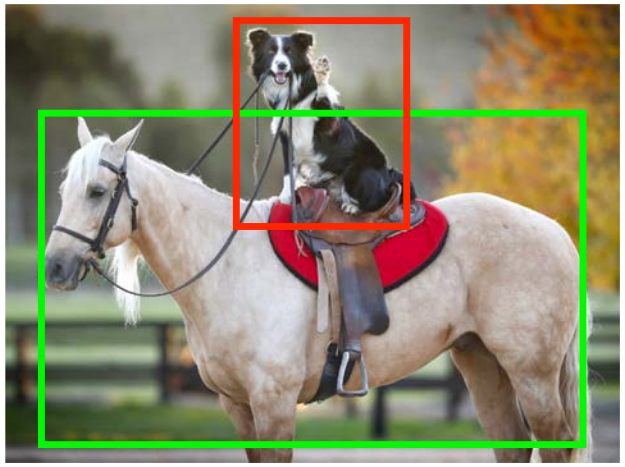
DOCK: Detecting Objects by transferring Common-sense Knowledge

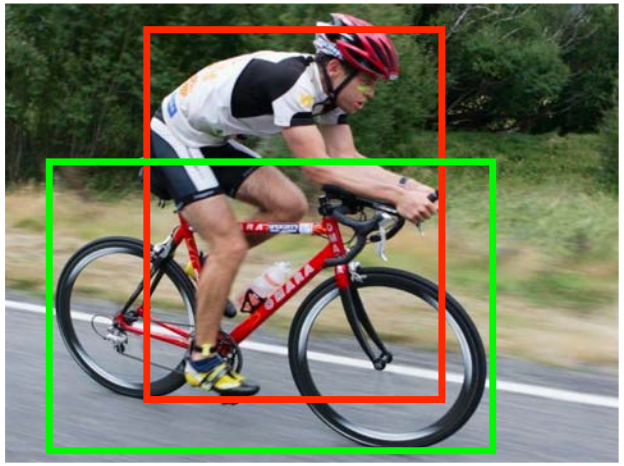
Krishna Kumar Singh, Santosh Divvala,
Ali Farhadi, Yong Jae Lee

Fully-supervised

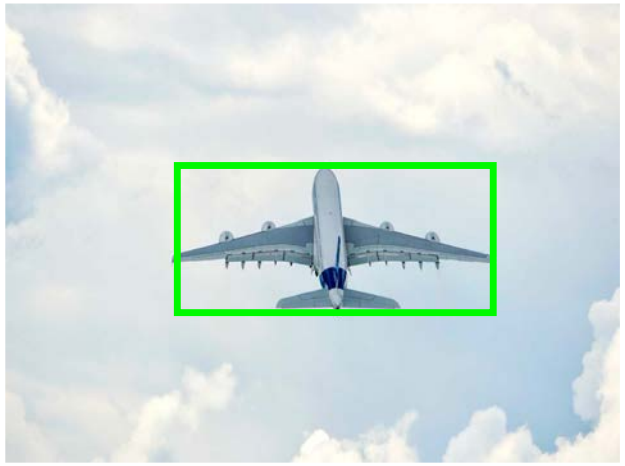
Car



Dog, Horse



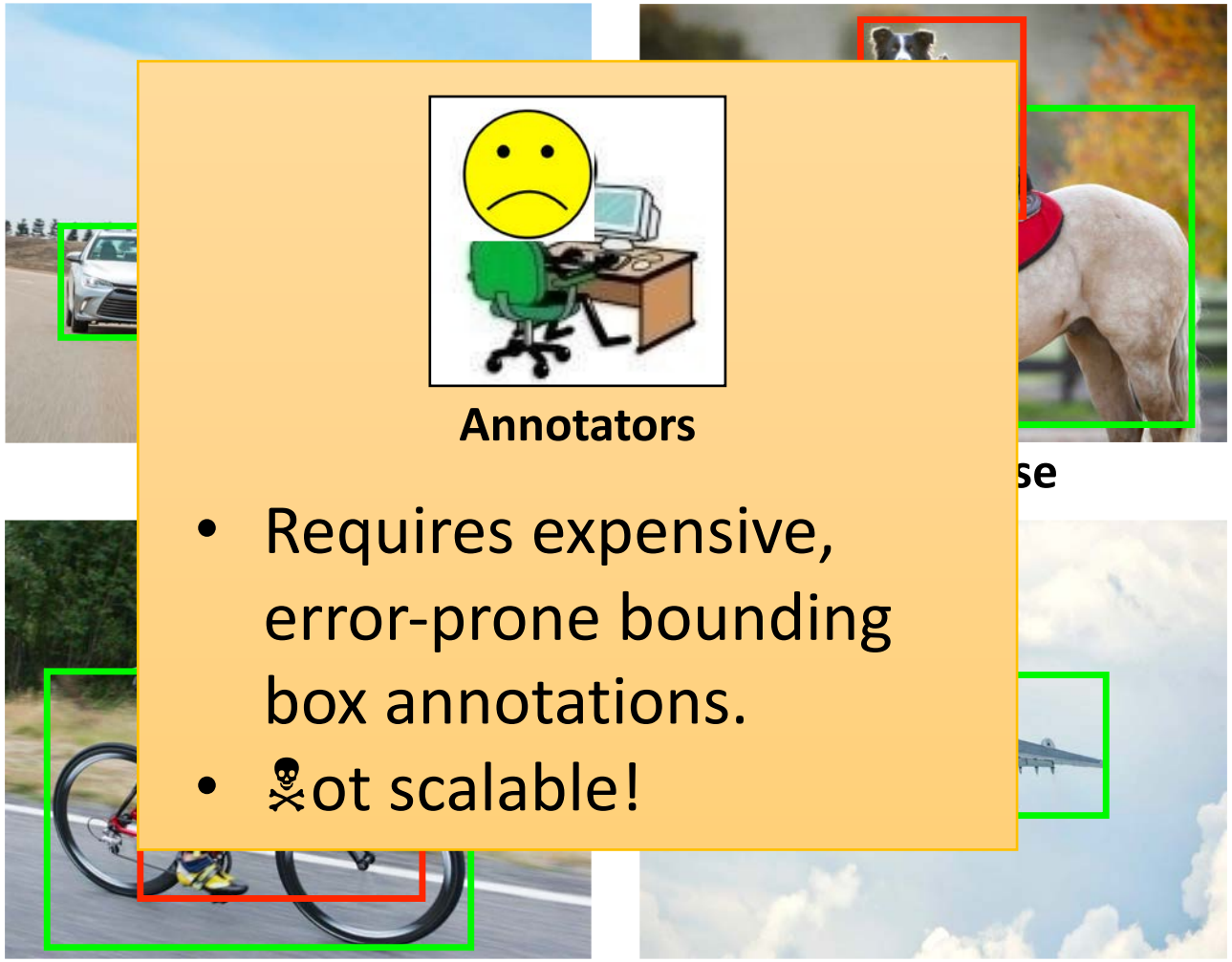
Person, Bike



Airplane

Training Images

Fully-supervised



Training Images

Fully-supervised

Annotators

- Requires expensive, error-prone bounding box annotations.
- ☠ot scalable!

Person, Bike

Airplane

Training Images

Weakly-supervised

Car

Dog, Horse

Person, Bike

Airplane

Training Images

Fully-supervised

Annotators

- Requires expensive, error-prone bounding box annotations.
- ☠ot scalable!

Person, Bike

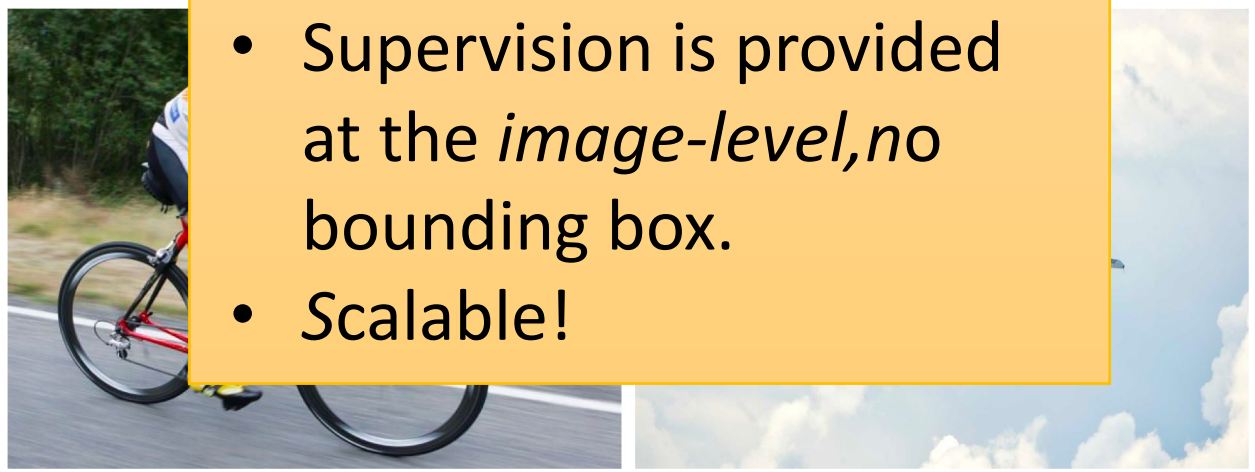
Airplane

Training Images

Weakly-supervised

Annotators

- Supervision is provided at the *image-level,no* bounding box.
- *S*calable!

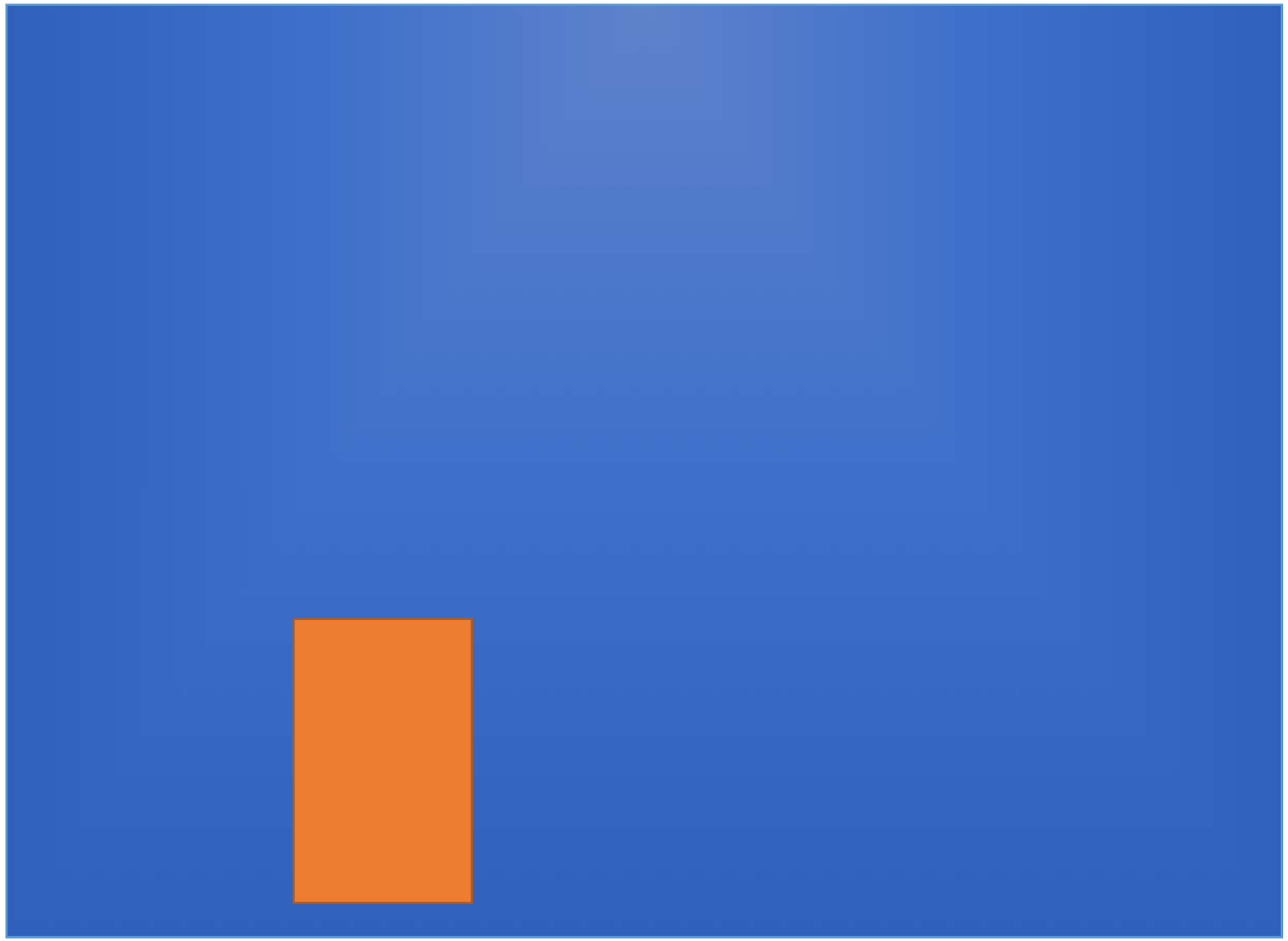


Person, Bike

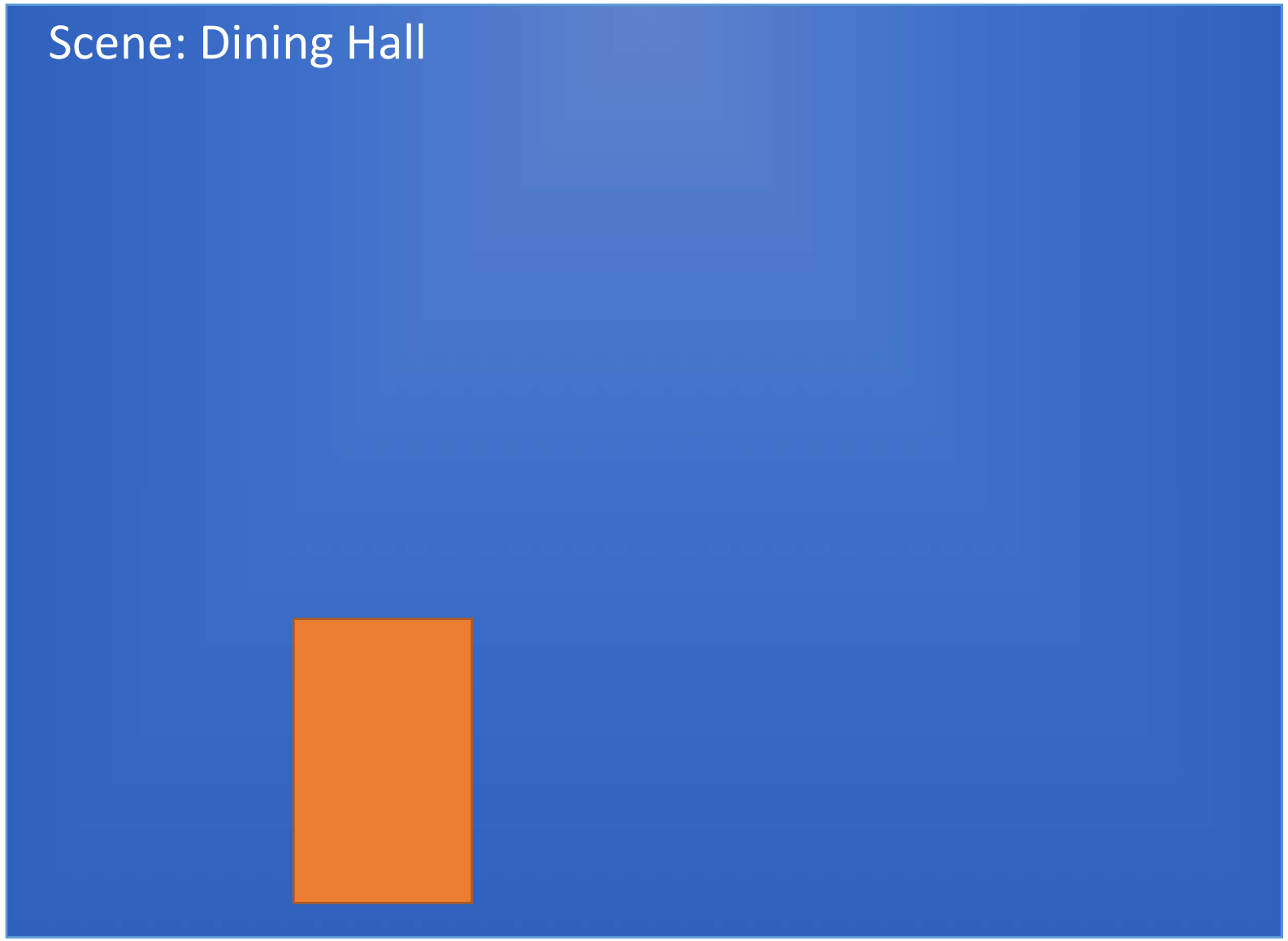
Airplane

Training Images

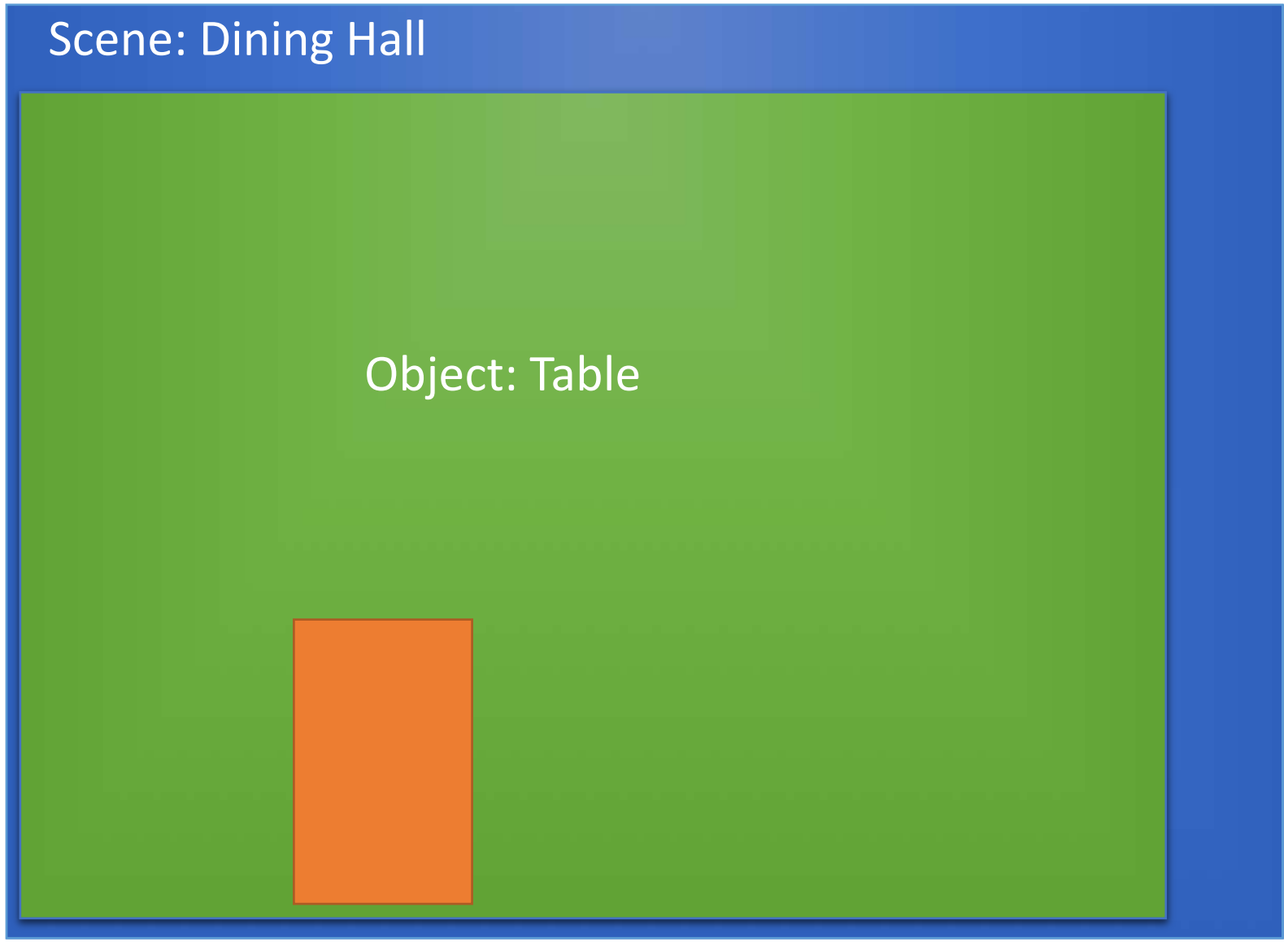
Can you guess the object in the region proposal?



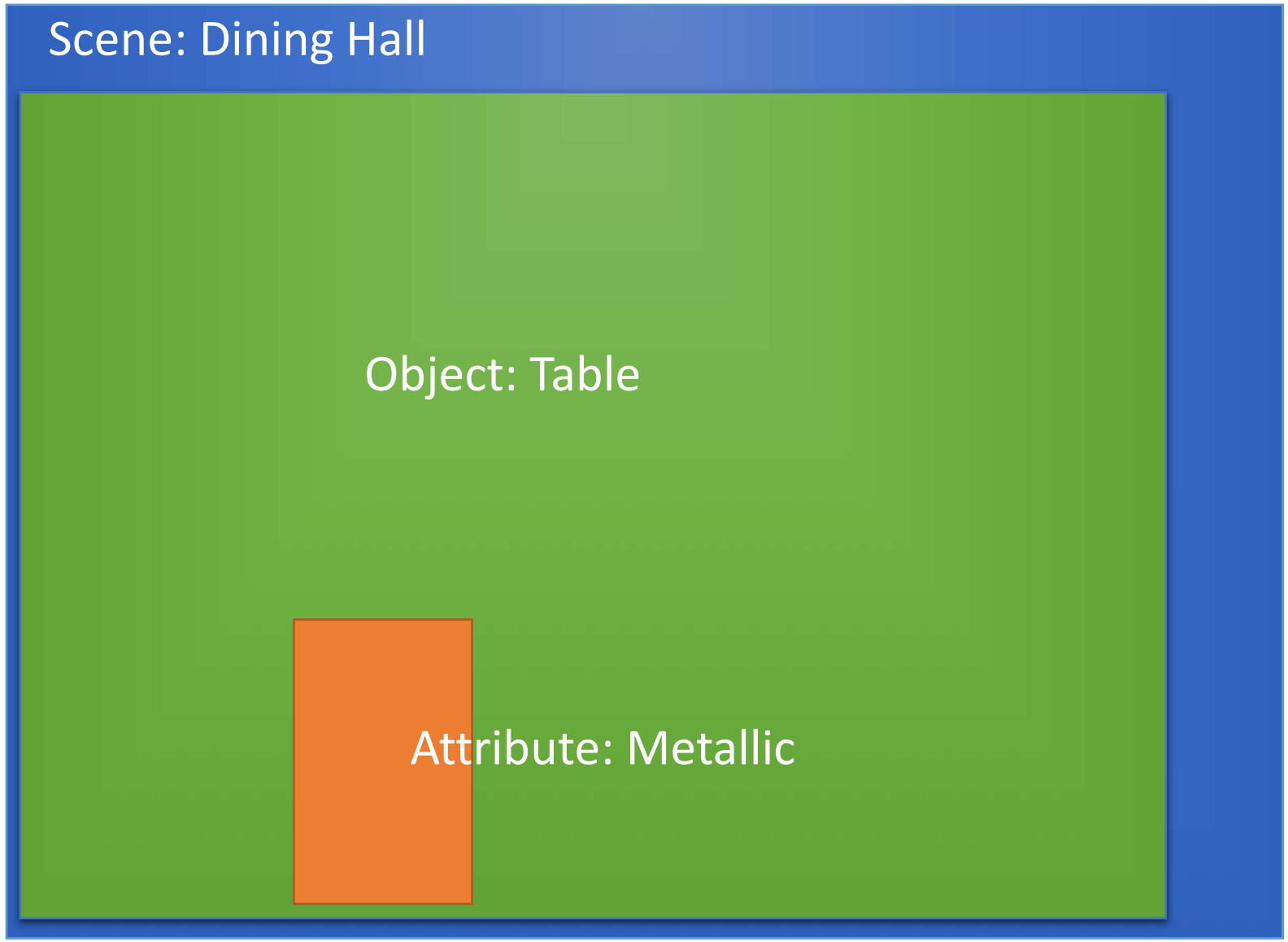
Can you guess the object in the region proposal?



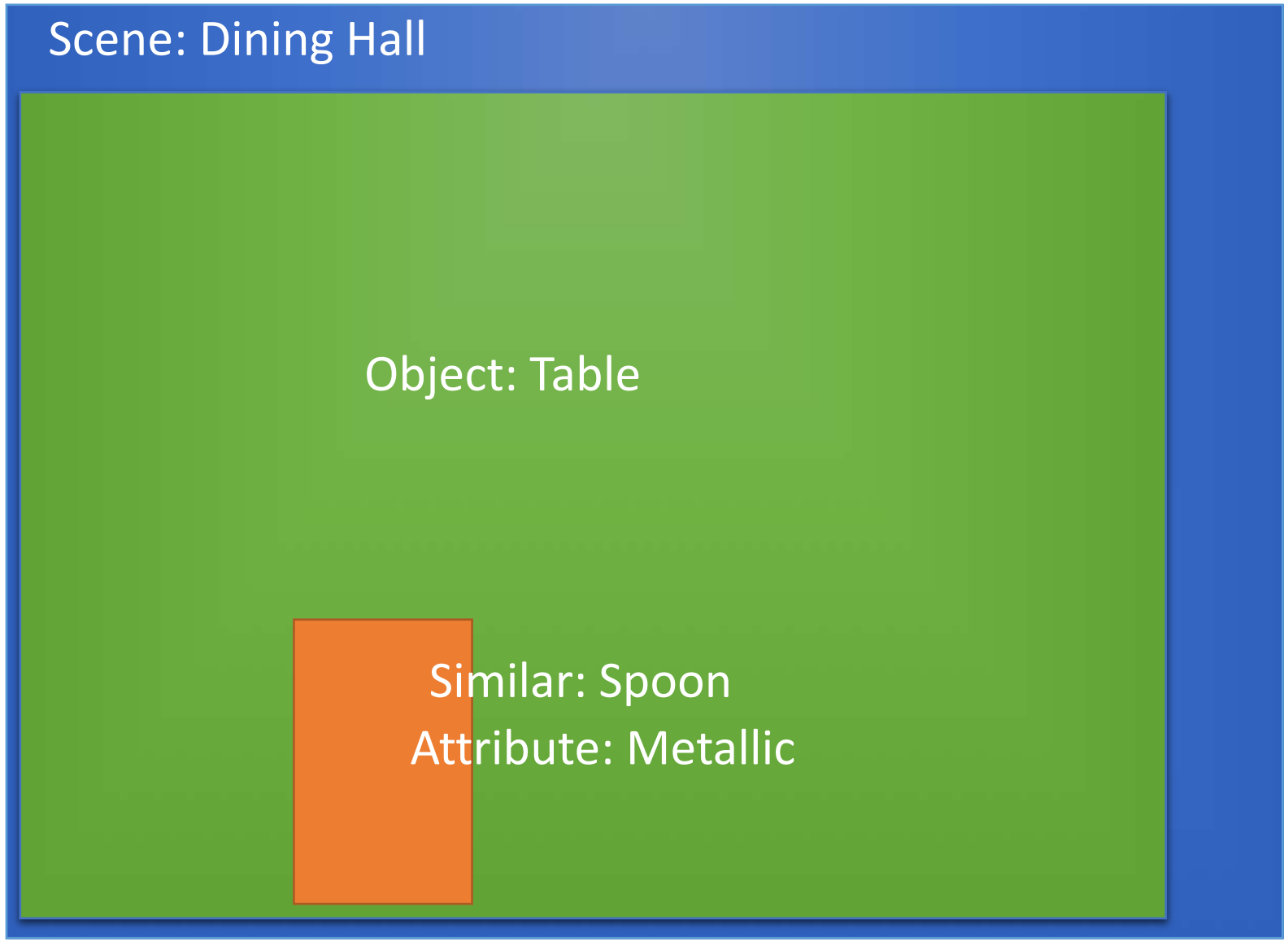
Can you guess the object in the region proposal?



Can you guess the object in the region proposal?



Can you guess the object in the region proposal?

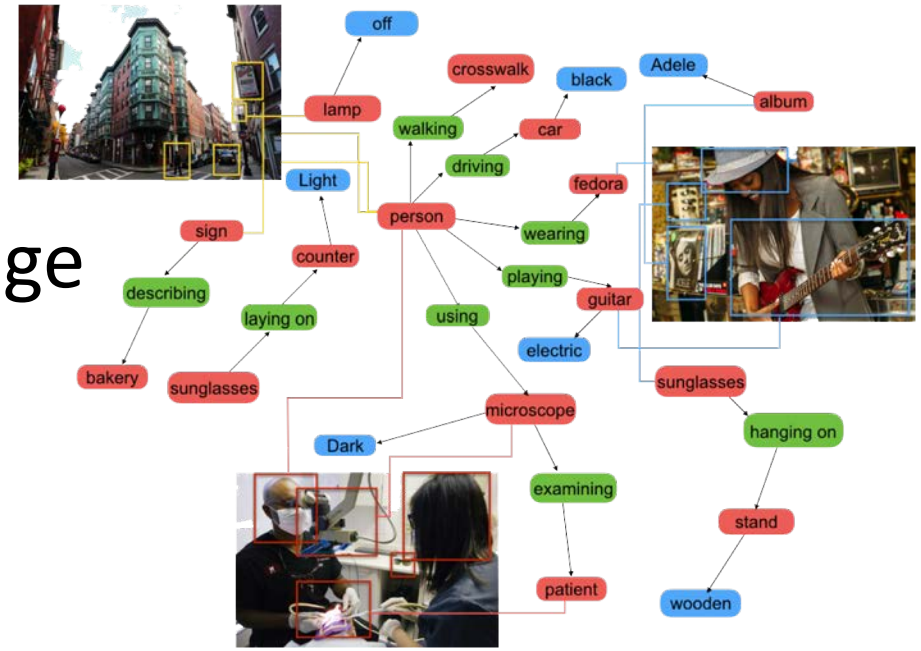


Fork

Problem Setup

- **Source categories:** *Bounding box annotations*
- **Target categories:** *Image-level annotations*
- Detectors are learned for **Target categories**

Knowledge Base



Common-sense knowledge will be used to improve detection

Types of common-sense knowledge
(**Target categories**)

Similarity to *source* objects

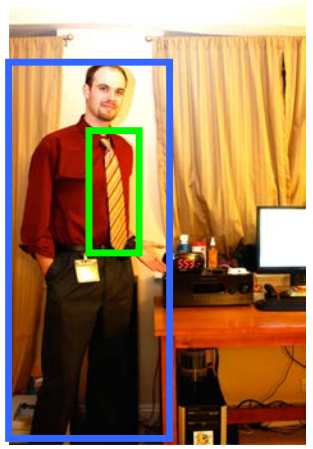
Degree of similarity

Attribute

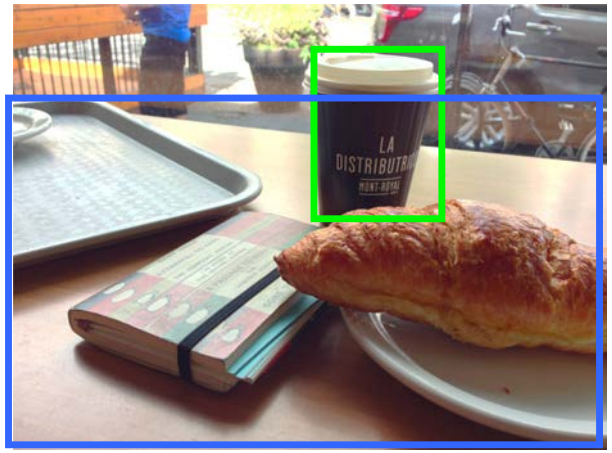
Car: *Shiny/Metallic*

Bowl: *Round*

Spatial relation to *source* objects



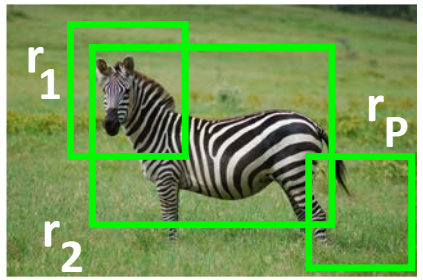
Person *wears* Tie



Cup *on* Table

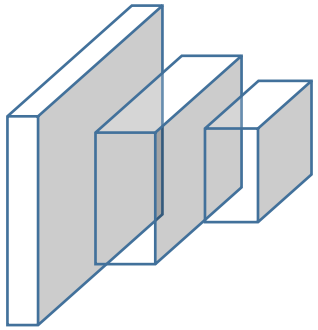
Approach

Transferring common-sense knowledge to improve weakly-supervised detection



Inputs:
Training Image,
Image Class-label (‘zebra’)
Region Proposals (r_1,…r_P)

Weakly-supervised Detection Network [Bilen&Vedaldi ‘15]

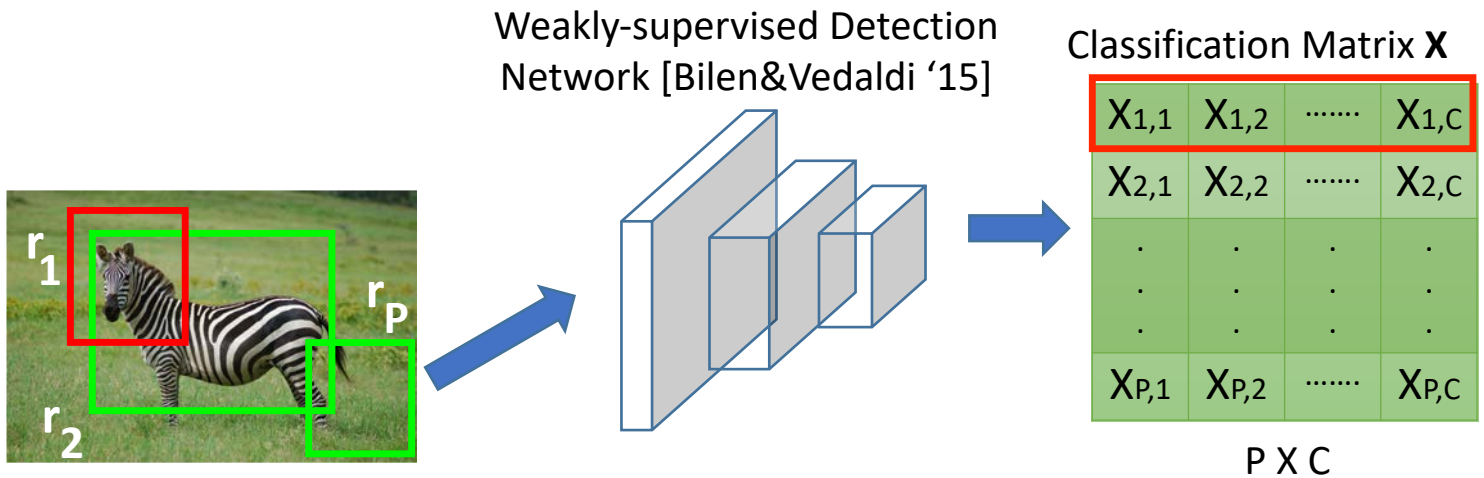


Classification Matrix **X**

$X_{1,1}$	$X_{1,2}$	.....	$X_{1,C}$
$X_{2,1}$	$X_{2,2}$	.....	$X_{2,C}$
. . .	. . .	. . .	. . .
$X_{P,1}$	$X_{P,2}$	.....	$X_{P,C}$

P X C

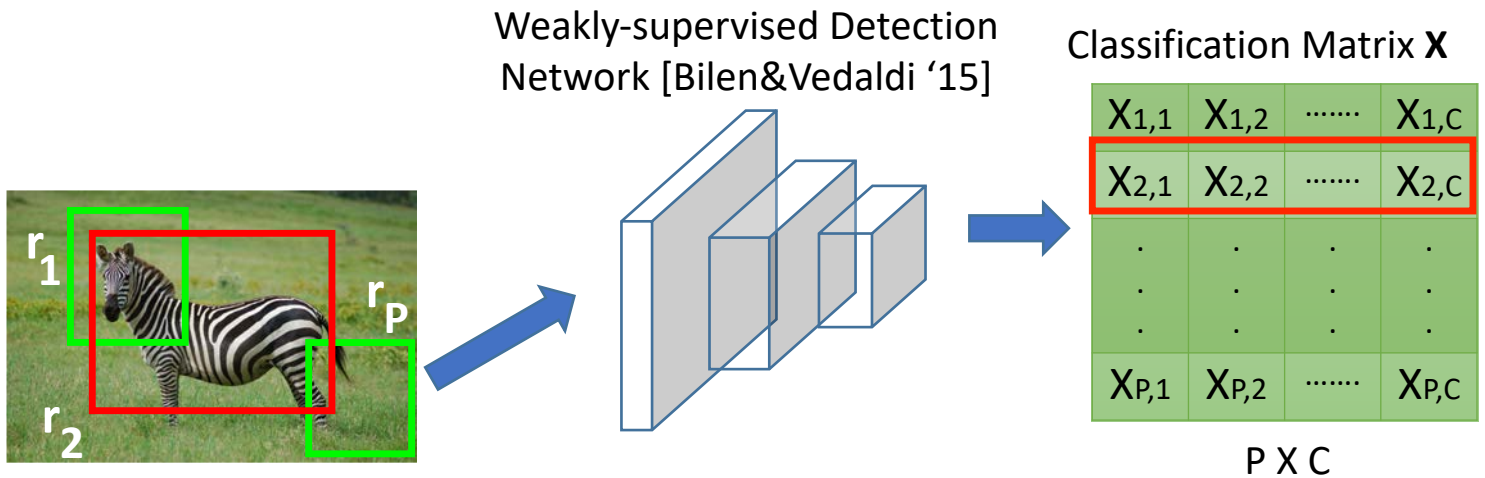
Transferring common-sense knowledge to improve weakly-supervised detection



Inputs:
Training Image,
Image Class-label ('zebra')
Region Proposals (r_1,…r_P)

Each row denotes class probabilities of a proposal

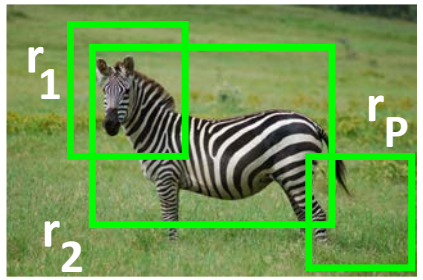
Transferring common-sense knowledge to improve weakly-supervised detection



Inputs:
Training Image,
Image Class-label ('zebra')
Region Proposals (r_1,…r_P)

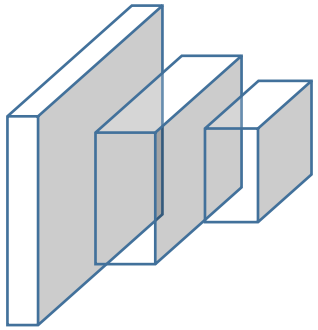
Each row denotes class probabilities of a proposal

Transferring common-sense knowledge to improve weakly-supervised detection



Inputs:
Training Image,
Image Class-label (‘zebra’)
Region Proposals (r_1,…r_P)

Weakly-supervised Detection Network [Bilen&Vedaldi ‘15]

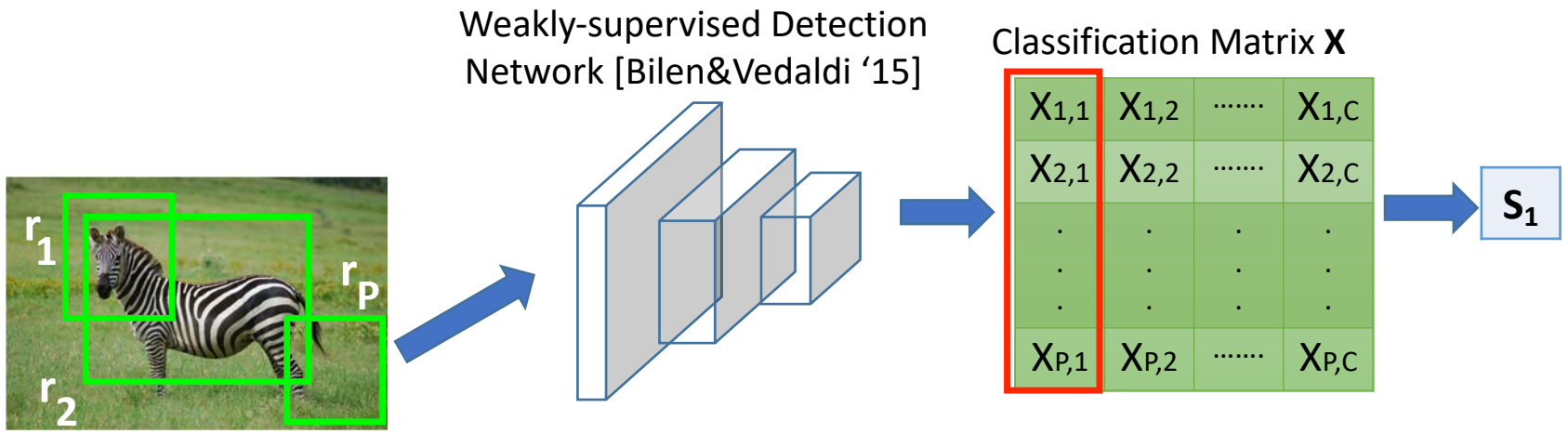


Classification Matrix **X**

$X_{1,1}$	$X_{1,2}$	.....	$X_{1,C}$
$X_{2,1}$	$X_{2,2}$	.....	$X_{2,C}$
. . .	. . .	. . .	. . .
$X_{P,1}$	$X_{P,2}$	.....	$X_{P,C}$

P X C

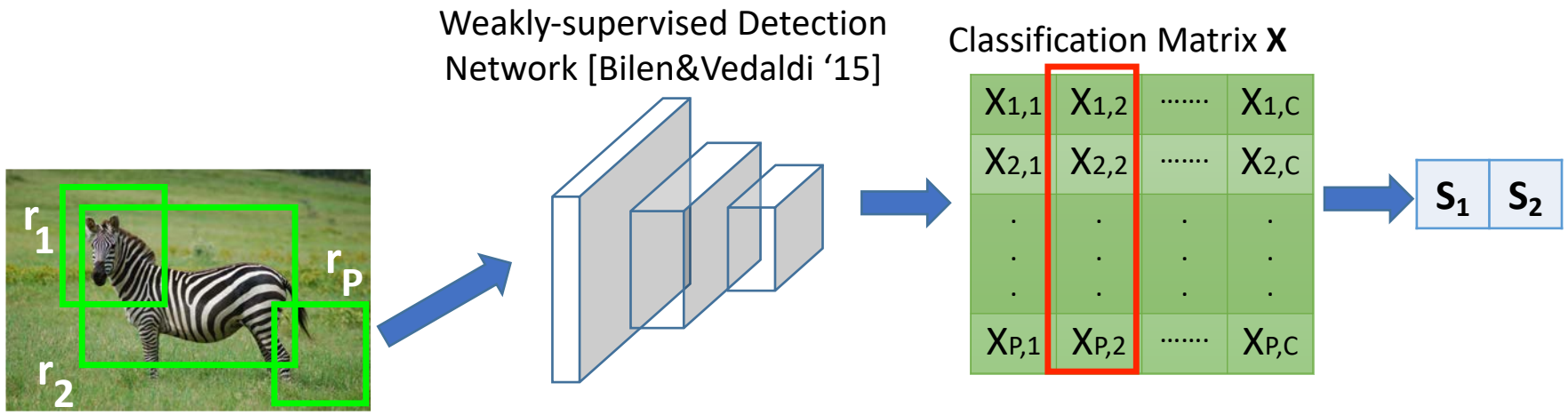
Transferring common-sense knowledge to improve weakly-supervised detection



Inputs:
Training Image,
Image Class-label ('zebra')
Region Proposals (r_1,…r_P)

Class score is computed by summing proposal probabilities over the column

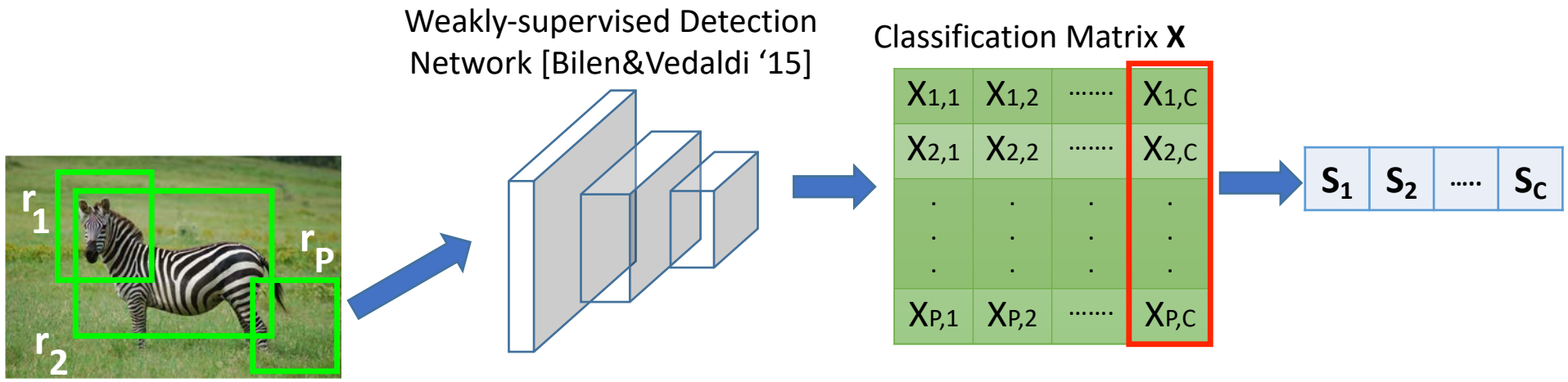
Transferring common-sense knowledge to improve weakly-supervised detection



Inputs:
Training Image,
Image Class-label ('zebra')
Region Proposals (r_1,…r_P)

Class score is computed by summing proposal probabilities over the column

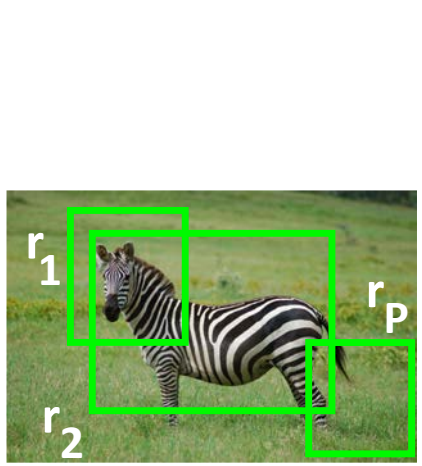
Transferring common-sense knowledge to improve weakly-supervised detection



Inputs:
Training Image,
Image Class-label ('zebra')
Region Proposals (r_1,…r_P)

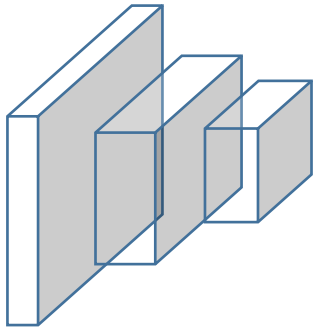
Class score is computed by summing proposal probabilities over the column

Transferring common-sense knowledge to improve weakly-supervised detection



Inputs:
Training Image,
Image Class-label ('zebra')
Region Proposals (r_1,…r_P)

Weakly-supervised Detection Network [Bilen&Vedaldi '15]



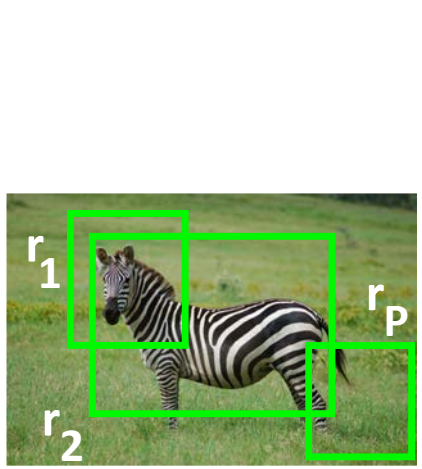
Classification Matrix **X**

$X_{1,1}$	$X_{1,2}$	.....	$X_{1,C}$
$X_{2,1}$	$X_{2,2}$	.....	$X_{2,C}$
.	.	.	.
.	.	.	.
.	.	.	.
$X_{P,1}$	$X_{P,2}$	.....	$X_{P,C}$

S_1	S_2	.....	S_C

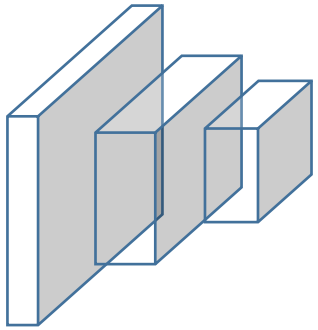
Image Classification Loss

Transferring common-sense knowledge to improve weakly-supervised detection



Inputs:
Training Image,
Image Class-label ('zebra')
Region Proposals (r_1,…r_P)

Weakly-supervised Detection Network [Bilen&Vedaldi '15]



Classification Matrix **X**

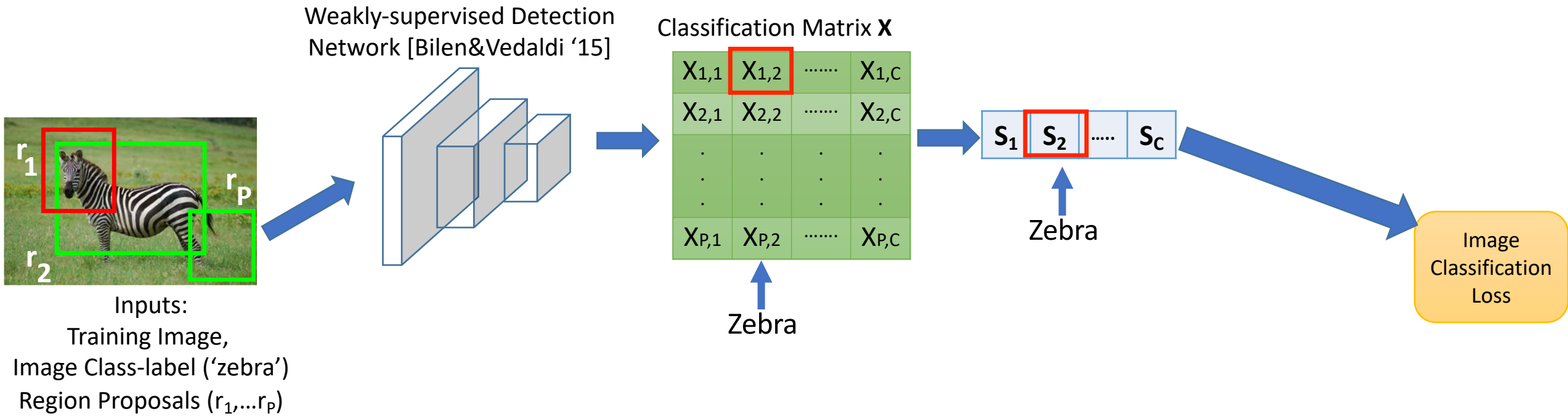
$X_{1,1}$	$X_{1,2}$	.....	$X_{1,C}$
$X_{2,1}$	$X_{2,2}$	.....	$X_{2,C}$
.	.	.	.
.	.	.	.
.	.	.	.
$X_{P,1}$	$X_{P,2}$	.....	$X_{P,C}$

S_1	S_2	.....	S_C

Image Classification Loss

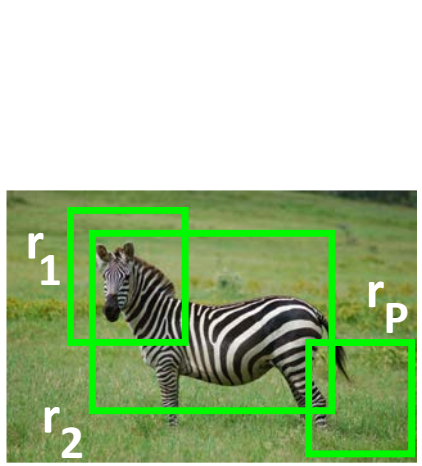
Can detect objects *without bounding box annotations*, but suffers from focusing on most discriminative parts

Transferring common-sense knowledge to improve weakly-supervised detection



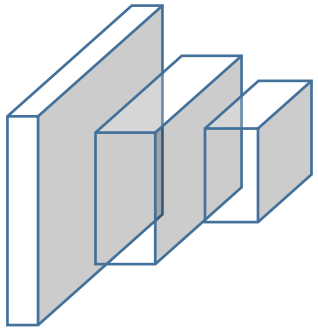
Can detect objects *without bounding box annotations*, but suffers from focusing on most discriminative parts

Transferring common-sense knowledge to improve weakly-supervised detection



Inputs:
Training Image,
Image Class-label ('zebra')
Region Proposals (r_1,…r_P)

Weakly-supervised Detection Network [Bilen&Vedaldi '15]



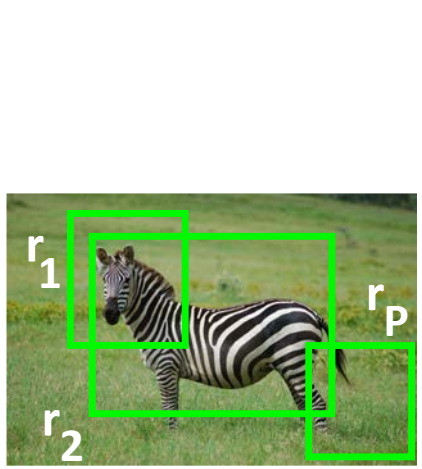
Classification Matrix **X**

$X_{1,1}$	$X_{1,2}$	.....	$X_{1,C}$
$X_{2,1}$	$X_{2,2}$	.....	$X_{2,C}$
.	.	.	.
.	.	.	.
.	.	.	.
$X_{P,1}$	$X_{P,2}$	.....	$X_{P,C}$

S_1	S_2	.....	S_C

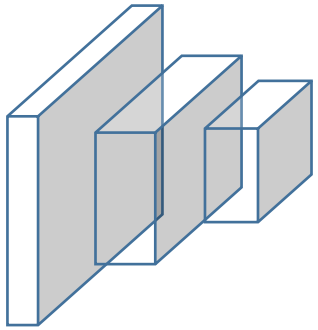
Image Classification Loss

Transferring common-sense knowledge to improve weakly-supervised detection



Inputs:
Training Image,
Image Class-label ('zebra')
Region Proposals (r_1,…r_P)

Weakly-supervised Detection Network [Bilen&Vedaldi '15]



Classification Matrix **X**

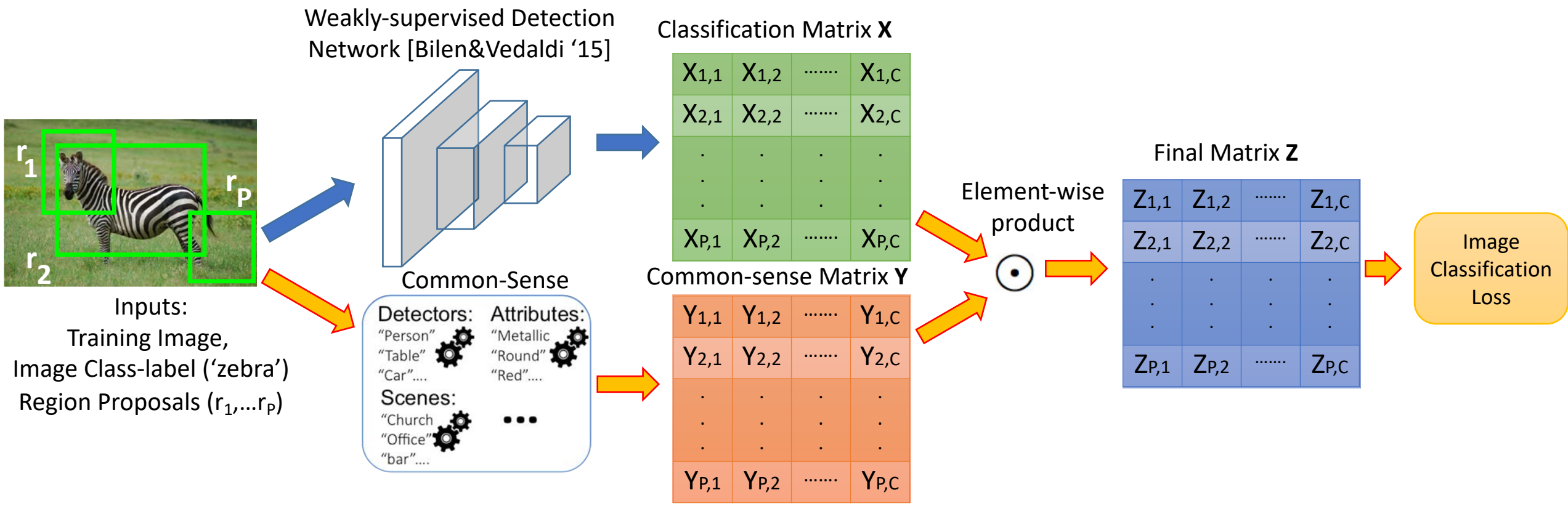
$X_{1,1}$	$X_{1,2}$	.....	$X_{1,C}$
$X_{2,1}$	$X_{2,2}$	.....	$X_{2,C}$
.	.	.	.
.	.	.	.
.	.	.	.
$X_{P,1}$	$X_{P,2}$	.....	$X_{P,C}$

S_1	S_2	.....	S_C

Image Classification Loss

Use *common-sense knowledge* to improve localization

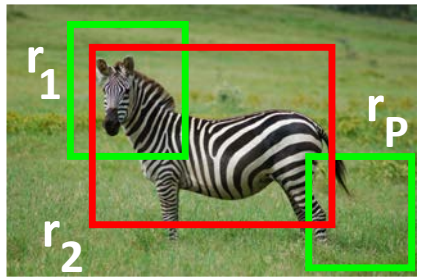
Transferring common-sense knowledge to improve weakly-supervised detection



Use *common-sense knowledge* to improve localization

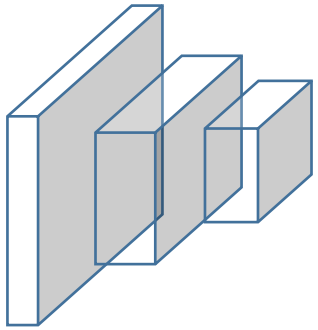
[K. Singh, S. Divvala, A. Farhadi, and Y. J. Lee, "DOCK: Detecting Objects by transferring Common-sense Knowledge", ECCV 2018]

Transferring common-sense knowledge to improve weakly-supervised detection



Inputs:
Training Image,
Image Class-label ('zebra')
Region Proposals (r_1,…r_P)

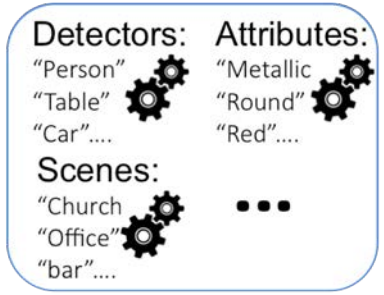
Weakly-supervised Detection Network [Bilen&Vedaldi '15]



Classification Matrix **X**

$X_{1,1}$	$X_{1,2}$	.....	$X_{1,C}$
$X_{2,1}$	$X_{2,2}$	.....	$X_{2,C}$
.	.	.	.
.	.	.	.
.	.	.	.
$X_{P,1}$	$X_{P,2}$	.....	$X_{P,C}$

Common-Sense



Common-sense Matrix **Y**

$Y_{1,1}$	$Y_{1,2}$	.....	$Y_{1,C}$
$Y_{2,1}$	$Y_{2,2}$	.....	$Y_{2,C}$
.	.	.	.
.	.	.	.
.	.	.	.
$Y_{P,1}$	$Y_{P,2}$	.....	$Y_{P,C}$

Zebra

Element-wise product

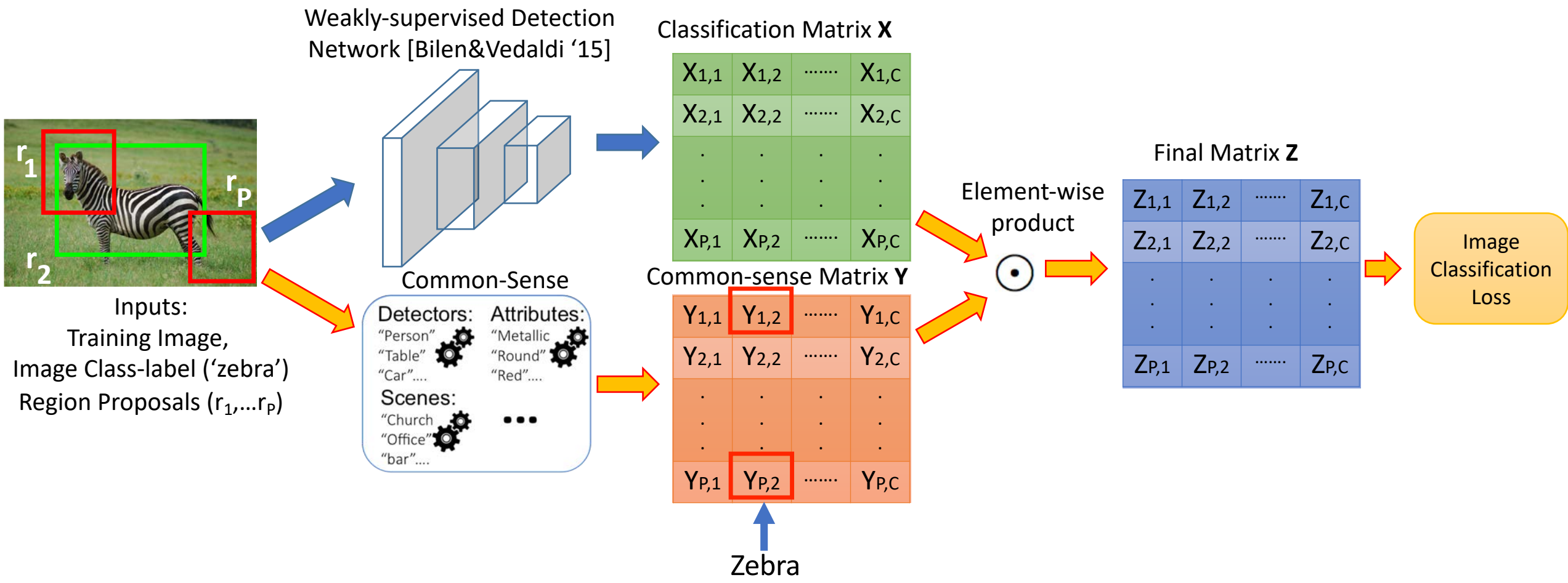
Final Matrix **Z**

$Z_{1,1}$	$Z_{1,2}$	.....	$Z_{1,C}$
$Z_{2,1}$	$Z_{2,2}$	.....	$Z_{2,C}$
.	.	.	.
.	.	.	.
.	.	.	.
$Z_{P,1}$	$Z_{P,2}$	.....	$Z_{P,C}$

Image Classification Loss

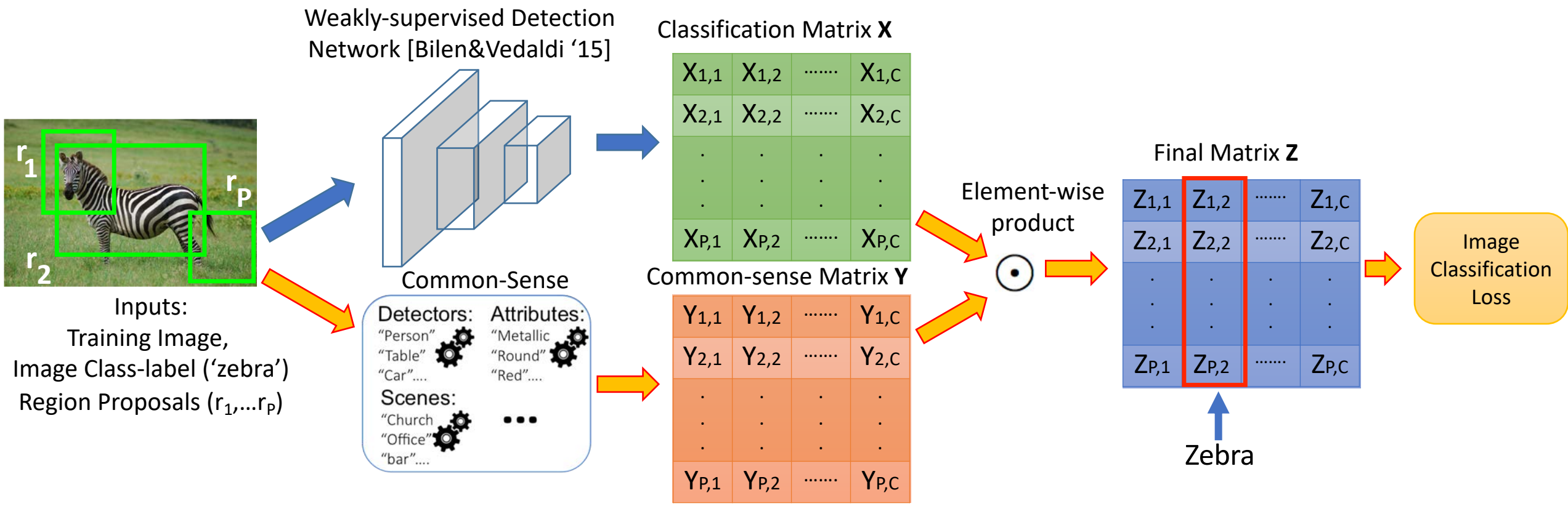
r_2 has high common-sense prior

Transferring common-sense knowledge to improve weakly-supervised detection

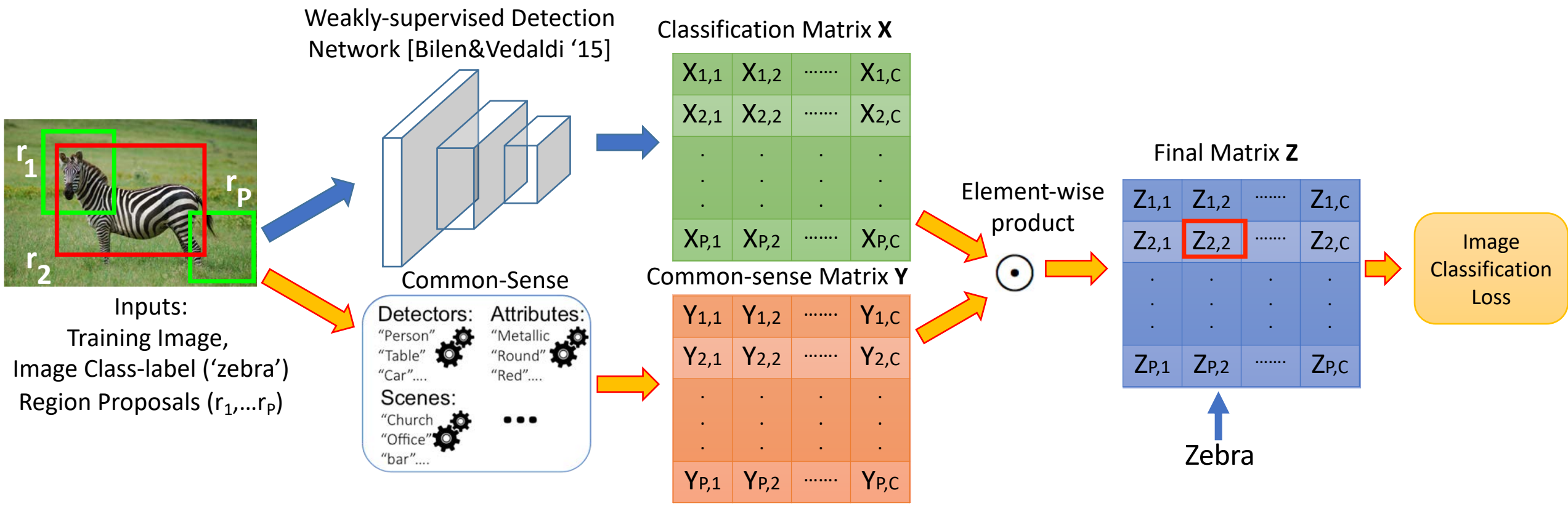


r_1 and r_P has low common-sense prior

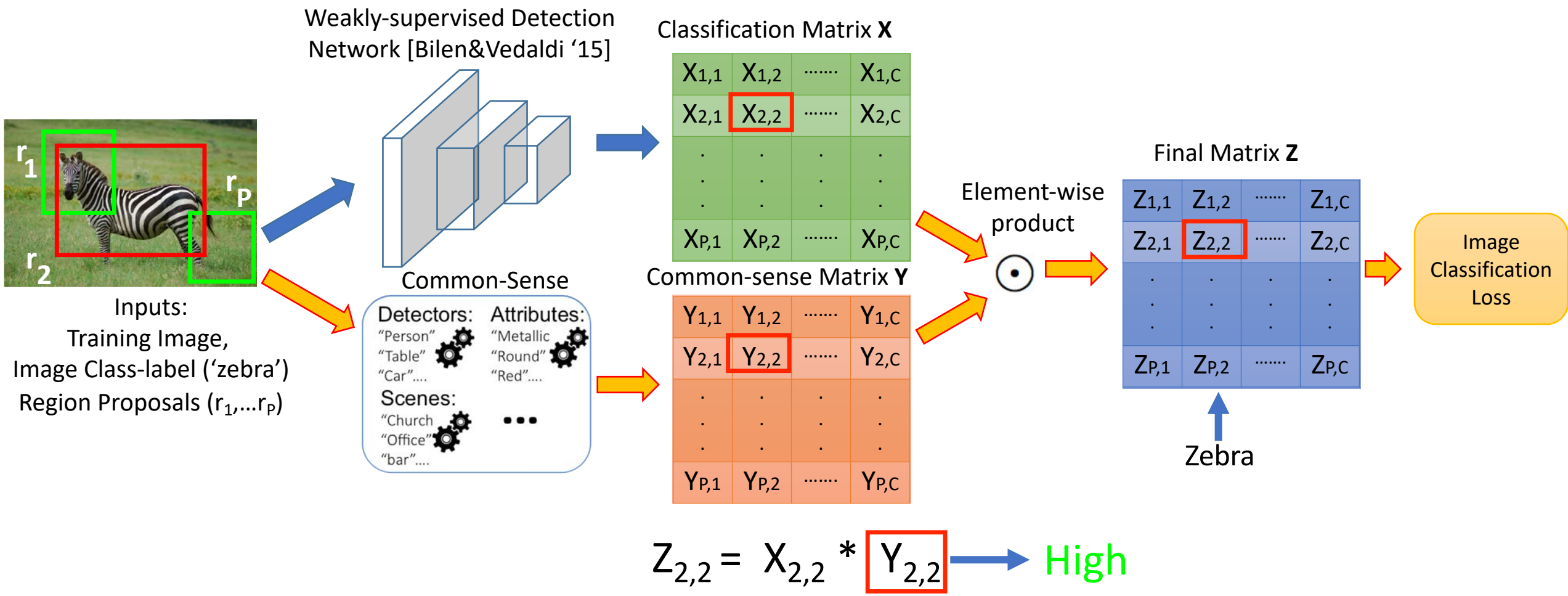
Transferring common-sense knowledge to improve weakly-supervised detection



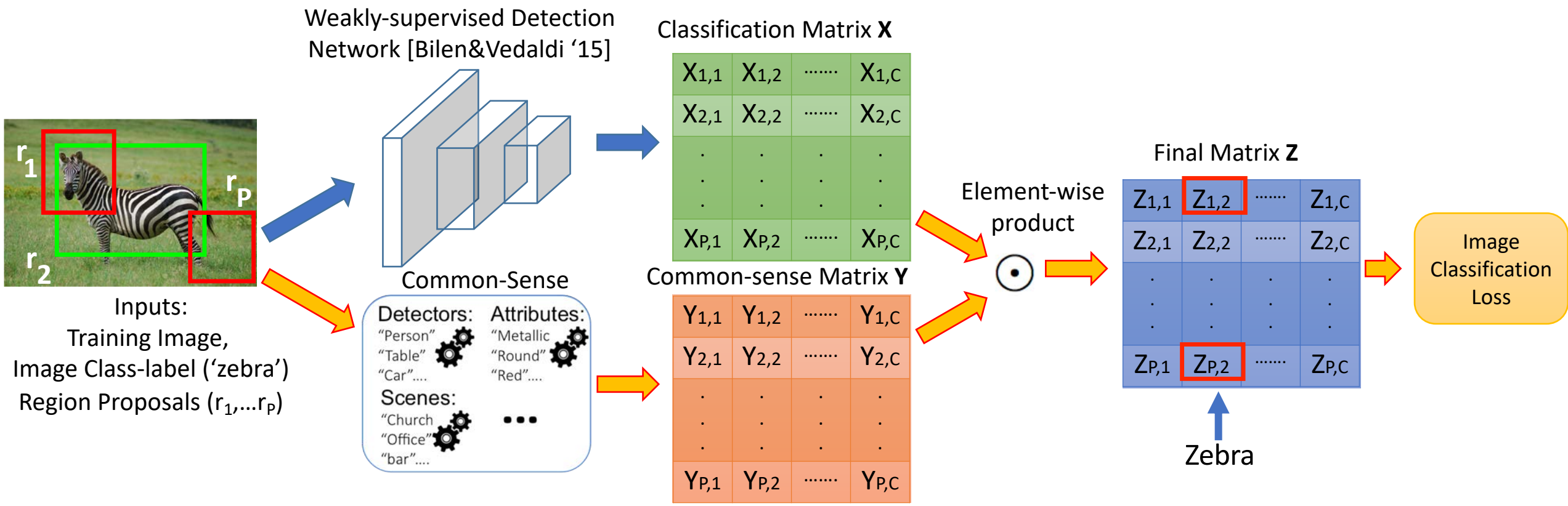
Transferring common-sense knowledge to improve weakly-supervised detection



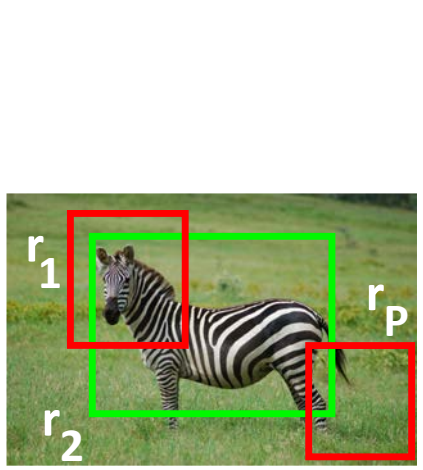
Transferring common-sense knowledge to improve weakly-supervised detection



Transferring common-sense knowledge to improve weakly-supervised detection

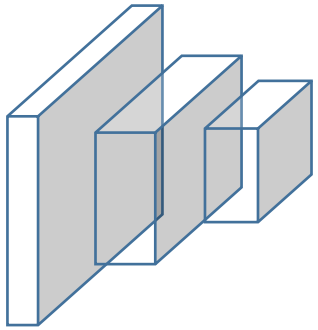


Transferring common-sense knowledge to improve weakly-supervised detection



Inputs:
Training Image,
Image Class-label ('zebra')
Region Proposals (r_1,…r_P)

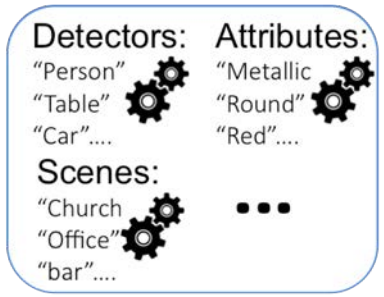
Weakly-supervised Detection Network [Bilen&Vedaldi '15]



Classification Matrix **X**

$X_{1,1}$	$X_{1,2}$	.....	$X_{1,C}$
$X_{2,1}$	$X_{2,2}$	.....	$X_{2,C}$
.	.	.	.
.	.	.	.
.	.	.	.
$X_{P,1}$	$X_{P,2}$	.....	$X_{P,C}$

Common-Sense



Common-sense Matrix **Y**

$Y_{1,1}$	$Y_{1,2}$	.....	$Y_{1,C}$
$Y_{2,1}$	$Y_{2,2}$	.....	$Y_{2,C}$
.	.	.	.
.	.	.	.
.	.	.	.
$Y_{P,1}$	$Y_{P,2}$	.....	$Y_{P,C}$

Element-wise product ⊙

Final Matrix **Z**

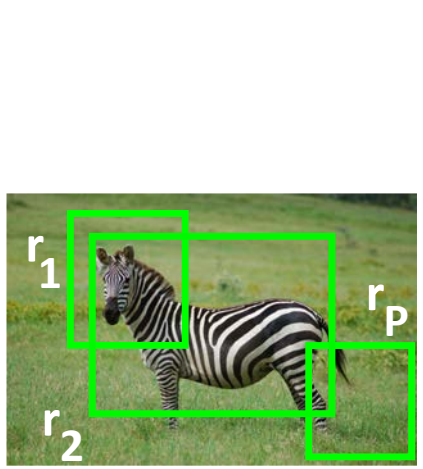
$Z_{1,1}$	$Z_{1,2}$	.....	$Z_{1,C}$
$Z_{2,1}$	$Z_{2,2}$	.....	$Z_{2,C}$
.	.	.	.
.	.	.	.
.	.	.	.
$Z_{P,1}$	$Z_{P,2}$	.....	$Z_{P,C}$

Zebra

Image Classification Loss

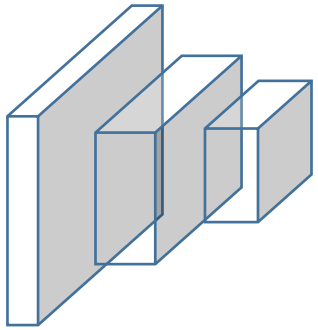
$$Z_{1,2} = X_{1,2} * Y_{1,2}$$
$$Z_{P,2} = X_{P,2} * Y_{P,2} \rightarrow \text{Low}$$

Transferring common-sense knowledge to improve weakly-supervised detection



Inputs:
Training Image,
Image Class-label ('zebra')
Region Proposals (r_1,…r_P)

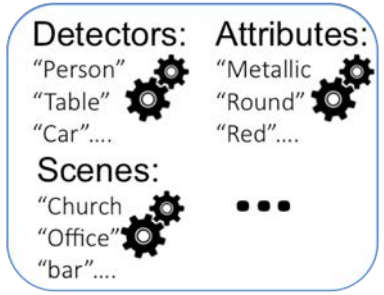
Weakly-supervised Detection Network [Bilen&Vedaldi '15]



Classification Matrix **X**

$X_{1,1}$	$X_{1,2}$	.....	$X_{1,C}$
$X_{2,1}$	$X_{2,2}$	.....	$X_{2,C}$
.	.	.	.
.	.	.	.
.	.	.	.
$X_{P,1}$	$X_{P,2}$	.....	$X_{P,C}$

Common-Sense



Common-sense Matrix **Y**

$Y_{1,1}$	$Y_{1,2}$	.....	$Y_{1,C}$
$Y_{2,1}$	$Y_{2,2}$	.....	$Y_{2,C}$
.	.	.	.
.	.	.	.
.	.	.	.
$Y_{P,1}$	$Y_{P,2}$	.....	$Y_{P,C}$

Element-wise product

Final Matrix **Z**

$Z_{1,1}$	$Z_{1,2}$	.....	$Z_{1,C}$
$Z_{2,1}$	$Z_{2,2}$	.....	$Z_{2,C}$
.	.	.	.
.	.	.	.
.	.	.	.
$Z_{P,1}$	$Z_{P,2}$	.....	$Z_{P,C}$

Image Classification Loss

Matrices of different common-sense are averaged to create single matrix

Similarity Common-sense




Visual/semantic similarity decreases

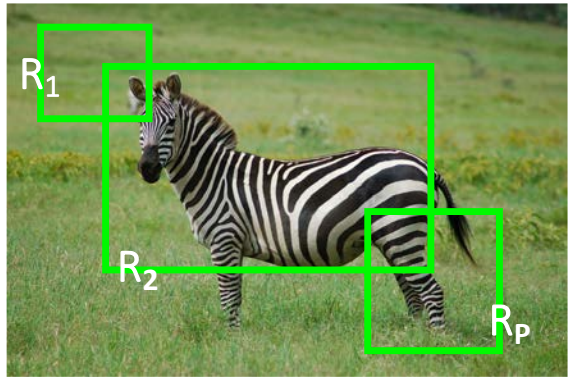
Similarity Common-sense




Visual/semantic similarity decreases

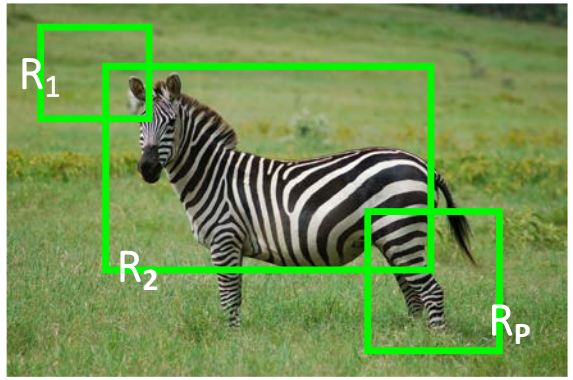
- Assumptions:
  - We have pre-trained detector for the source classes
  - Semantic similarity of source classes with target classes is known (Word2Vec based)
  - Ex: Horse->Zebra, Car->Truck, TV->Laptop …..

Similarity Common-sense



Input Image
+
Region proposals
+
Horse ~ Zebra

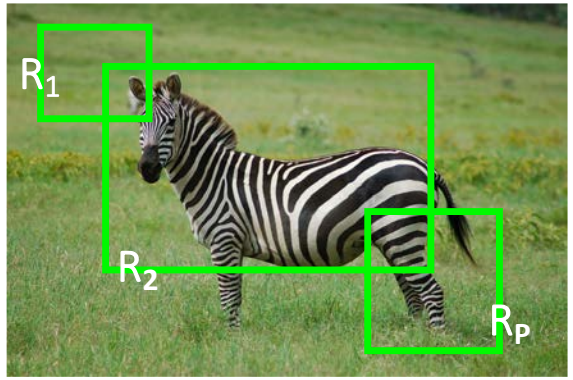
Similarity Common-sense



Input Image
+
Region proposals
+
Horse ~ Zebra

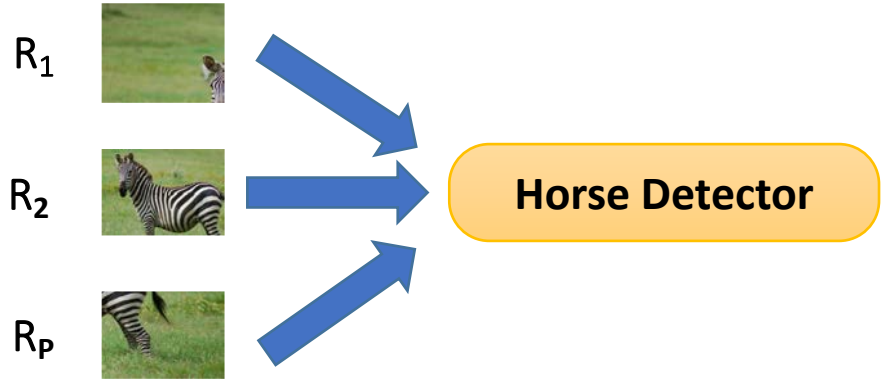
	Cup	**Zebra**		**Kite**
	Y_{11}	Y_{12}	.....	Y_{1C}
	Y_{21}	Y_{22}	.....	Y_{2C}
	.	.	.	.
	.	.	.	.
	.	.	.	.
	Y_{P1}	Y_{P2}	.....	Y_{PC}

Similarity Common-sense

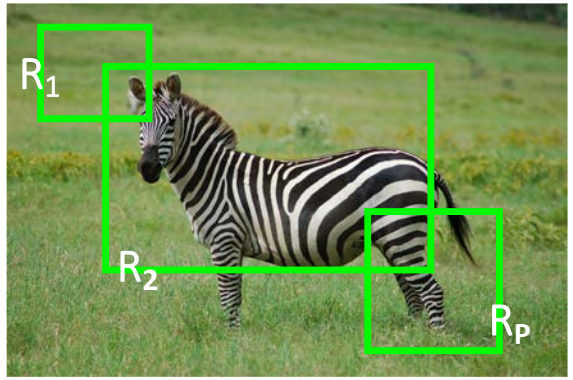


Input Image
+
Region proposals
+
Horse ~ Zebra

	Cup	**Zebra**		**Kite**
	Y_{11}	Y_{12}	.....	Y_{1C}
	Y_{21}	Y_{22}	.....	Y_{2C}
	.	.	.	.
	.	.	.	.
	.	.	.	.
	Y_{P1}	Y_{P2}	.....	Y_{PC}



Similarity Common-sense



Input Image
+
Region proposals
+
Horse ~ Zebra

	Cup	Zebra	.....	Kite
	Y_{11}	0.1	.....	Y_{1C}
	Y_{21}	0.9	.....	Y_{2C}
	.	.	.	.
	.	.	.	.
	.	.	.	.
	Y_{P1}	0.3	.....	Y_{PC}

Attribute Common-sense

Car

Bowl

Dog

Attribute Common-sense

Car: Shiny/Metallic

Bowl: Round

Dog: Furry

Attribute Common-sense

Car: Shiny/Metallic

Bowl: Round

Dog: Furry

- Assumptions:
  - Three types of attributes are assigned using knowledge base:
    - **Color**, ex: Banana -> Yellow
    - **Shape**, ex: Apple -> Round
    - **Physical properties**, ex: Spoon -> Shiny
  - Pretrained attribute classifier

Attribute Common-sense

Car: Shiny/Metallic

Bowl: Round

Dog: Furry

- Assumptions:
  - Three types of attributes are assigned using knowledge base:
    - **Color**, ex: Banana -> Yellow
    - **Shape**, ex: Apple -> Round
    - **Physical properties**, ex: Spoon -> Shiny
  - Pretrained attribute classifier

Red

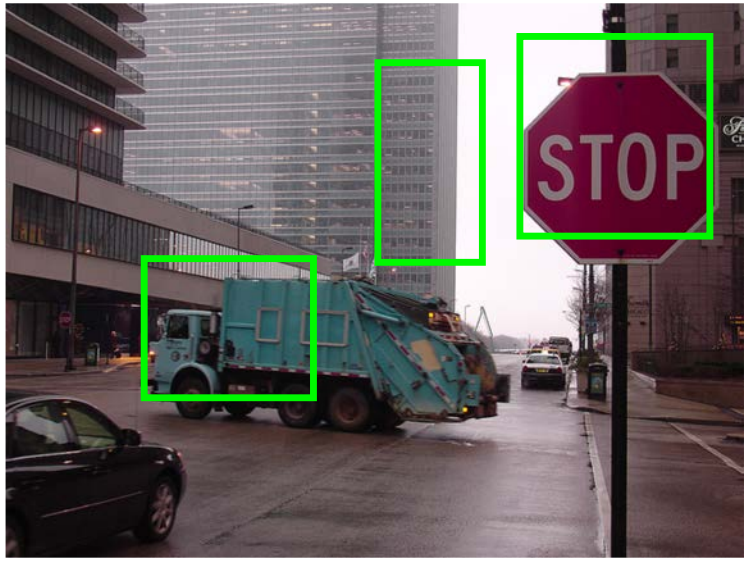
Round

Wooden

Attribute Common-sense

Class: Stop Sign

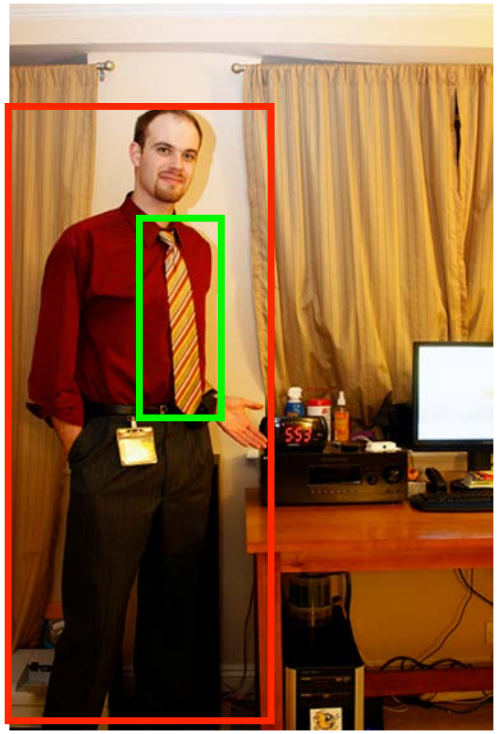
Attributes: Red, Round, Shiny



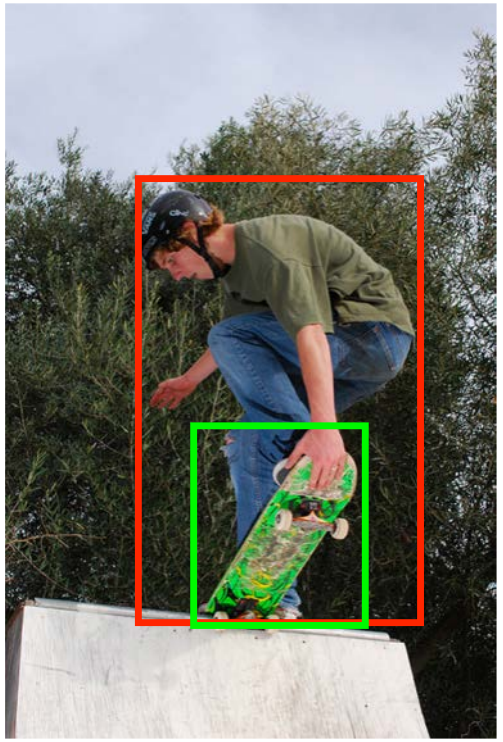
	Cup	Stop	.....	Kite
	Y_{11}	0.3	.....	Y_{1C}
	Y_{21}	0.8	.....	Y_{2C}
	.	.	.	.
	.	.	.	.
	.	.	.	.
	Y_{P1}	0.1	.....	Y_{PC}

Spatial Common-sense

Spatial Common-sense



Person-> Tie



Person-> Skateboard

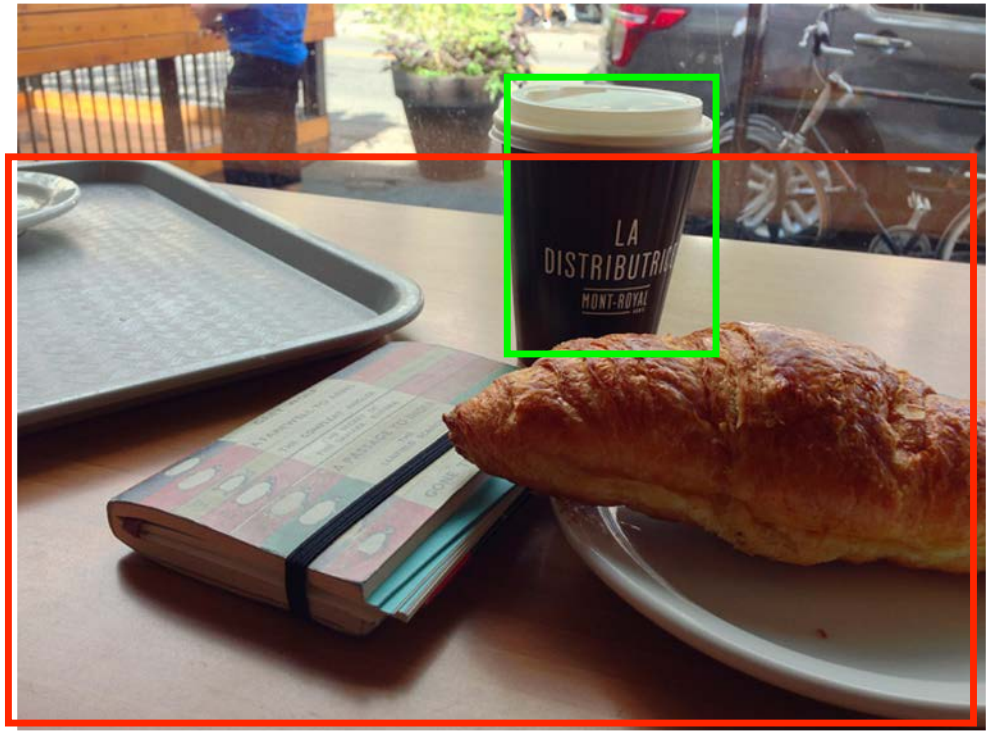


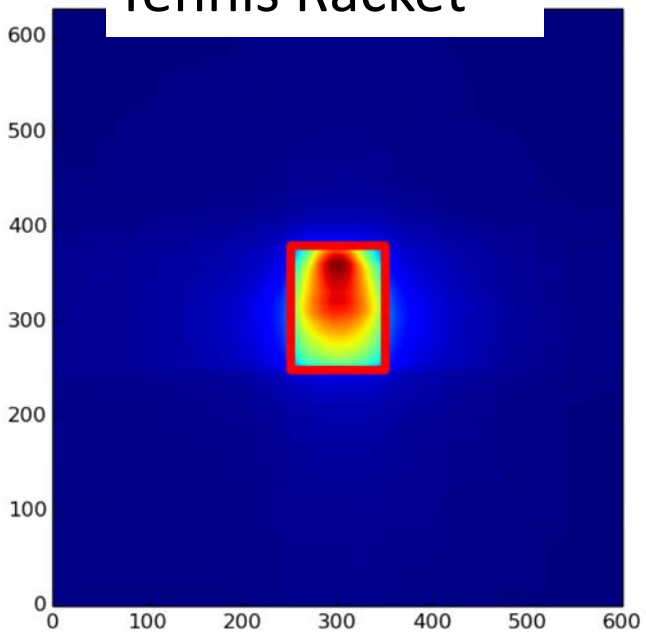
Table-> Cup

- Assumption:
  - Spatial relationship of source and target classes are known through visual genome

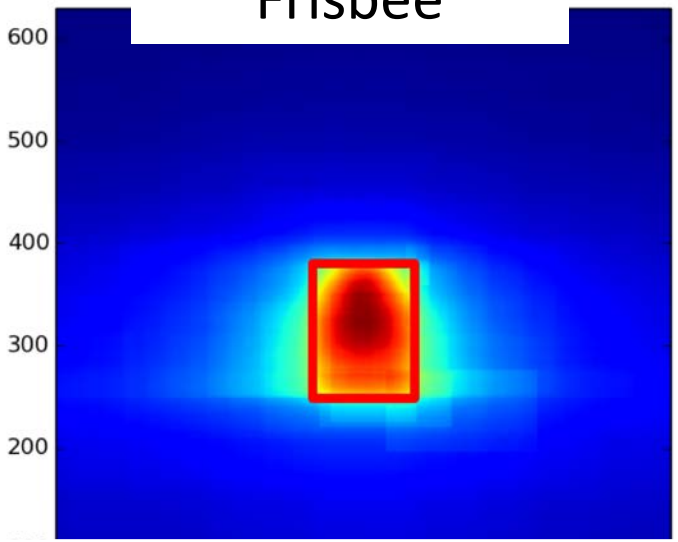
Creating Spatial Relation Maps

Person

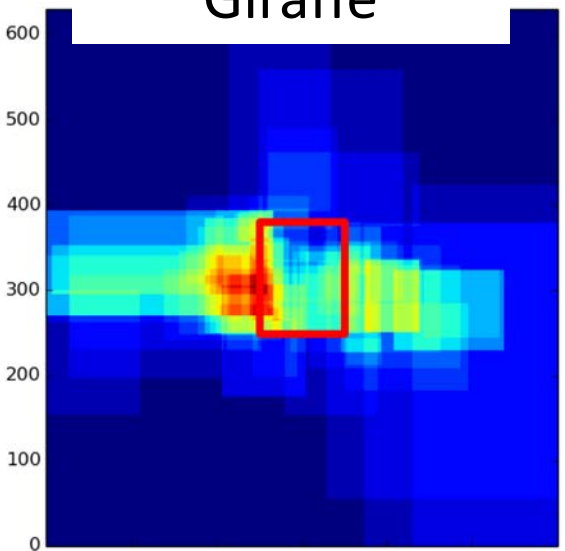
Tennis Racket



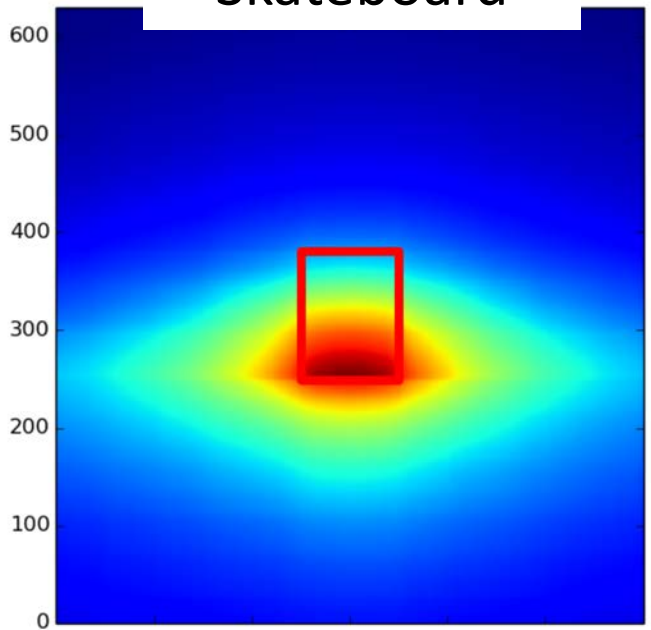
Frisbee



Giraffe



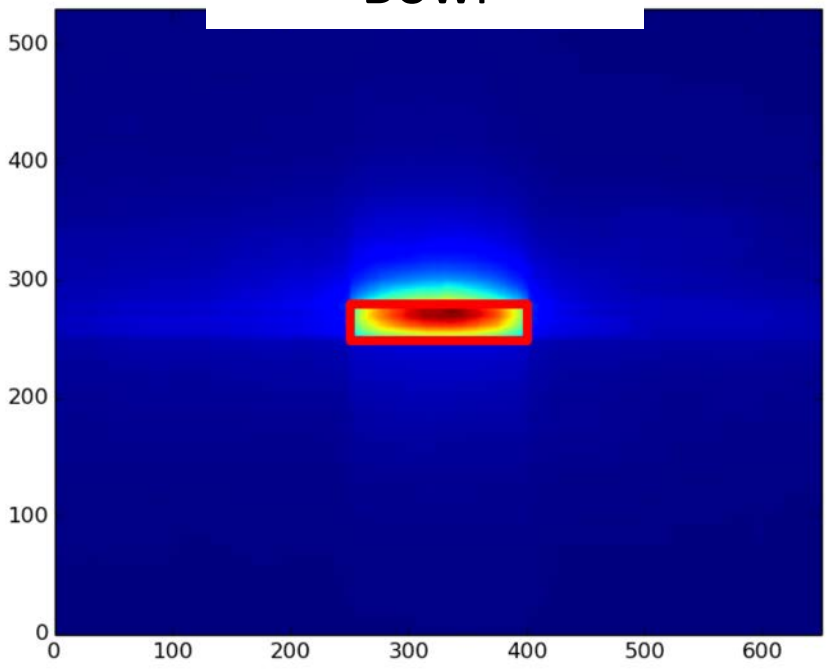
Skateboard



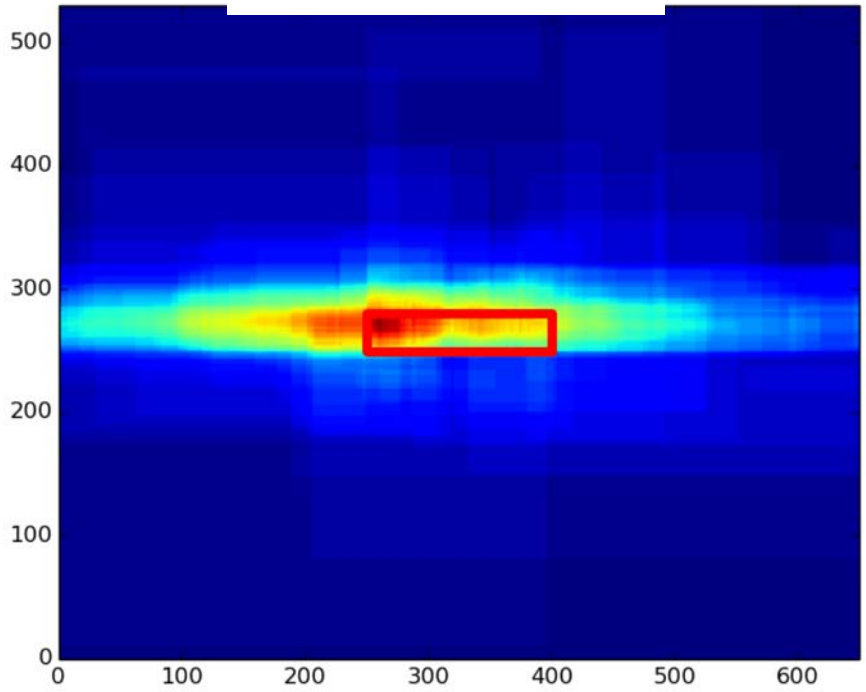
Creating Spatial Relation Maps

Table

Bowl

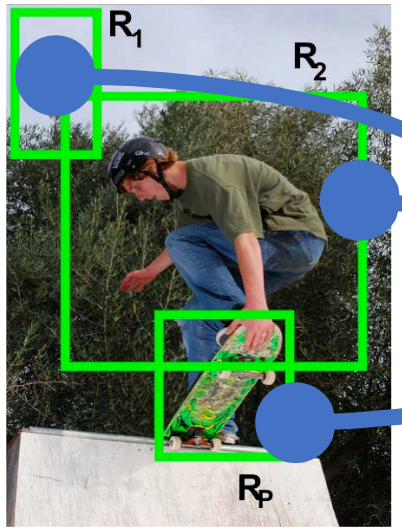


Bed



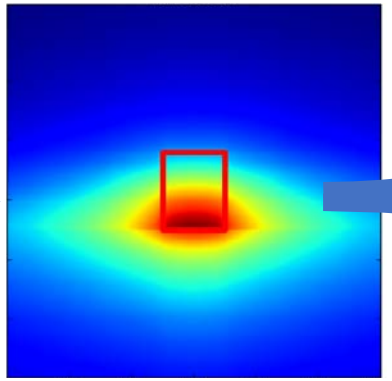
Spatial Common-sense

Image with region proposals



C_{vis} = {'person'}

rel('person', 'skateboard') = 'along'



γ('person', 'along')

Spatial Y_{SP}

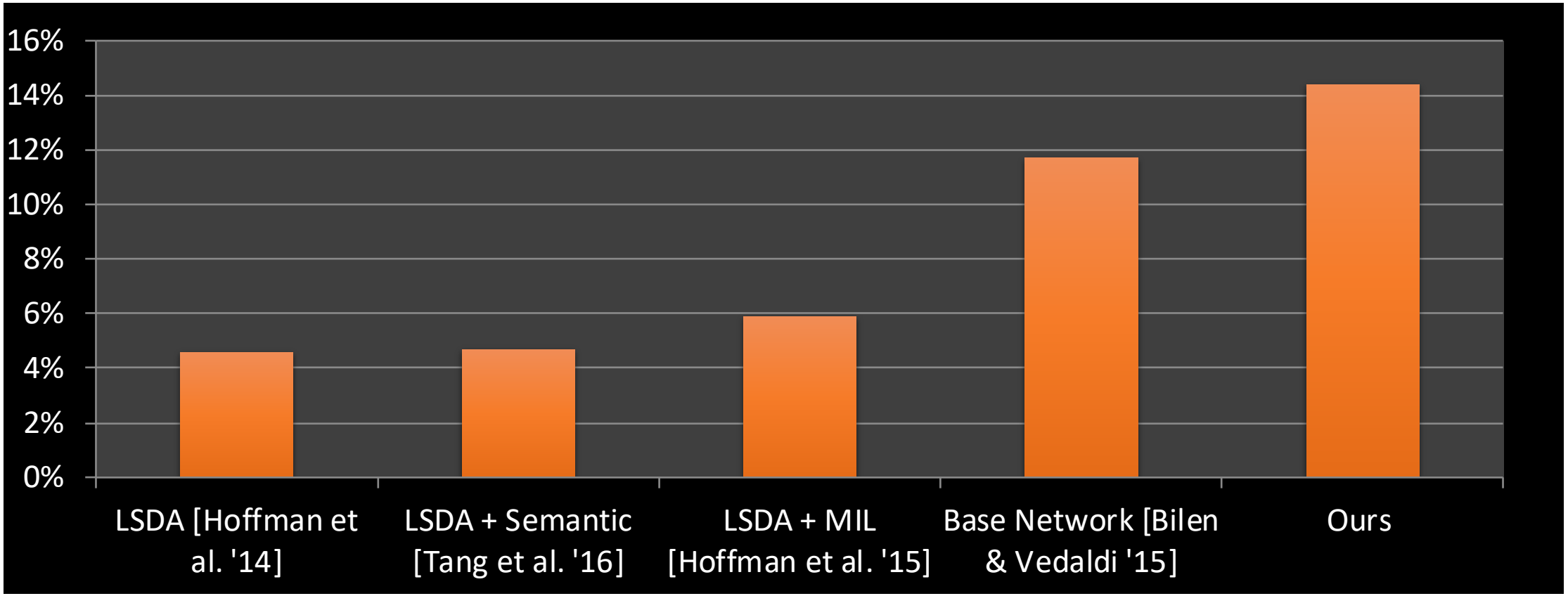
$Y_{1,1}$	.....	0.1	.....	$Y_{1,C}$
$Y_{2,1}$	.....	0.4	.....	$Y_{2,C}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$Y_{P,1}$	.....	1.0	.....	$Y_{P,C}$

skateboard

Quantitative Results

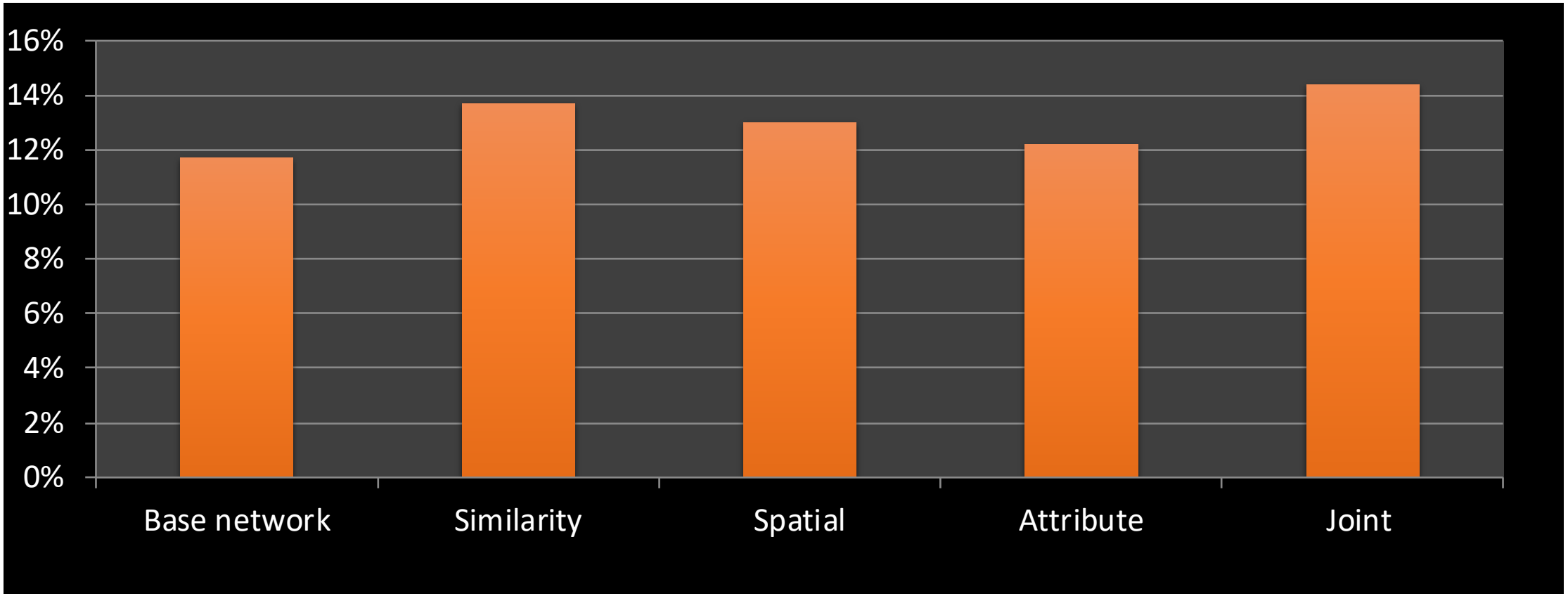
- **Source categories:** 20 PASCAL classes; *bounding box annotations*
- **Target categories:** 60 MS COCO classes; *image-level annotations*
- Use WSDDN as baseline (without common-sense) [Bilen et al. 2016]
- Compare with approaches based on LSDA [Hoffman et al. 2014]

MS COCO Object Detection mAP



- Significant boost over existing approaches

Ablation Study



- Each type of common-sense contributes and they are complementary

Qualitative Results

LSDA + semantic [Tang ‘16]

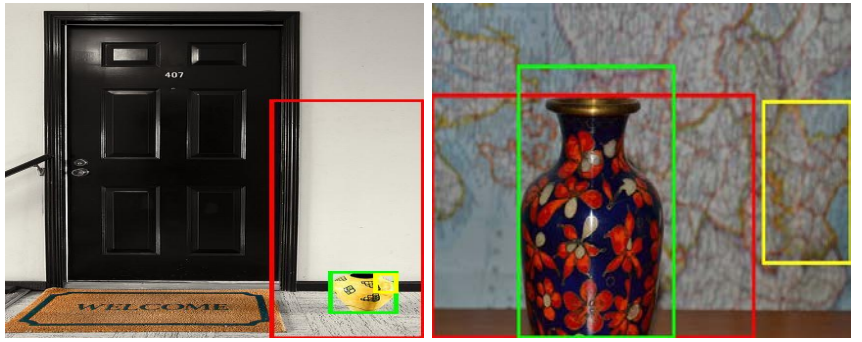
Appearance-only WSDDN [Bilen ‘15]

Ours using common-sense

Similarity

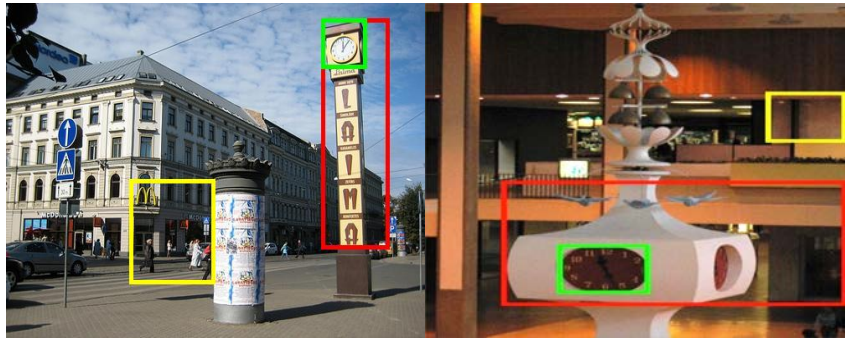
Giraffe

Vase



Attribute

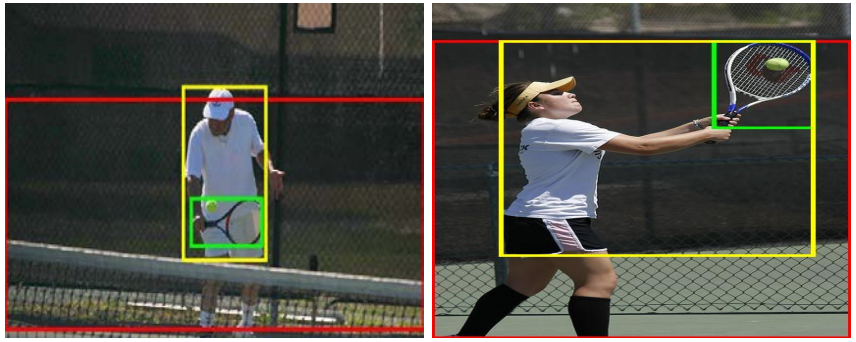
Clock



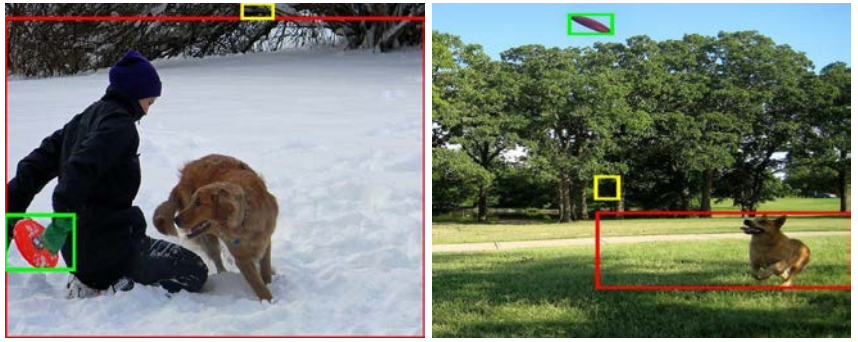
Oven

Spatial

Tennis Racket



Frisbee



Failure Cases

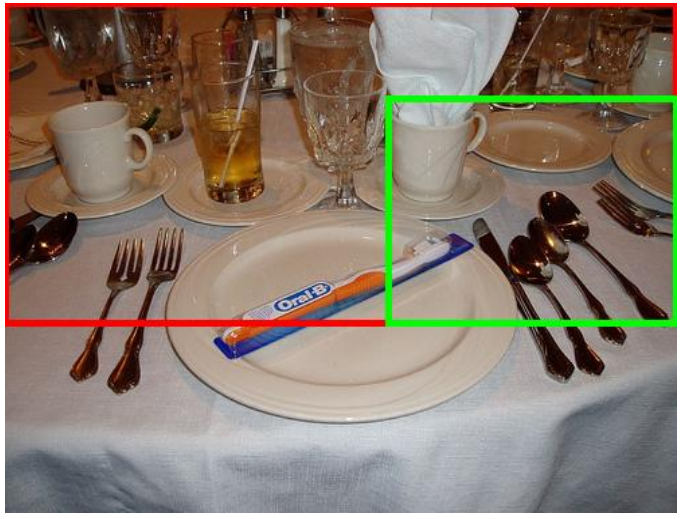
Laptop

Cup

Failure Cases

Laptop

Spoon



Cup

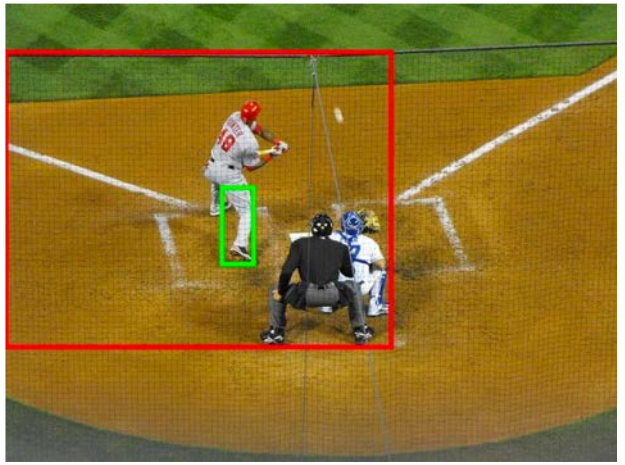
Handbag

Failure Cases

Laptop

Spoon

Baseball Bat



Cup

Handbag

Baseball Bat

Questions?