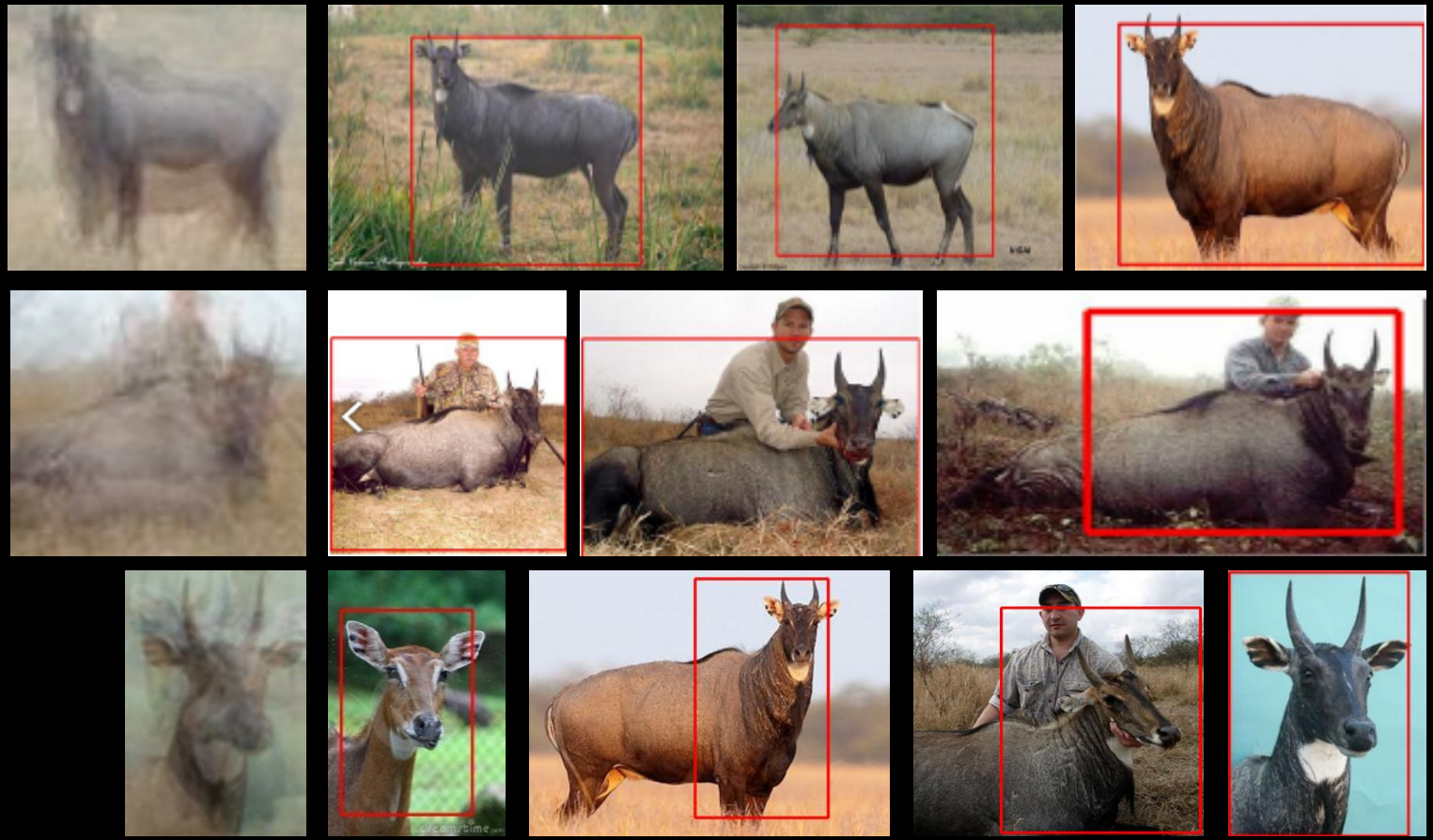
car

Affinity Graph



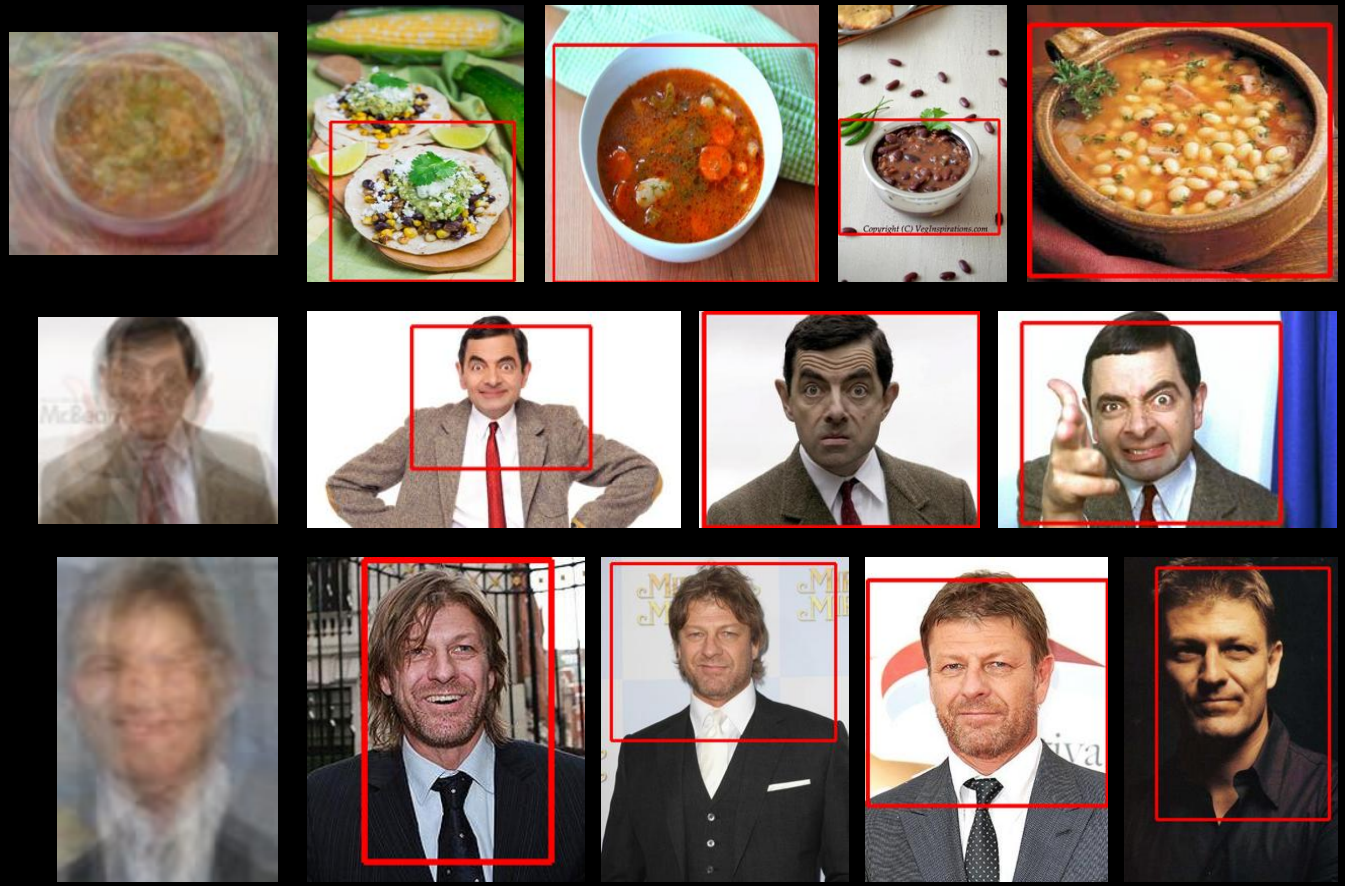
nilgai

Subcategories (HOG)



bean

Subcategories (Polysemy)

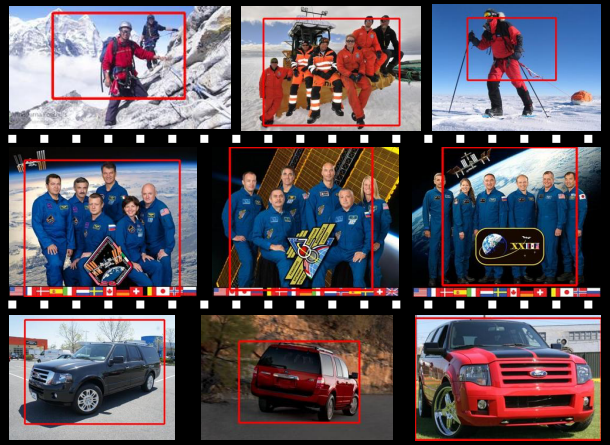


Subcategories (HOG)

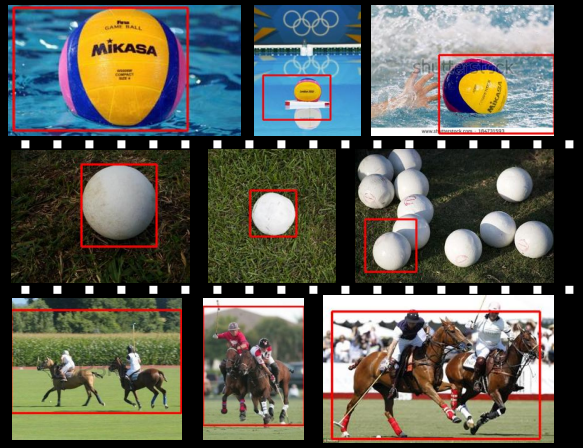
Subcategories (web fc7)

alligator lizard

expedition



polo ball



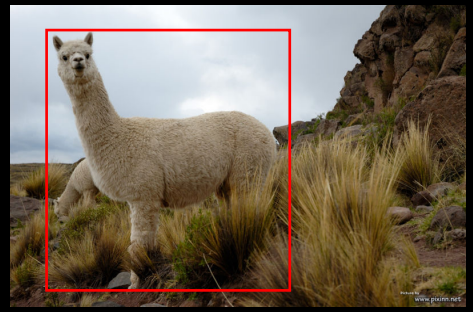
lychee

R-CNN AlexNet from Web on VOC 2007

(Zitnick & Dollar, 2014) (Felzenszwalb et al., 2010)

R-CNN AlexNet from Web on VOC 2007

alpaca

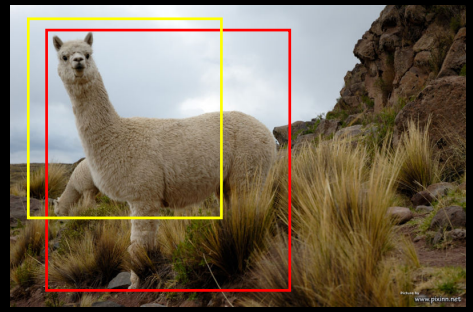


- Basic, (FlickrG)

(Zitnick & Dollar, 2014) (Felzenszwalb et al., 2010)

R-CNN AlexNet from Web on VOC 2007

alpaca

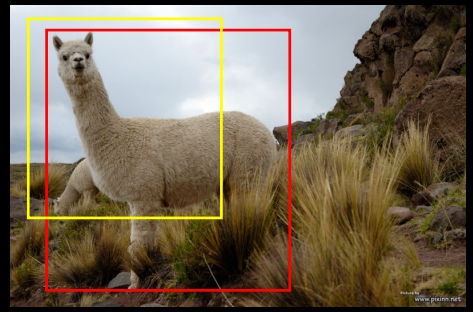


- Basic, (FlickrG)
- with More Data (MD)

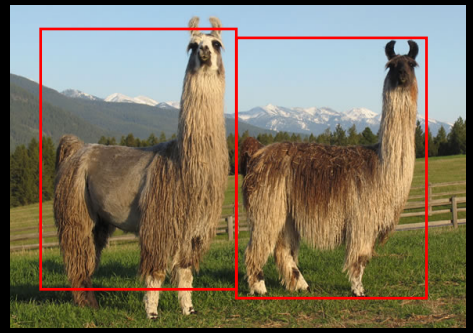
(Zitnick & Dollar, 2014) (Felzenszwalb et al., 2010)

R-CNN AlexNet from Web on VOC 2007

alpaca



llama

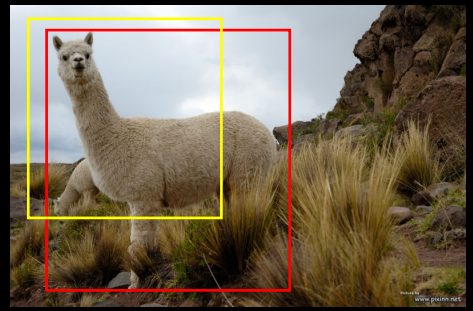


- Basic, (FlickrG)
- with More Data (MD)
- with More related Categories (MC)

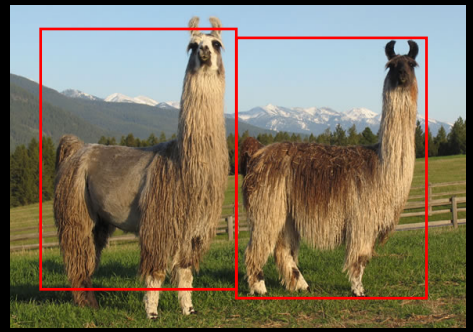
(Zitnick & Dollar, 2014) (Felzenszwalb et al., 2010)

R-CNN AlexNet from Web on VOC 2007

alpaca



llama



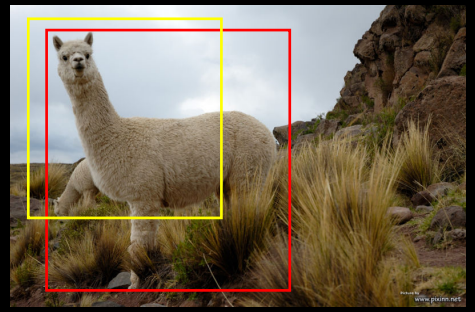
- Basic, (FlickrG)
- with More Data (MD)
- with More related Categories (MC)

FlickrG	MD	MC
22.9	23.0	**24.4**

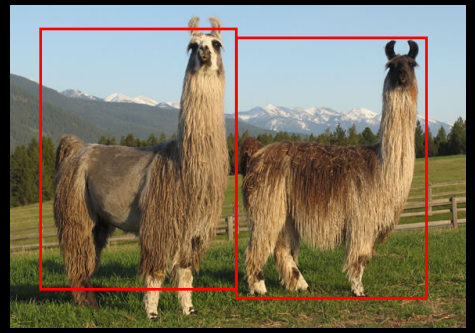
(Zitnick & Dollar, 2014) (Felzenszwalb et al., 2010)

R-CNN AlexNet from Web on VOC 2007

alpaca

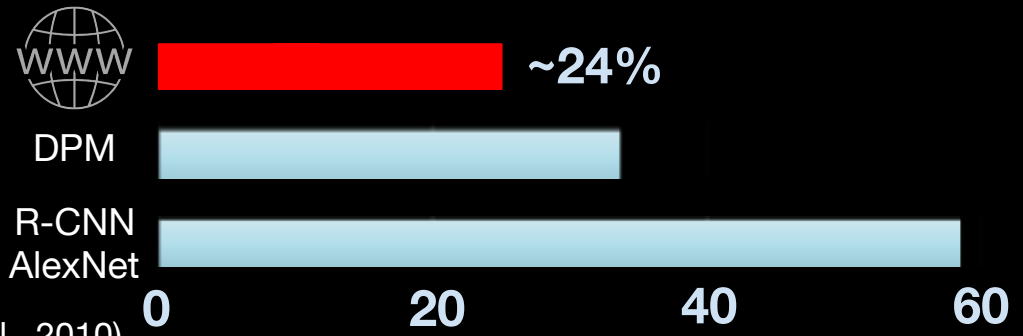


llama



- Basic, (FlickrG)
- with More Data (MD)
- with More related Categories (MC)

FlickrG	MD	MC
22.9	23.0	**24.4**

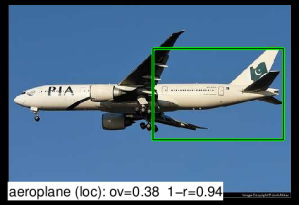


(Zitnick & Dollar, 2014) (Felzenszwalb et al., 2010)

False Positives (MC)

plane


aeroplane (loc): ov=0.44 1−r=0.99



aeroplane (loc): ov=0.38 1−r=0.94

aeroplane (loc): ov=0.32 1−r=0.93


aeroplane (loc): ov=0.28 1−r=0.93


aeroplane (loc): ov=0.47 1−r=0.93

bottle

bottle (loc): ov=0.29 1−r=0.99

bottle (loc): ov=0.30 1−r=0.99


bottle (loc): ov=0.32 1−r=0.99

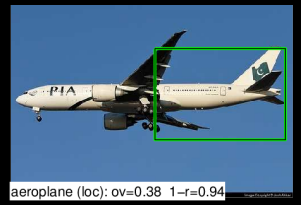

bottle (loc): ov=0.28 1−r=0.99

bottle (loc): ov=0.39 1−r=0.99

False Positives (MC)

plane


aeroplane (loc): ov=0.44 1−r=0.99



aeroplane (loc): ov=0.38 1−r=0.94

aeroplane (loc): ov=0.32 1−r=0.93


aeroplane (loc): ov=0.28 1−r=0.93


aeroplane (loc): ov=0.47 1−r=0.93

bottle

bottle (loc): ov=0.29 1−r=0.99

bottle (loc): ov=0.30 1−r=0.99

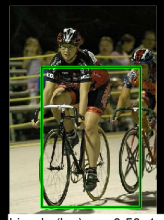

bottle (loc): ov=0.32 1−r=0.99

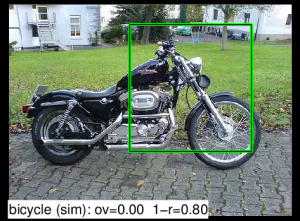

bottle (loc): ov=0.28 1−r=0.99

bottle (loc): ov=0.39 1−r=0.99

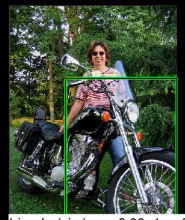
bike


bicycle (loc): ov=0.41 1−r=0.85


bicycle (loc): ov=0.50 1−r=0.83


bicycle (sim): ov=0.00 1−r=0.80

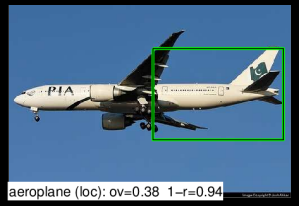

bicycle (loc): ov=0.38 1−r=0.79


bicycle (sim): ov=0.00 1−r=0.77

False Positives (MC)

plane


aeroplane (loc): ov=0.44 1−r=0.99



aeroplane (loc): ov=0.38 1−r=0.94

aeroplane (loc): ov=0.32 1−r=0.93


aeroplane (loc): ov=0.28 1−r=0.93


aeroplane (loc): ov=0.47 1−r=0.93

bottle

bottle (loc): ov=0.29 1−r=0.99

bottle (loc): ov=0.30 1−r=0.99

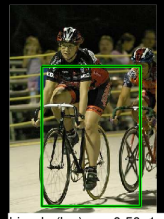

bottle (loc): ov=0.32 1−r=0.99

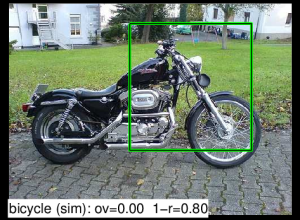

bottle (loc): ov=0.28 1−r=0.99

bottle (loc): ov=0.39 1−r=0.99

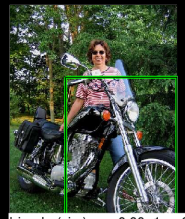
bike


bicycle (loc): ov=0.41 1−r=0.85

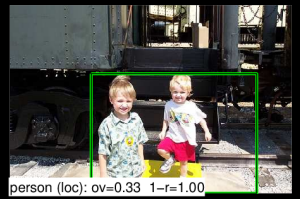

bicycle (loc): ov=0.50 1−r=0.83

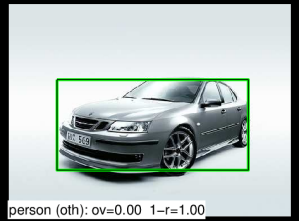

bicycle (sim): ov=0.00 1−r=0.80


bicycle (loc): ov=0.38 1−r=0.79


bicycle (sim): ov=0.00 1−r=0.77

person


person (loc): ov=0.33 1−r=1.00

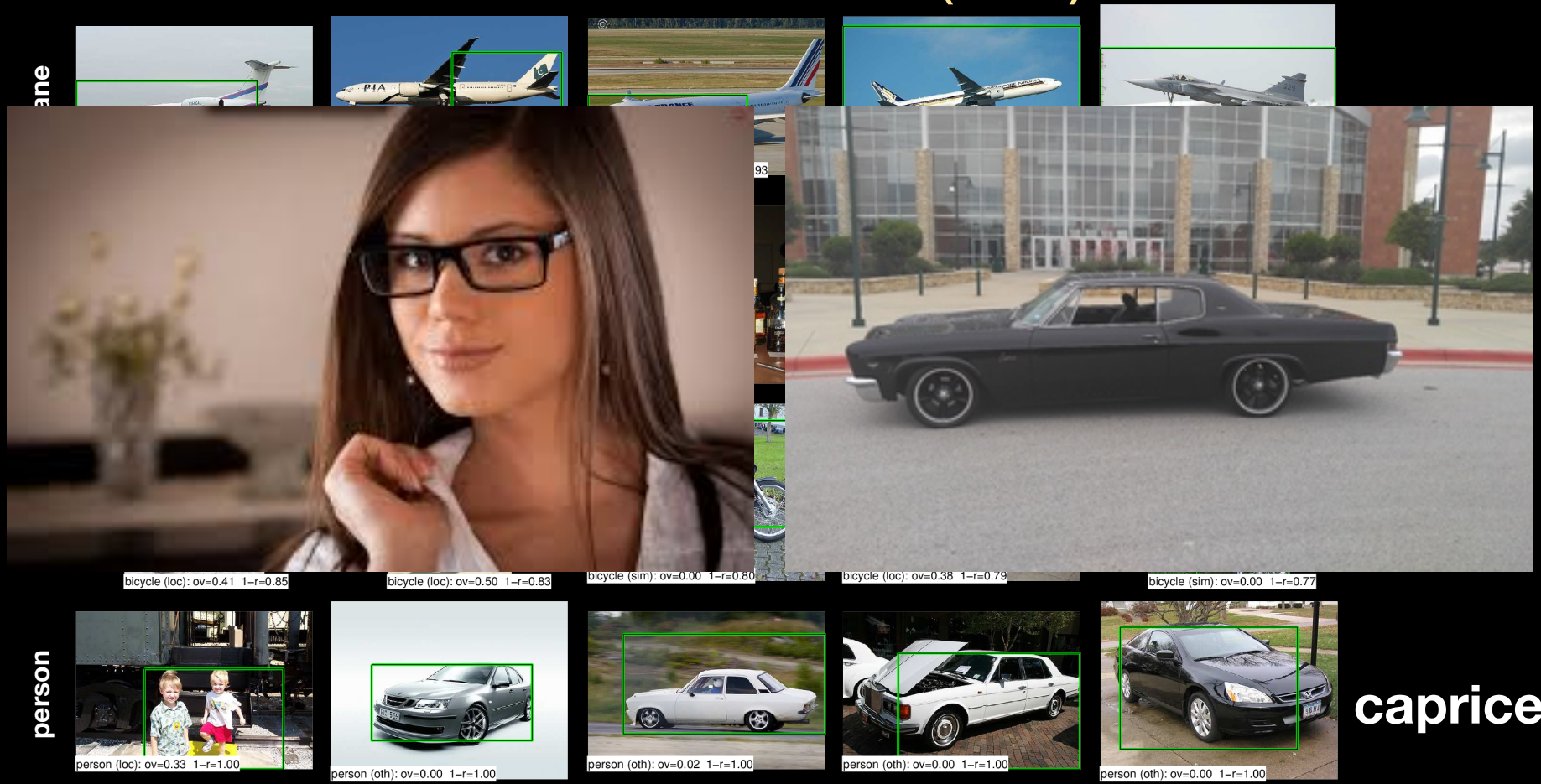

person (oth): ov=0.00 1−r=1.00

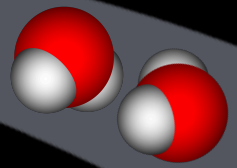

person (oth): ov=0.02 1−r=1.00


person (oth): ov=0.00 1−r=1.00


person (oth): ov=0.00 1−r=1.00

False Positives (MC)





- **Never Ending Image Learner [ICCV 2013]**
- Spatial Memory Network [ICCV 2017]

(II) Build Relationships

1. Relationships Help Single Image Understanding

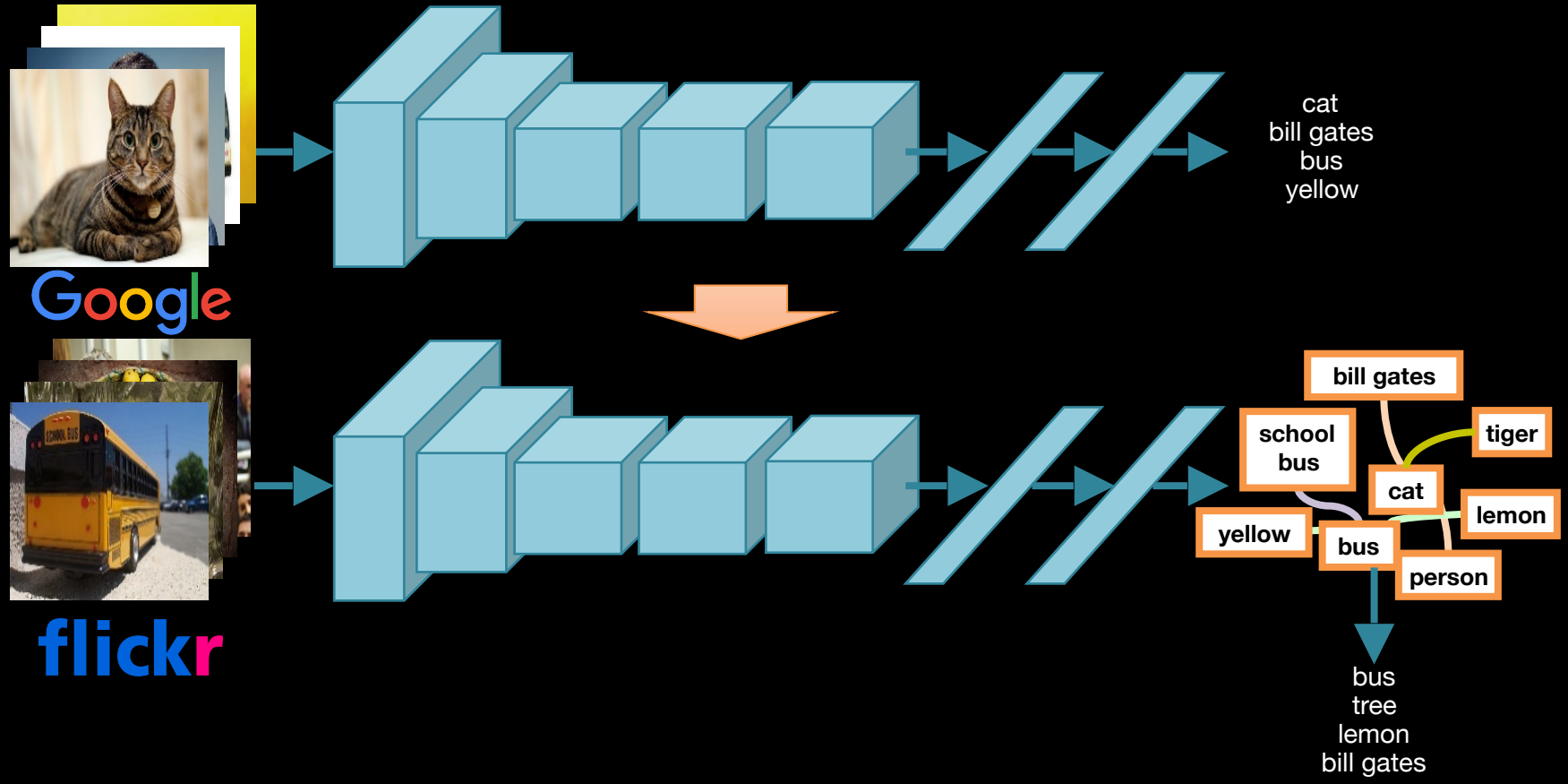
1. Relationships Help Single Image Understanding

1. Relationships Help Single Image Understanding

computer

monitor
is part of
computer

2. Relationships Help Learning Process



How to Acquire Relationships?

How to Acquire Relationships?

Cycorp®

Home of smarter solutions

How to Acquire Relationships?

Cycorp®
Home of smarter solutions

~7M rules
~30 years

tree is a **plant**
London is capital of **UK**

How to Acquire Relationships?

tree is a **plant**
London is capital of **UK**

~7M rules
~30 years

How many exist?
How many are needed?

(Li & Fei-Fei, 2010) (Chen et al., 2013) (Divvala et al., 2014)

NEIL

@2013

Never Ending Image Learner

Trying to understand images on the **web** and build a structured visual knowledge base automatically...

(Chen et al, 2013)

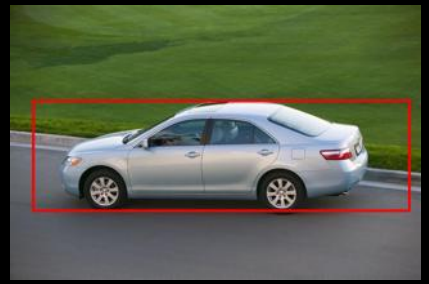
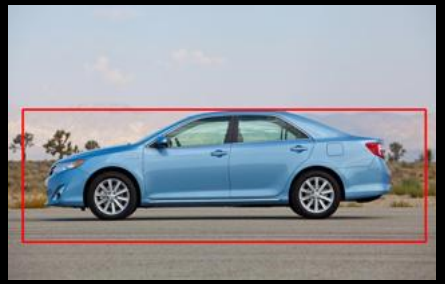
NEIL’s Knowledge Base

Concepts

Relationships

Objects

Camry







Scenes

parking lot




raceway




Attributes

round shape

crowded

Relationships

Object-Object

Taxonomy

or

Similarity

Corolla is a kind of/looks similar to **car**

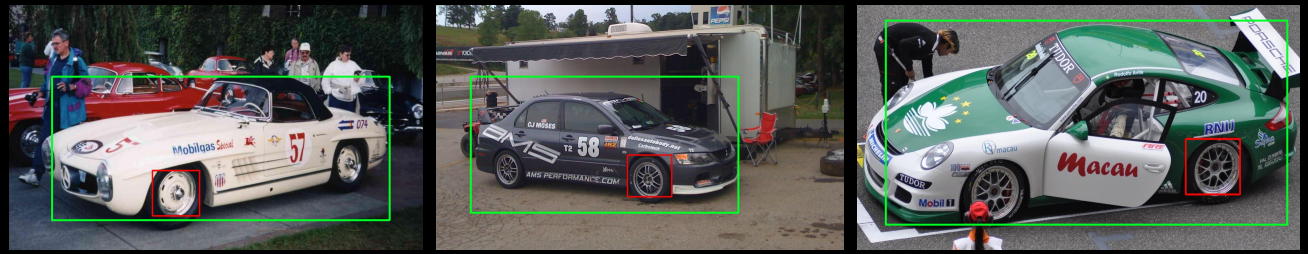
Relationships

Object-Object

Taxonomy or Similarity

Corolla is a kind of/looks similar to **car**

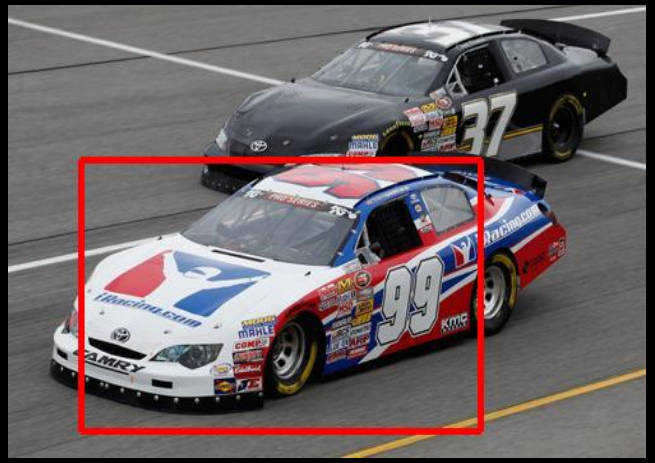
Partonomy

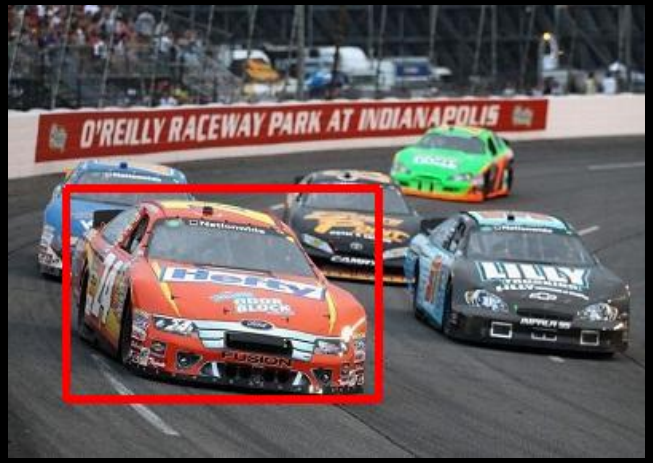


wheel is a part of **car**

Relationships

Object-Scene

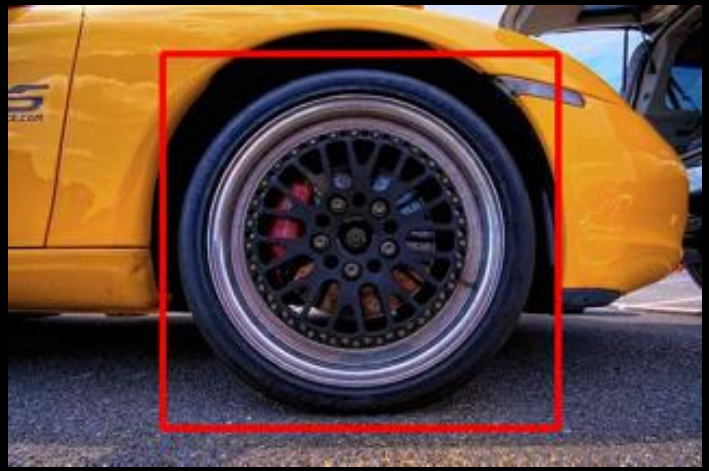




car is found in **raceway**

Relationships

Object-Attribute



wheel has **round shape**

Relationships

Scene-Attribute

trading floor is **crowded**

NEIL's Knowledge Base

Concepts

- **Objects**
- **Scenes**
- **Attributes**

Relationships

- **Object-Object**
  - Partonomy
  - Taxonomy/Similarity
- **Object-Scene**
- **Object-Attribute**
- **Scene-Attribute**

NEIL at Work:
Relationship Constrained Learning

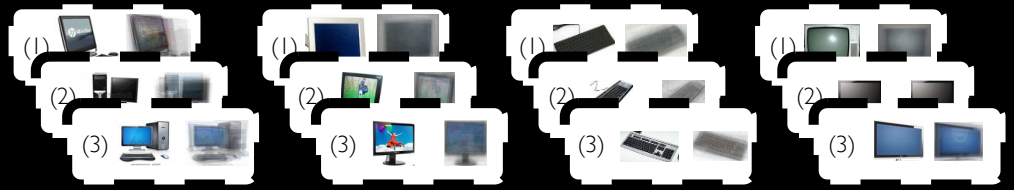
(0) Web Images

computer

monitor

keyboard

computer monitor keyboard television



(1) Subcategory Discovery

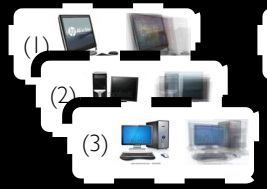
(0) Web Images

computer

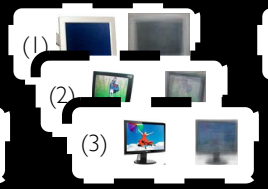
monitor

keyboard

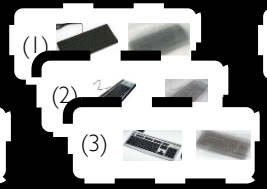
computer



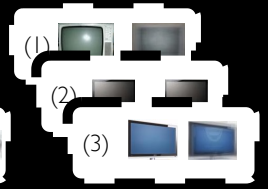
monitor



keyboard



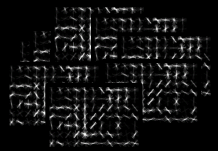
television



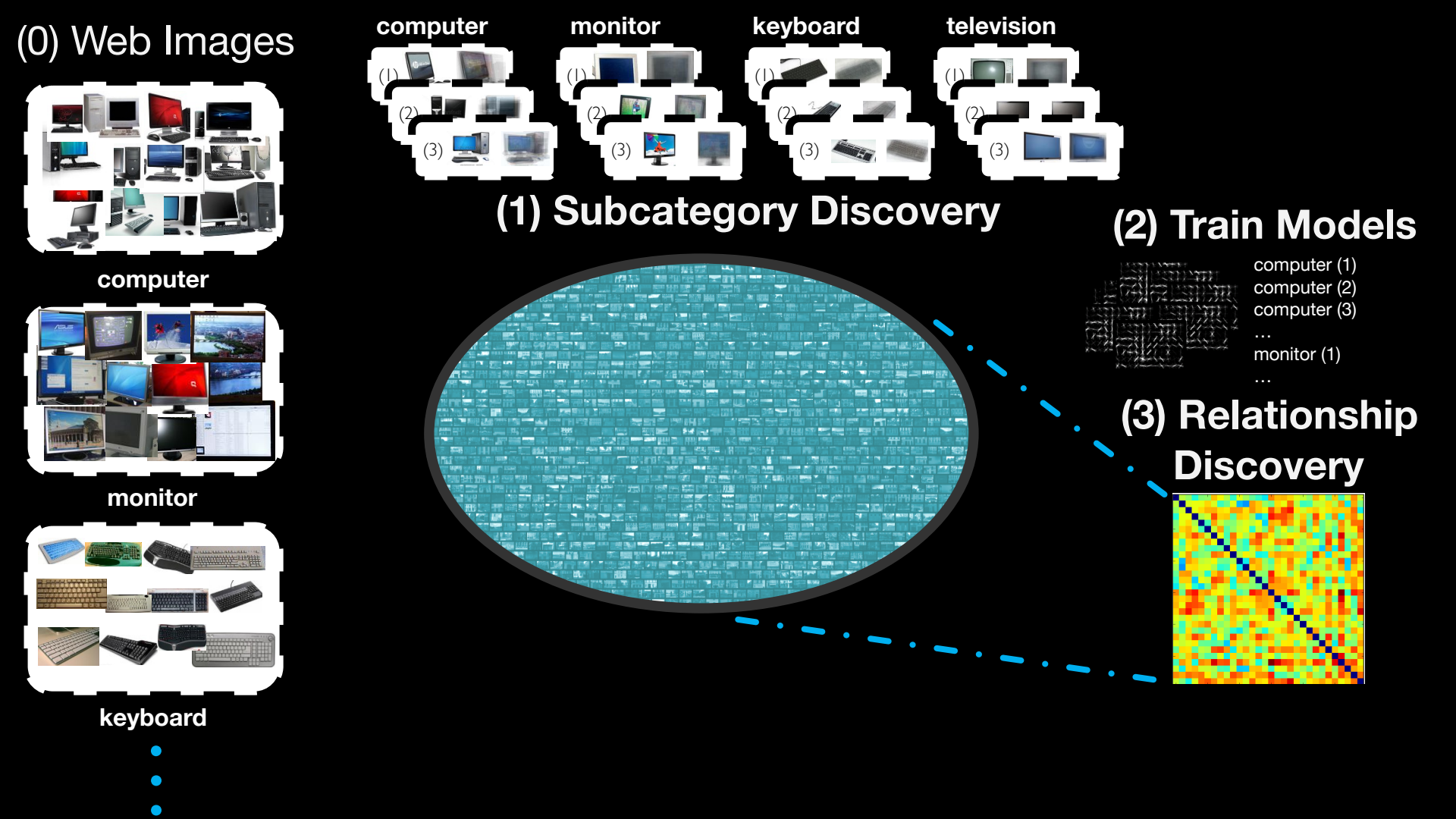
(1) Subcategory Discovery

@2013

(2) Train Models



computer (1)
computer (2)
computer (3)
…
monitor (1)
…


(0) Web Images
computer
monitor
keyboard
computer
monitor
keyboard
television
(1)
(2)
(3)
(1) Subcategory Discovery
(2) Train Models
computer (1)
computer (2)
computer (3)
...
monitor (1)
...
(3) Relationship Discovery

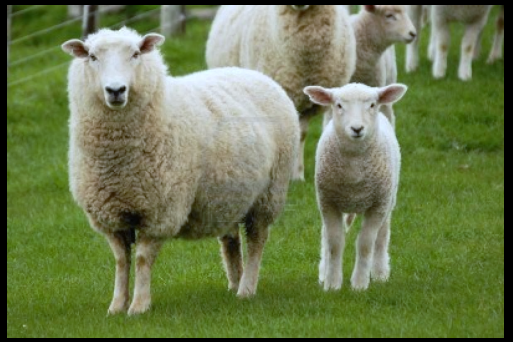
Micro-vision

Macro-vision

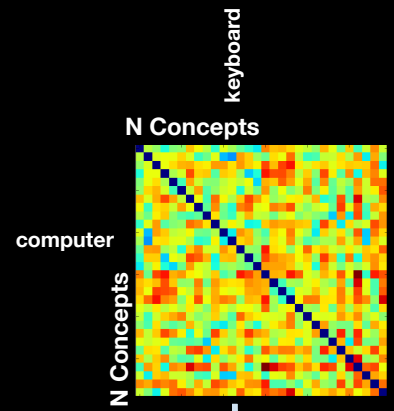
Structured Visual World



car is found on **road**

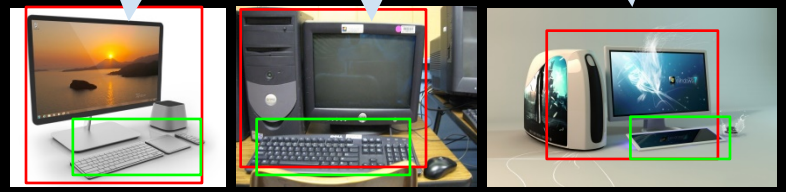



sheep is **white**



Learned relationships:

- **keyboard** is a part of **computer**
- **monitor** is a part of **computer**
- **television** looks similar to **monitor**



computer

keyboard

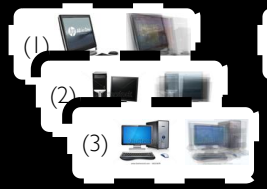
(0) Web Images

computer

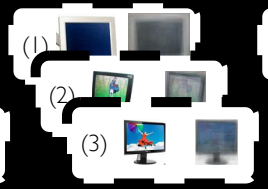
monitor

keyboard

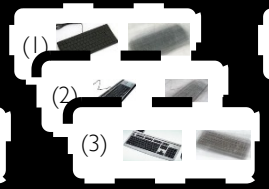
computer



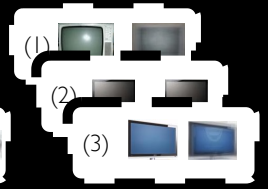
monitor



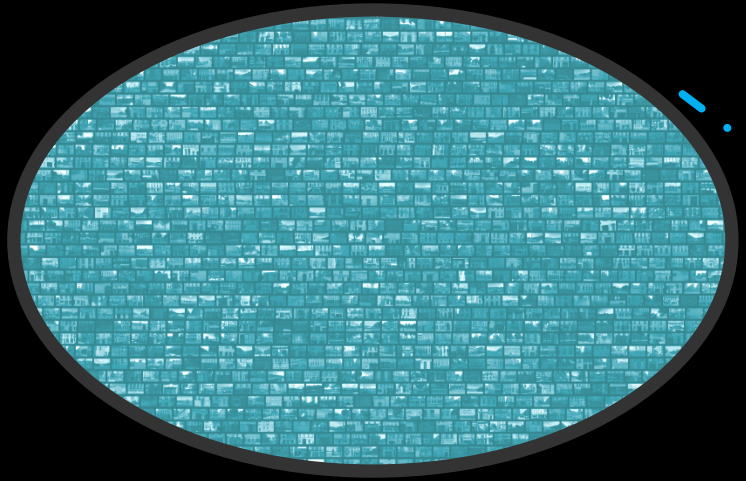
keyboard



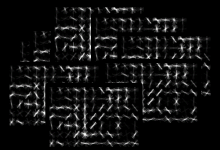
television



(1) Subcategory Discovery

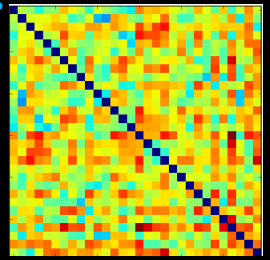


(2) Train Models

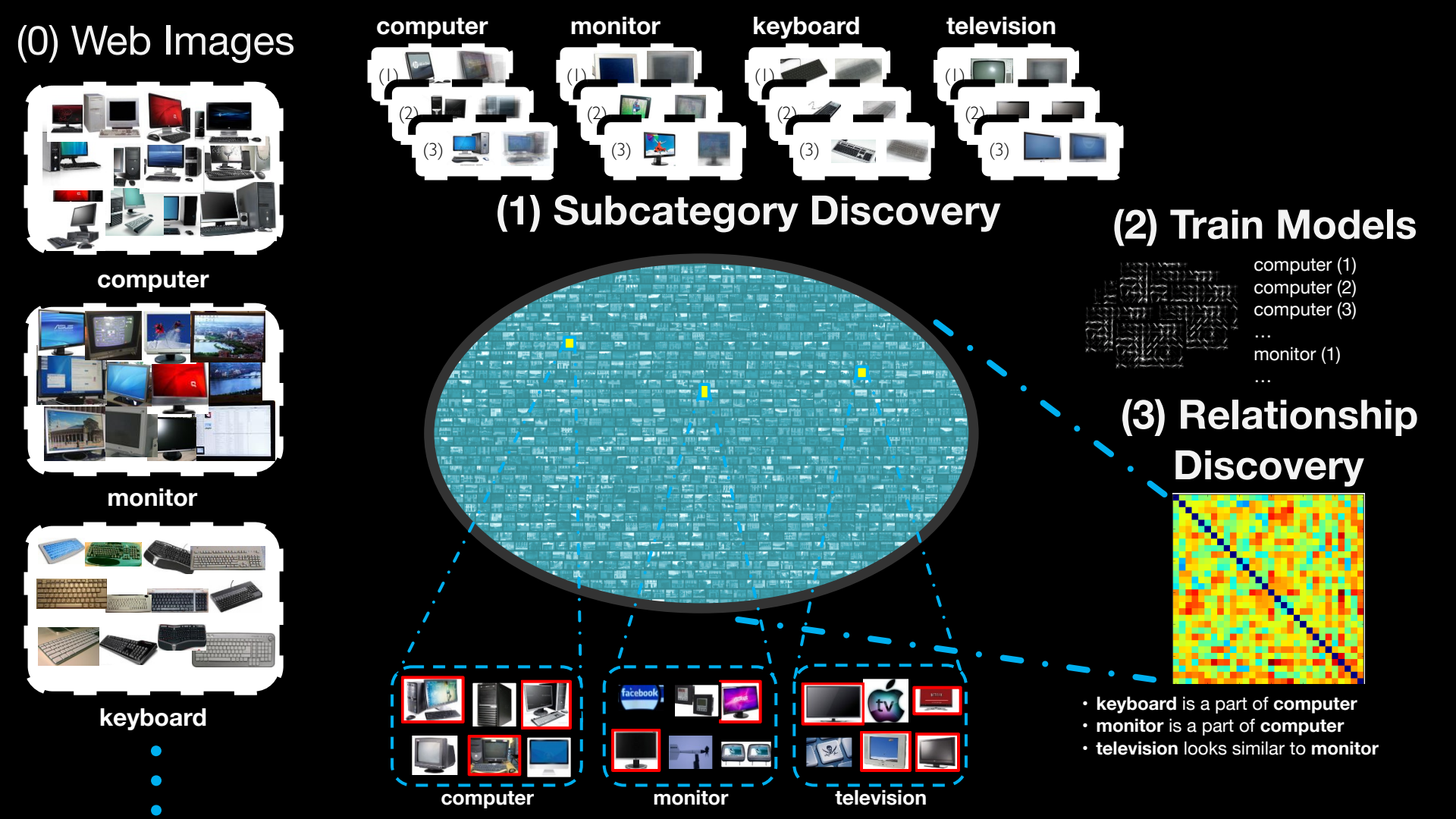


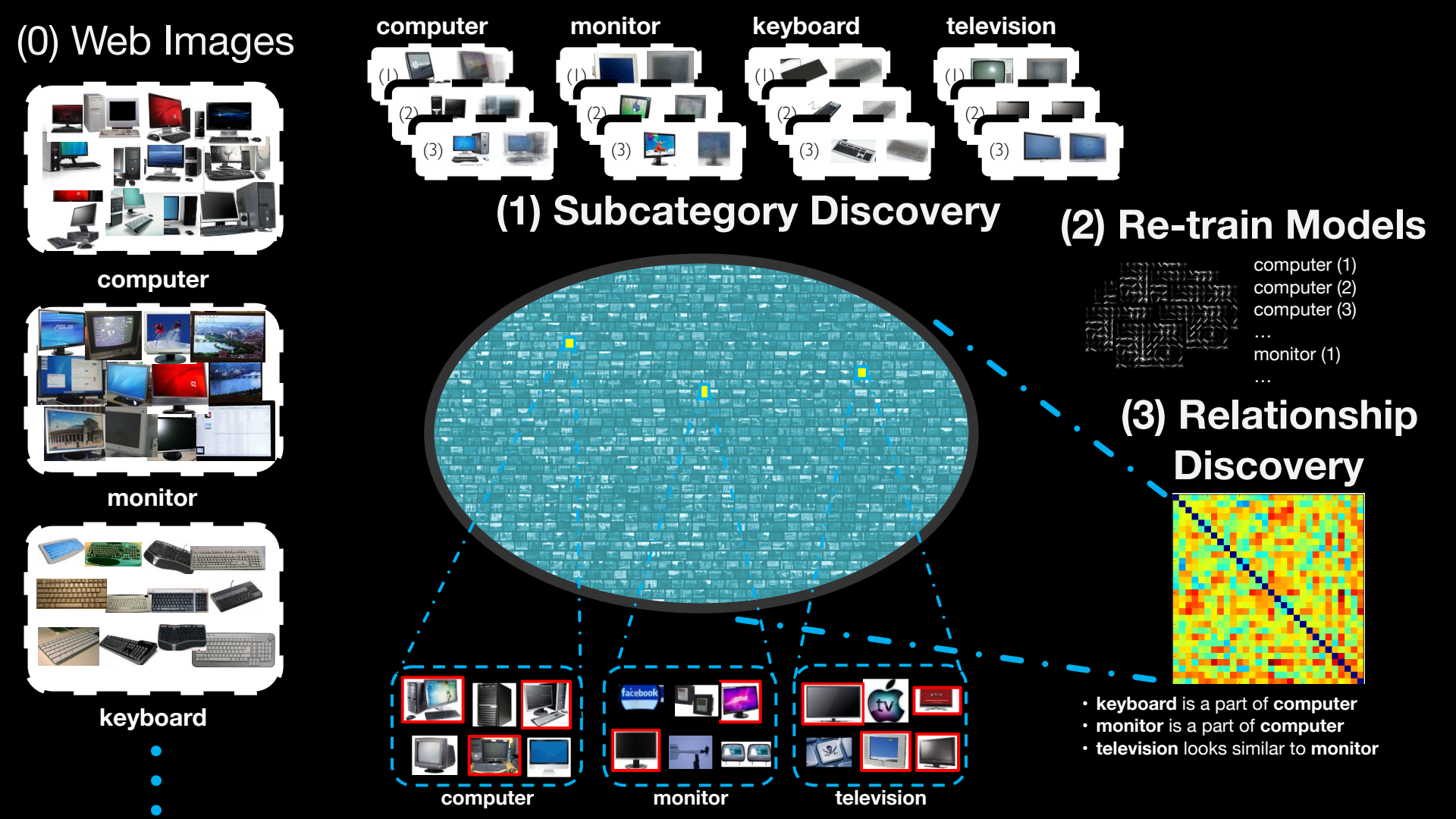
computer (1)
computer (2)
computer (3)
…
monitor (1)
…

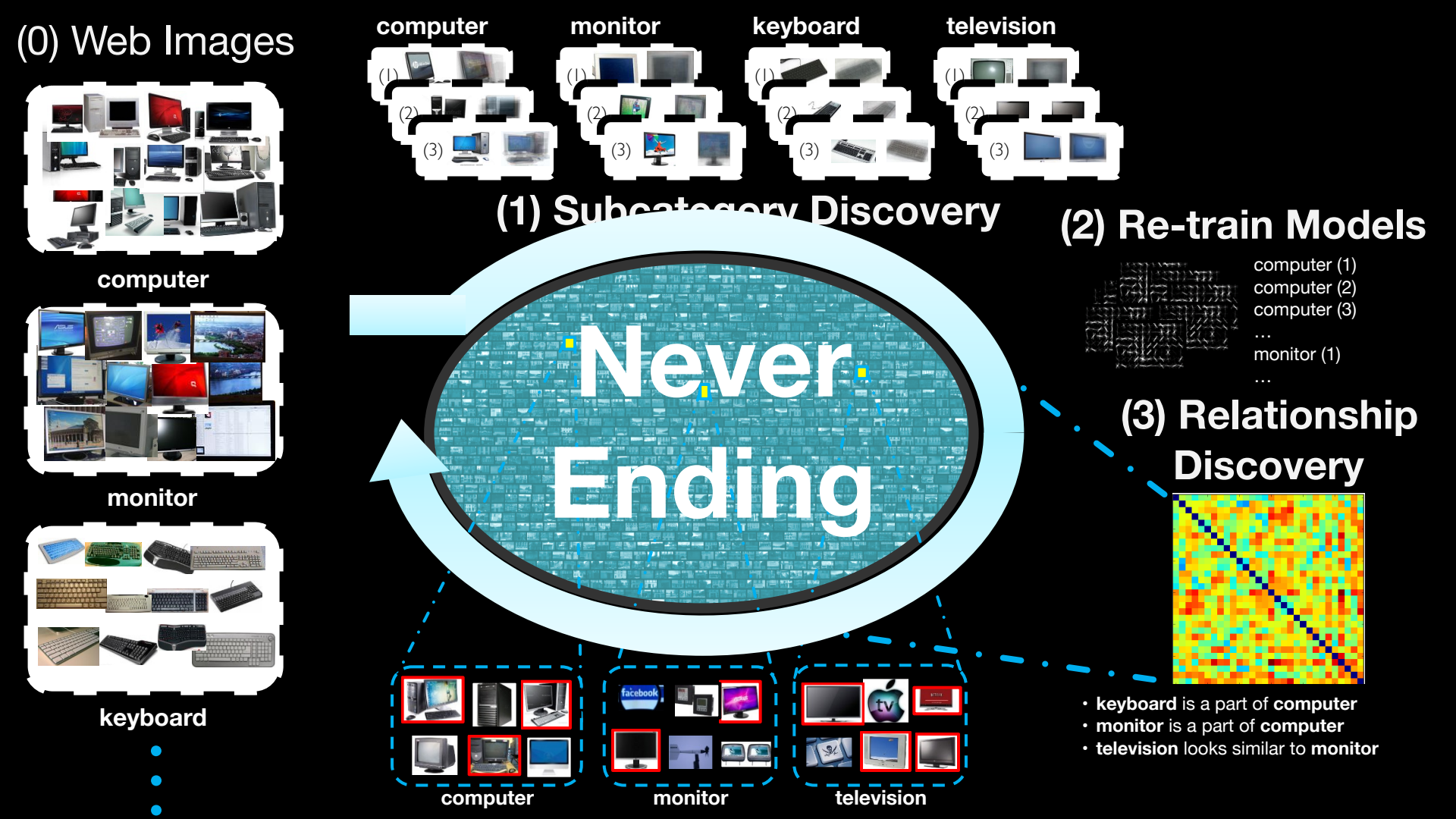
(3) Relationship Discovery



- **keyboard** is a part of **computer**
- **monitor** is a part of **computer**
- **television** looks similar to **monitor**

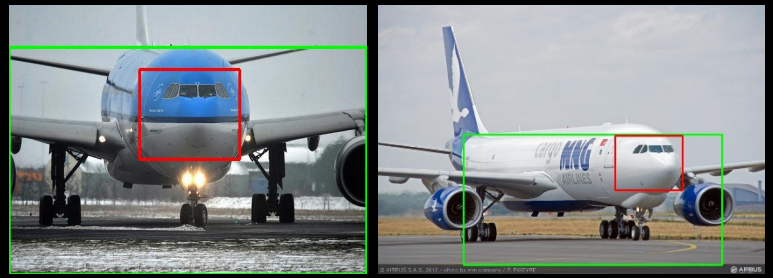

(0) Web Images
computer
monitor
keyboard
computer
monitor
keyboard
television
(1)
(2)
(3)
(1) Subcategory Discovery
(2) Train Models
computer (1)
computer (2)
computer (3)
...
monitor (1)
...
(3) Relationship Discovery
computer
monitor
television
• keyboard is a part of computer
• monitor is a part of computer
• television looks similar to monitor


(0) Web Images
computer
monitor
keyboard
computer
monitor
keyboard
television
(1)
(2)
(3)
(1) Subcategory Discovery
(2) Re-train Models
computer (1)
computer (2)
computer (3)
...
monitor (1)
...
(3) Relationship Discovery
• keyboard is a part of computer
• monitor is a part of computer
• television looks similar to monitor
computer
monitor
television


(0) Web Images
computer
monitor
keyboard
computer
monitor
keyboard
television
(1)
(2)
(3)
(1) Subcategory Discovery
(2) Re-train Models
computer (1)
computer (2)
computer (3)
...
monitor (1)
...
Never Ending
(3) Relationship Discovery
• keyboard is a part of computer
• monitor is a part of computer
• television looks similar to monitor
facebook
tv
computer
monitor
television

More Relationship Examples

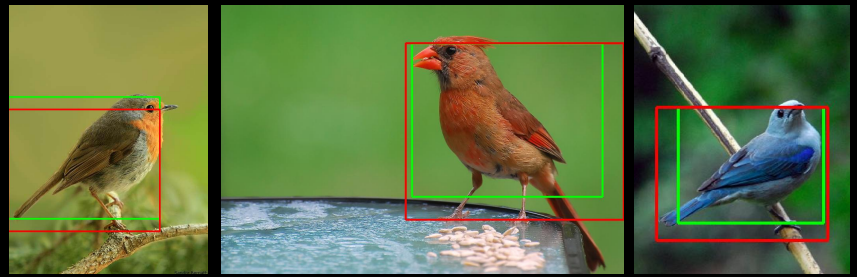
Object-Object



airplane nose is a part of **airbus 330**

eye is a part of **baby**

van is a kind of **ambulance**

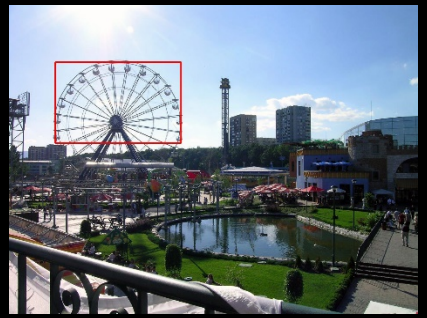


sparrow is a kind of **bird**

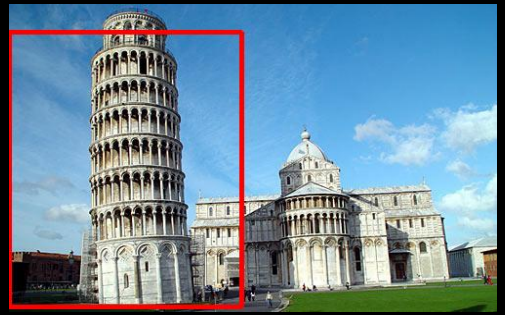
More Relationship Examples

Object-Scene

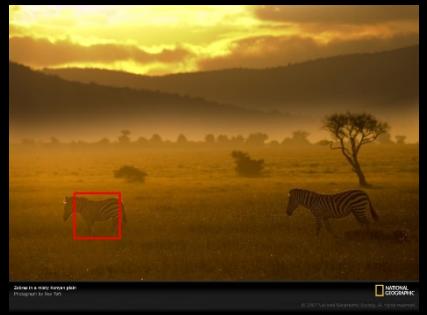

helicopter is found in **airfield**




ferris wheel is found in **amusement park**




leaning tower is found in **Pisa**

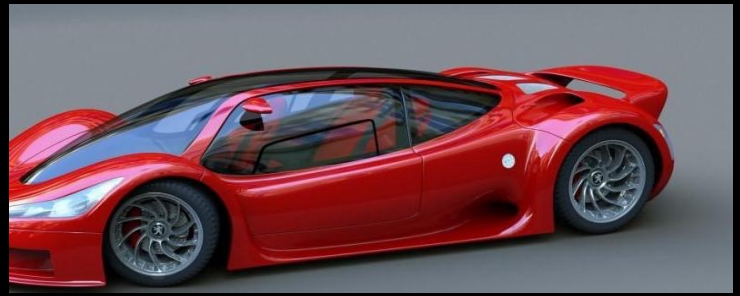
zebra is found in **savanna**

car

The Role of Relationships

car

The Role of Relationships



25th Iteration

Egypt

The Role of Relationships

Egypt

The Role of Relationships

25th Iteration

trench

The Role of Relationships




trench

The Role of Relationships

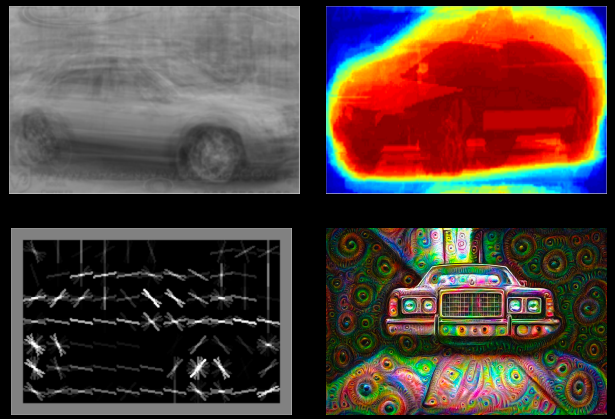




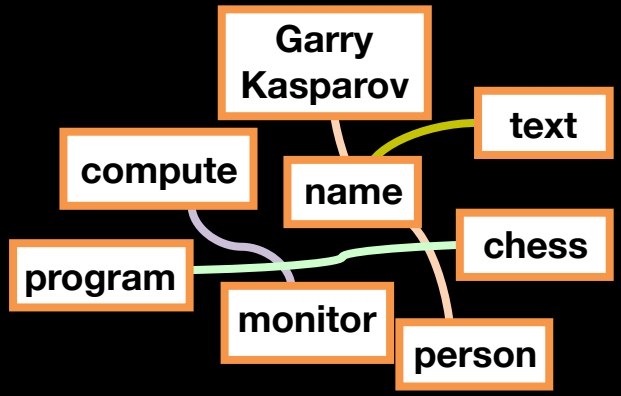

25th Iteration

The Story So Far

(I) Expand Vocabulary

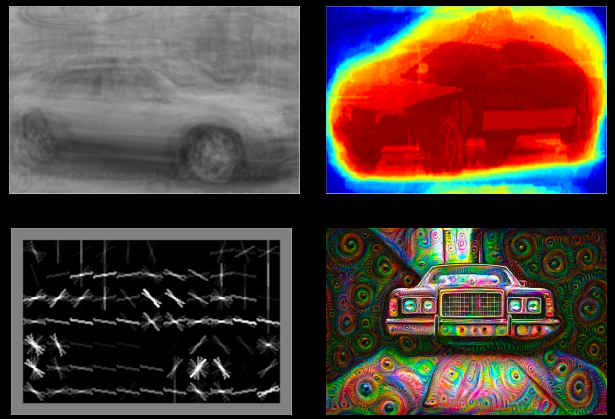


(II) Build Relationships

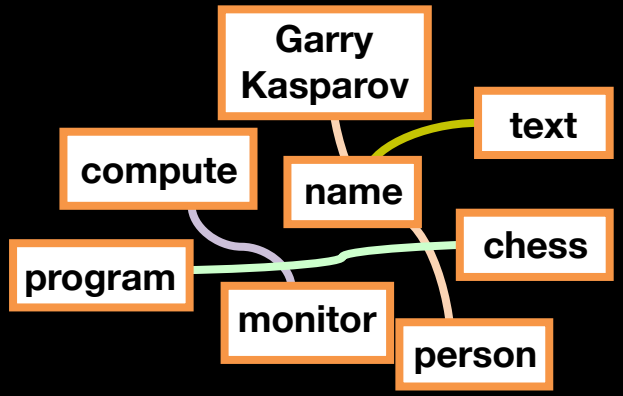


The Story So Far

(I) Expand Vocabulary

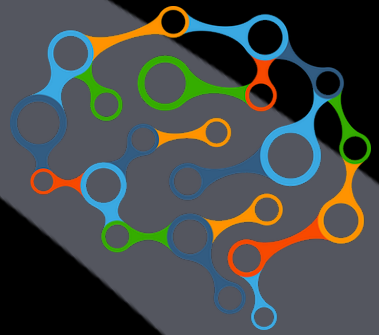


(II) Build Relationships



Learn Knowledge Automatically

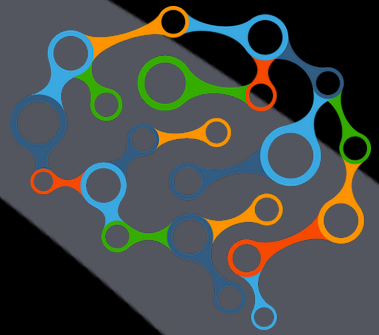
(III) Reasoning



- **Iterative Reasoning [submitted]**

Use Knowledge Effectively

(III) Reasoning



- **Iterative Reasoning [submitted]**

Task for Evaluation

Background: Faster RCNN Object Detector

(Ren et al., 2015) (Chen & Gupta, 2017)

Background: Faster RCNN Object Detector

Region Proposal

(Ren et al., 2015) (Chen & Gupta, 2017)

Background: Faster RCNN Object Detector

Region Proposal

Region of Interest (RoI)

(Ren et al., 2015) (Chen & Gupta, 2017)

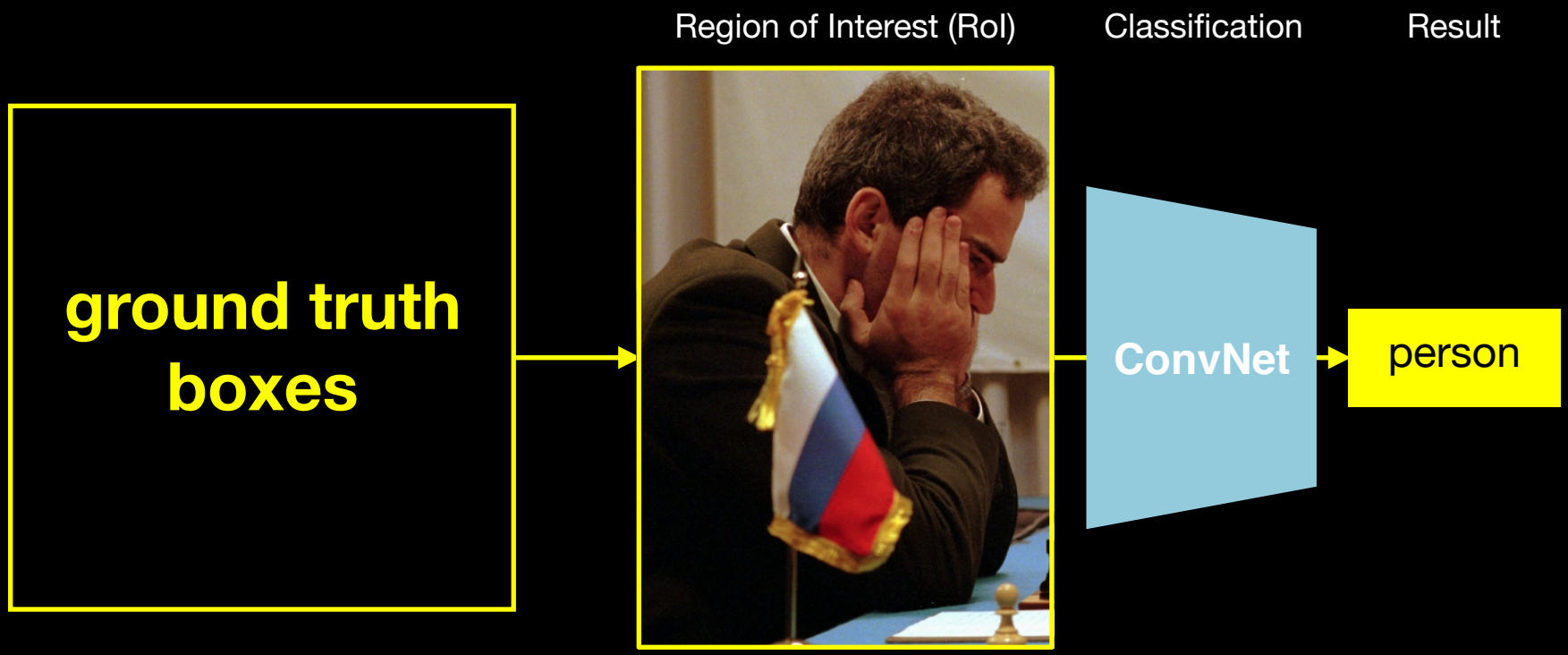
Background: Faster RCNN Object Detector

(Ren et al., 2015) (Chen & Gupta, 2017)

Background: Faster RCNN Object Detector

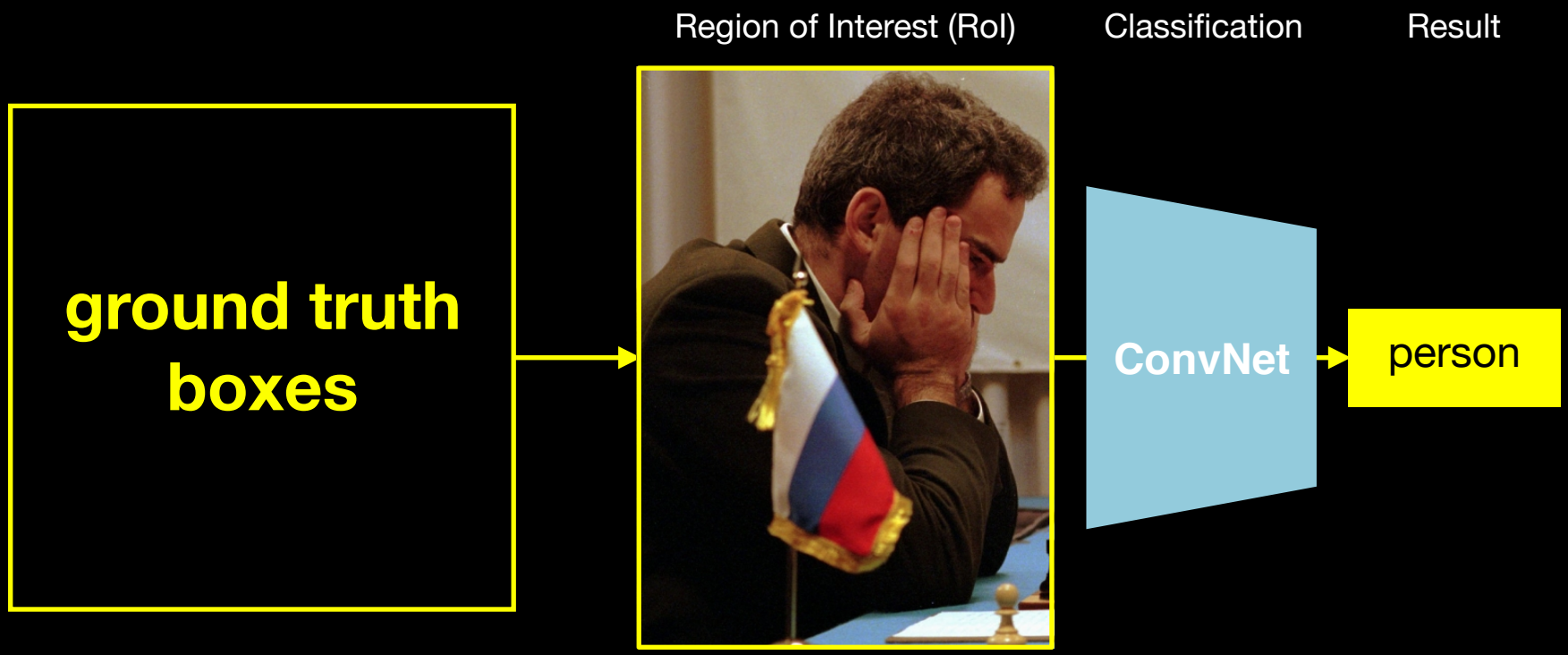
(Ren et al., 2015) (Chen & Gupta, 2017)

Our Task: Region Classification



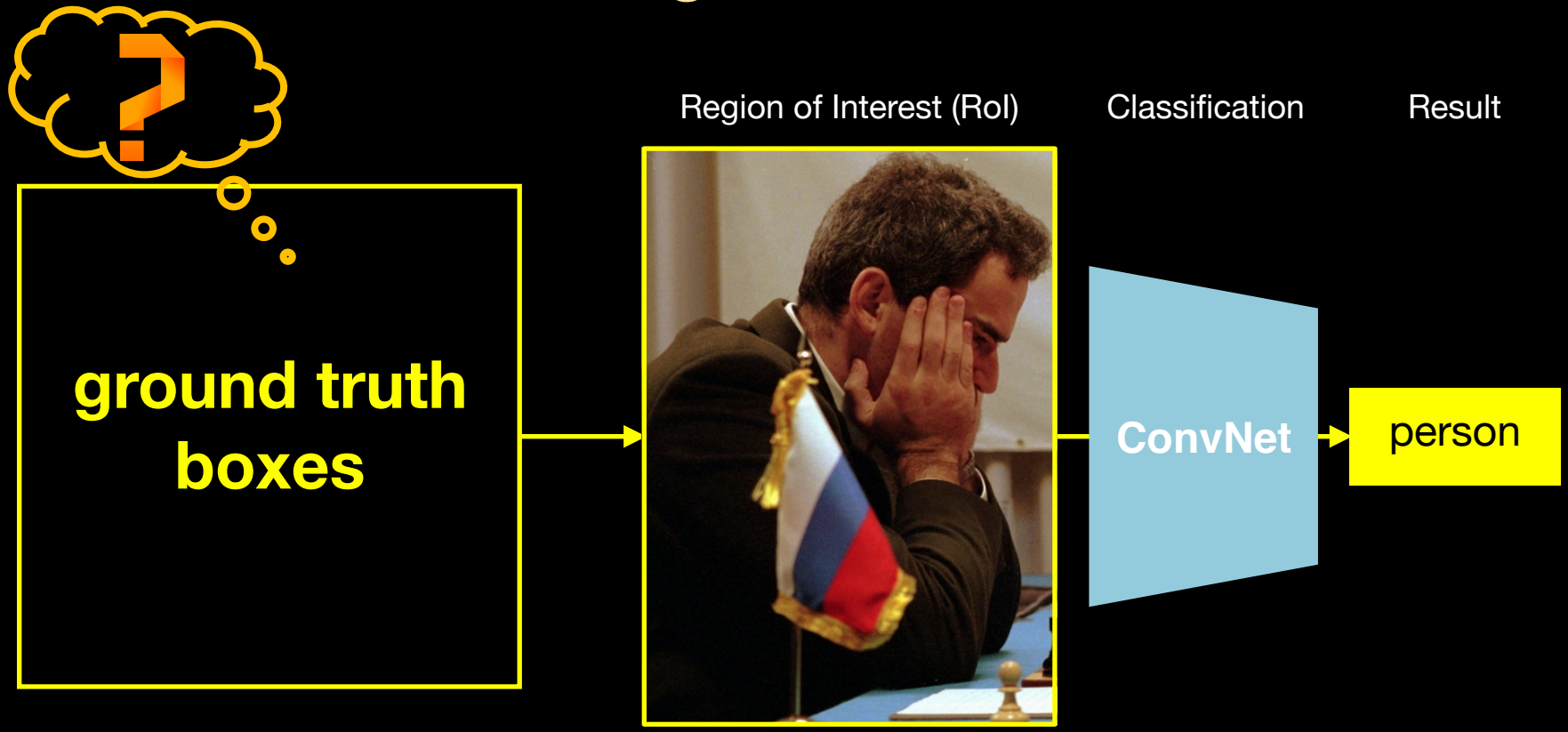
(Girshick, 2015)

Our Task: Region Classification



(Girshick, 2015)

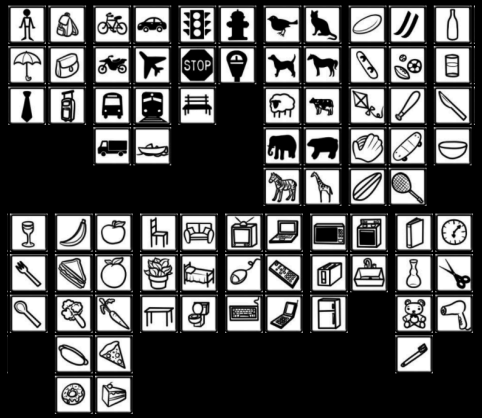
Our Task: Region Classification



(Girshick, 2015)

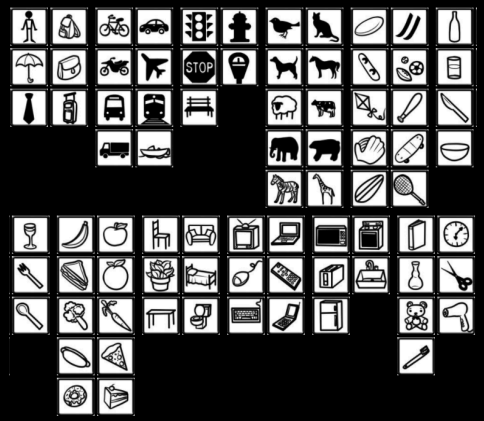
Reasoning Calls for More —

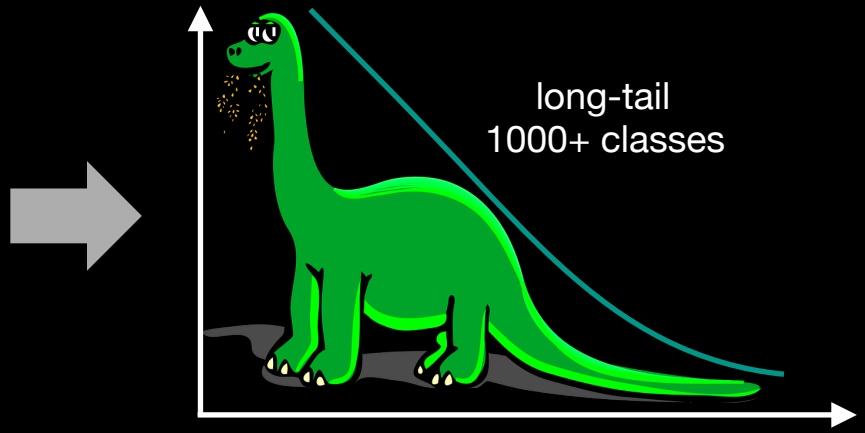
classes



Reasoning Calls for More —

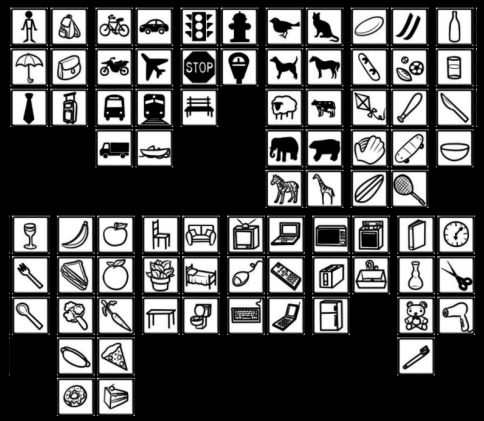
classes

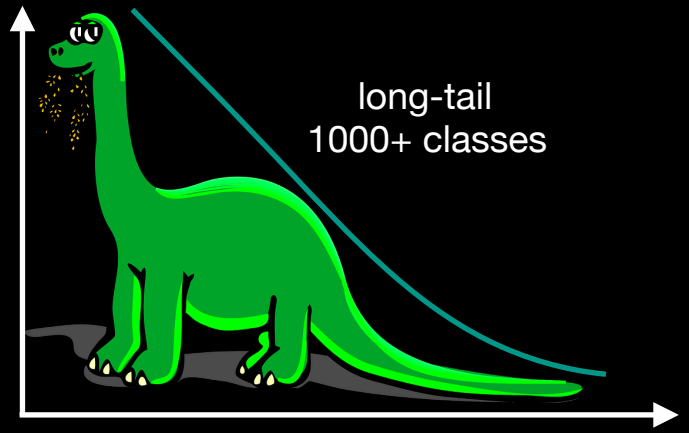




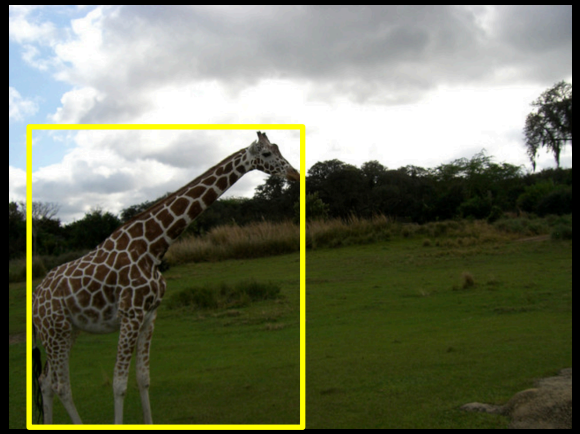
Reasoning Calls for More —

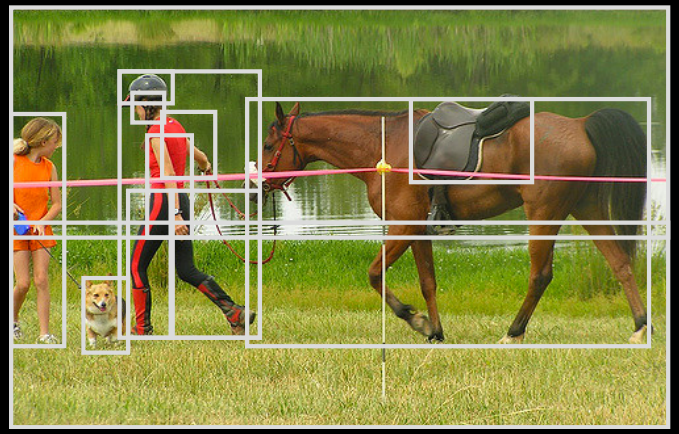
classes



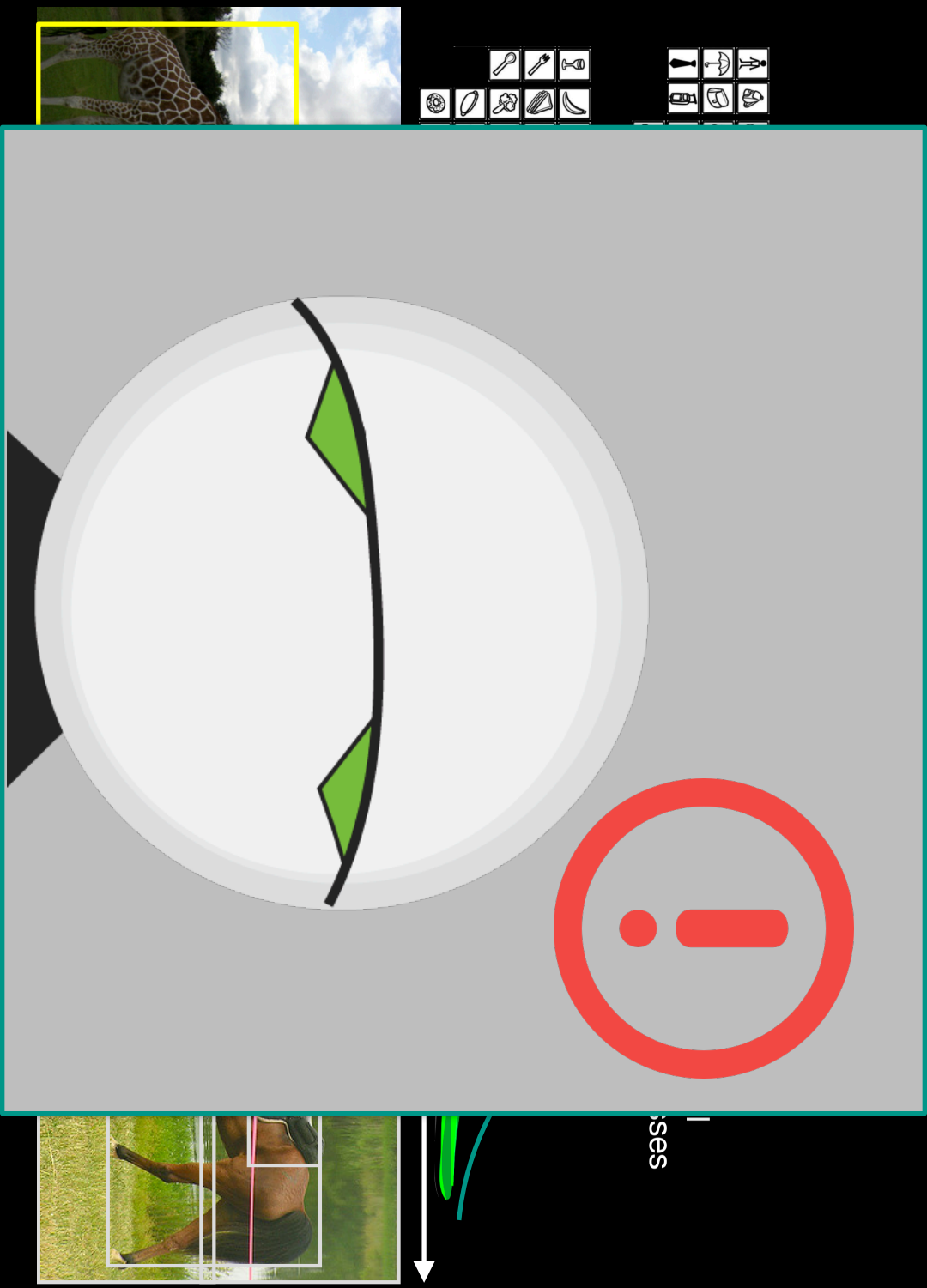


coverage

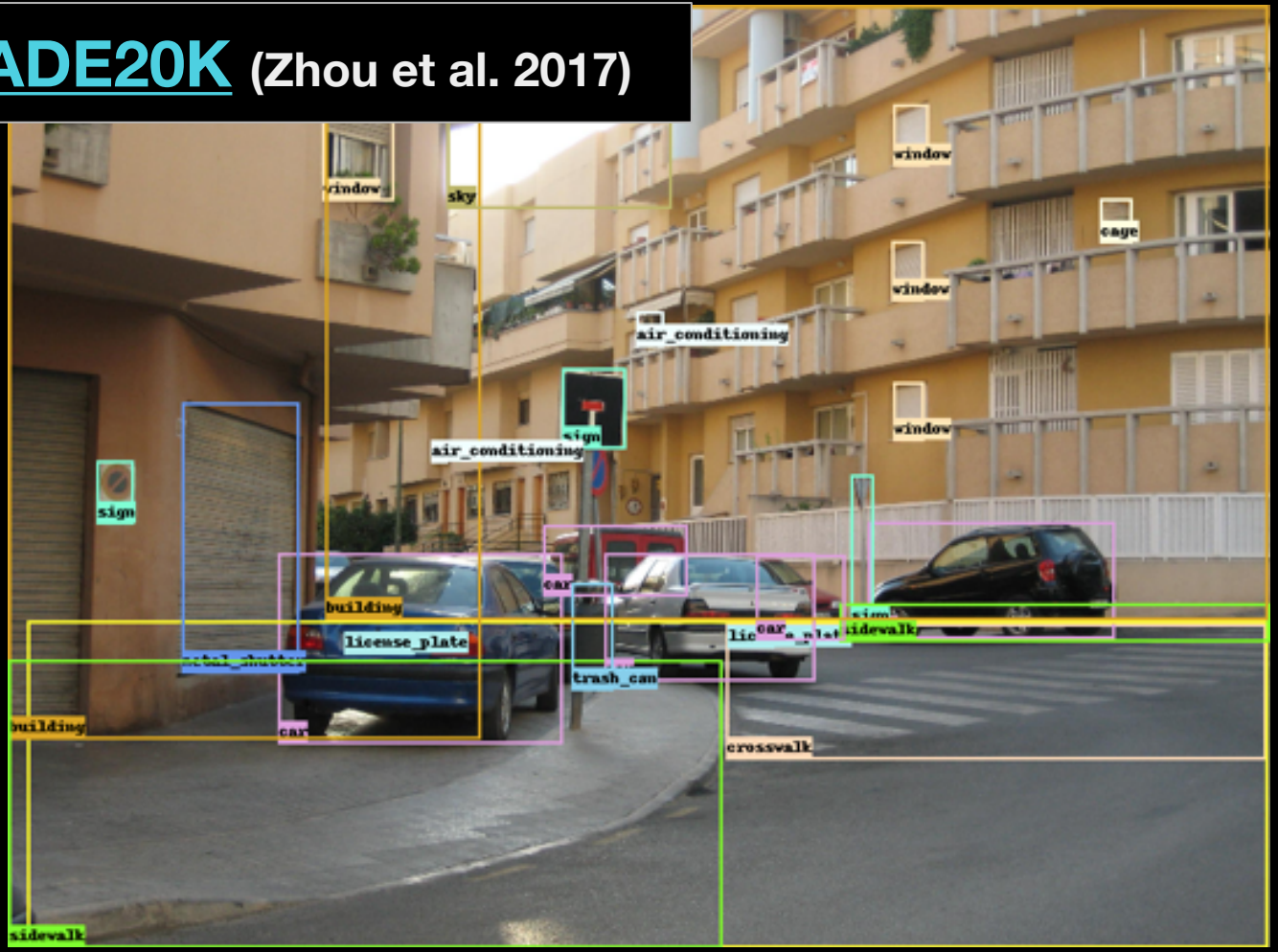




coverage classes



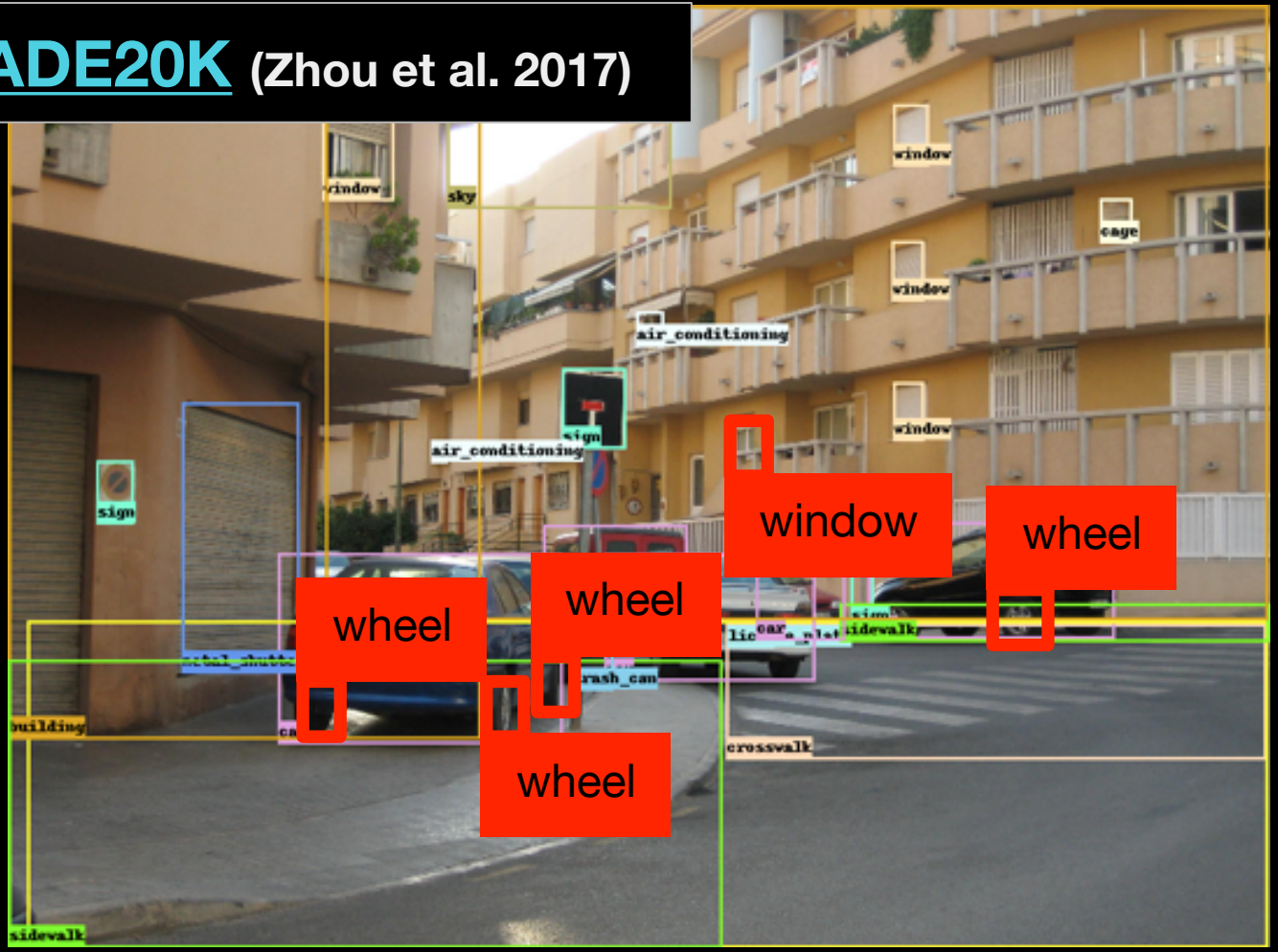
ADE20K (Zhou et al. 2017)



Missing Labels

(Krishna et al., 2016)

ADE20K (Zhou et al. 2017)



Missing Labels

(Krishna et al., 2016)

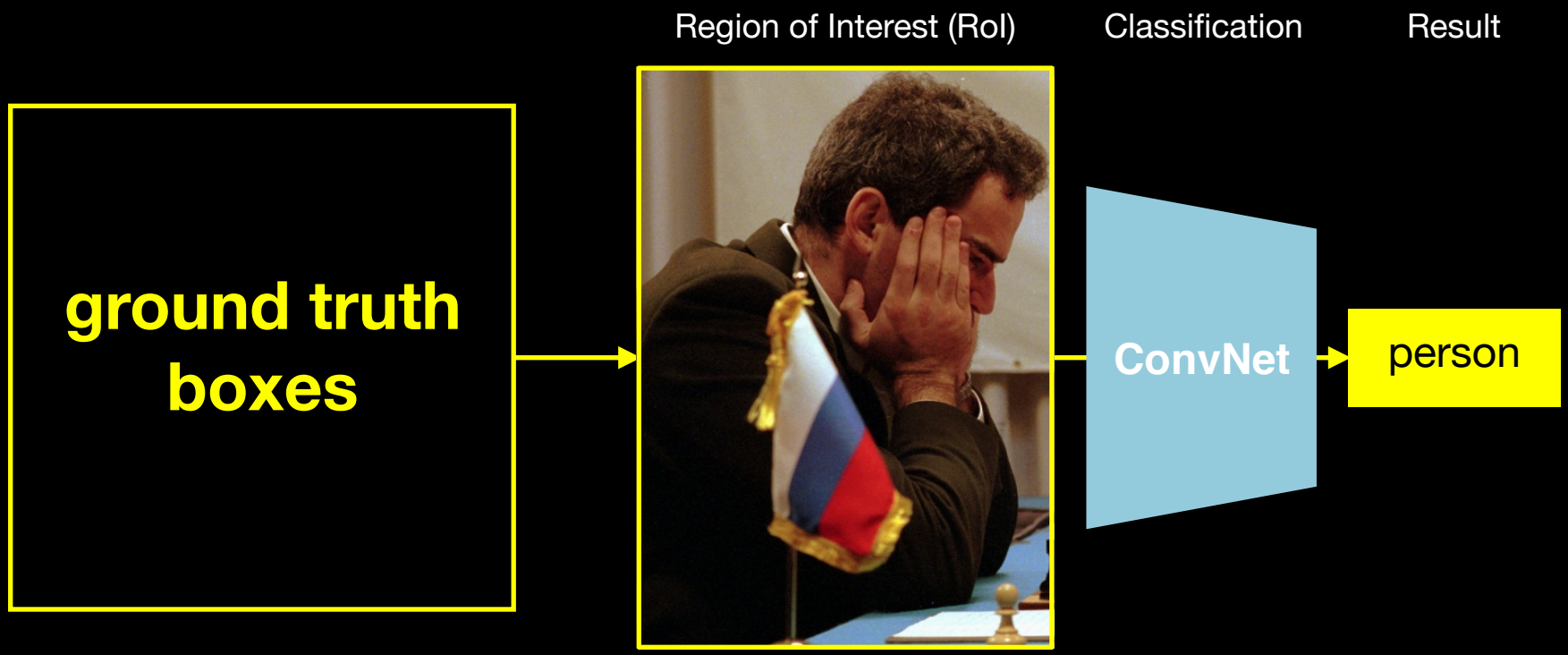
Stuff: Cannot Apply Detector

image

labels (segmentation)

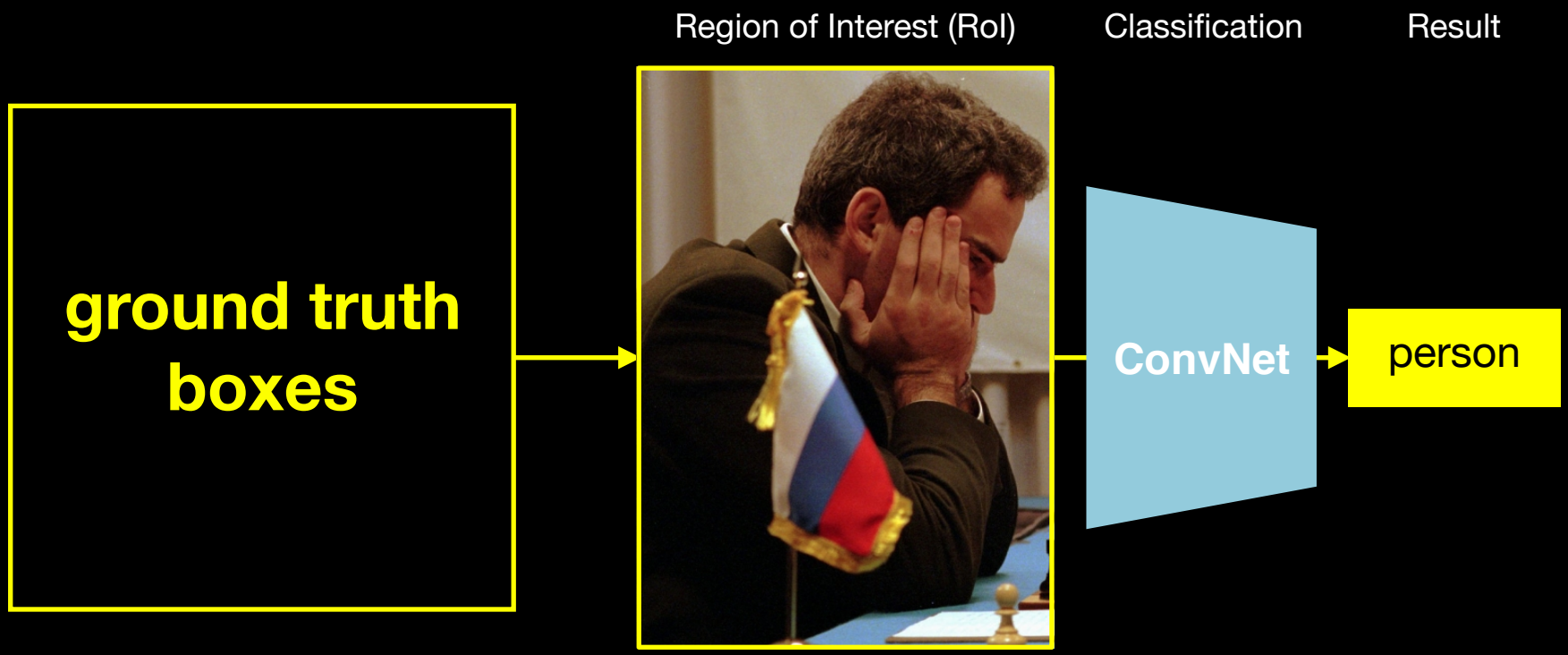
(Mottaghi et al., 2014) (Caesar et al., 2017) (Zhou et al., 2017)

Region Classification



(Girshick, 2015)

Region Classification



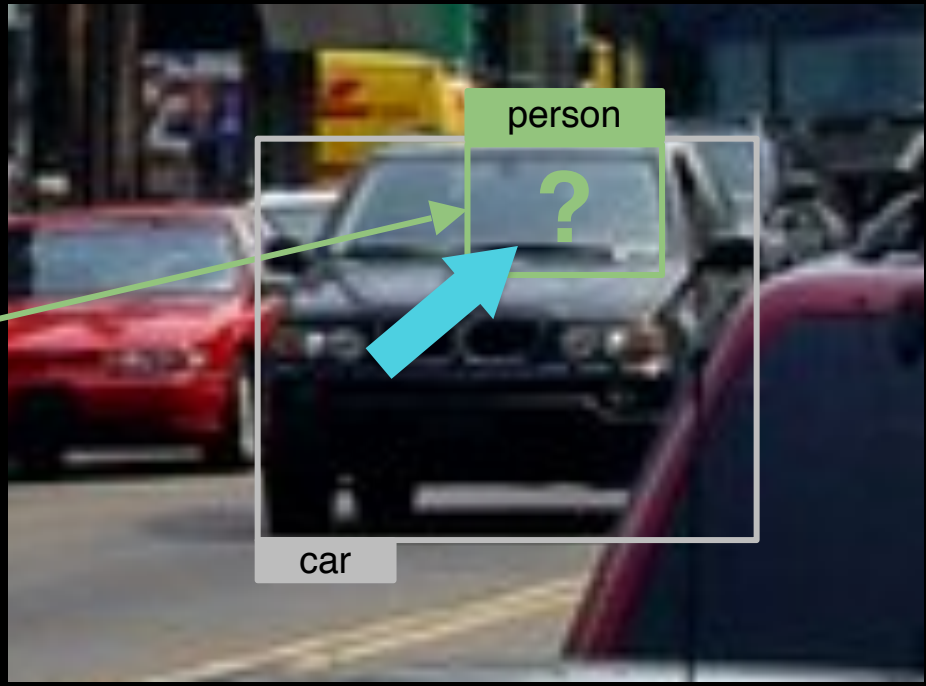
(Girshick, 2015)

Reasoning Framework

What do We Mean by Reasoning Here?

What do We Mean by Reasoning Here?

What do We Mean by Reasoning Here?

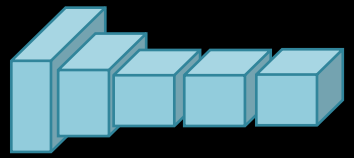


What do We Mean by Reasoning Here?

Reasoning — Easy ones help understand **Hard** ones!

Required Components Breakdown

Base Classifier

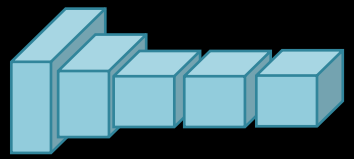


recognize

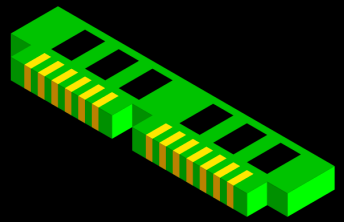
car

Required Components Breakdown

Base Classifier



Memory



recognize

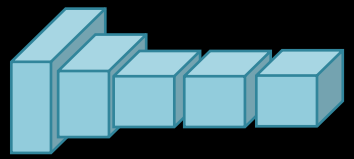
car

store

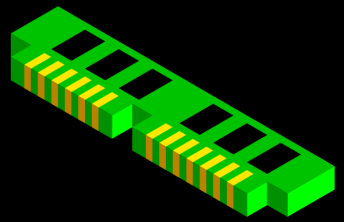
car

Required Components Breakdown

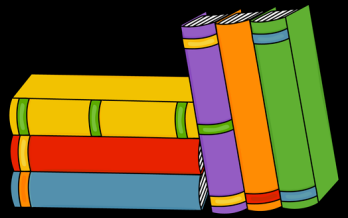
Base Classifier



Memory



Knowledge

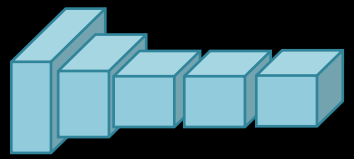
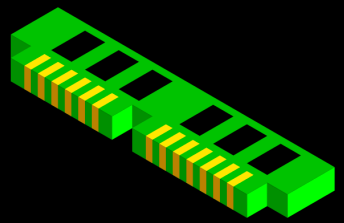
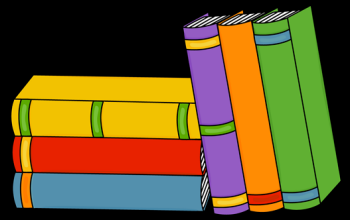


recognize
car

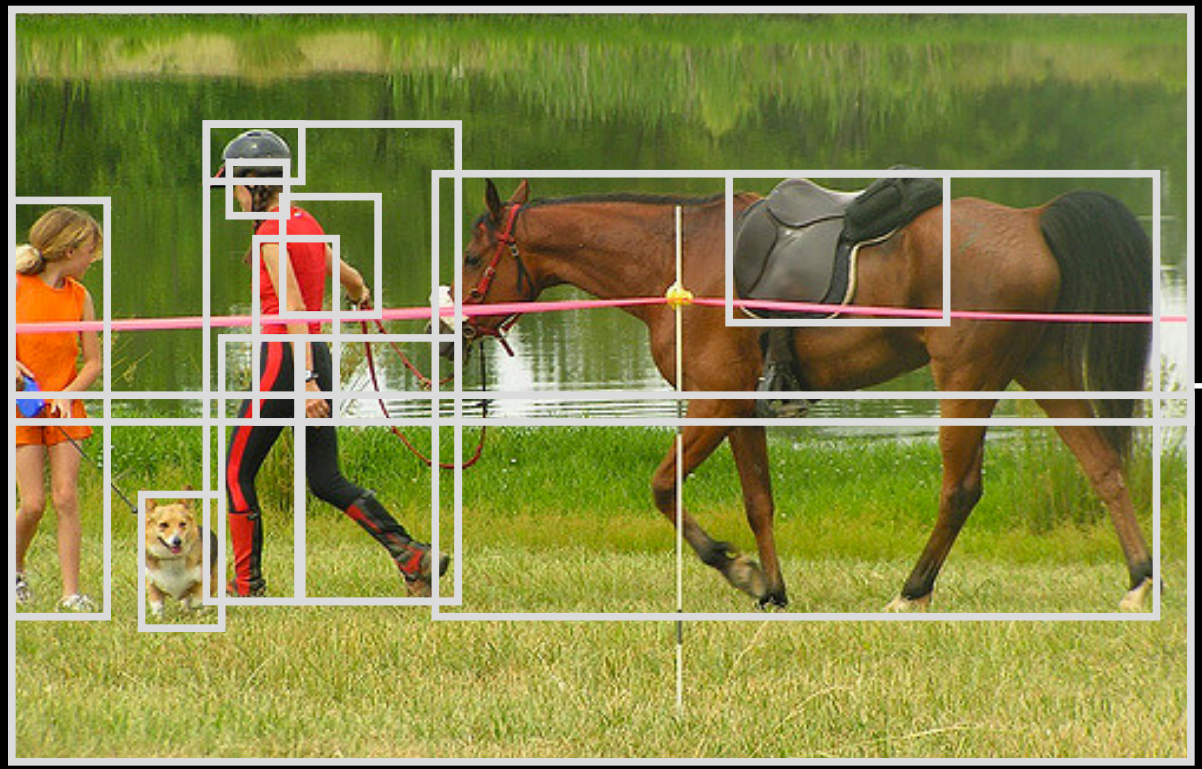
store
car

know
person drives **car**

Required Components Breakdown

Base Classifier	Memory	Knowledge	Iteration
			
recognize **car**	store **car**	know **person** drives **car**	recognize **person**

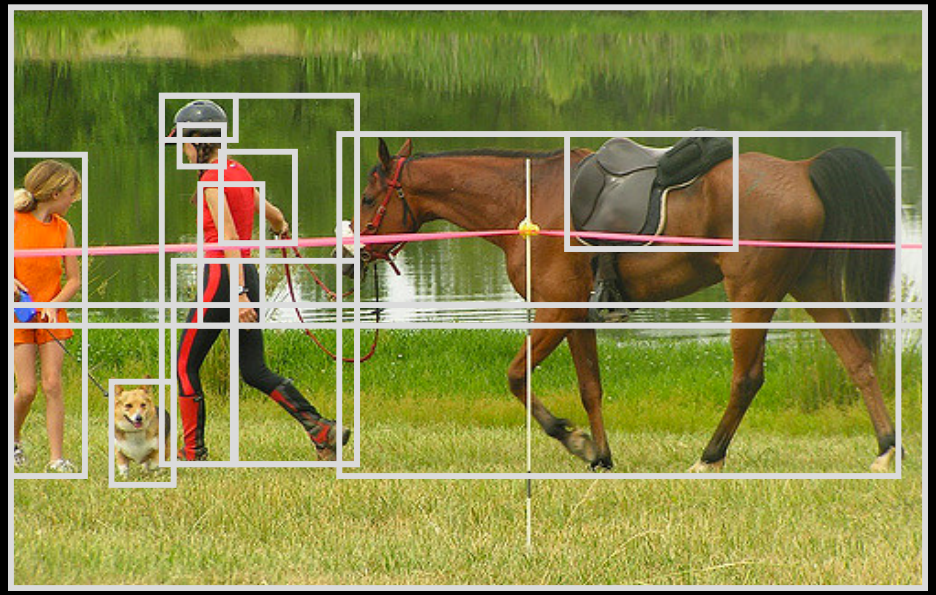
Base Classifier

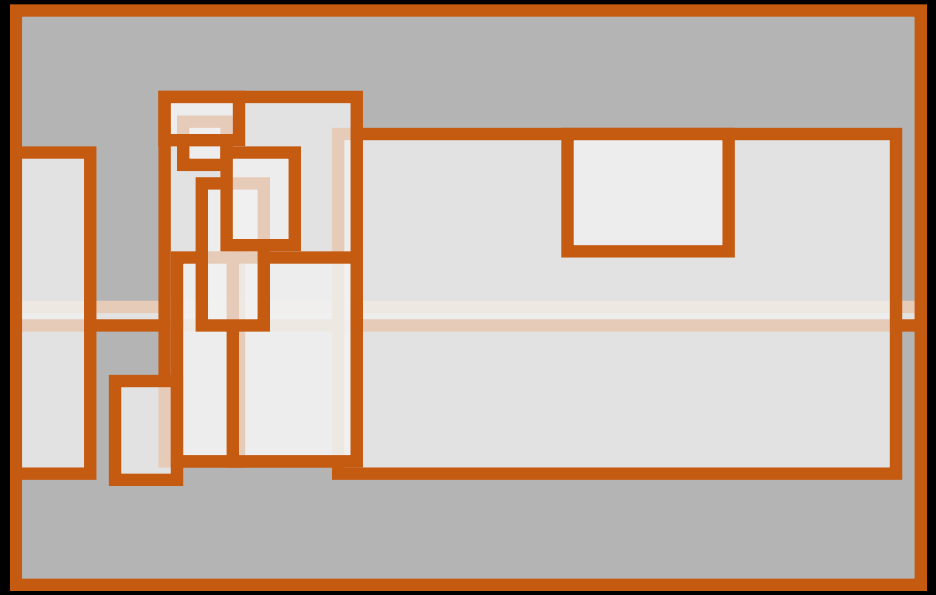


ConvNet → f_0

(Ren et al., 2015) (Chen & Gupta, 2017)

Spatial Memory: Preserves Layout

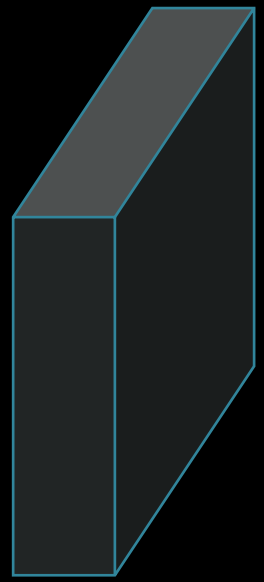




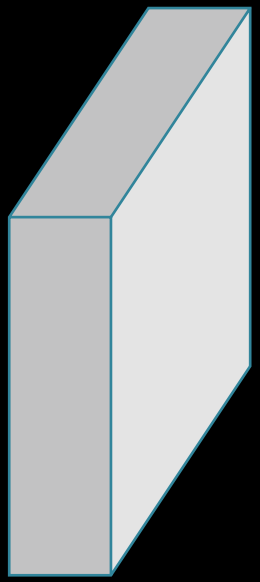
(Chen & Gupta, 2017)

(Chung et al., 2014)

Things to Put into Spatial Memory



memory t

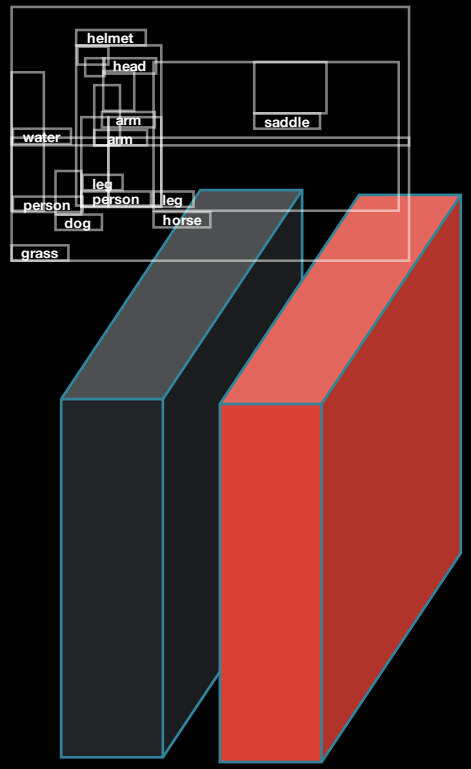


memory $t+1$

(Chung et al., 2014)

Things to Put into Spatial Memory

high-level prediction



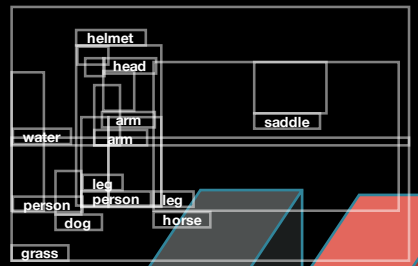
memory t

memory $t+1$

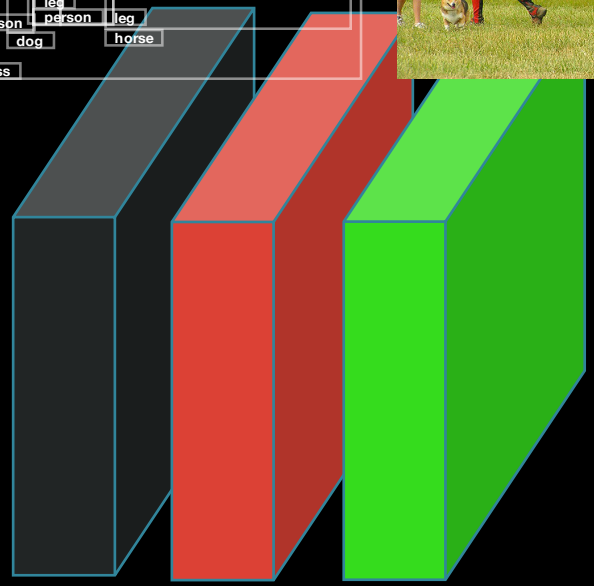
(Chung et al., 2014)

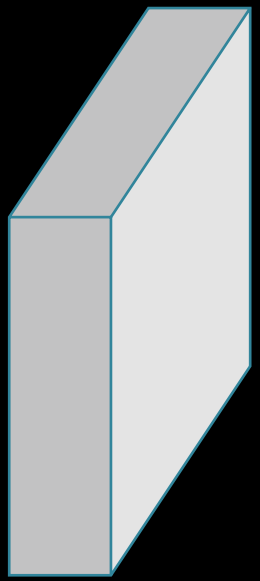
Things to Put into Spatial Memory

high-level prediction



mid-level feature



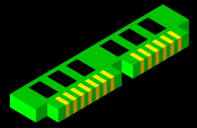


memory t

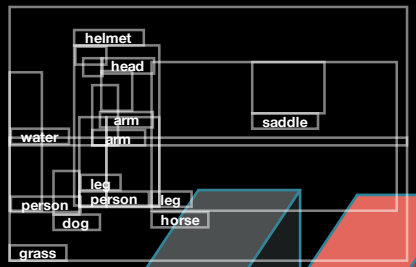
memory $t+1$

(Chung et al., 2014)

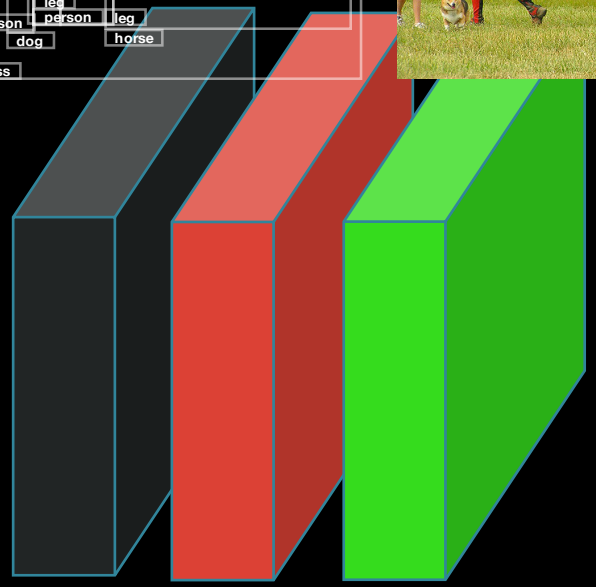
Things to Put into Spatial Memory



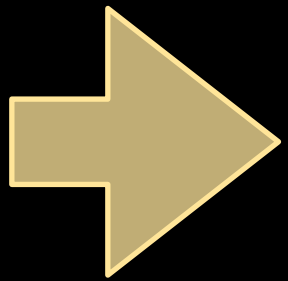
high-level prediction



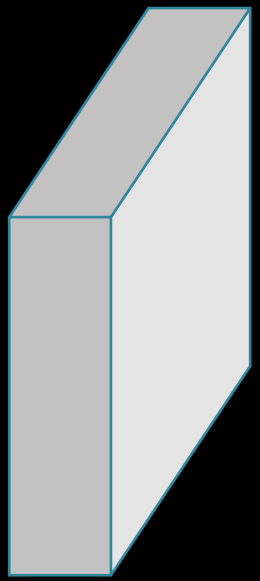
mid-level feature



memory t

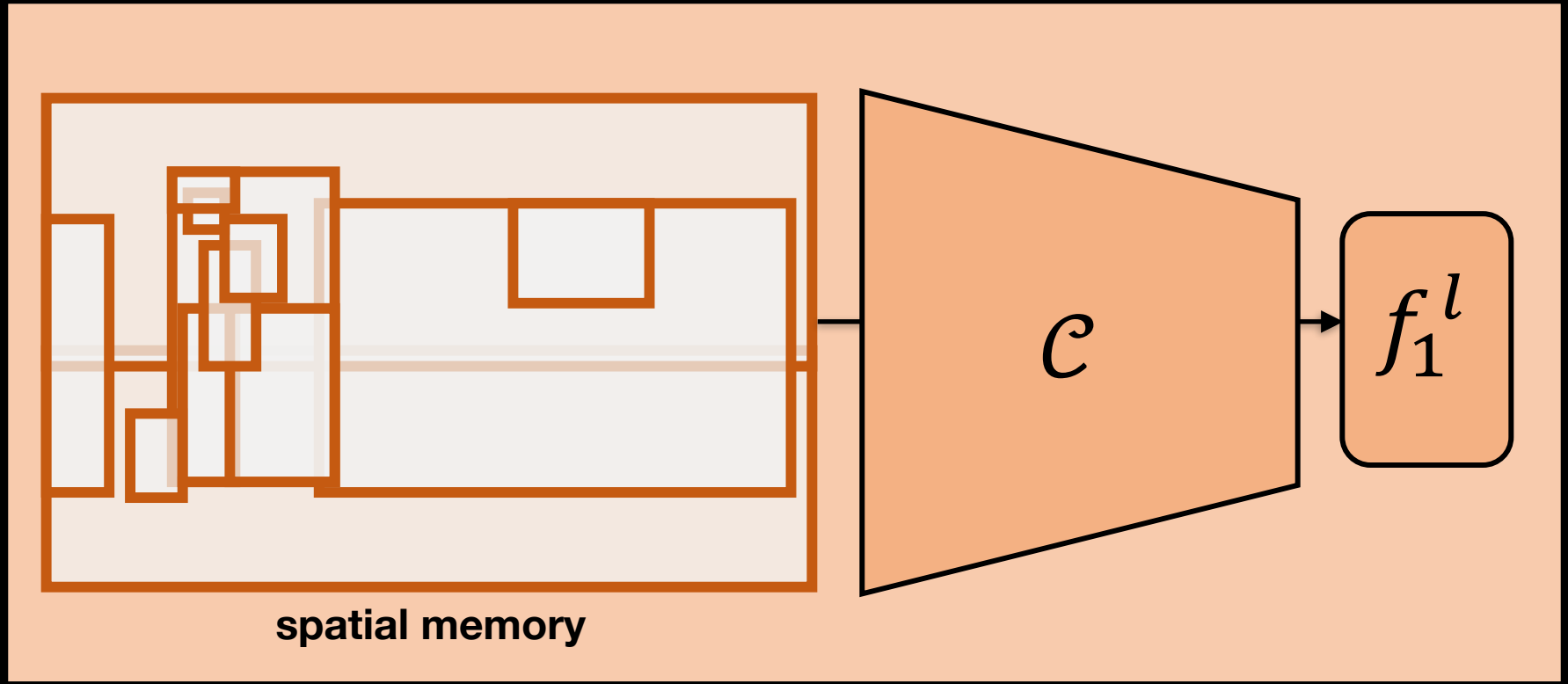


Gated Recurrent Unit



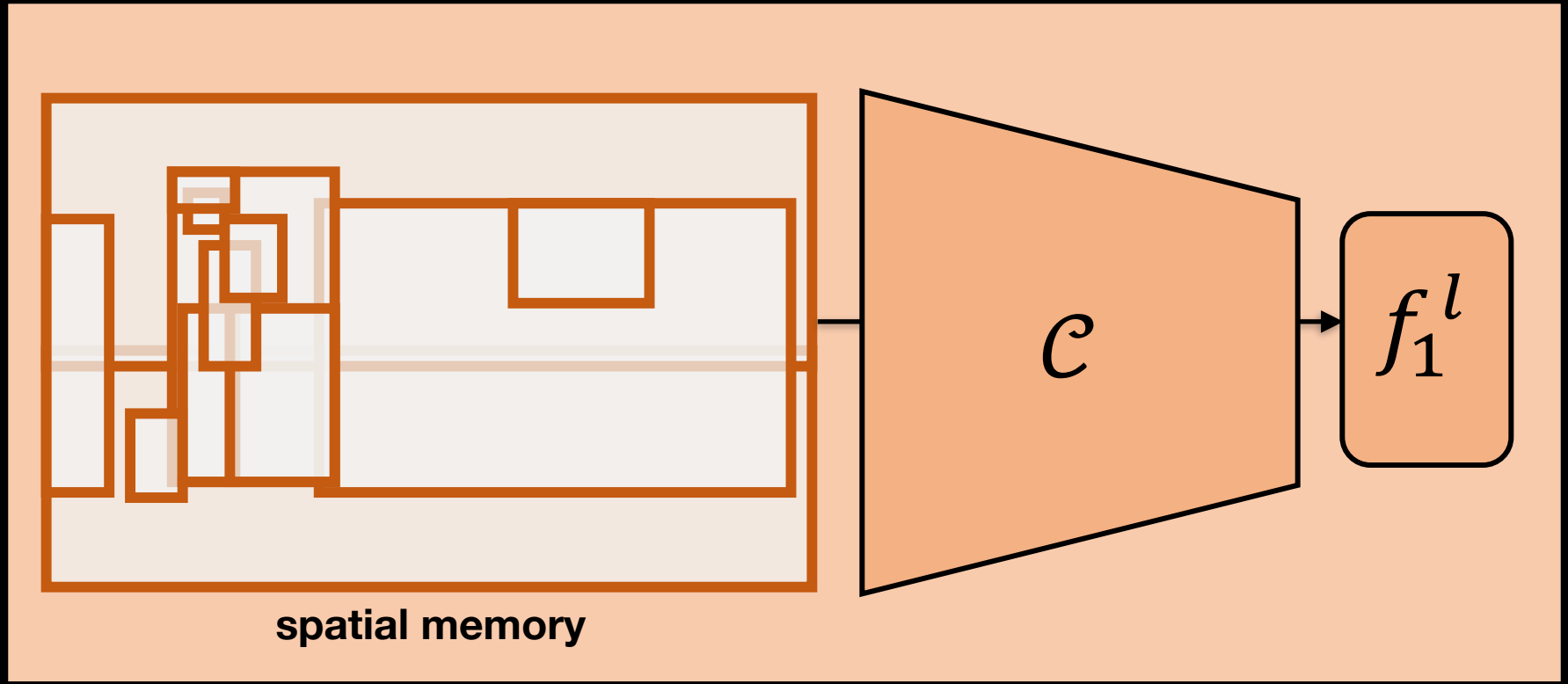
memory $t+1$

Local Module: Convolution Reasoning



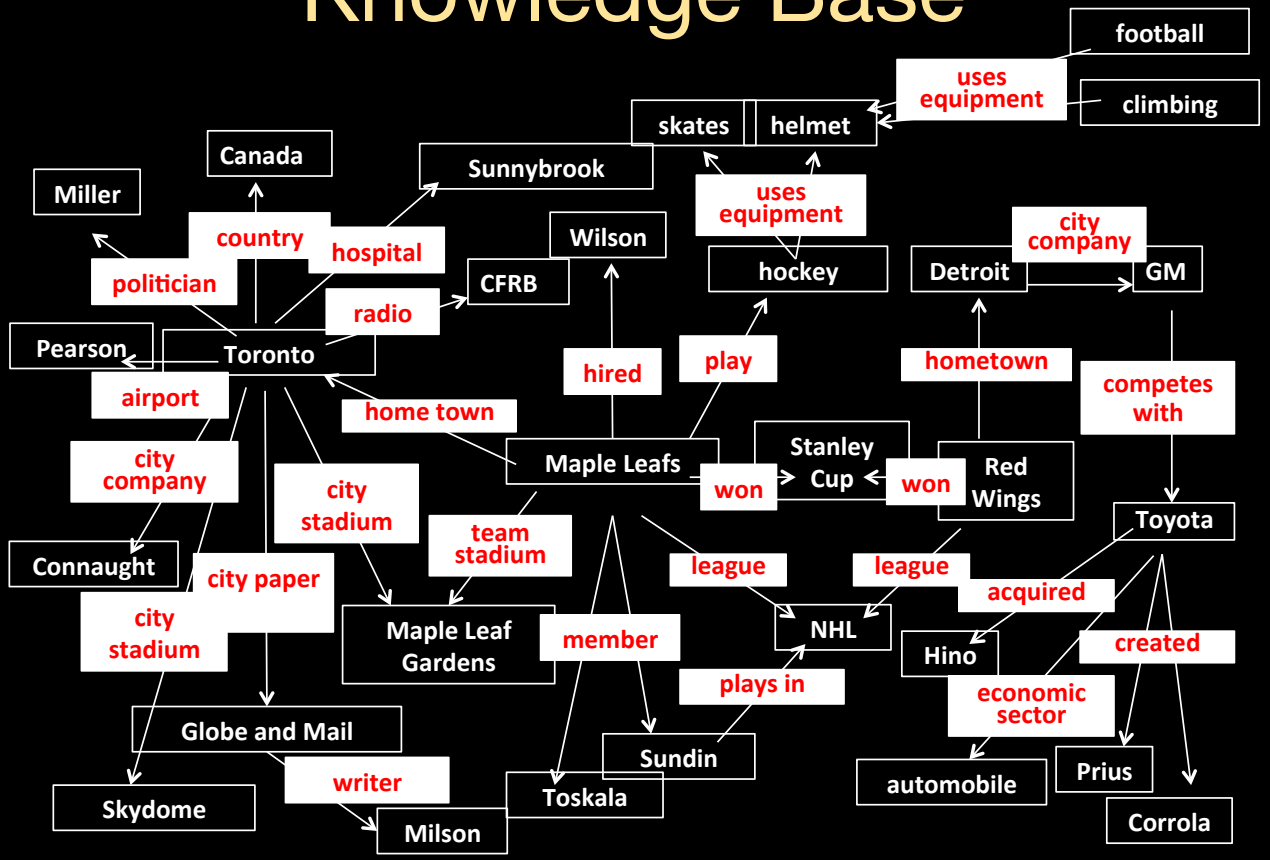
(Chen & Gupta, 2017)

Local Module: Convolution Reasoning



(Chen & Gupta, 2017)

Knowledge Base



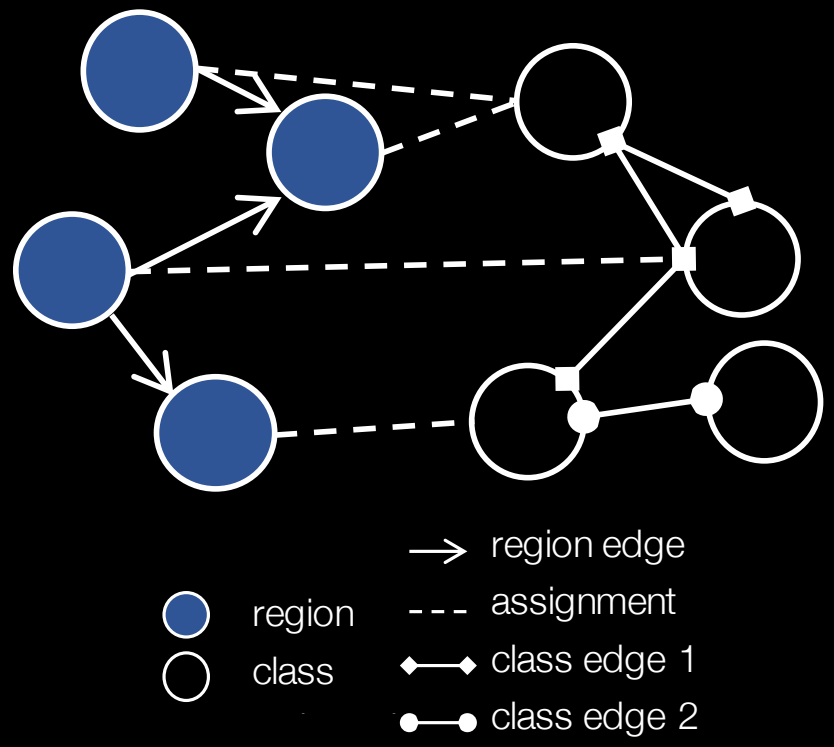
(Carlson et al., 2010)

Graph!

(Carlson et al., 2010)

Our Graph Structure

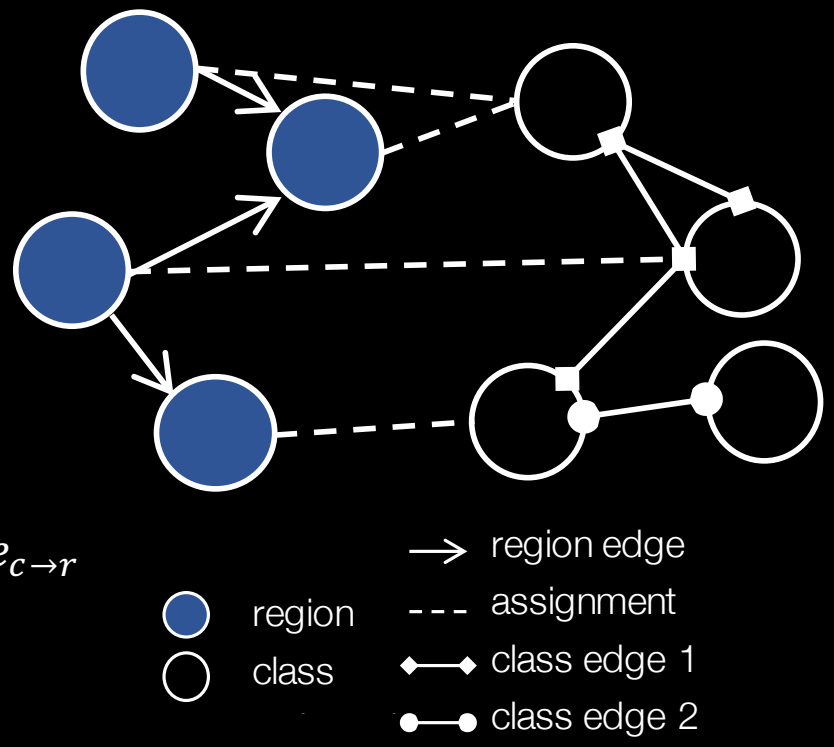
- **Nodes:**
  - region: M_r
  - class: M_c



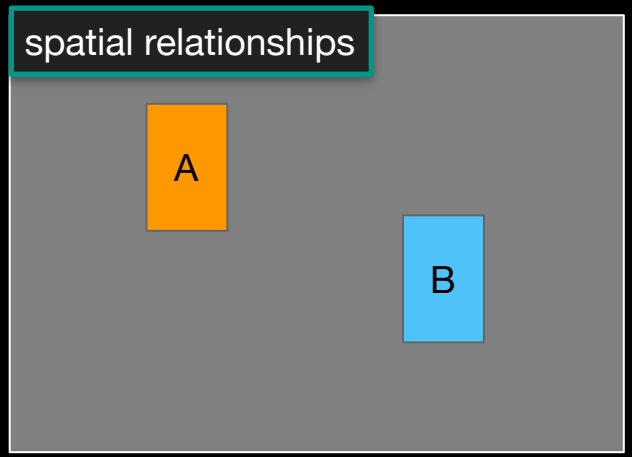
Our Graph Structure

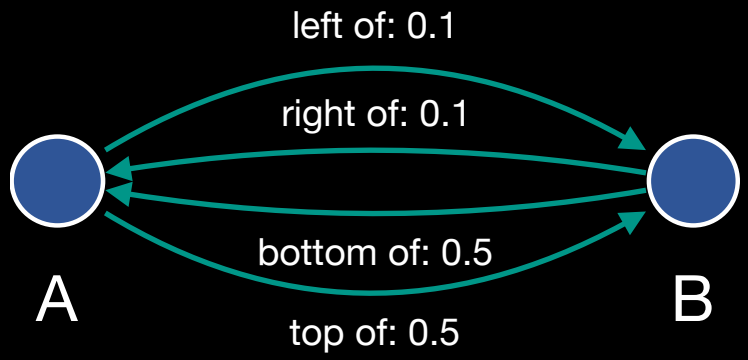
- **Nodes:**
  - region: M_r
  - class: M_c

- **Edges:**
  - region graph: $\mathcal{E}_{r \to r}$
  - region assignment: $e_{r \to c}$ & $e_{c \to r}$
  - knowledge graph: $\mathcal{E}_{c \to c}$

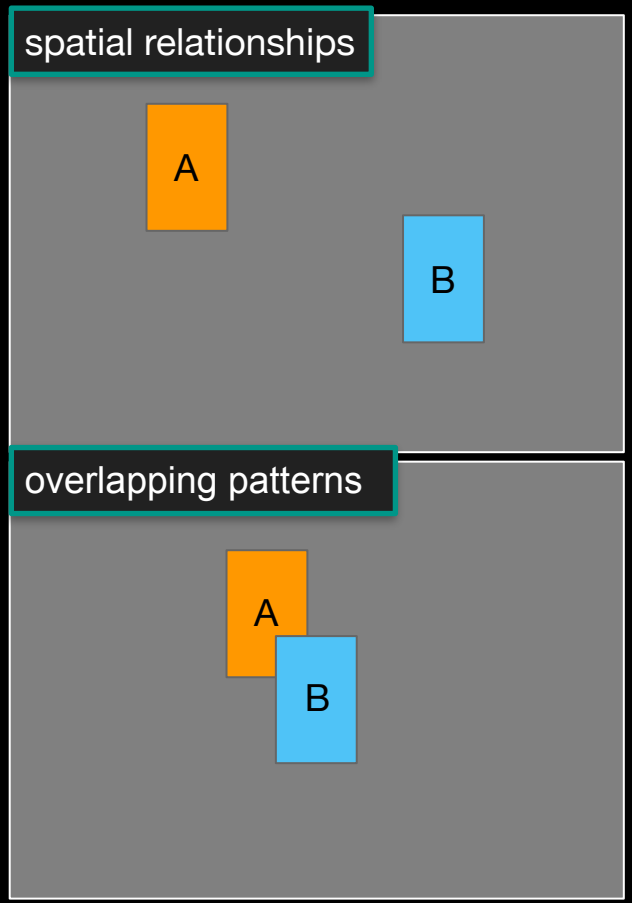


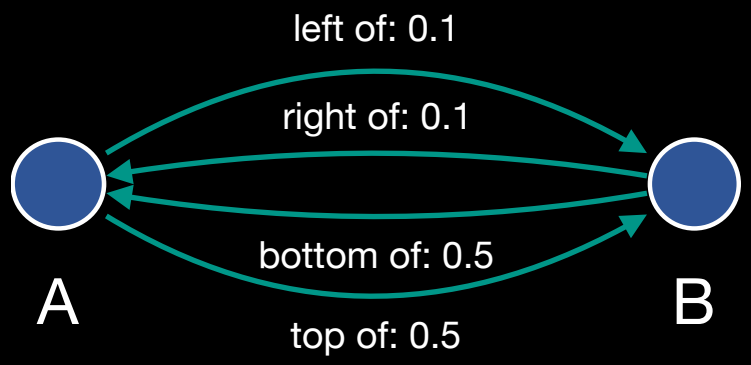
Region Graph $\mathcal{E}_{r \to r}$

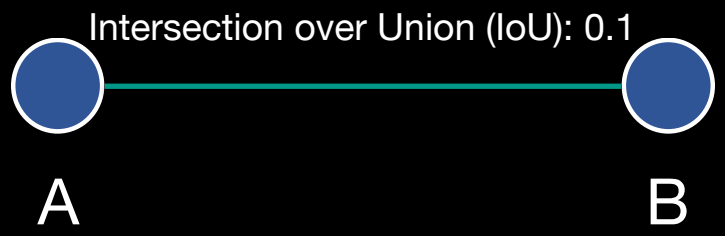




Region Graph $\mathcal{E}_{r \to r}$







Knowledge Graph $\mathcal{E}_{c \to c}$

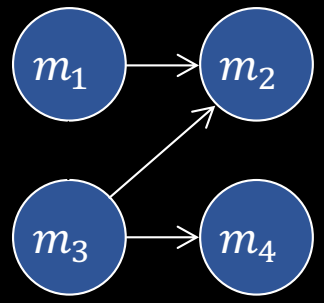
- “commonsense"
  - **Similarity**: cat *vs* tiger
  - **Is-kind-of**: BMW *vs* car
  - **Is-part-of**: wheel *vs* car
  - **Plural form**: person *vs* people
  - **Left-right**: left arm *vs* right arm

Knowledge Graph $\mathcal{E}_{c \to c}$

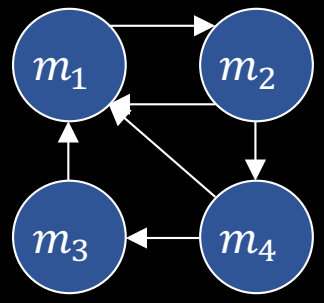
- “commonsense"
  - **Similarity**: cat *vs* tiger
  - **Is-kind-of**: BMW *vs* car
  - **Is-part-of**: wheel *vs* car
  - **Plural form**: person *vs* people
  - **Left-right**: left arm *vs* right arm
- more image-specific
  - **Spatial configurations**: near-by
  - **Actions**: ride, hit

Reasoning: Message Passing w/ Edges

e_1:



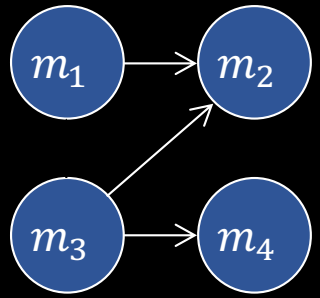
e_2:



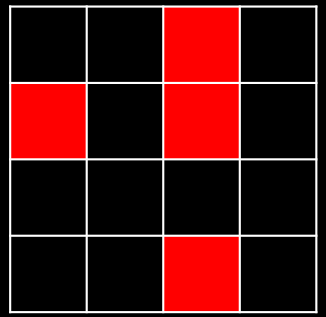
Edge Type

Reasoning: Message Passing w/ Edges

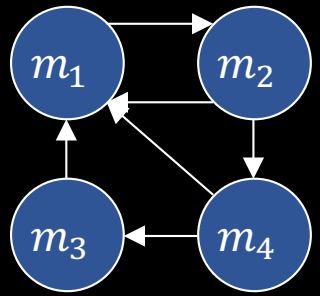
e_1:



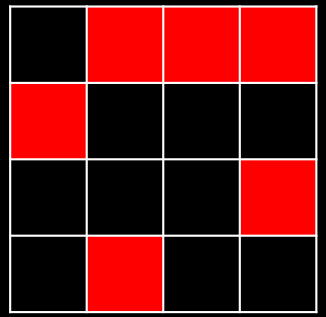
$G_1 =$ A_1



e_2:



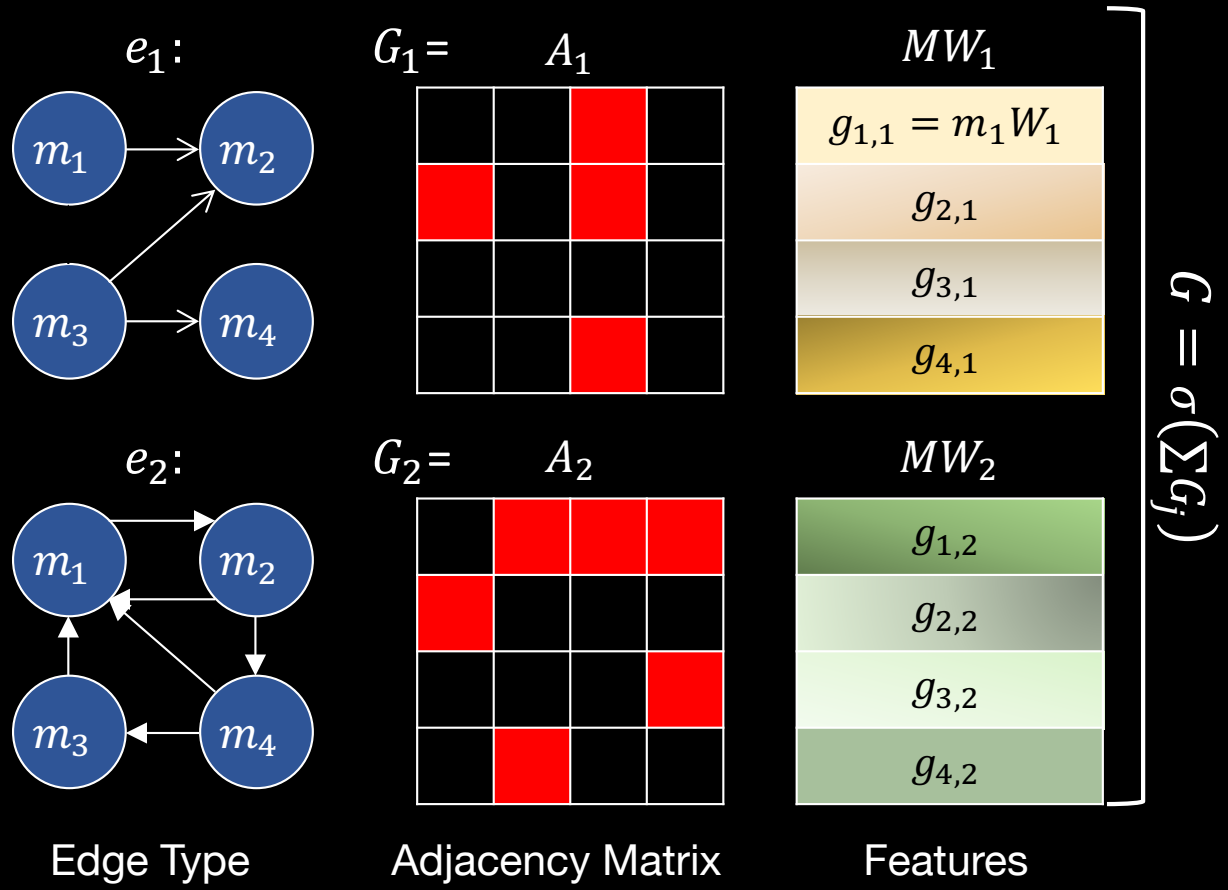
$G_2 =$ A_2



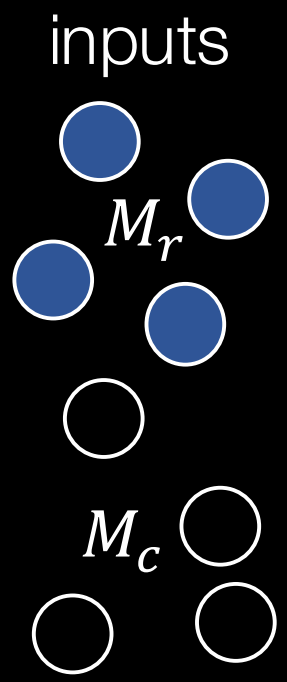
Edge Type

Adjacency Matrix

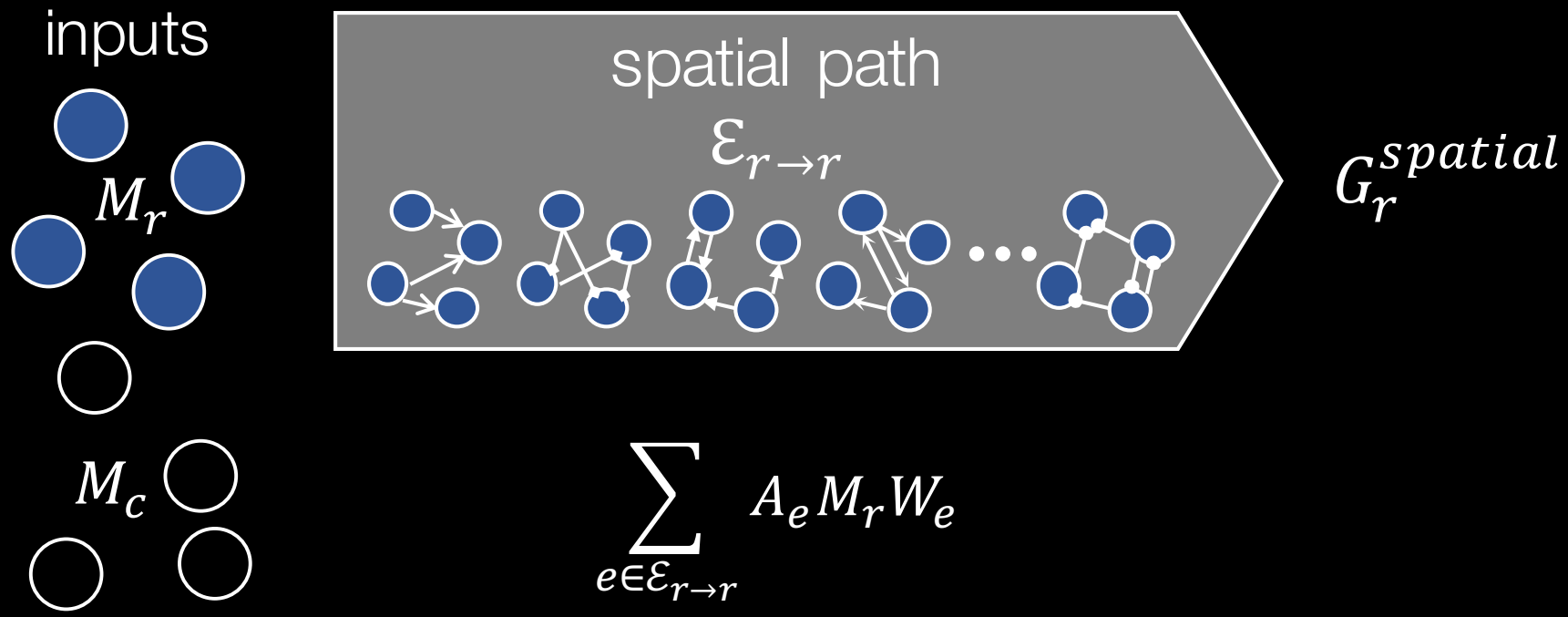
Reasoning: Message Passing w/ Edges



Spatial Path: within Region Graph

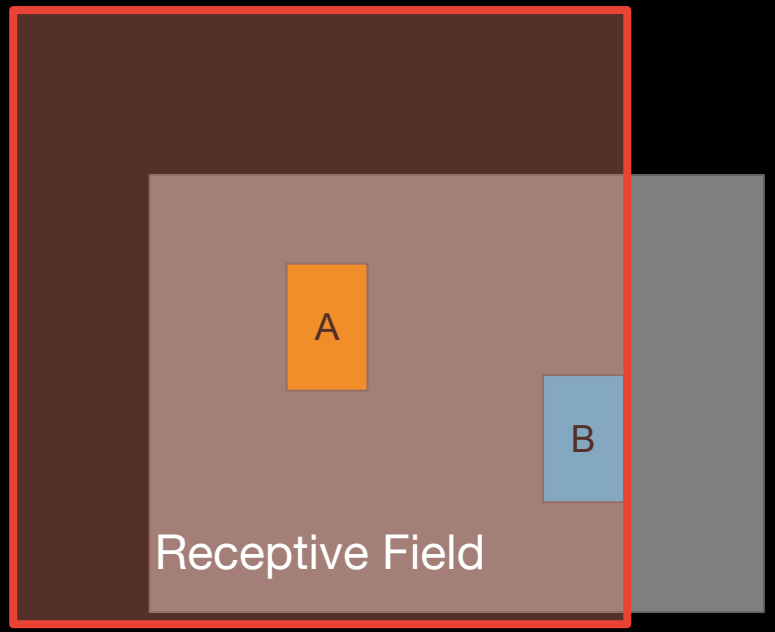


Spatial Path: within Region Graph

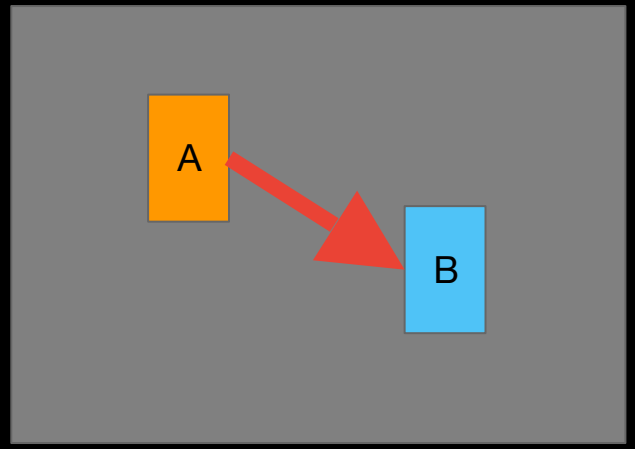


$$\sum_{e \in \mathcal{E}_{r \to r}} A_e M_r W_e$$

Spatial Reasoning: a Comparison

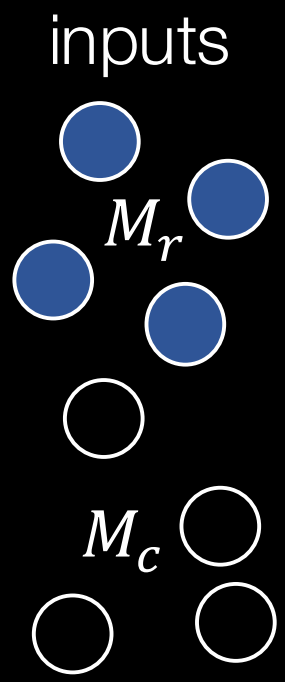


ConvNet based



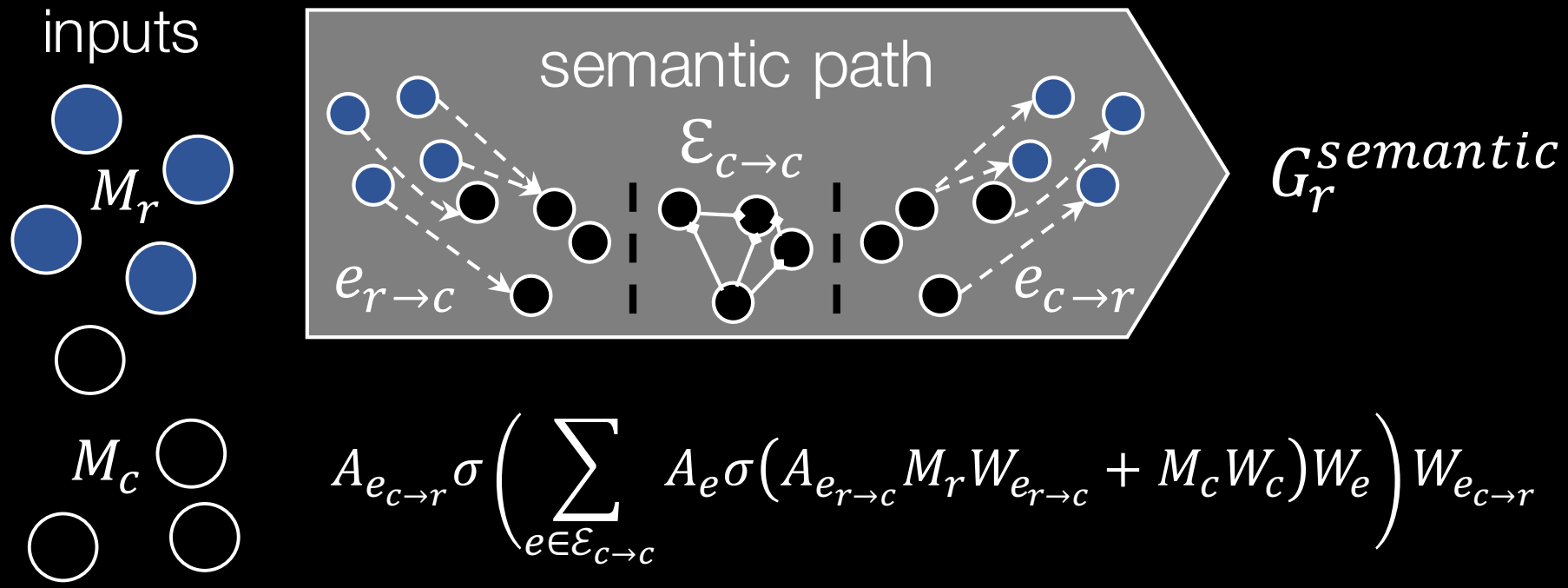
Graph based

Semantic Path: w/ Knowledge



activation function ReLU

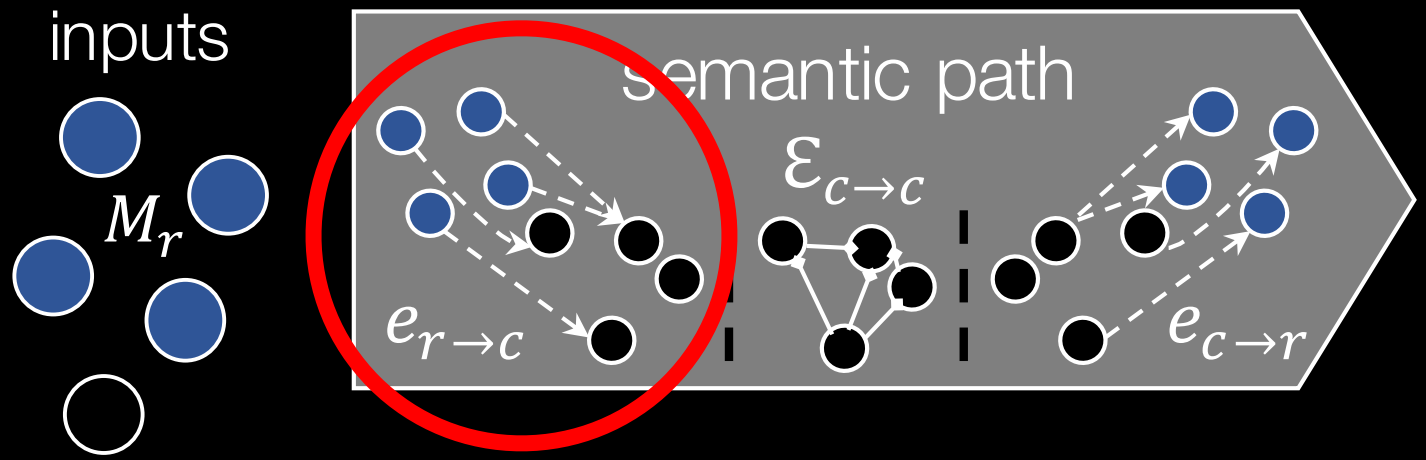
Semantic Path: w/ Knowledge



$$A_{e_{c\to r}}\sigma\left(\sum_{e\in\mathcal{E}_{c\to c}} A_e\sigma\left(A_{e_{r\to c}}M_rW_{e_{r\to c}} + M_cW_c\right)W_e\right)W_{e_{c\to r}}$$

activation function ReLU

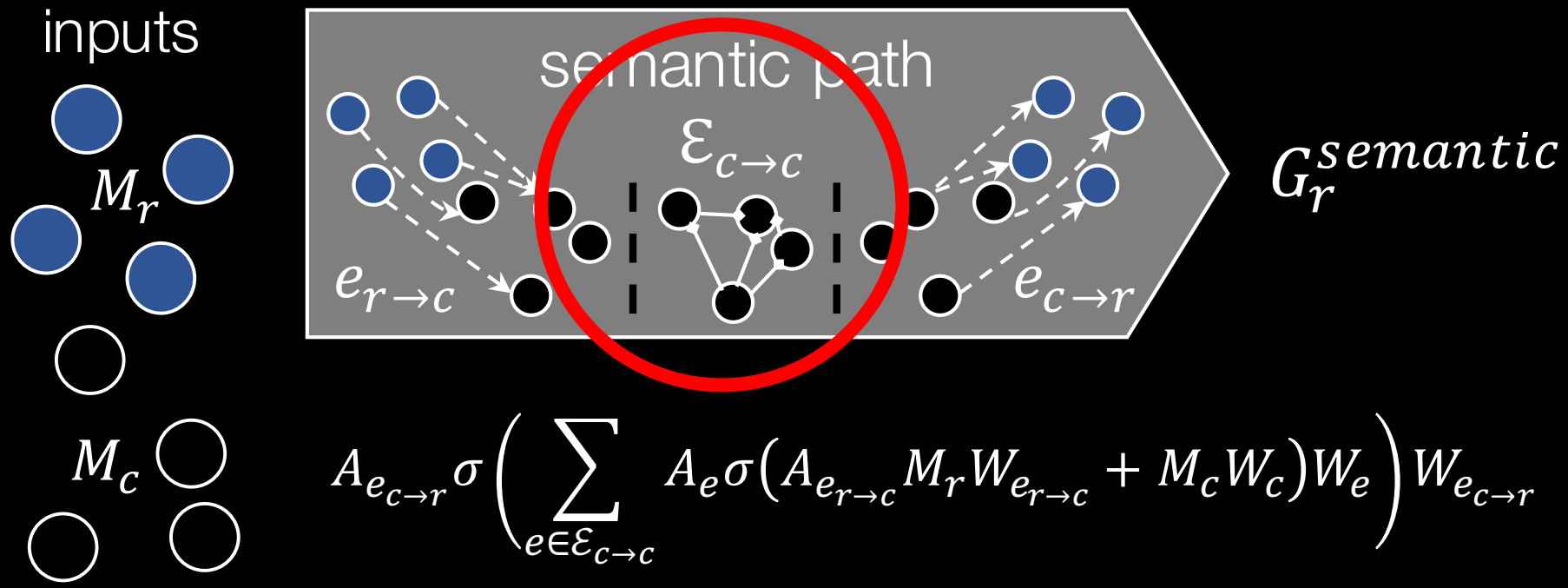
Semantic Path: w/ Knowledge



$$A_{e_{c\to r}}\sigma\left(\sum_{e\in\mathcal{E}_{c\to c}} A_e\sigma\left(A_{e_{r\to c}}M_rW_{e_{r\to c}} + M_cW_c\right)W_e\right)W_{e_{c\to r}}$$

activation function ReLU

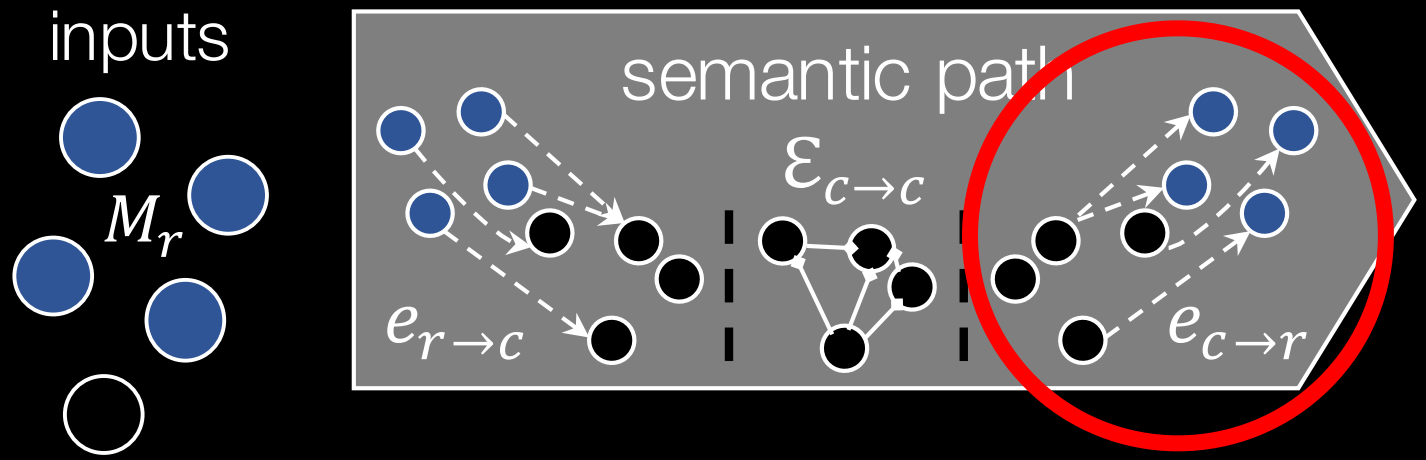
Semantic Path: w/ Knowledge



$$A_{e_{c\to r}}\sigma\left(\sum_{e\in\mathcal{E}_{c\to c}} A_e\sigma\left(A_{e_{r\to c}}M_rW_{e_{r\to c}} + M_cW_c\right)W_e\right)W_{e_{c\to r}}$$

activation function ReLU

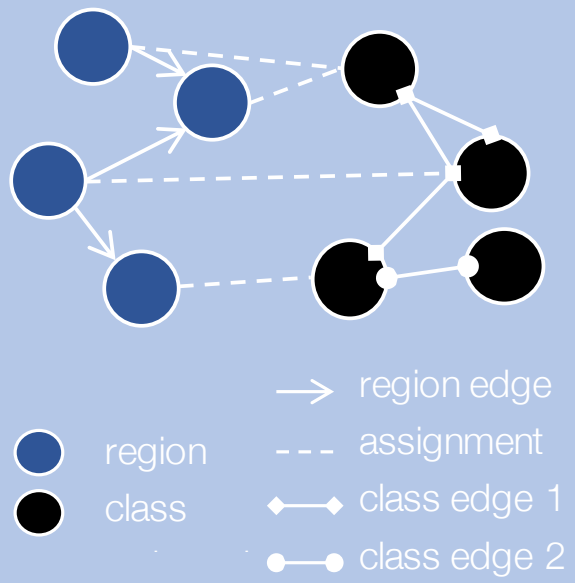
Semantic Path: w/ Knowledge



$$A_{e_{c\to r}}\sigma\left(\sum_{e\in\mathcal{E}_{c\to c}} A_e\sigma\left(A_{e_{r\to c}}M_rW_{e_{r\to c}} + M_cW_c\right)W_e\right)W_{e_{c\to r}}$$

activation function ReLU

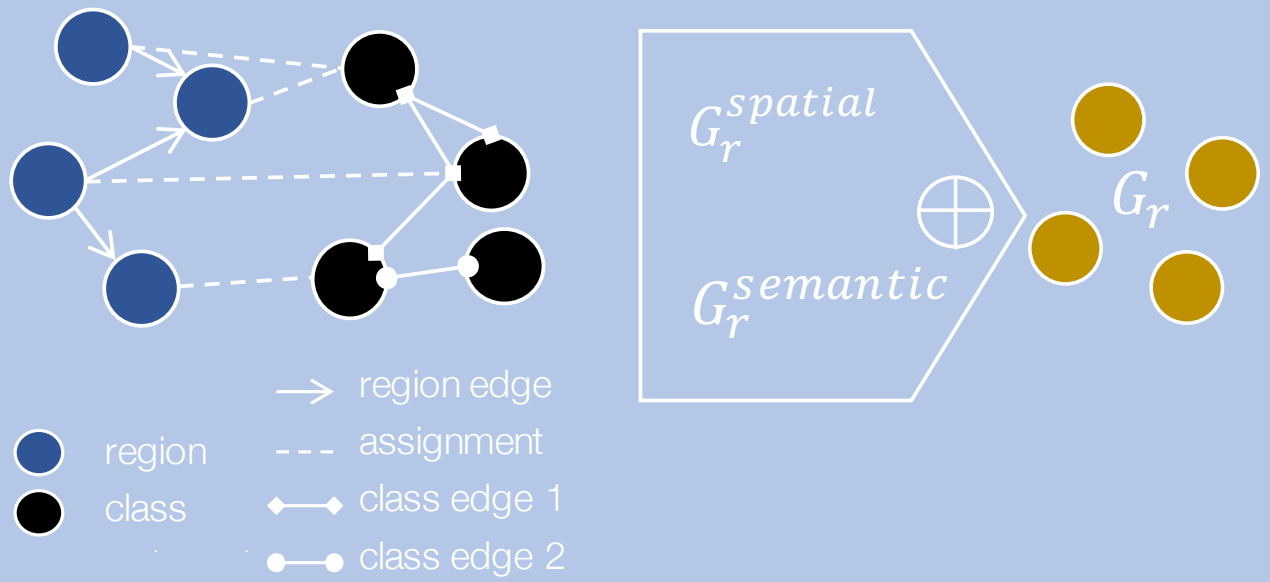
Global Module: Graph Reasoning



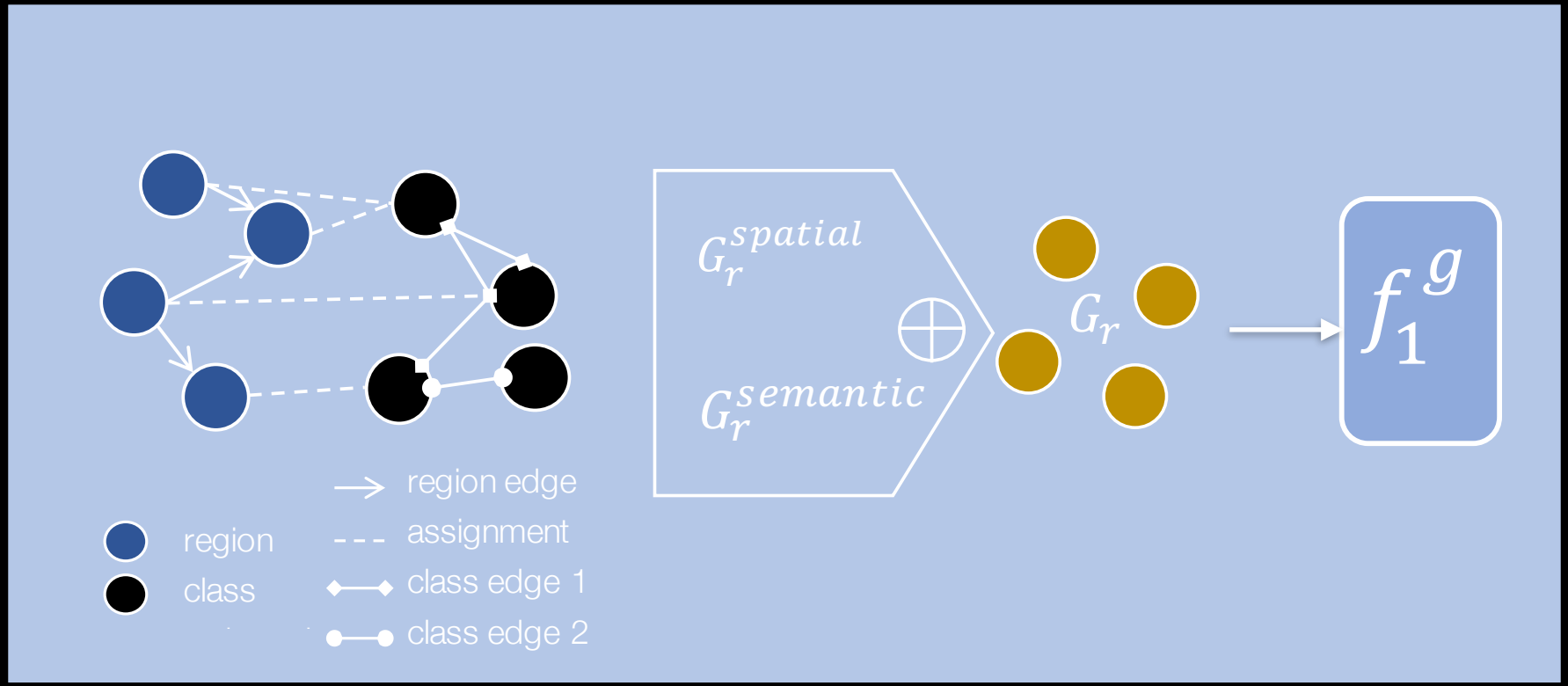
$G_r^{spatial}$

$G_r^{semantic}$

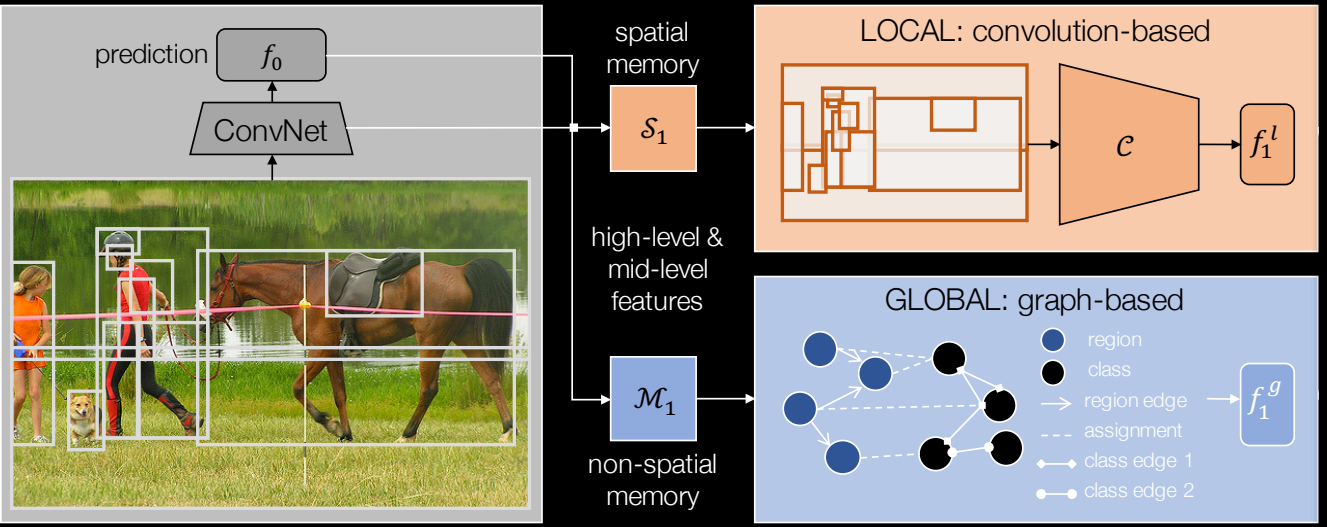
Global Module: Graph Reasoning



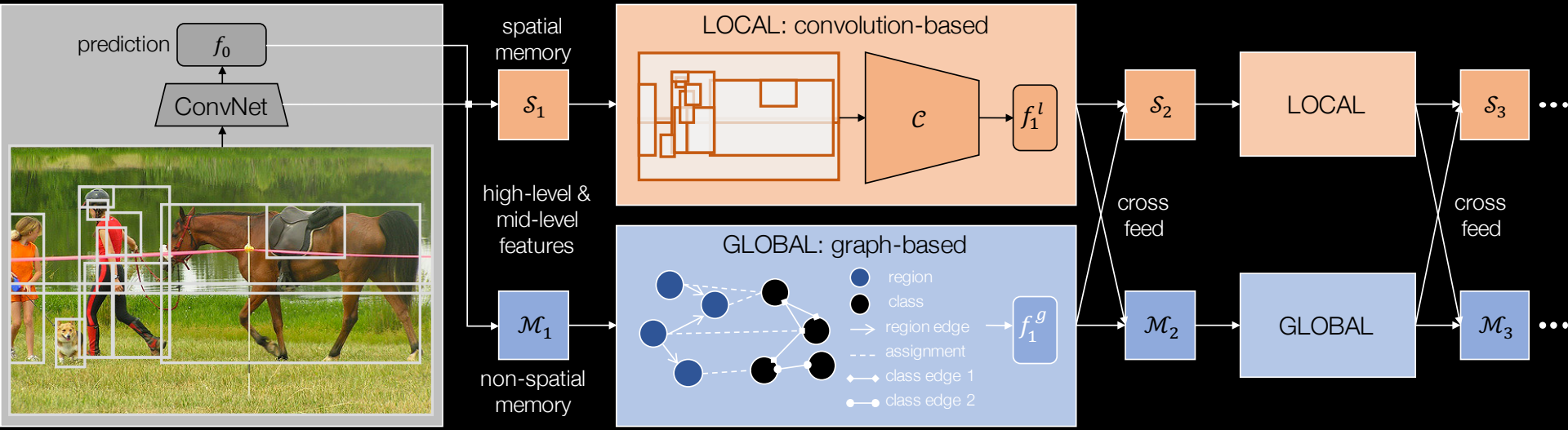
Global Module: Graph Reasoning



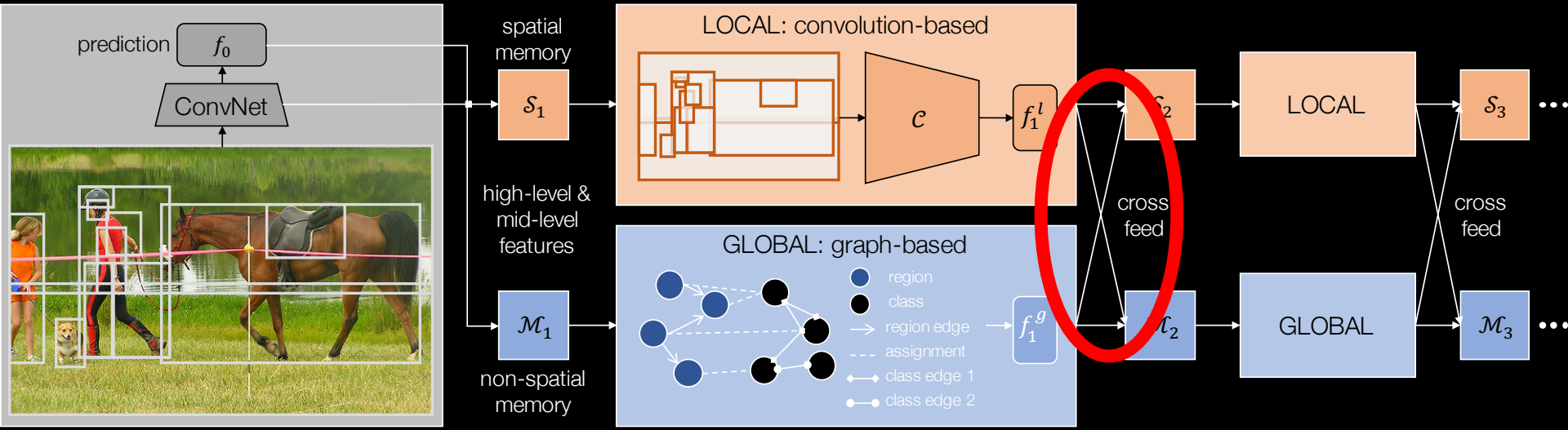
Iterative Predictions



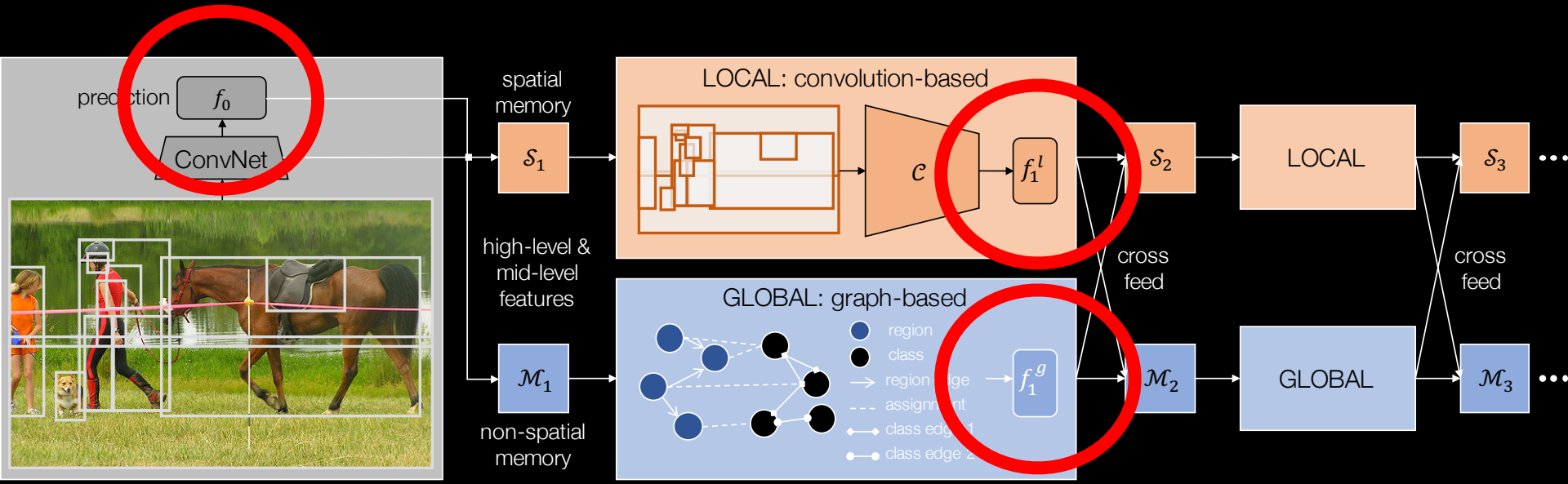
Iterative Predictions



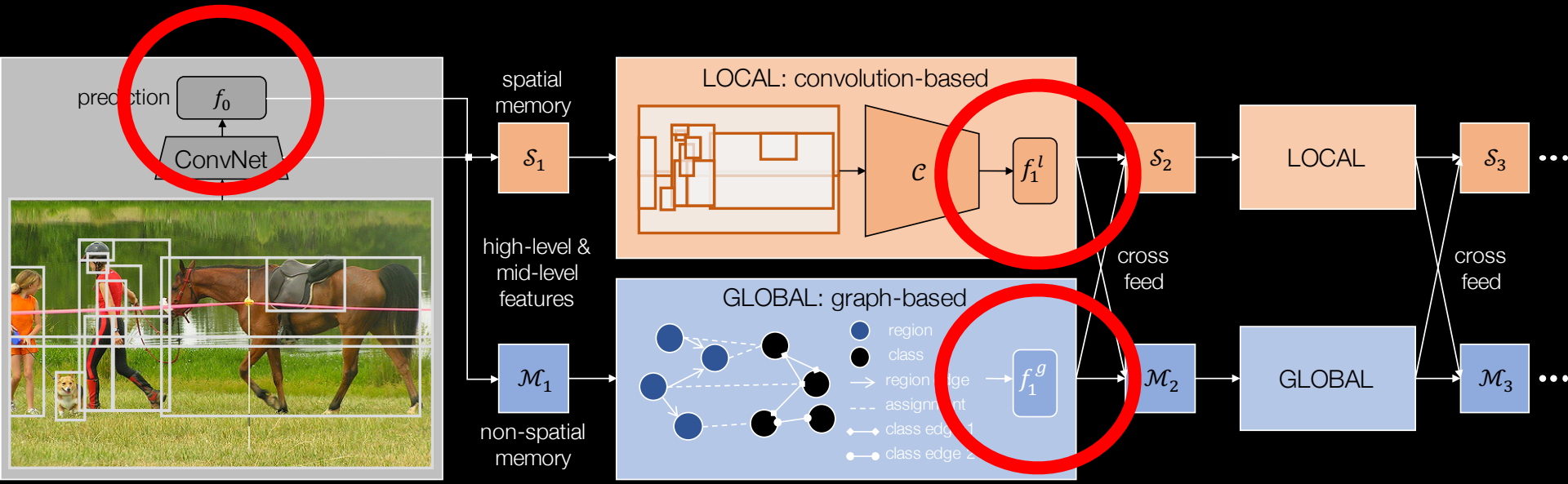
Iterative Predictions



Iterative Predictions

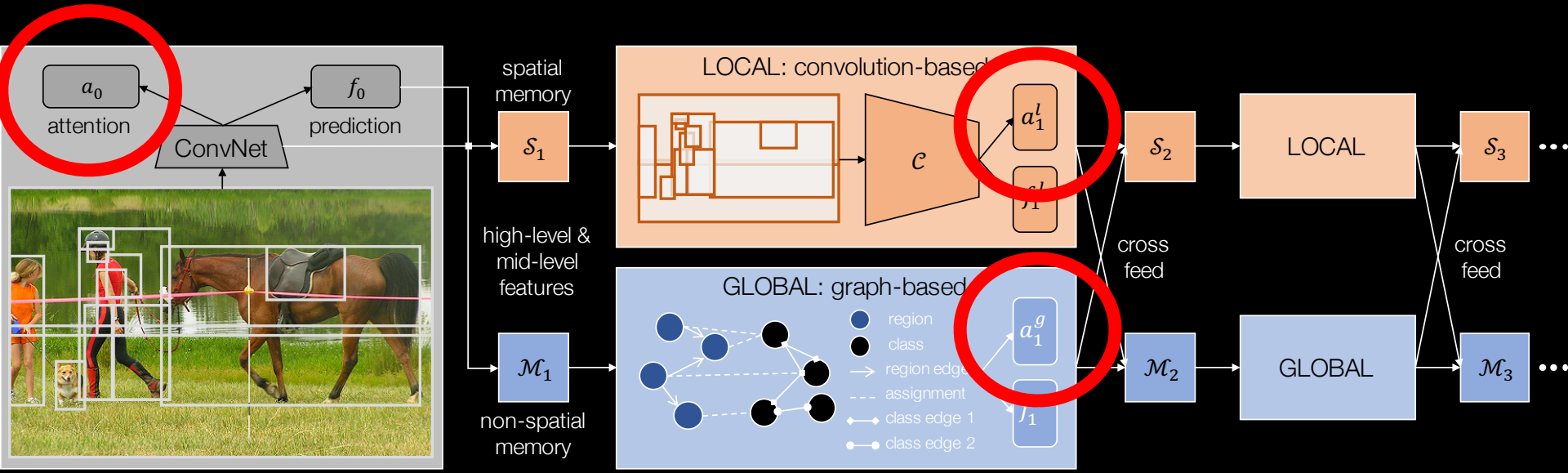


Iterative Predictions

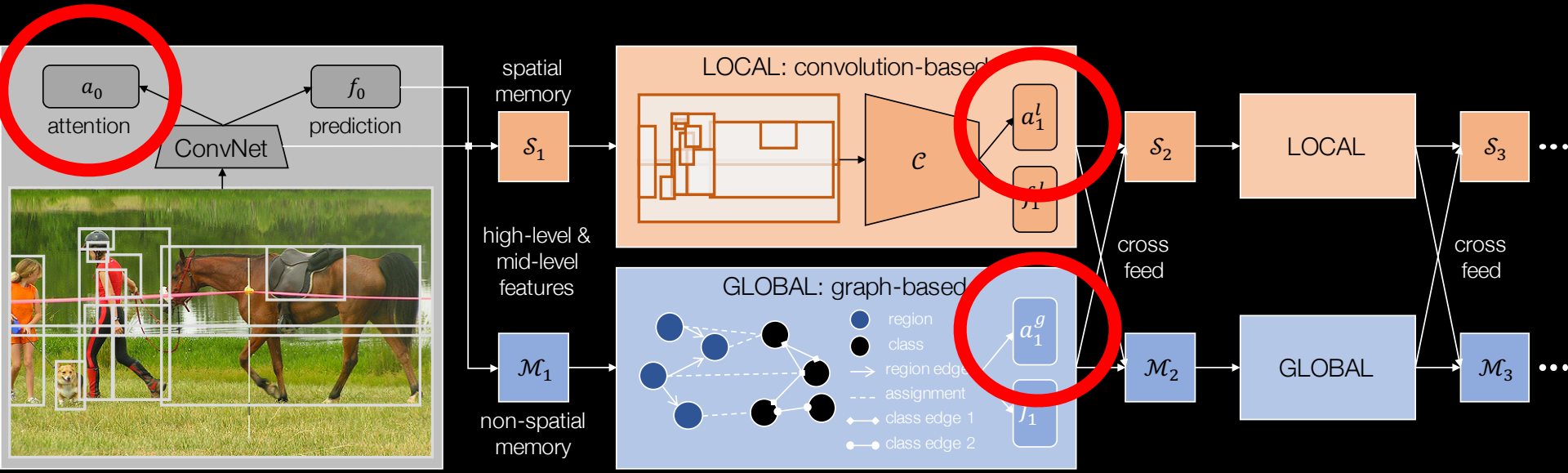


How to Combine Predictions

Combine Predictions: Attention



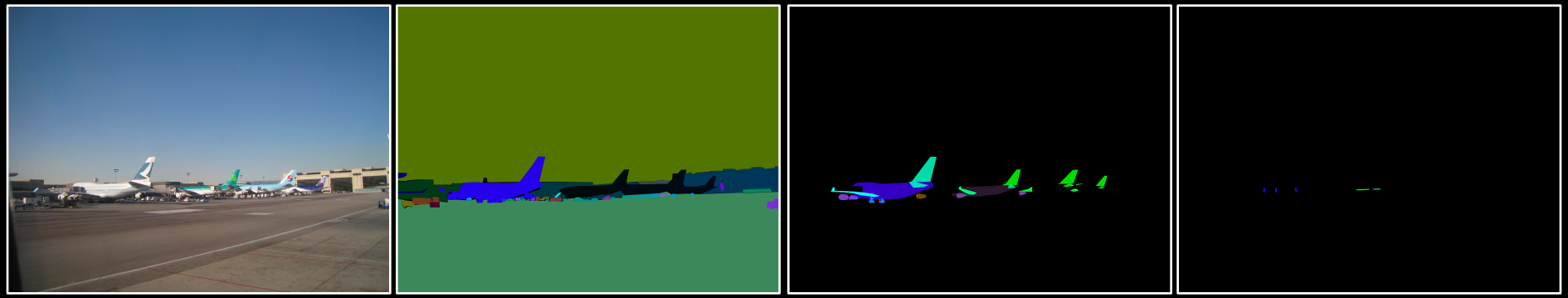
Combine Predictions: Attention



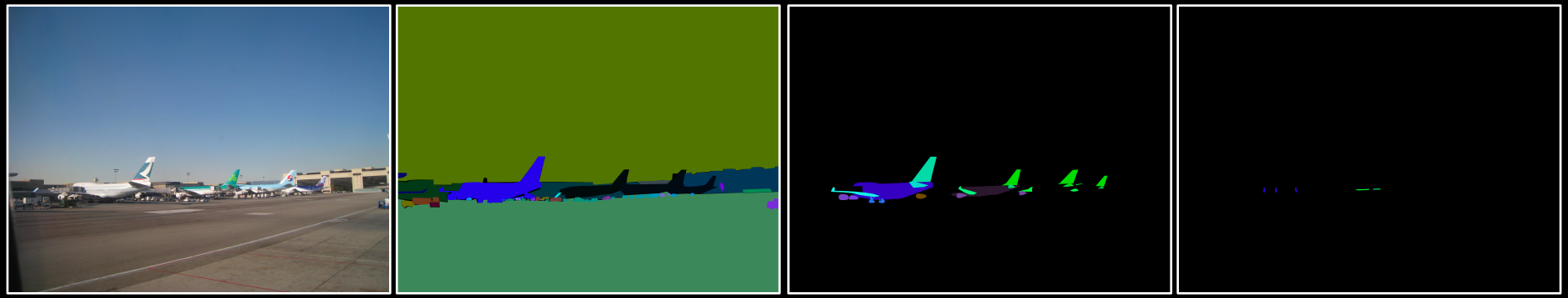
$$f = \sum_t w_t f_t, \qquad w_t = \frac{\exp(-a_t)}{\sum_{t'} \exp(-a_{t'})}$$

Experimental Results

ADE20K, Converted

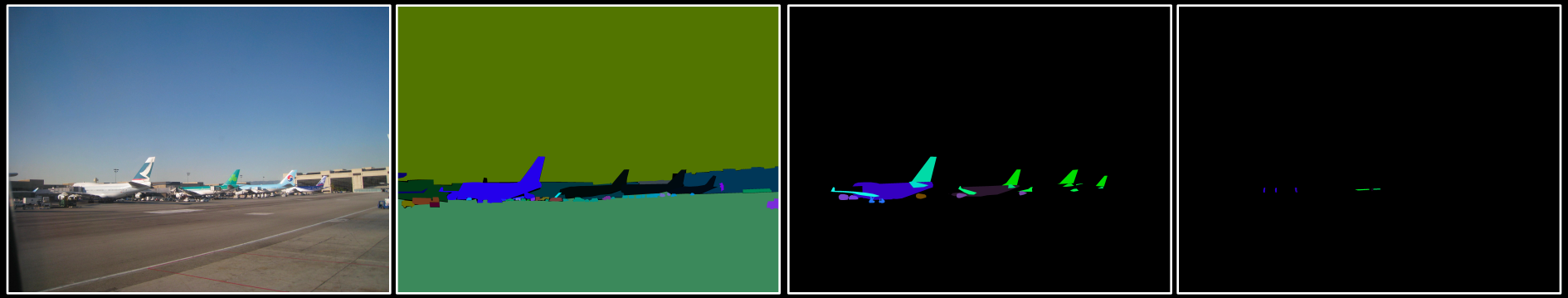


ADE20K, Converted



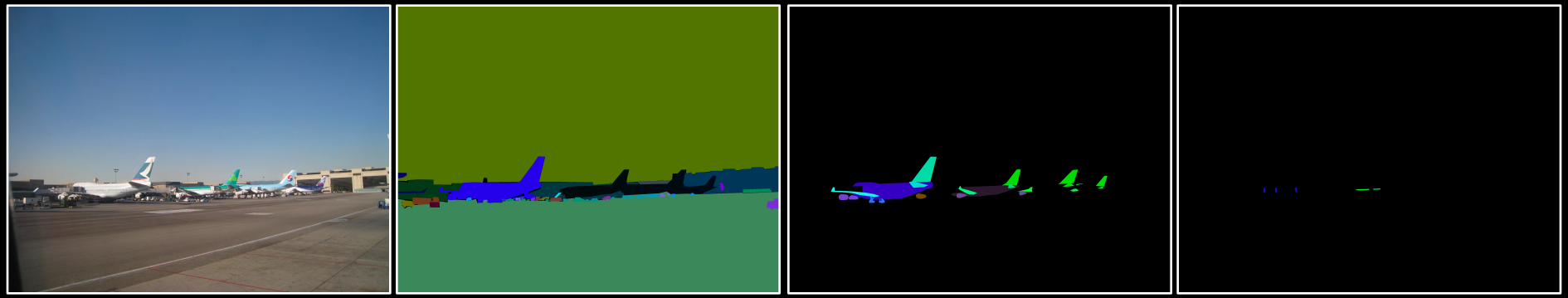
- stuff, object, part, part-of-part,…

ADE20K, Converted



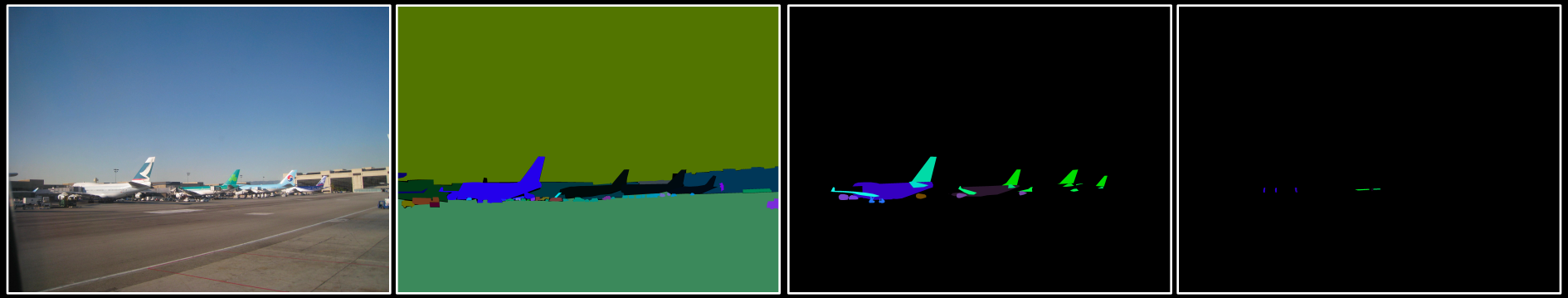
- stuff, object, part, part-of-part,…
- segments converted to **bounding boxes**

ADE20K, Converted



- stuff, object, part, part-of-part,…
- segments converted to **bounding boxes**
- 20.2K training, 1K validation, 1K testing, 1.5K classes

ADE20K, Converted



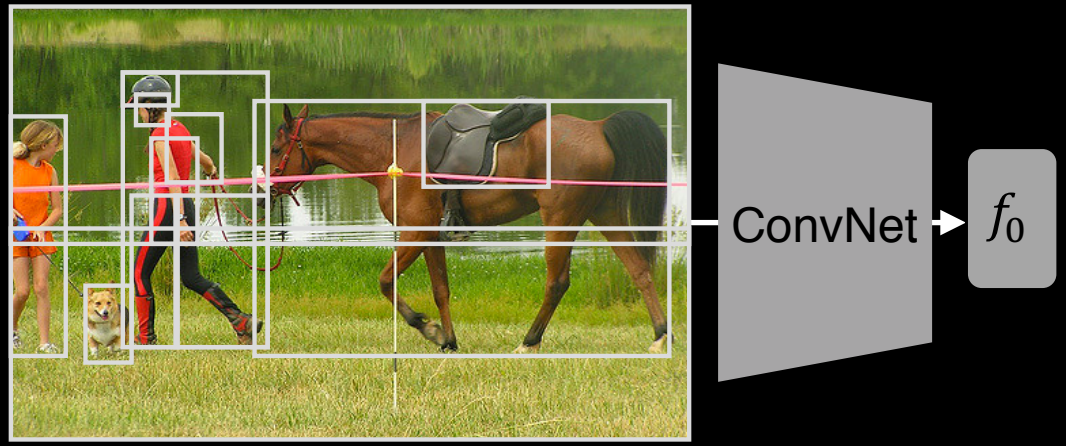
- stuff, object, part, part-of-part,…
- segments converted to **bounding boxes**
- 20.2K training, 1K validation, 1K testing, 1.5K classes
- relationships
  - is-a, is-kind-of, is-part-of, *etc*

Quantitative Results

Quantitative Results

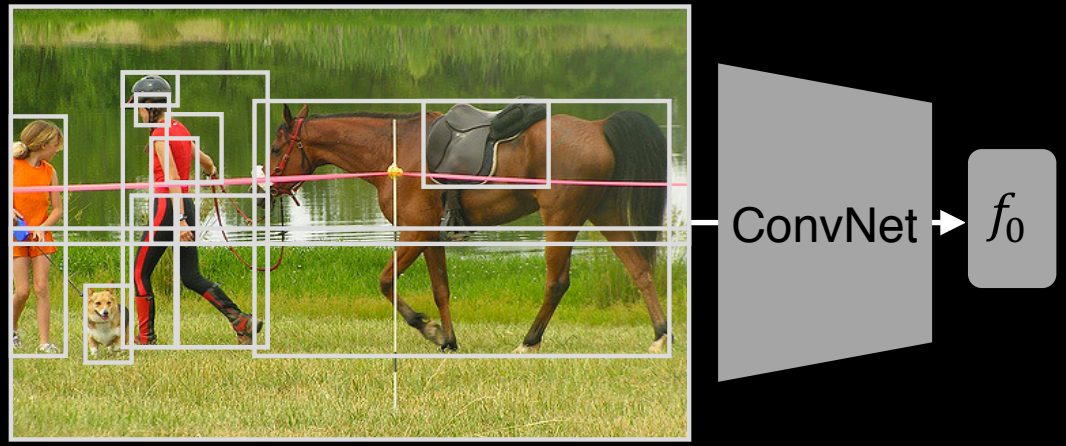
AP
per-class
per-instance

Quantitative Results



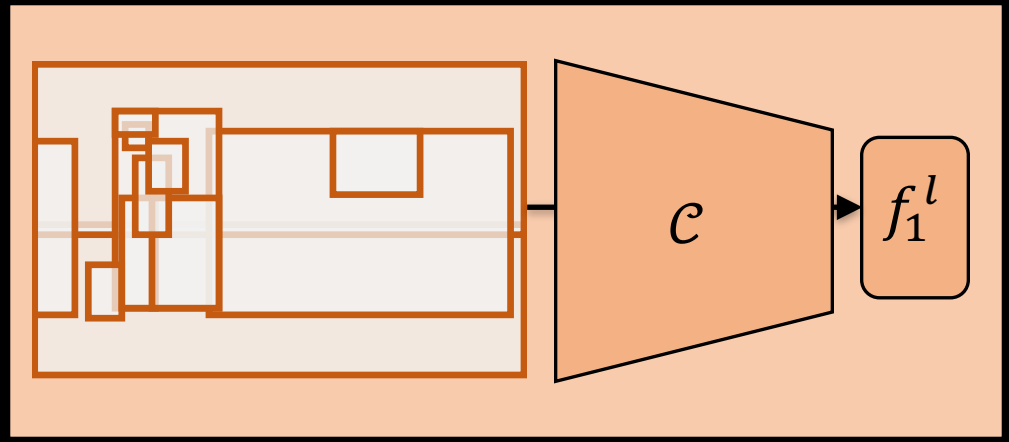
AP	Res-50
per-class	40.1
per-instance	67.0

Quantitative Results



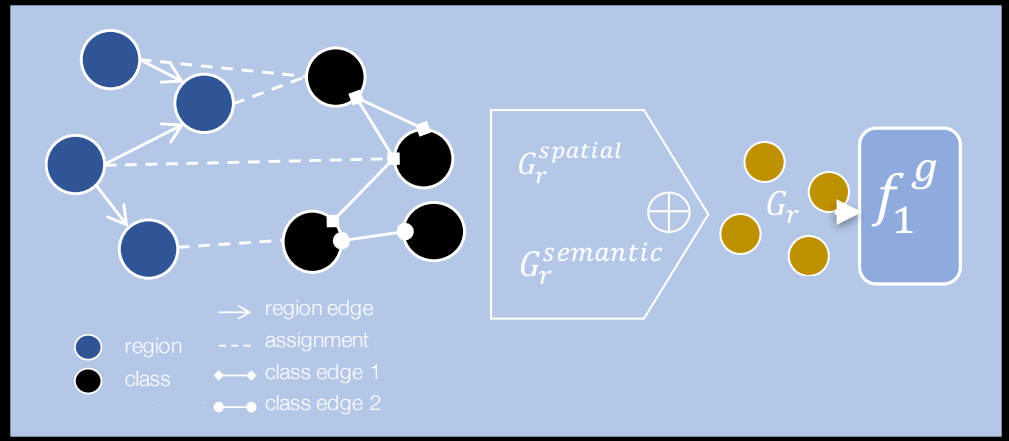
AP	Res-50	Res-101	High-Res
per-class	40.1	40.8	41.0
per-instance	67.0	68.2	68.2

Quantitative Results



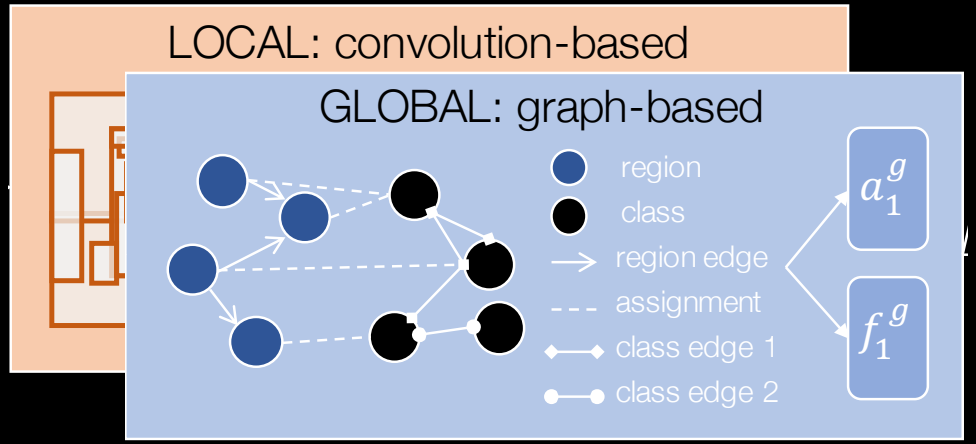
AP	Res-50	Res-101	High-Res	Local
per-class	40.1	40.8	41.0	47.9
per-instance	67.0	68.2	68.2	71.6

Quantitative Results



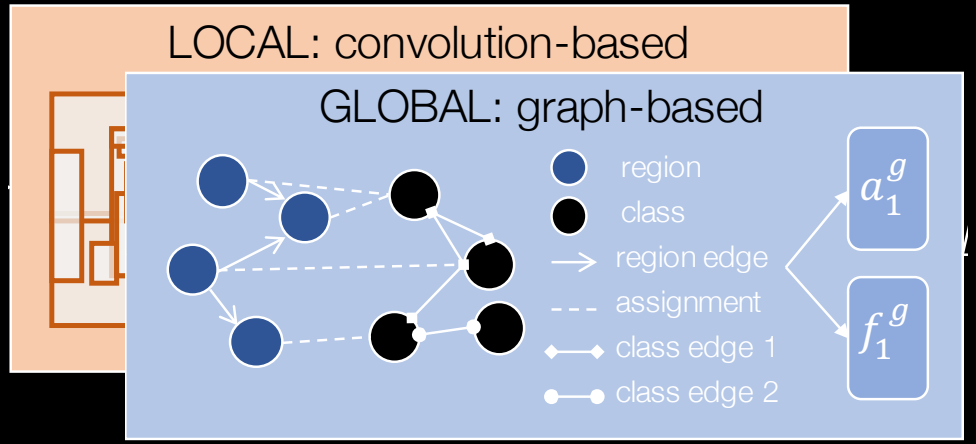
AP	Res-50	Res-101	High-Res	Local	Global
per-class	40.1	40.8	41.0	47.9	44.5
per-instance	67.0	68.2	68.2	71.6	69.8

Quantitative Results



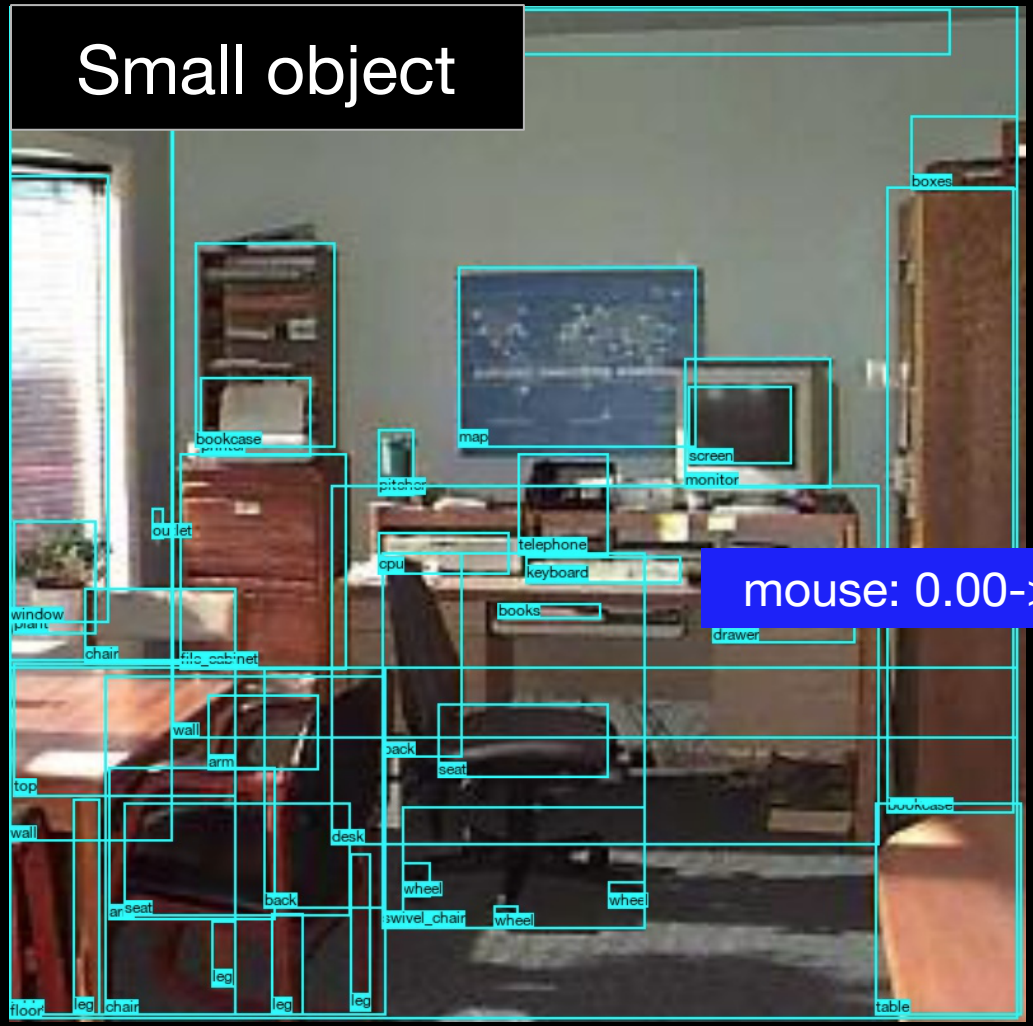
AP	Res-50	Res-101	High-Res	Local	Global	Ours
per-class	40.1	40.8	41.0	47.9	44.5	**48.5**
per-instance	67.0	68.2	68.2	71.6	69.8	**72.6**

Quantitative Results

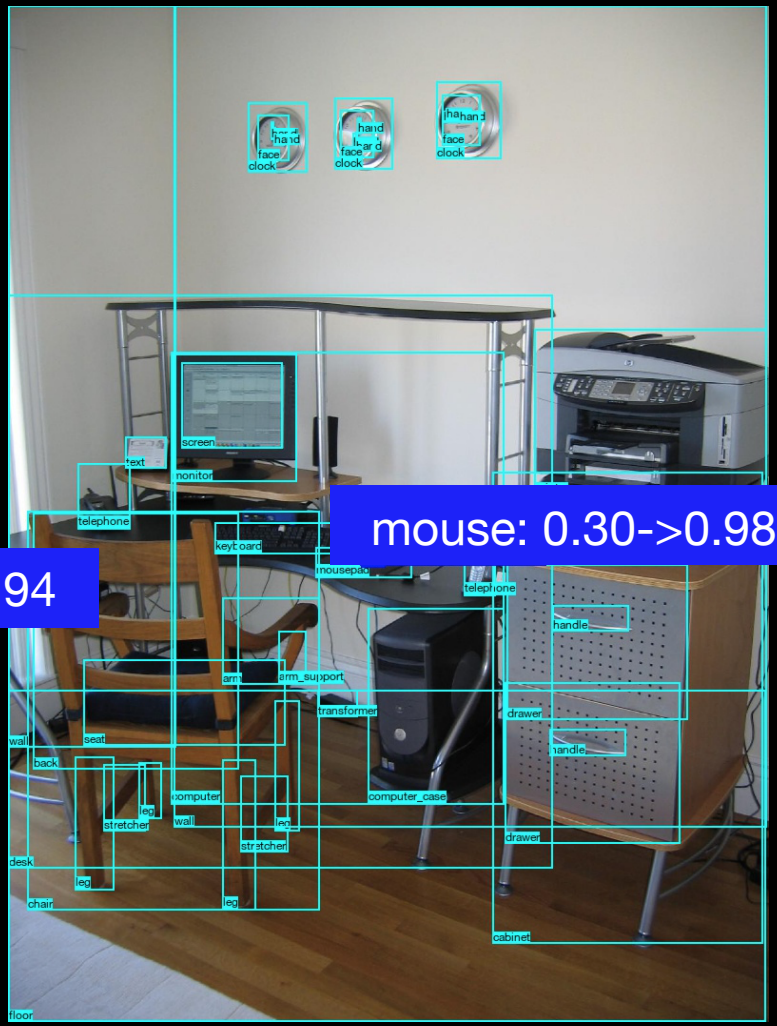


AP	Res-50	Res-101	High-Res	Local	Global	Ours
per-class	40.1	40.8	41.0	47.9	44.5	**48.5**
per-instance	67.0	68.2	68.2	71.6	69.8	**72.6**

Small object

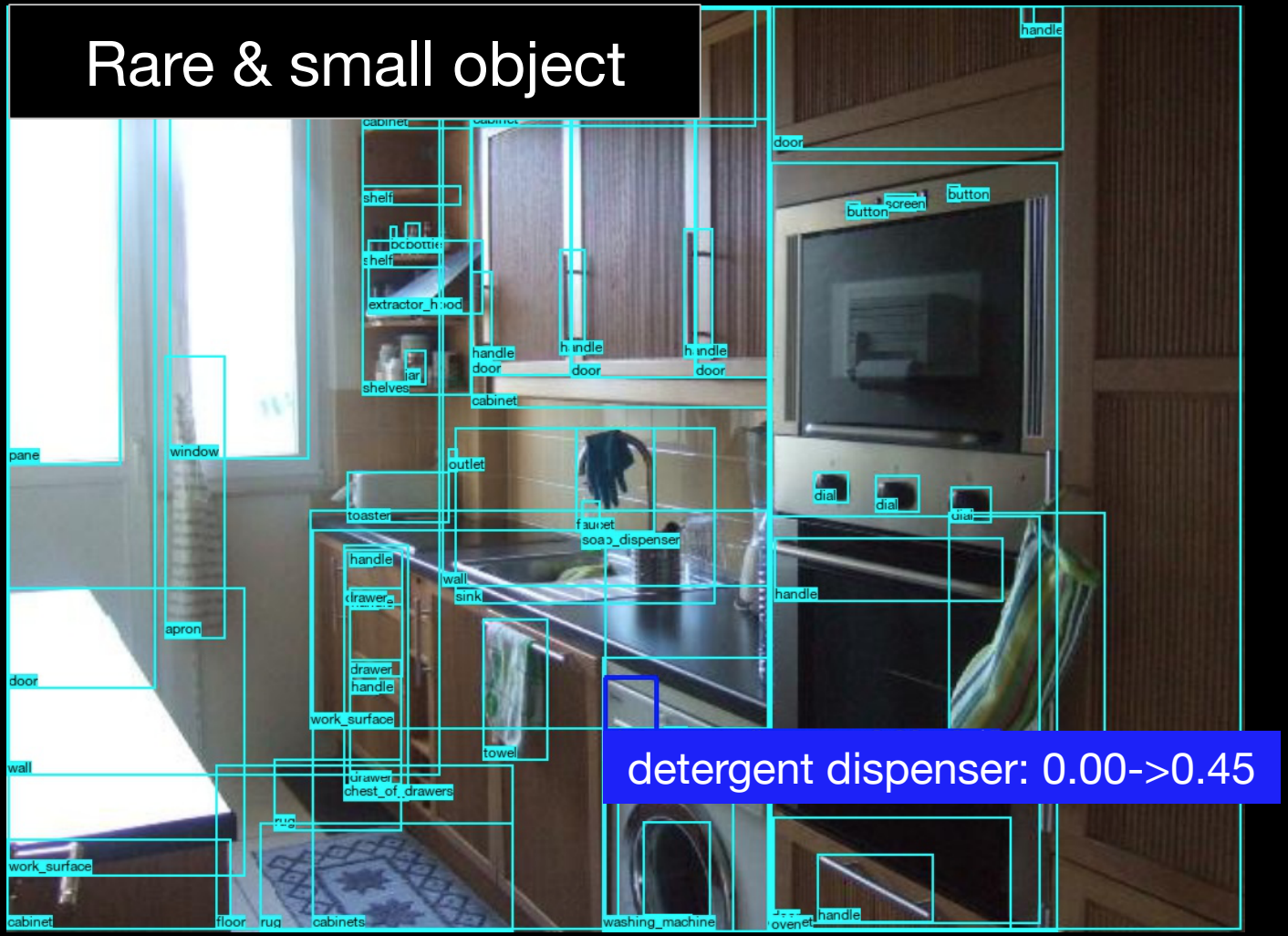


mouse: 0.00->0.94

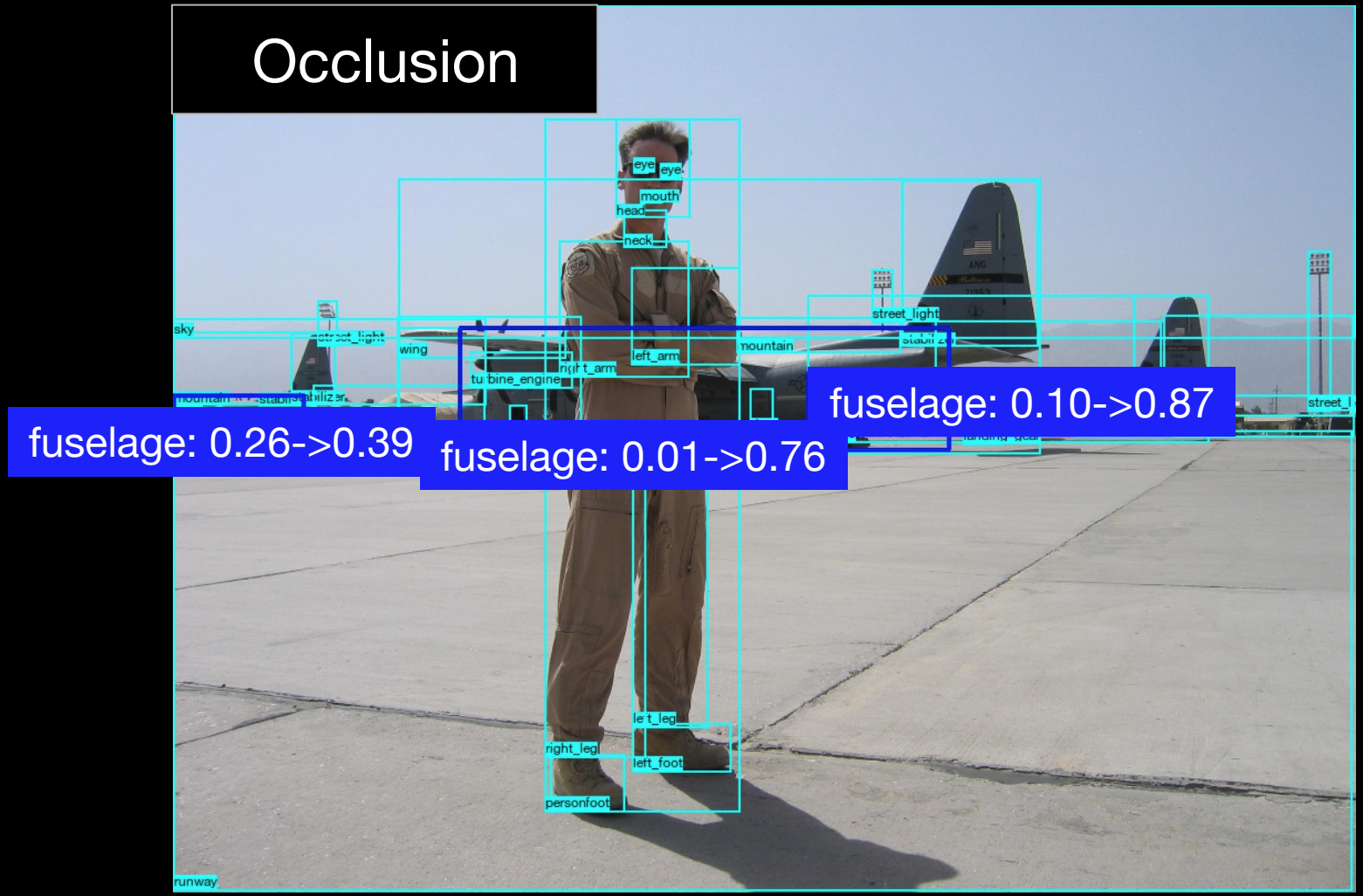


mouse: 0.30->0.98

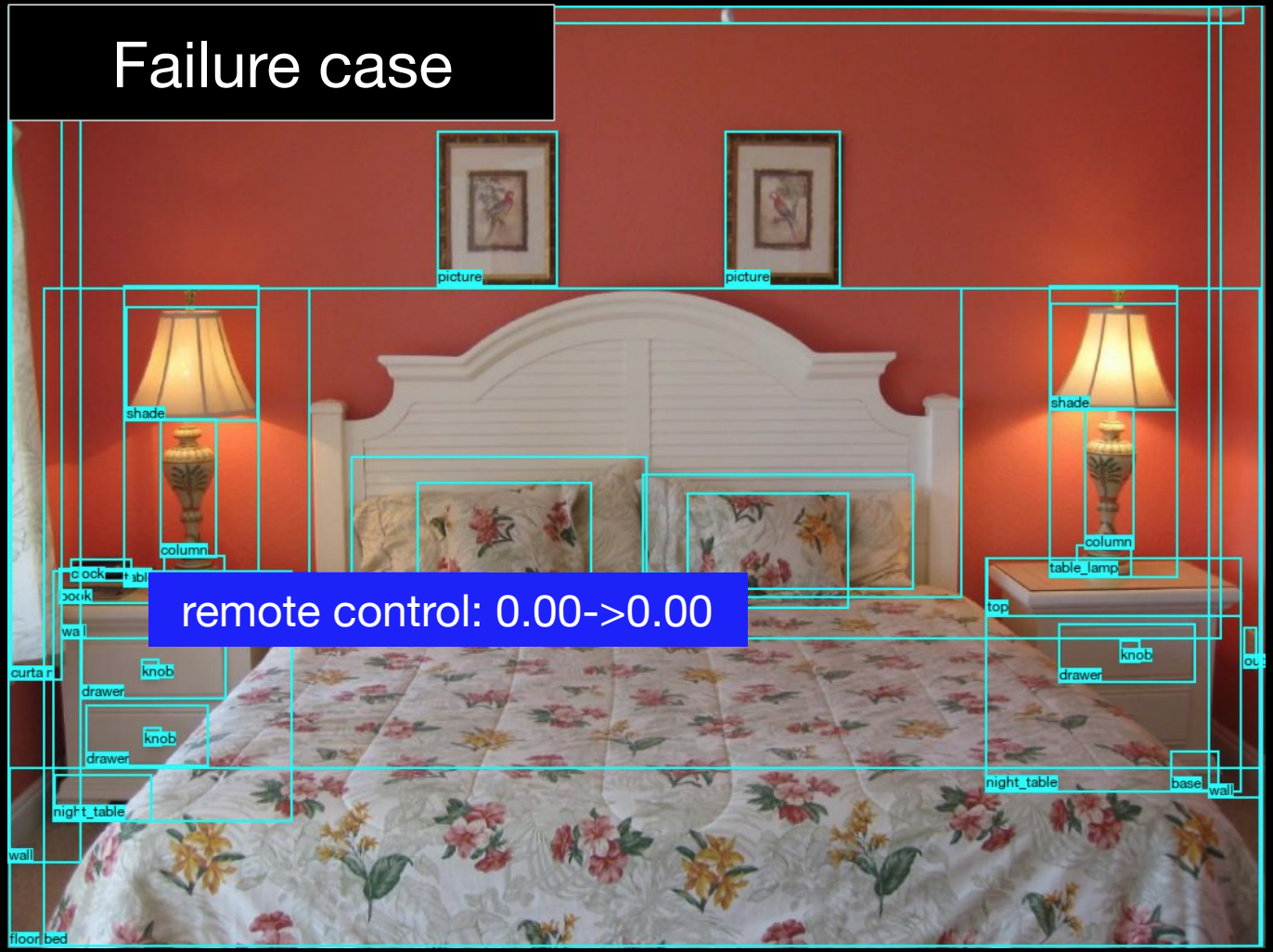
Rare & small object



Occlusion

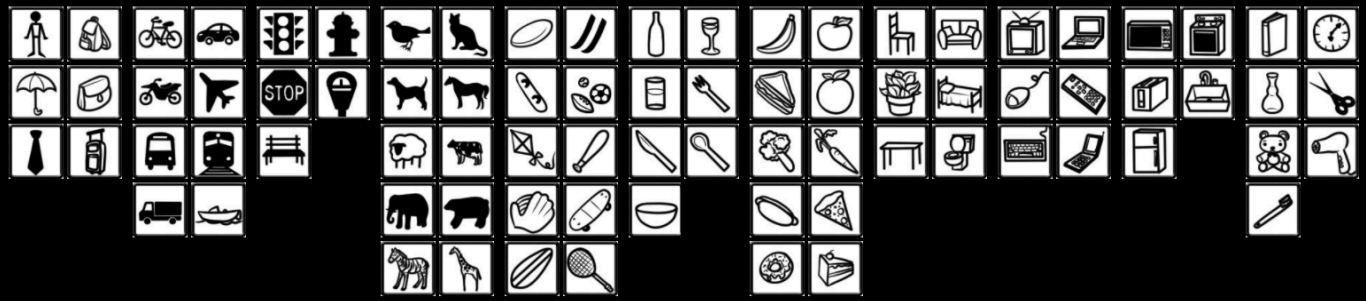

fuselage: 0.26->0.39
fuselage: 0.01->0.76
fuselage: 0.10->0.87
eye
eye
mouth
head
neck
sky
wing
left_arm
mountain
street_light
tu bine_engine
right_arm
left_leg
right_leg
left_foot
personfoot
runway

Failure case

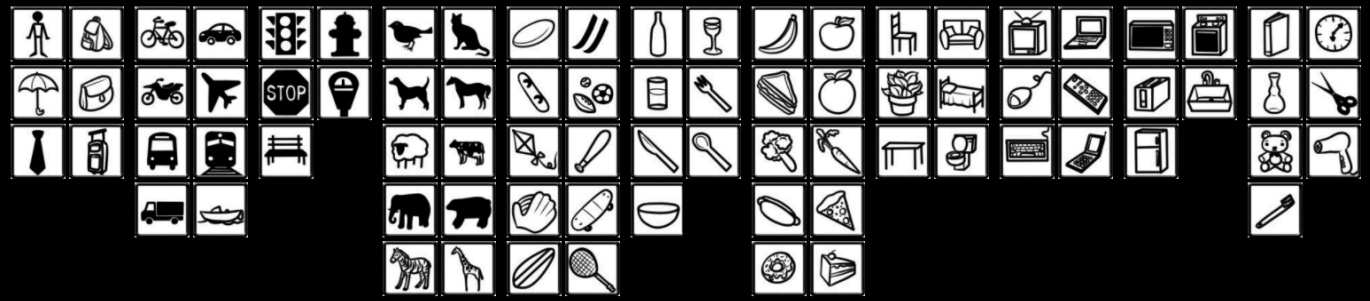


remote control: 0.00->0.00

COCO Detection Dataset

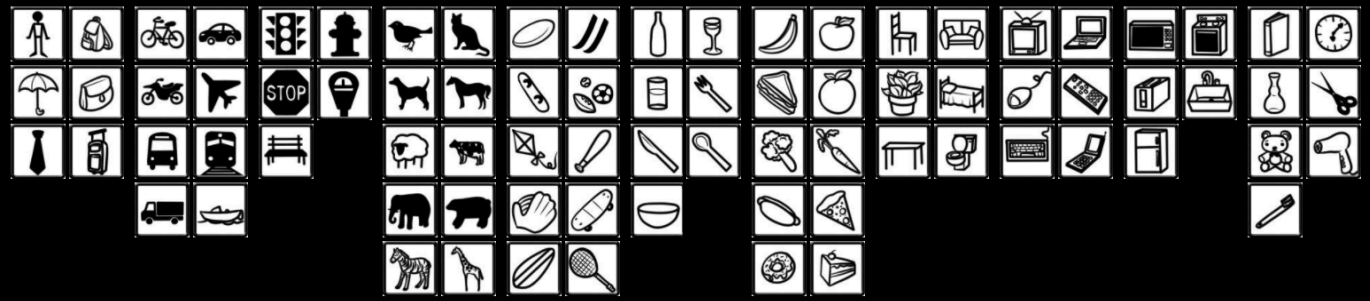


COCO Detection Dataset



- **NO** knowledge graph

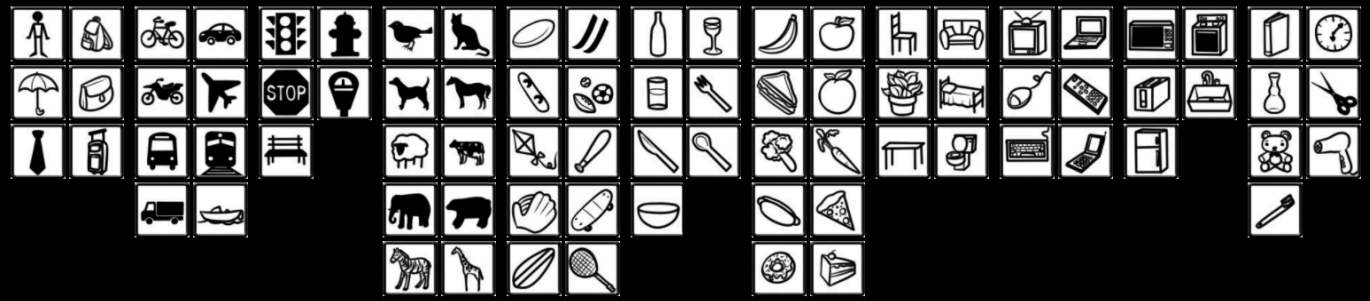
COCO Detection Dataset



- **NO** knowledge graph

AP	Res-50
per-class	83.7
per-instance	83.2

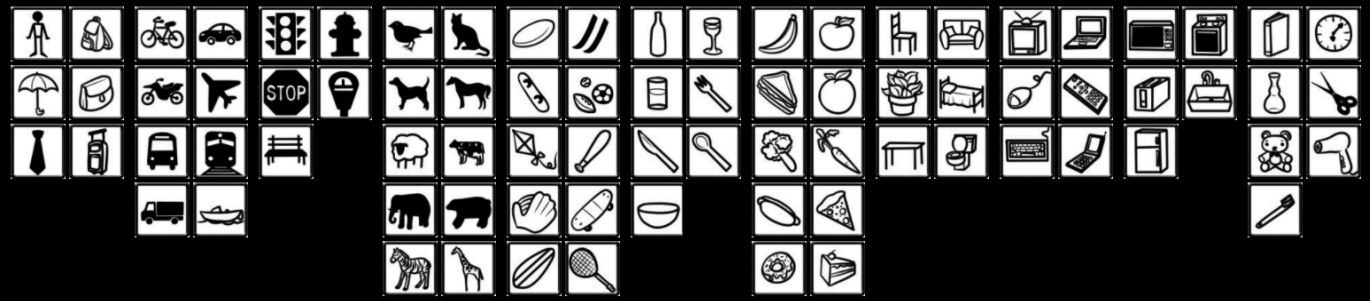
COCO Detection Dataset



- **NO** knowledge graph

AP	Res-50	Local
per-class	83.7	85.8
per-instance	83.2	84.9

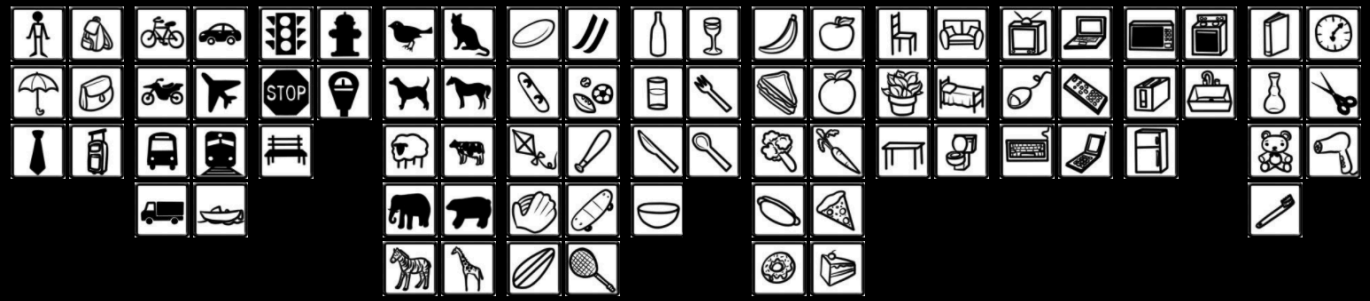
COCO Detection Dataset



- **NO** knowledge graph

AP	Res-50	Local	Global
per-class	83.7	85.8	86.9
per-instance	83.2	84.9	85.6

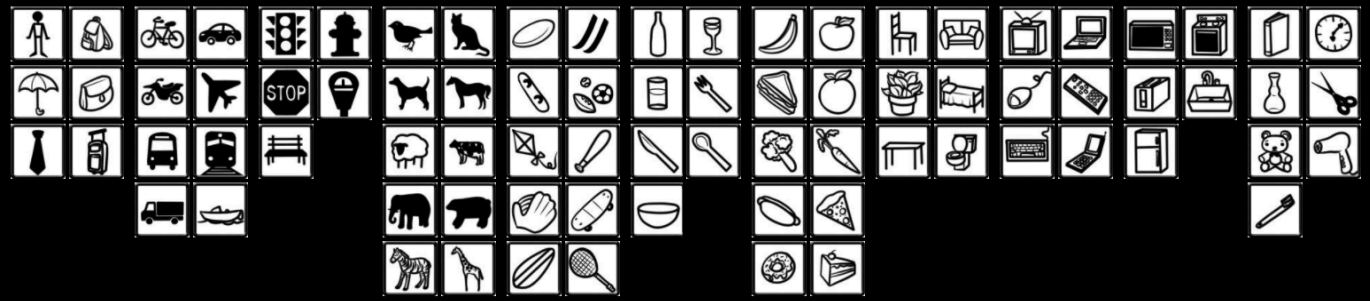
COCO Detection Dataset



- **NO** knowledge graph

AP	Res-50	Local	Global	Ours
per-class	83.7	85.8	86.9	**87.4**
per-instance	83.2	84.9	85.6	**86.0**

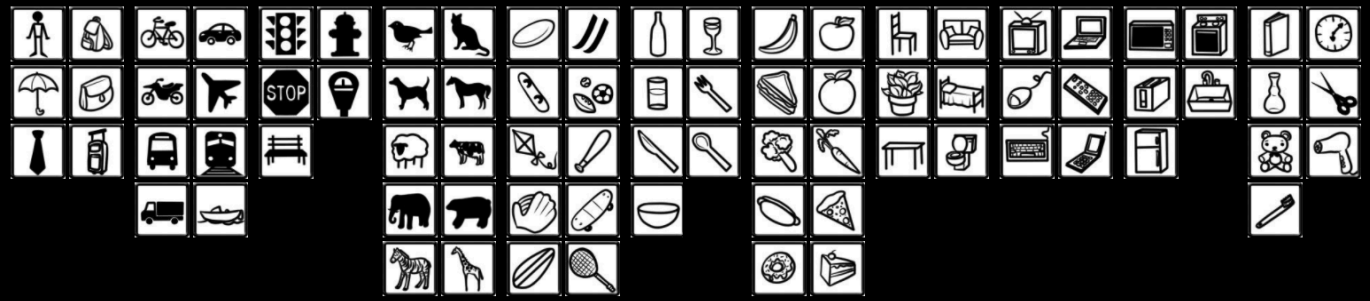
COCO Detection Dataset



- **NO** knowledge graph

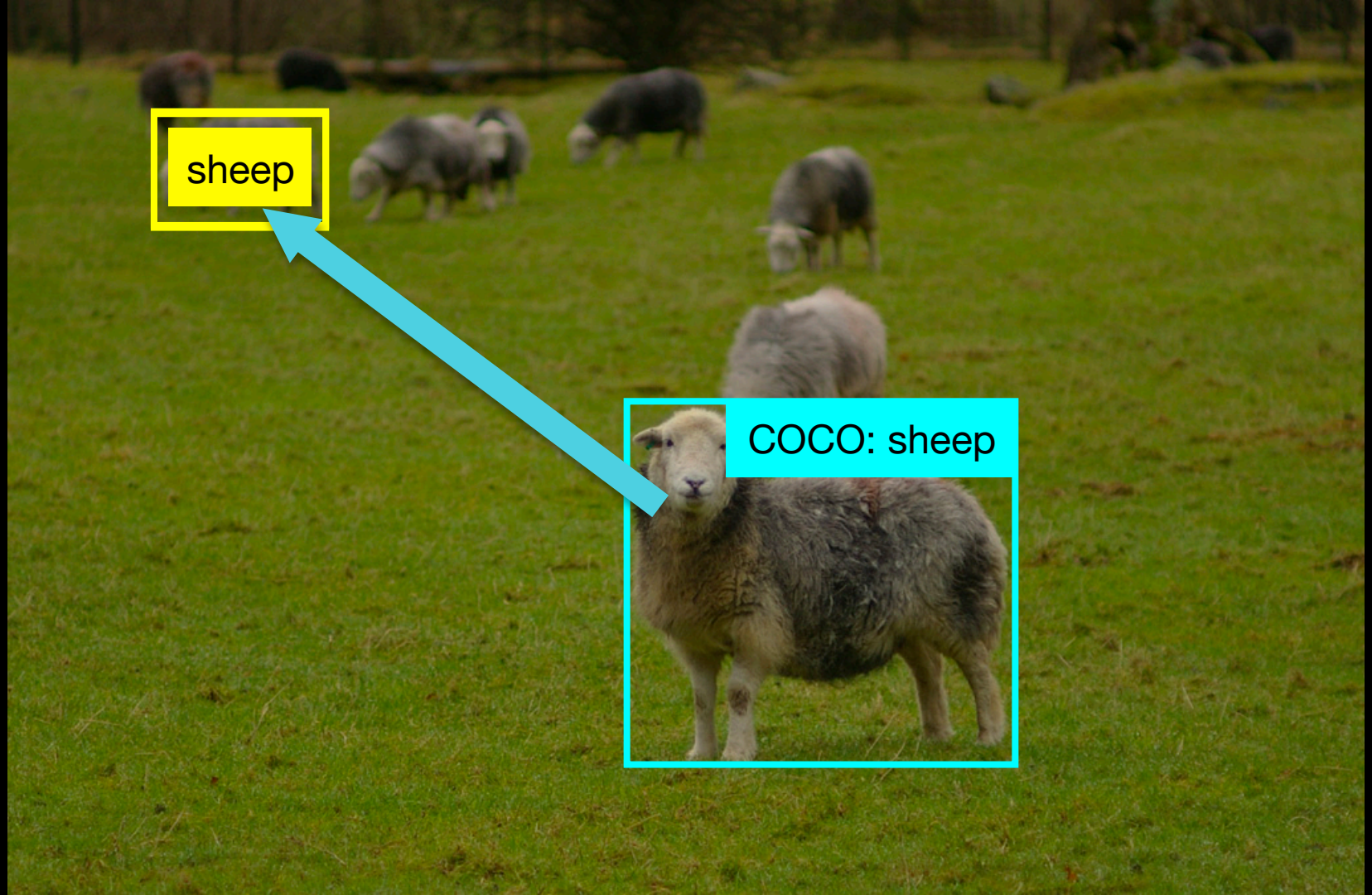
AP	Res-50	Local	Global	Ours
per-class	83.7	85.8	86.9	**87.4**
per-instance	83.2	84.9	85.6	**86.0**

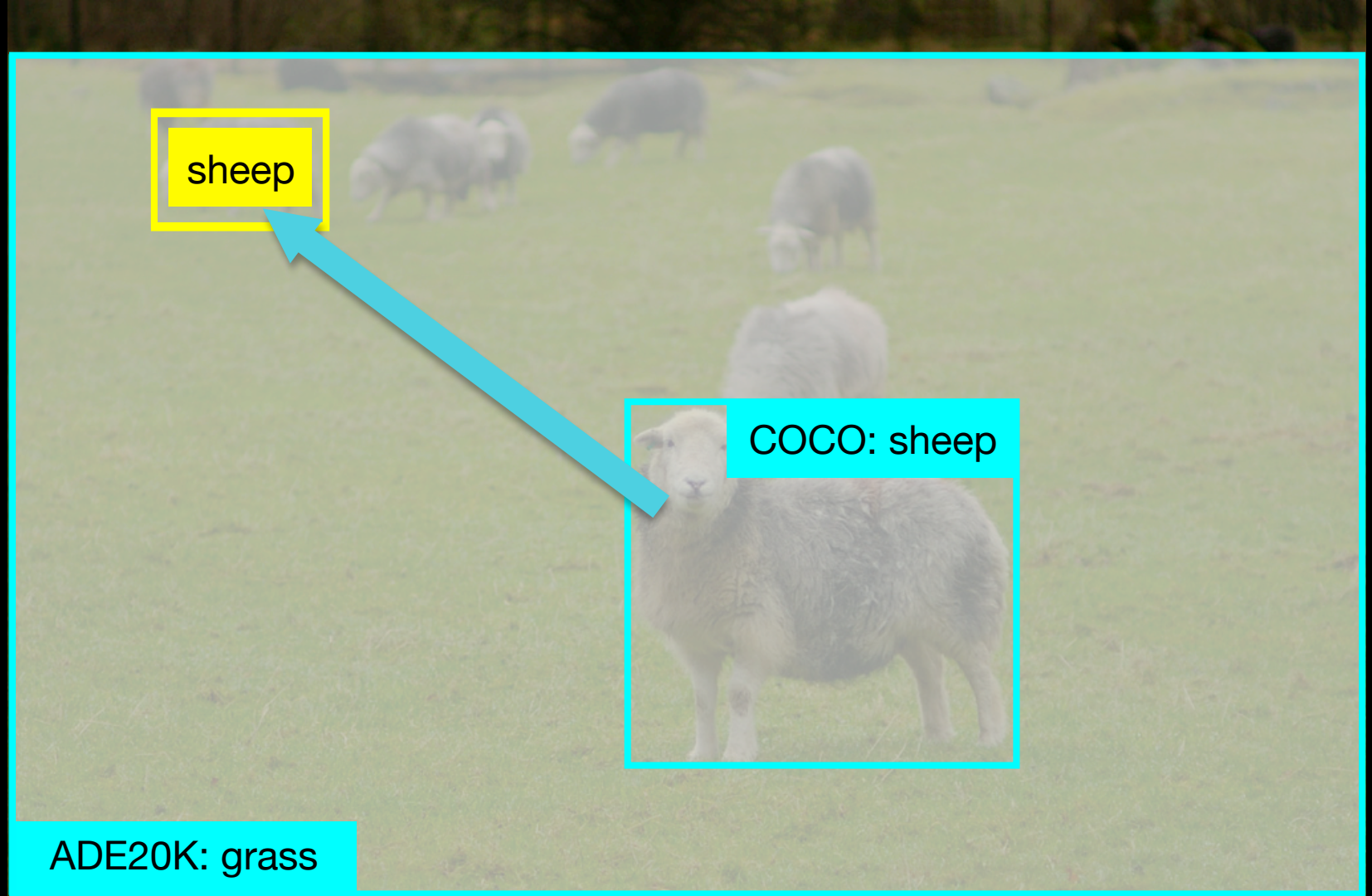
COCO Detection Dataset



- **NO** knowledge graph

AP	Res-50	Local	Global	Ours
per-class	83.7	85.8	86.9	**87.4**
per-instance	83.2	84.9	85.6	**86.0**


sheep
COCO: sheep


sheep
COCO: sheep
ADE20K: grass

Real-World Scenario: Missing Regions

region proposal methods can fail…

Real-World Scenario: Missing Regions

region proposal methods can fail…

Is our reasoning framework **ROBUST** to missing regions in current detectors?

Region Dropping

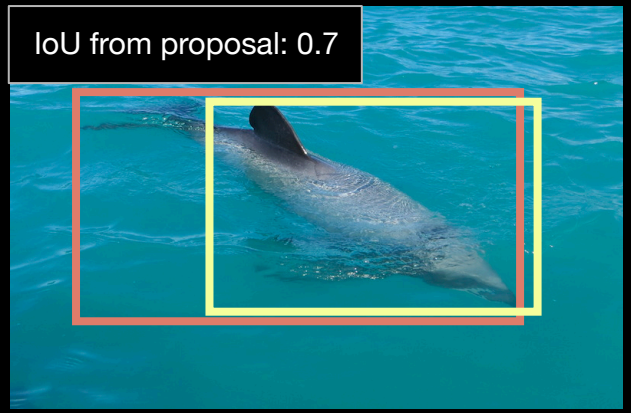
- Idea
  - filter out **hard** ground truth regions

Region Dropping

- Idea
  - filter out **hard** ground truth regions
- Hardness metric
  - IoU from **best** proposal

Region Dropping

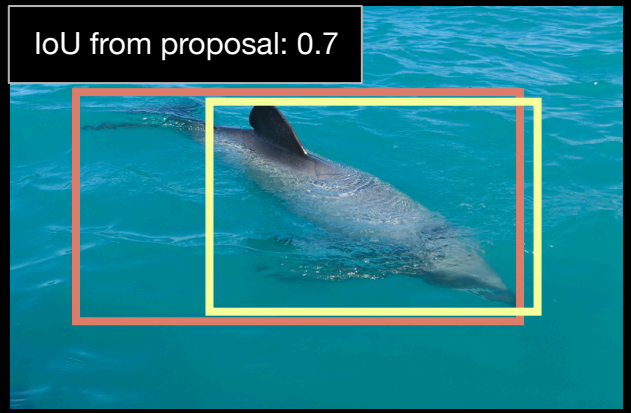
- Idea
  - filter out **hard** ground truth regions
- Hardness metric
  - IoU from **best** proposal



Region Dropping

>=0.8!

- Idea
  - filter out **hard** ground truth regions
- Hardness metric
  - IoU from **best** proposal



Region Dropping

>=0.8!

- Idea
  - filter out **hard** ground truth regions
- Hardness metric
  - IoU from **best** proposal

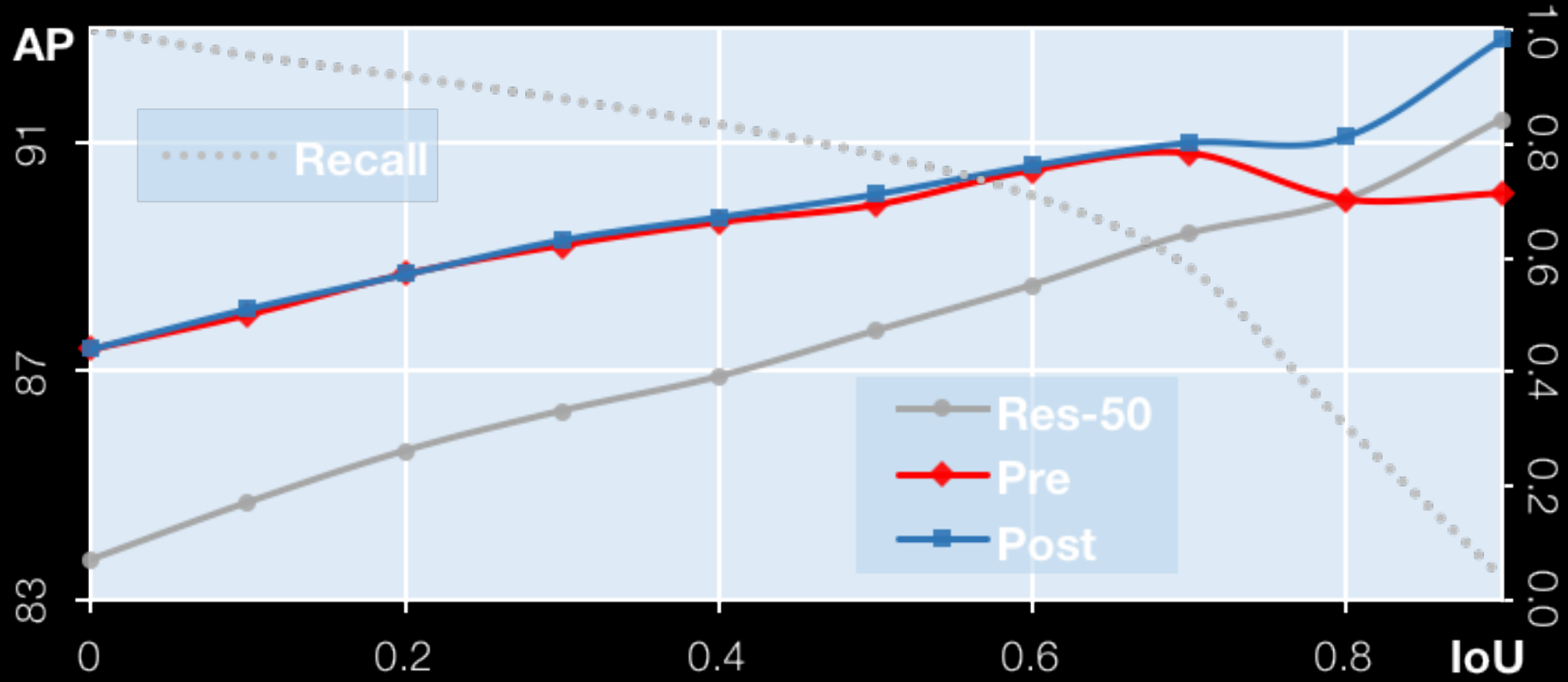
Region Dropping

>=0.8!

- Idea
  - filter out **hard** ground truth regions
- Hardness metric
  - IoU from **best** proposal
- Settings
  - **Pre:** filter **before** reasoning
  - **Post**: filter **after** reasoning

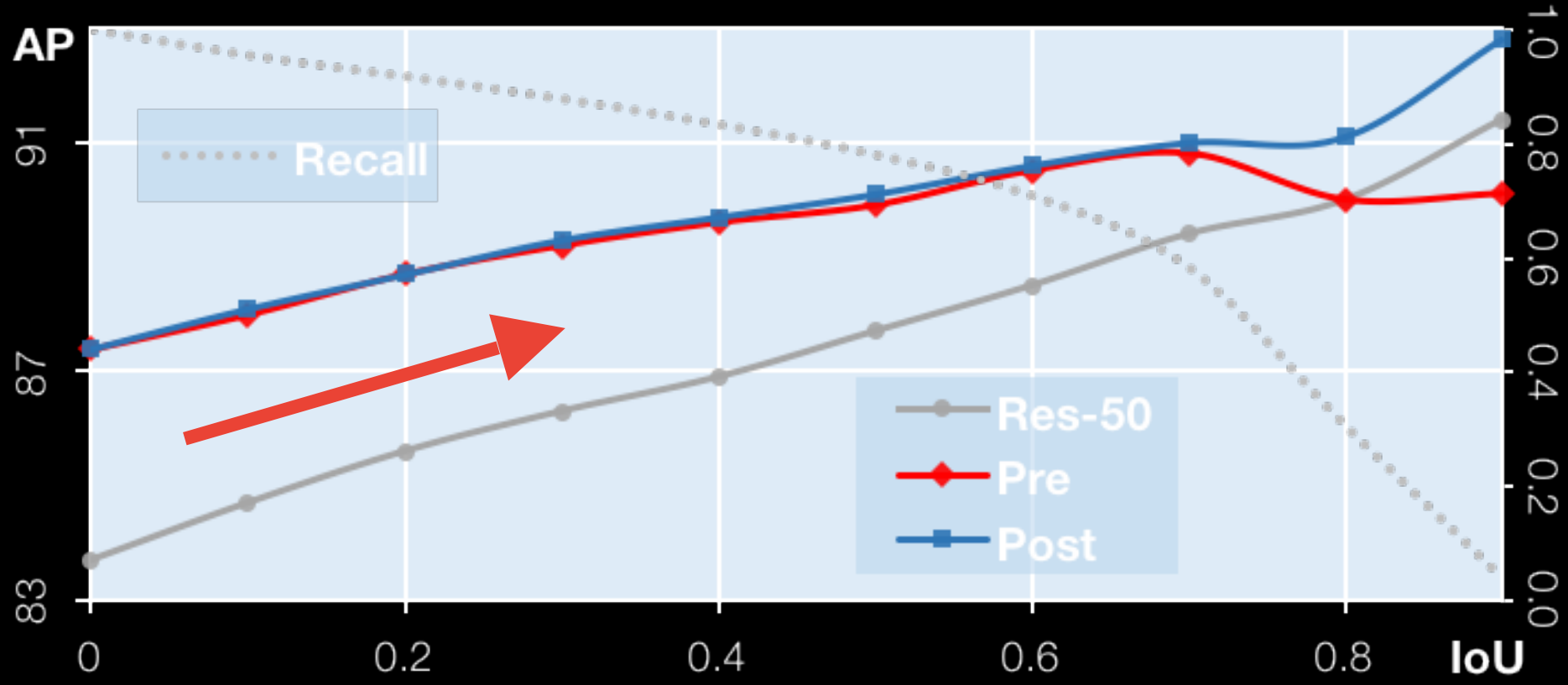
  - Same for baseline Res-50

Trends as More Regions are Dropped



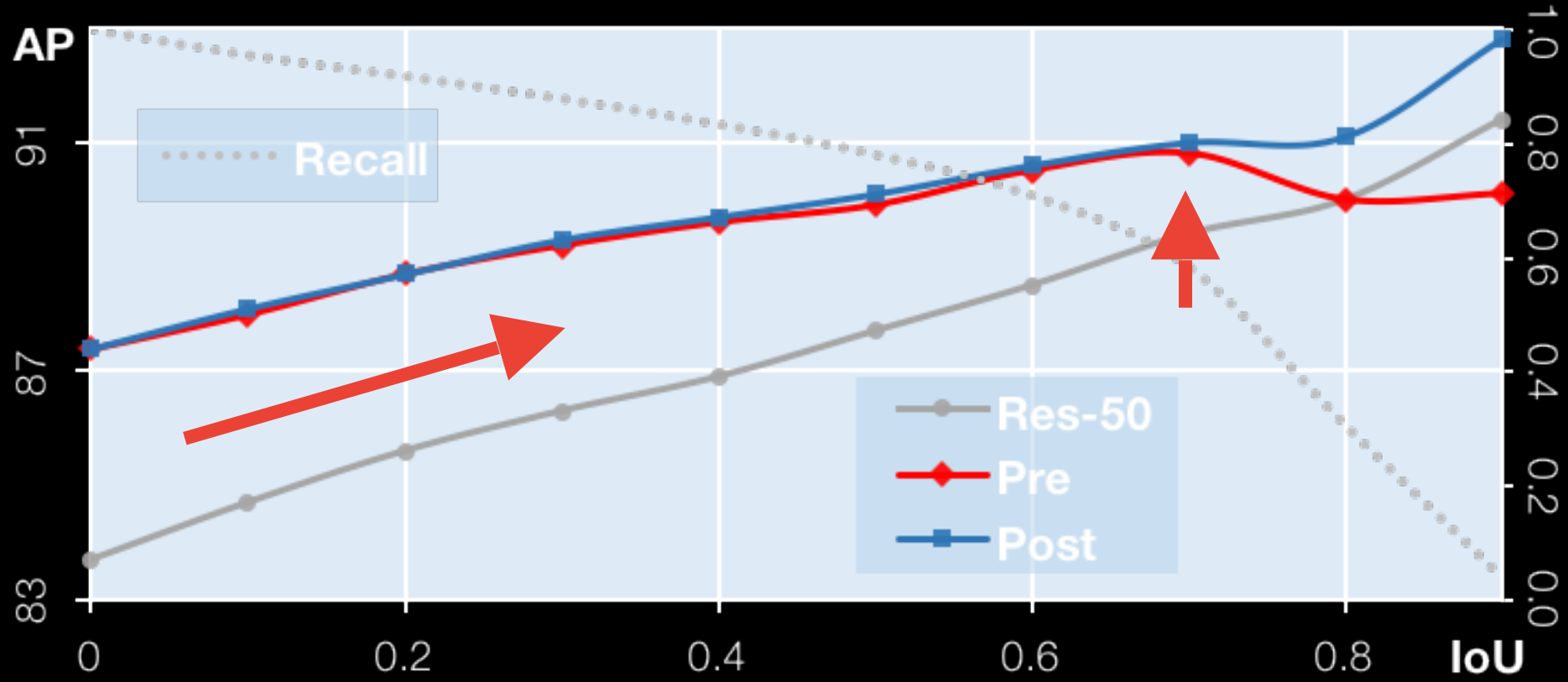
region proposal network from a detector with COCO detection AP 32.4%

Trends as More Regions are Dropped



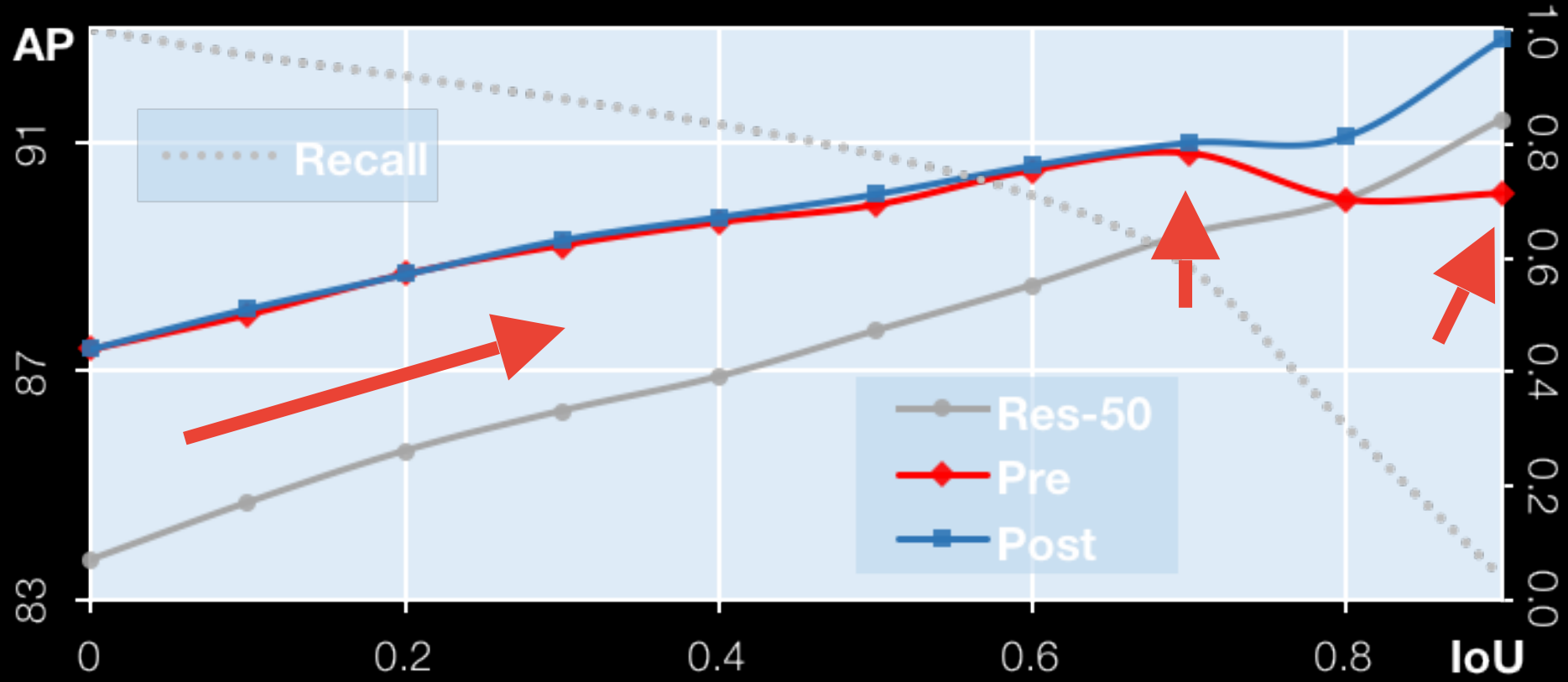
region proposal network from a detector with COCO detection AP 32.4%

Trends as More Regions are Dropped



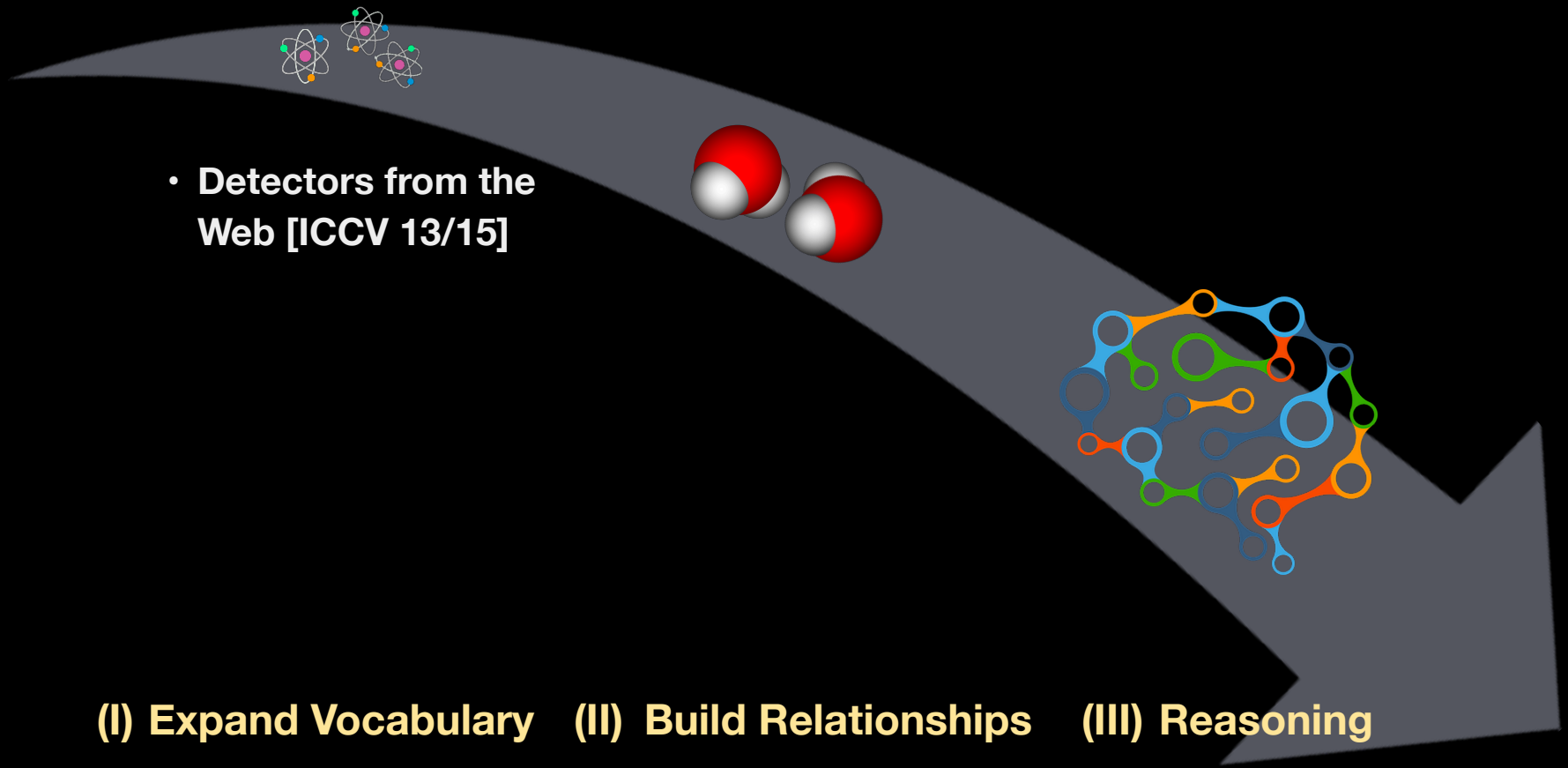
region proposal network from a detector with COCO detection AP 32.4%

Trends as More Regions are Dropped



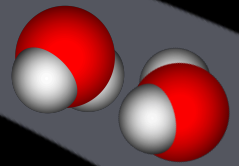
region proposal network from a detector with COCO detection AP 32.4%

- Detectors from the Web [ICCV 13/15]

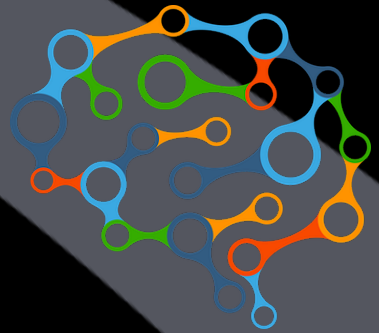


(I) Expand Vocabulary (II) Build Relationships (III) Reasoning

- **Detectors from the Web [ICCV 13/15]**
- Pixel-Level Labeling [CVPR 2014]
- Sense Discovery [CVPR 2015]



- **Never Ending Image Learner [ICCV 2013]**
- Spatial Memory Network [ICCV 2017]

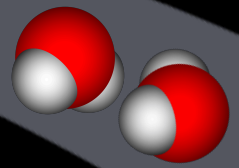


- **Iterative Reasoning [submitted]**

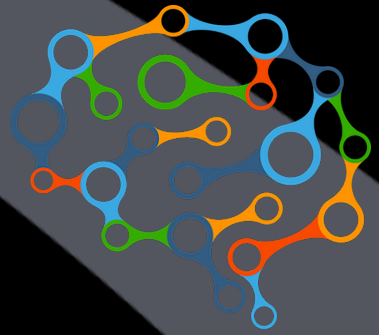
(I) Expand Vocabulary **(II) Build Relationships** **(III) Reasoning**

Thank You!

- **Detectors from the Web [ICCV 13/15]**
- Pixel-Level Labeling [CVPR 2014]
- Sense Discovery [CVPR 2015]



- **Never Ending Image Learner [ICCV 2013]**
- Spatial Memory Network [ICCV 2017]



- **Iterative Reasoning [submitted]**

(I) Expand Vocabulary **(II) Build Relationships** **(III) Reasoning**