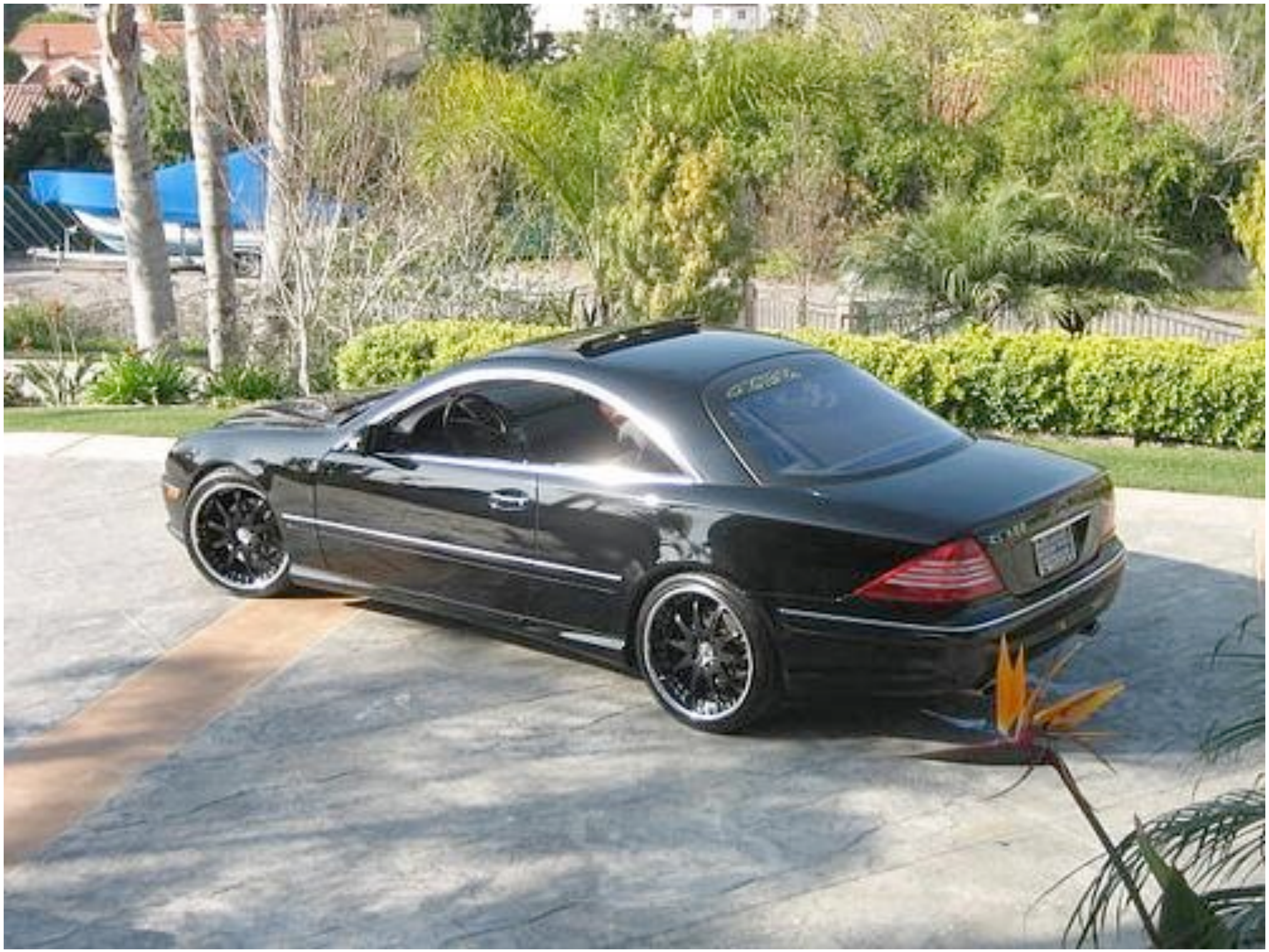
Microsoft Research - Inria
JOINT CENTRE

Willow project-team

Learning and transferring mid-level image representions using convolutional neural networks

Maxime Oquab,
Léon Bottou, Ivan Laptev, Josef Sivic

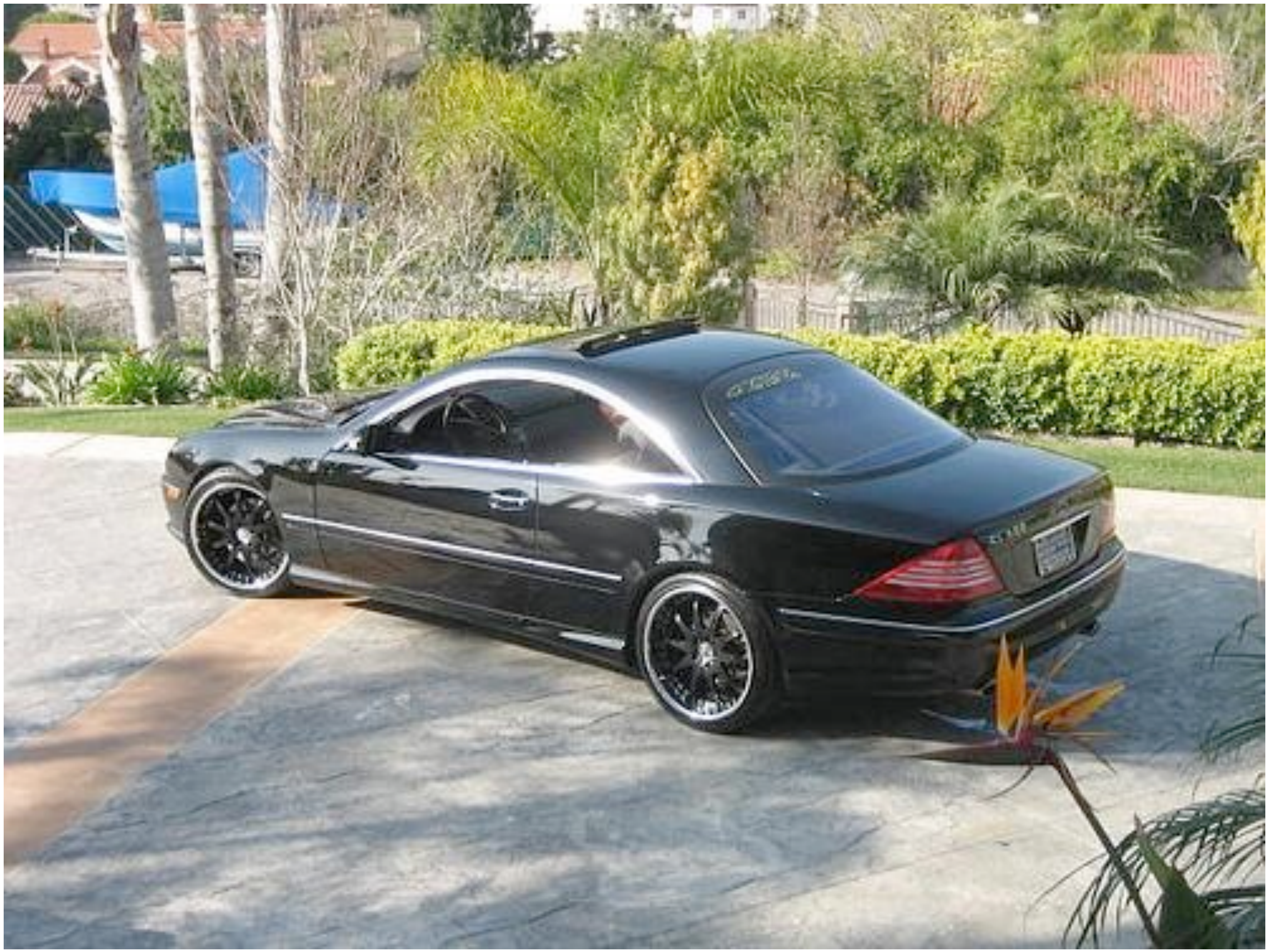
Image classification (easy)



Is there a **car** ?

Source : Pascal VOC dataset

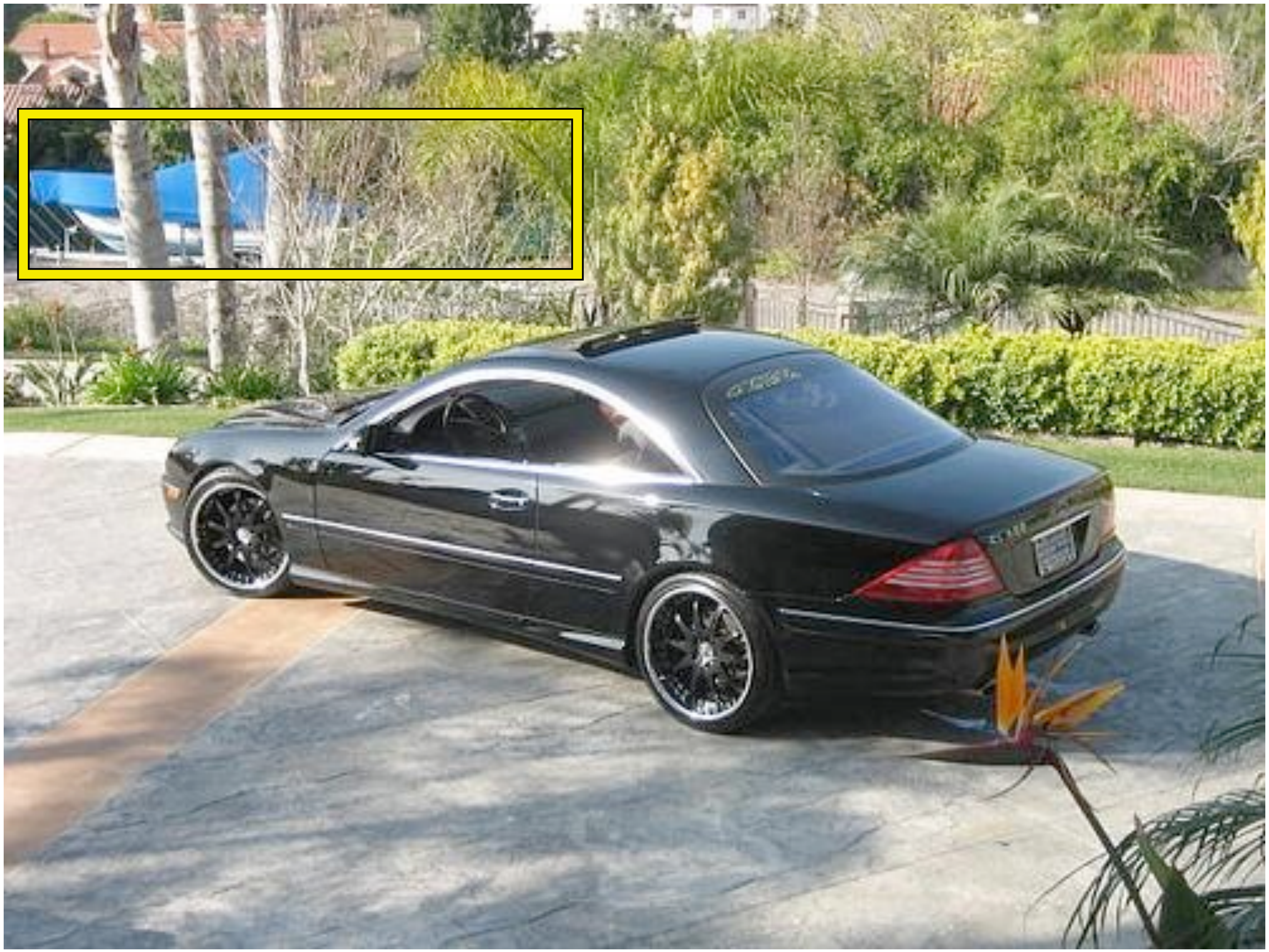
Image classification (harder)



Is there a **boat** ?

Source : Pascal VOC dataset

Image classification (harder)



Is there a **boat** ?

Source : Pascal VOC dataset

Image classification (v.hard)

Is there
a **person** ?

Source : Pascal VOC dataset

Image classification (v.hard)

Source : Pascal VOC dataset

Pascal VOC vs. ImageNet classification

Pascal VOC :
complex scenes
20 object classes
10k images

ImageNet :
object-centric
1000 object classes
1.2M images

Image classification

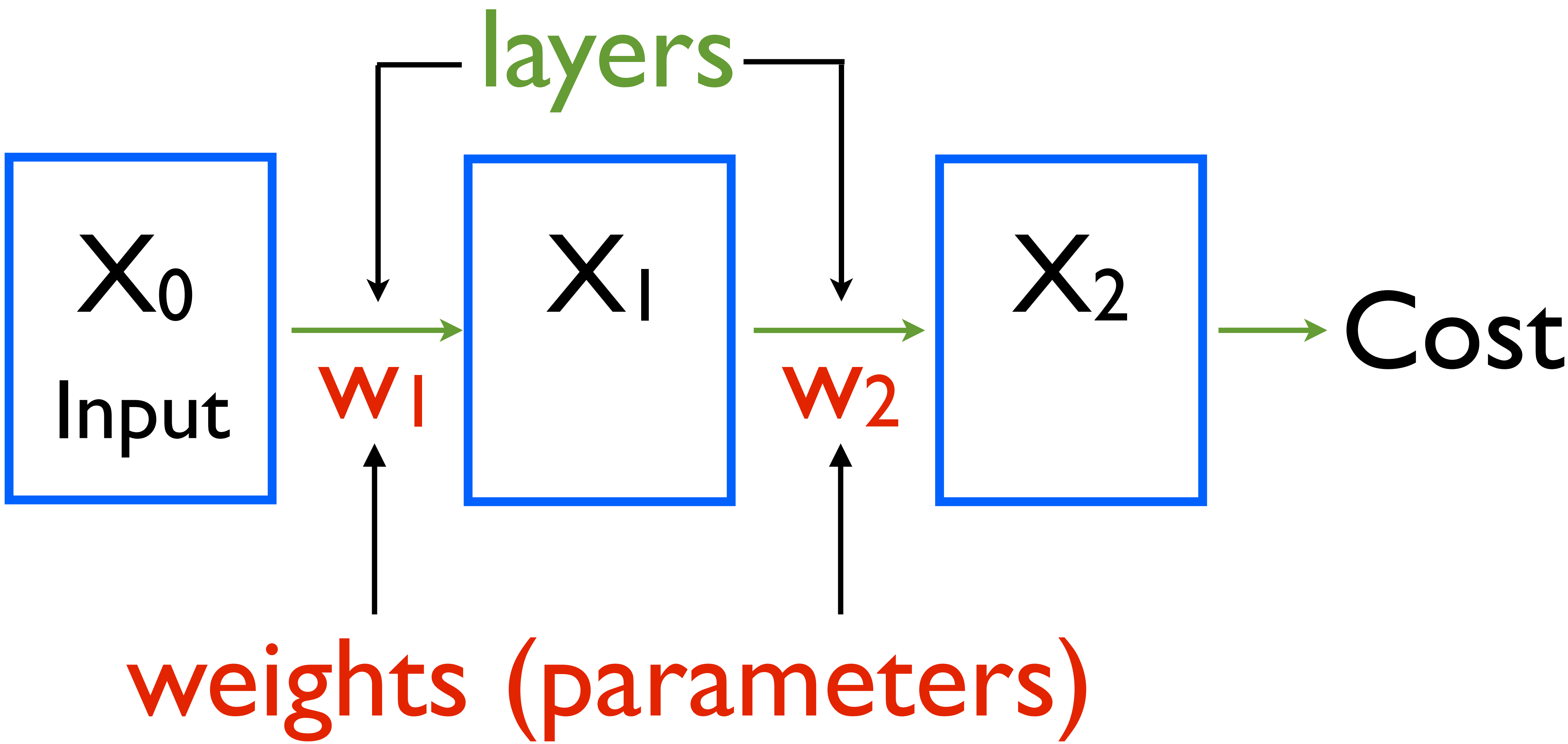
- Traditional methods: HOG, SIFT, FV, SVMs, DPM, k-Means, GMM...
  [Csurka et al.'04], [Lowe'04], [Sivic & Zisserman'03], [Perronin et al.'10], [Lazebnik et al.'06], [Zhang et al. '07], [Boureau et al.'10], [Singh et al.'12], [Juneja et al.'13], [Chatfield et al. '11], [van Gemert et al. '08], [Wang et al. '10], [Zhou et al. '10], [Dong et al. '13], [Feifei et al. '05], [Shotton et al. '05], [Moosmann et al.'05], [Grauman & Darrell '05] [Harzallah et al. '09], [...]

- Convolutional neural networks
  ImageNet challenge
  [Krizhevsky et al. 2012]

Brief history of CNNs

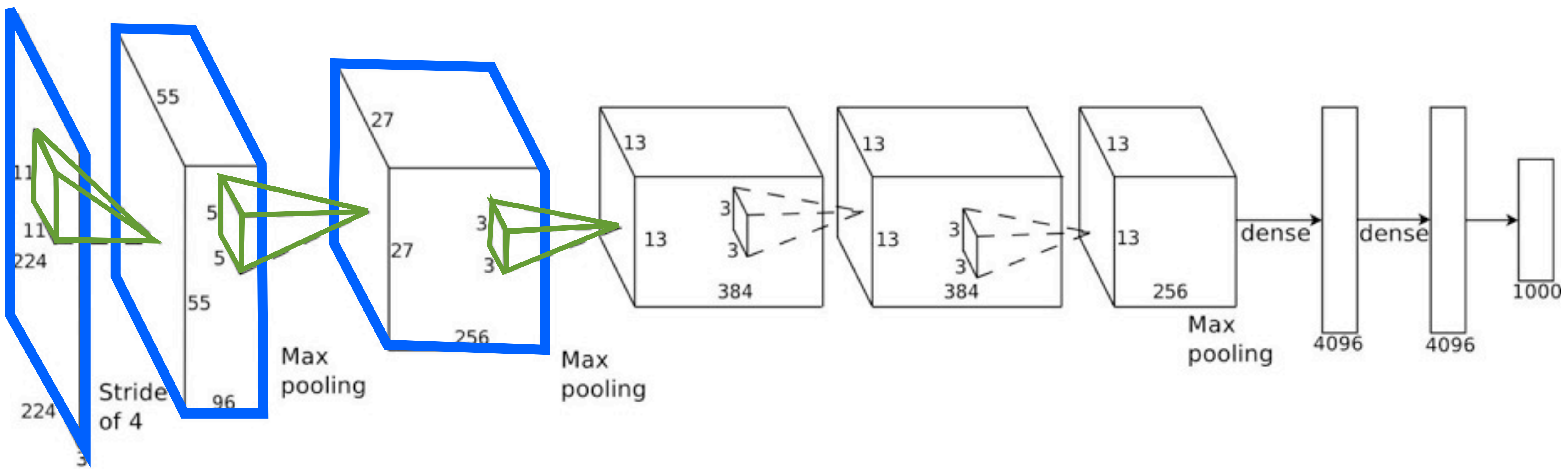
- **Rosenblatt, 1957 : *The perceptron : a perceiving and recognizing automaton.***
- Hubel & Wiesel 1959 : *Receptive fields of single neurons in the cat's striate cortex*
- Fukushima 1980 : *Neocognition*
- Rumelhart et al. 1986 : Learning representations by back-propagating errors
- **LeCun et al. 1989 : *Backpropagation applied to handwritten zip code recognition.***
- LeCun et al. 1998 : *Efficient Backprop*
- LeCun et al. 1998 *: Gradient-based learning applied to document recognition*
- Hinton & Salakhutdinov, 2006 : *Reducing the Dimensionality of Data with Neural Networks*
- **Krizhevsky et al. 2012 : *ImageNet classification with deep convolutional neural networks.***
- Zeiler & Fergus, 2013 : *Visualizing and understanding neural networks*
- Sermanet et al. 2013 : *Overfeat,*
- Donahue et al. 2013 : *Decaf*
- Girshick et al. 2014 : *Rich feature hierarchies for accurate object detection and semantic segmentation*
- Razavian et al. 2014 : *CNN features off-the-shelf, an astounding baseline for recognition*
- Chatfield et al. 2014 : *Return of the devil in the details*

Neural Networks



Differentiable operations :
weights trained by gradient descent.

8-layer NN
[Krizhevsky et al.]



60 million parameters :

- ImageNet (1.2M images) : OK
- Pascal VOC (10k images) : ?

Pascal VOC : different task

Car examples from
Pascal VOC

Typical car examples
from ImageNet

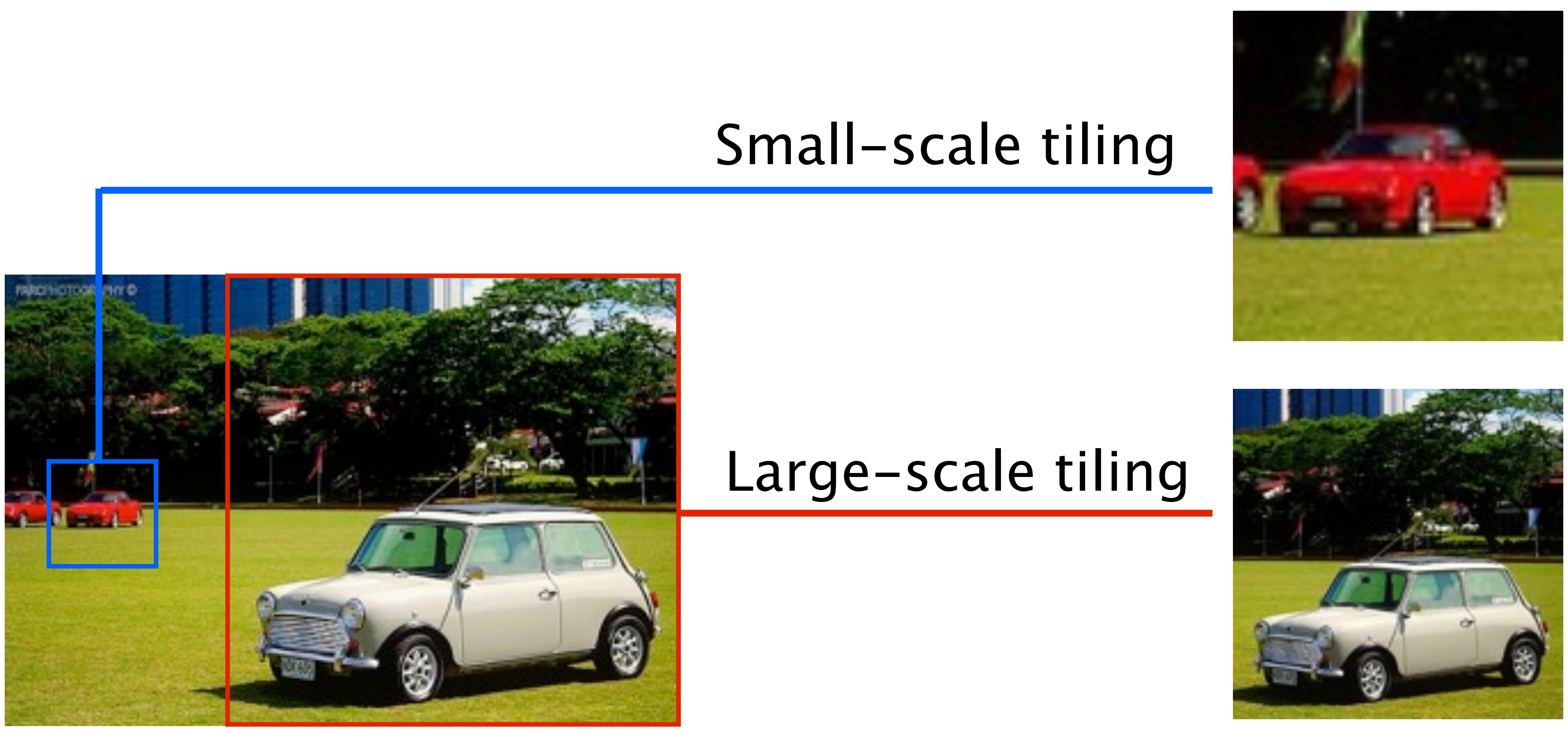
Pascal VOC : different task

Car examples from
Pascal VOC

Typical car examples
from ImageNet

Solution : multi-scale patch tiling

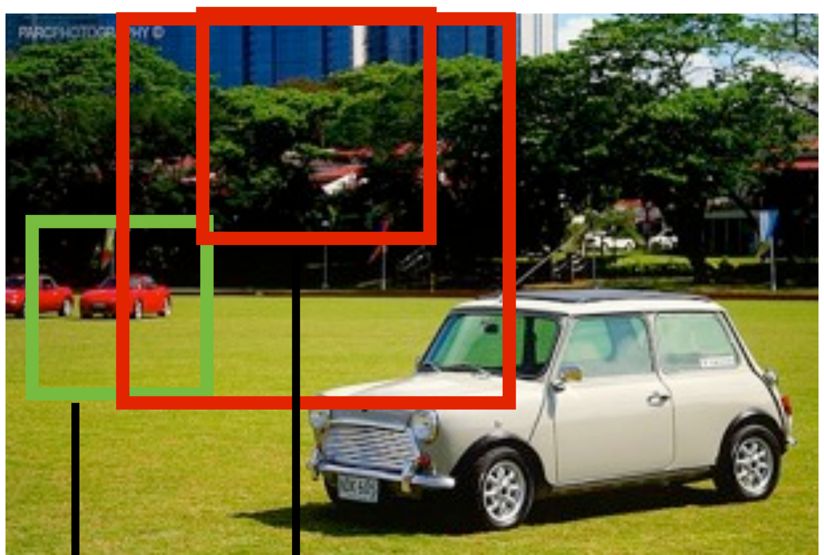
- Goal : obtain a **dataset that looks like ImageNet.**



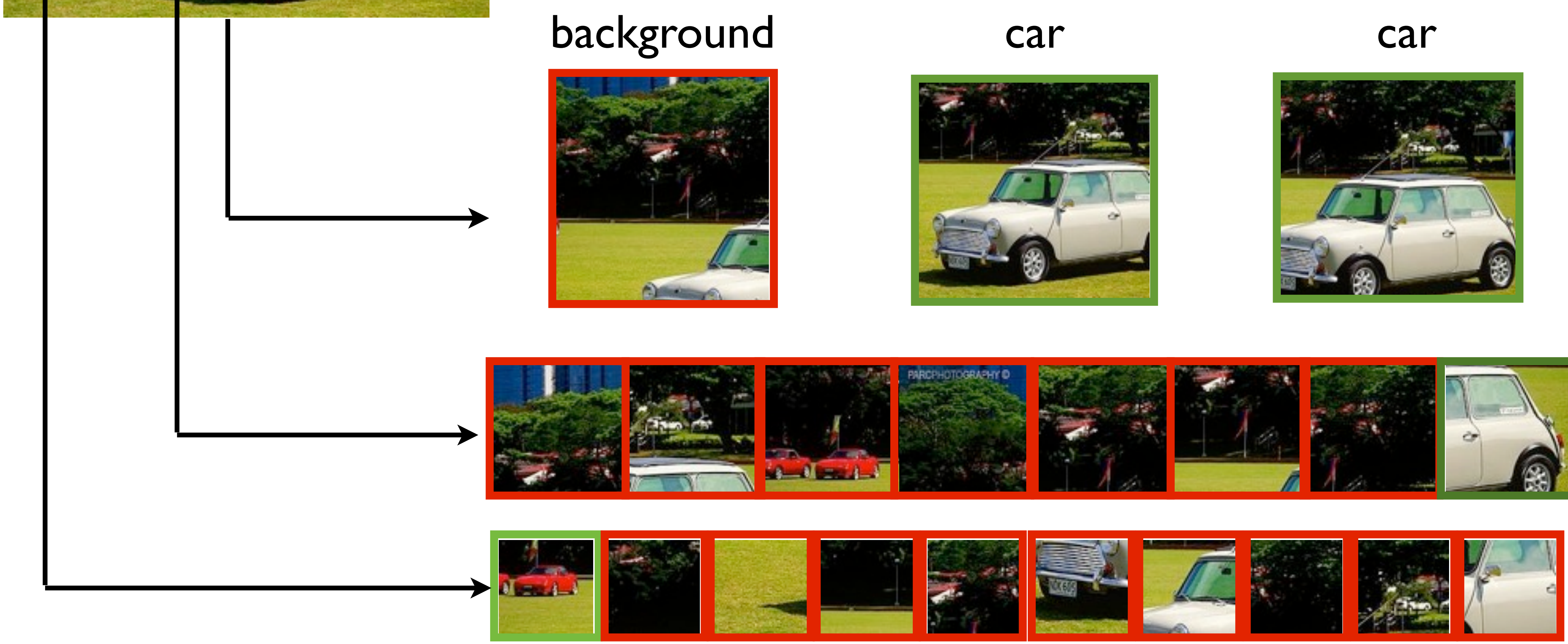
Typical Pascal VOC car example ... **... in disguise**

Typical car examples from ImageNet

Solution : multi-scale patch tiling



- Around 500 tiles per image.
- Multiple scales and positions.
- Label depending on overlap.

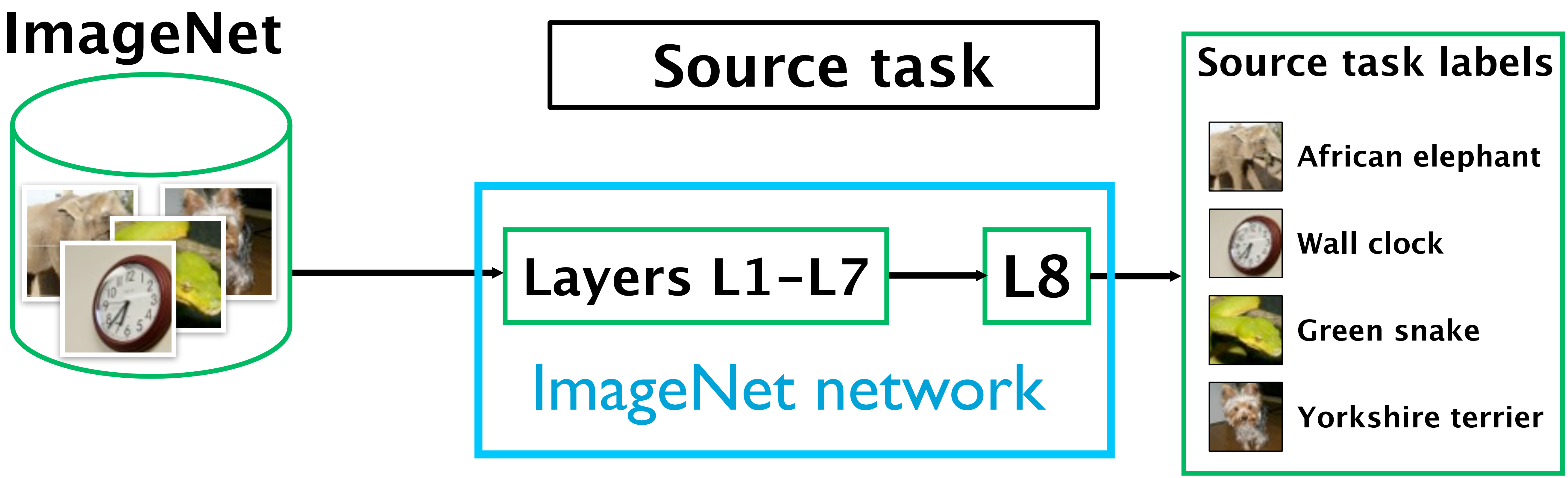


First attempt

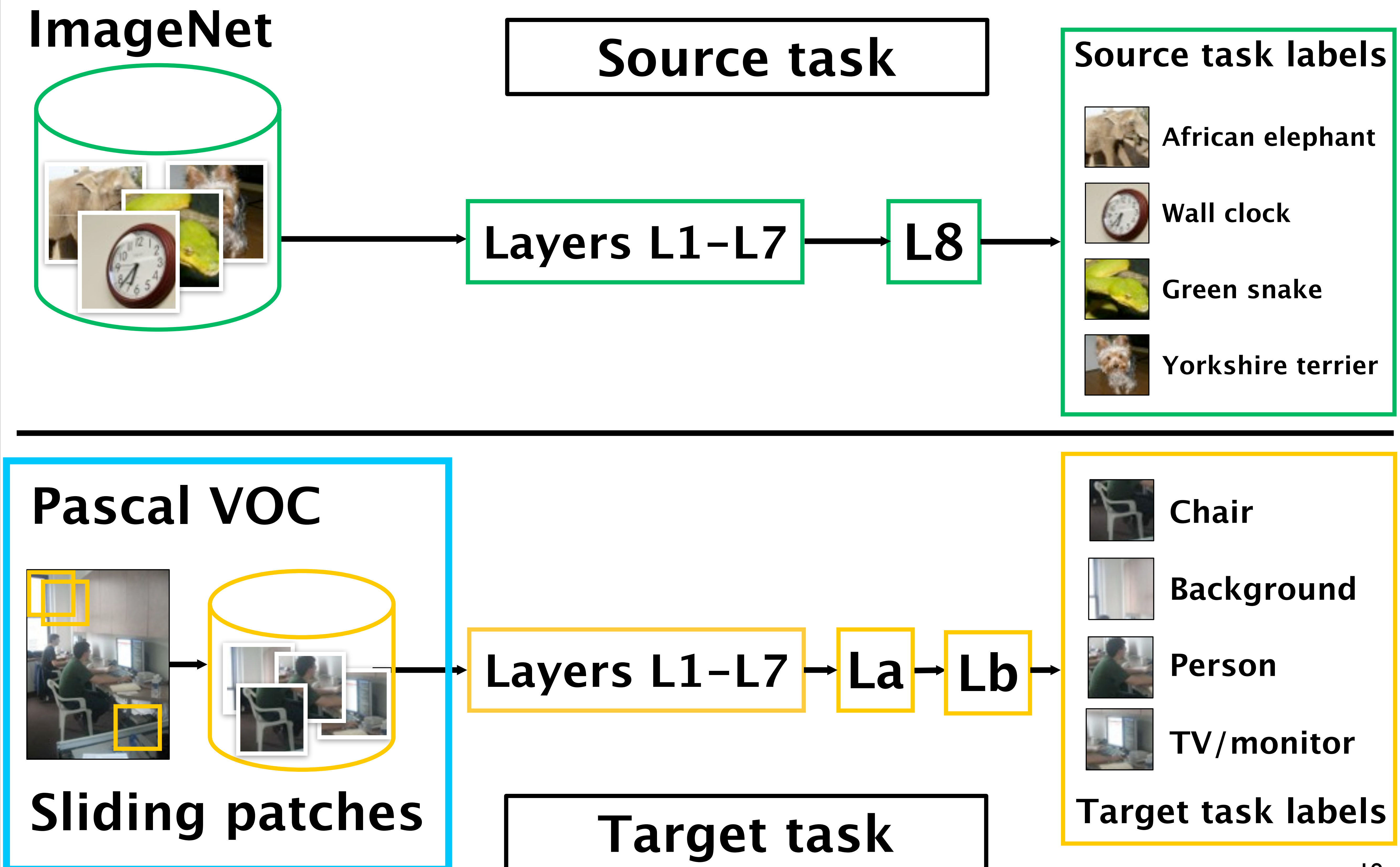
- Train CNN on Pascal VOC patches :
  - Result : 70.9% mAP.
  - We observe **overfitting**.
  - State of the art : 82.2% mAP (NUS-PSL).
- How to benefit from the power of neural networks ?

  We propose **transfer learning**.

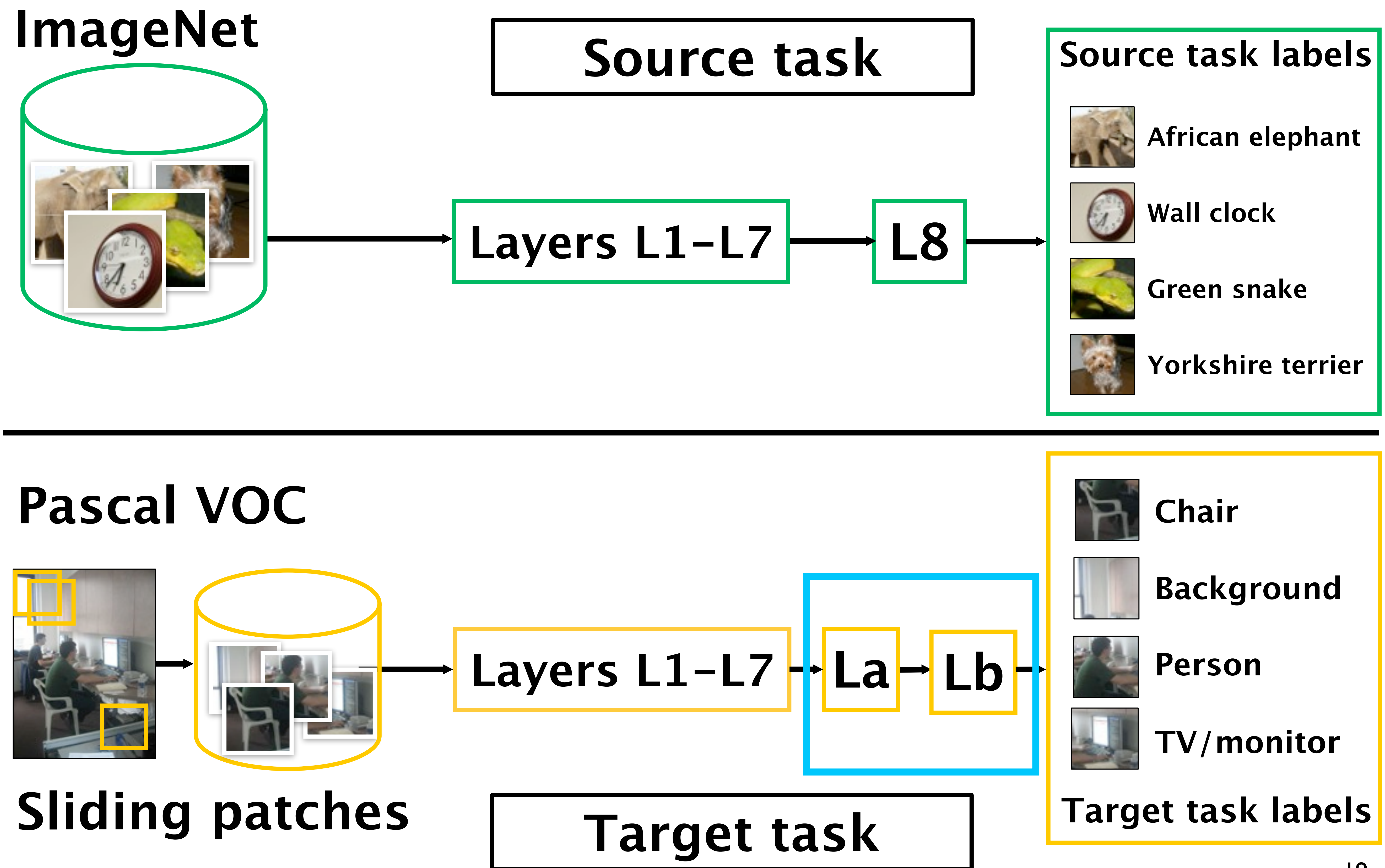
Transfer learning



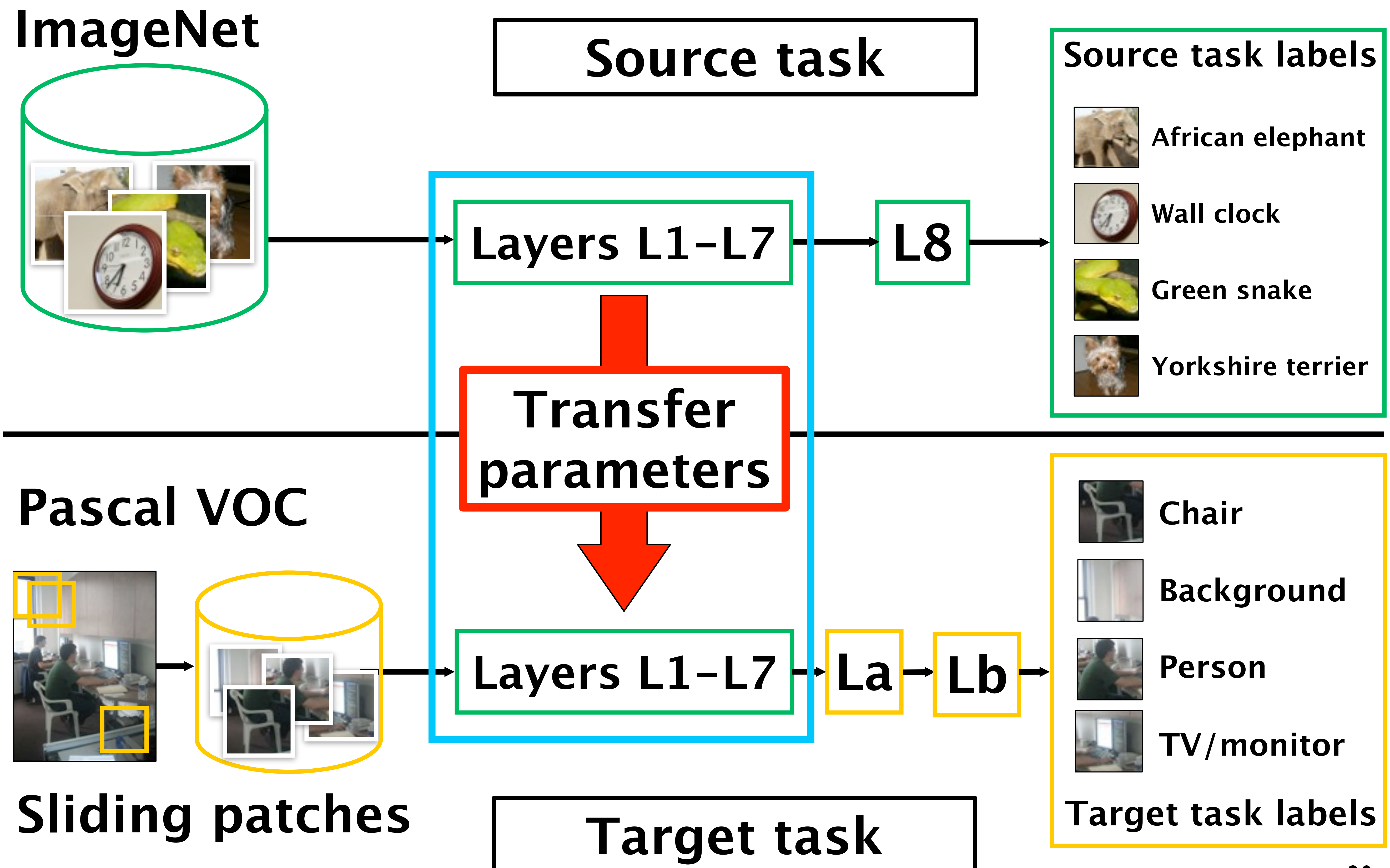
Transfer learning



Transfer learning



Transfer learning



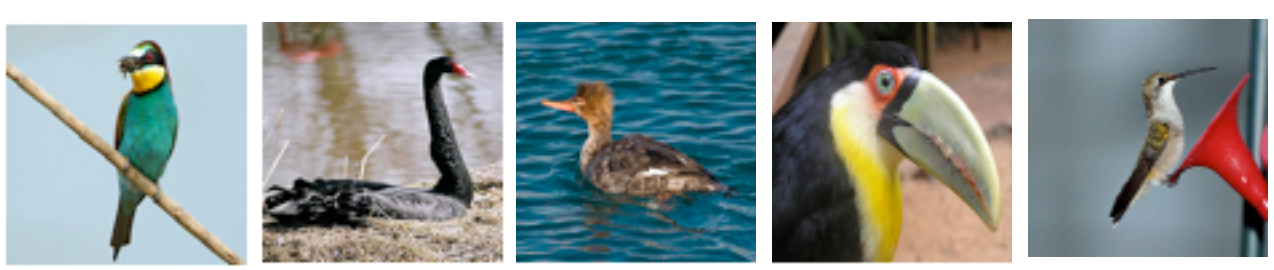
Second attempt (with pre-training)

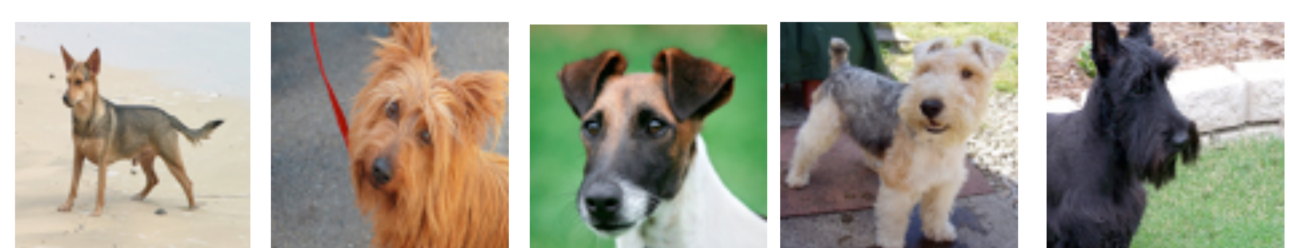
- After pre-training on the ILSVRC-2012 dataset, we obtain 78.7% mean AP (no pre-train : 70.9%).
- Significantly better but can we improve more ?

	plane	bike	bird	boat	btl	bus	car	cat	chair	cow	table	dog	horse	moto	pers	plant	sheep	sofa	train	tv	mAP
NUS-PSL [48]	**97.3**	**84.2**	80.8	**85.3**	**60.8**	**89.9**	**86.8**	89.3	**75.4**	77.8	**75.1**	83.0	87.5	**90.1**	95.0	57.8	79.2	**73.4**	**94.5**	**80.7**	82.2
NO PRETRAIN	85.2	75.0	69.4	66.2	48.8	82.1	79.5	79.8	62.4	61.9	49.8	75.9	71.4	82.7	93.1	59.1	69.7	49.3	80.0	76.7	70.9
PRE-1000C	93.5	78.4	*87.7*	80.9	57.3	85.0	81.6	*89.4*	66.9	73.8	62.0	*89.5*	83.2	87.6	*95.8*	*61.4*	79.0	54.3	88.0	78.3	78.7

+18 % (bird) +14 % (dog)

- Observe large boosts for dog and bird classes.
- Well-represented groups in ILSVRC-2012.

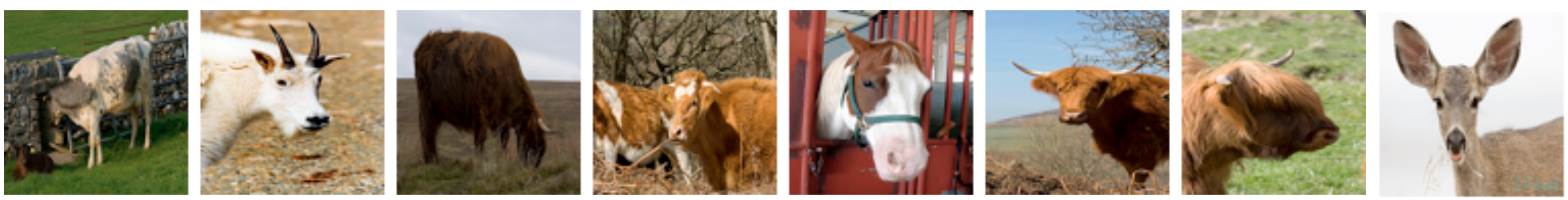




Pre-training data

- Inspect 22k classes of the ImageNet tree:
  - «furniture» subtree contains **chairs, dining tables, sofas**

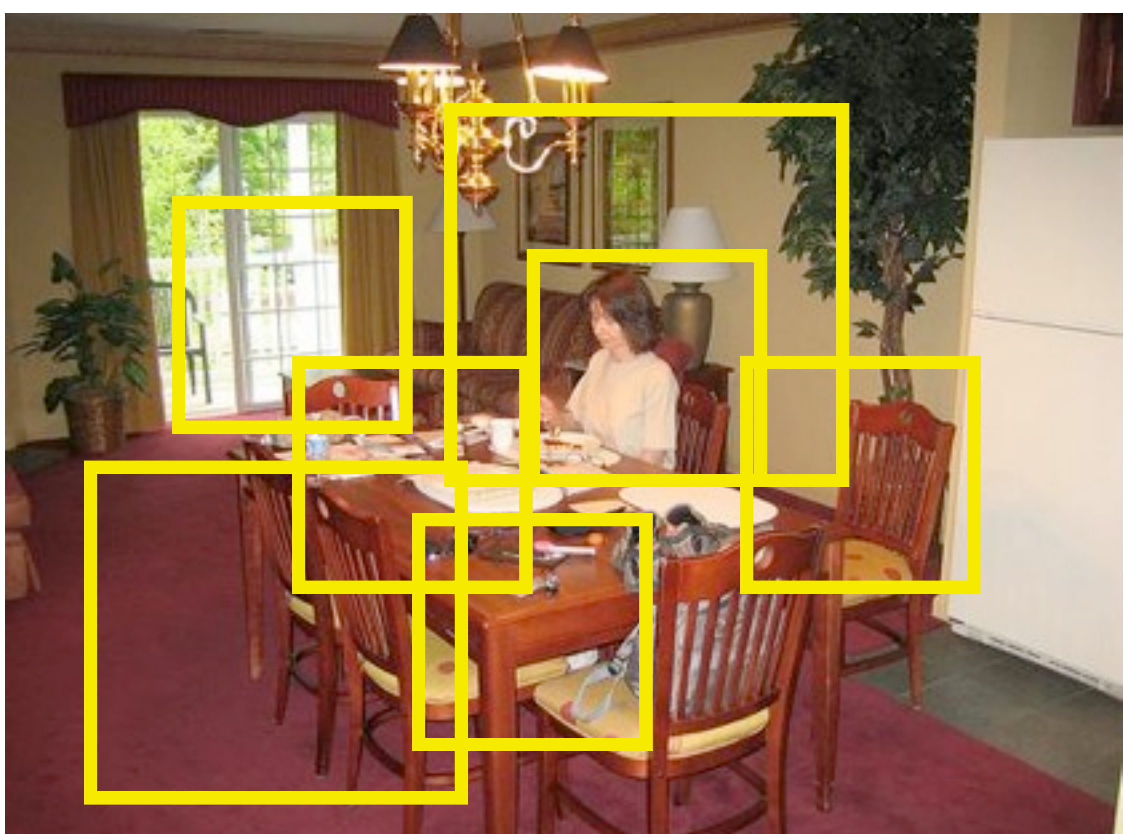
  - «hoofed mammal» subtree contains **sheep, horses, cows**



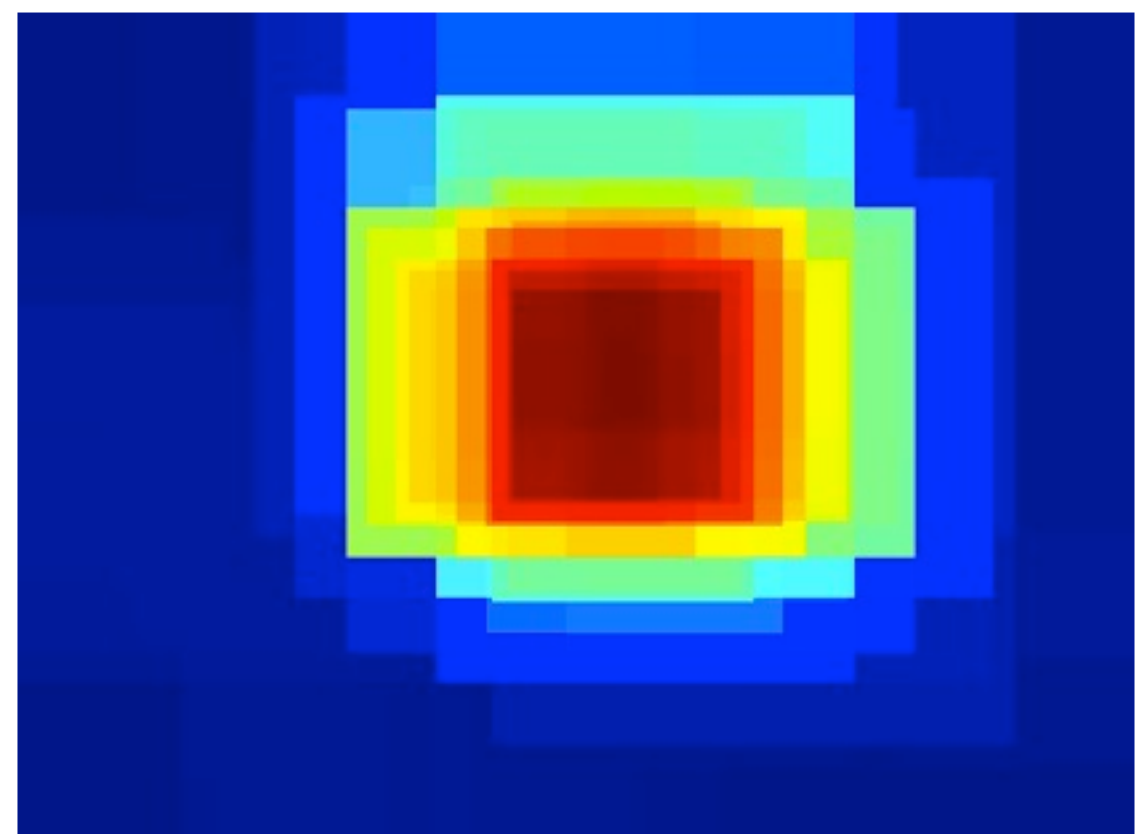
  - ...
- Add 512 classes to the pre-training,
- Result improves from 78.8% to **82.8%** mAP.
- All scores increase, targeted classes improve more.

Computing scores at test time

- We extract 500 multi-scale patches.
- **Image score = sum of all patch scores.**
- **Pixel score = sum of overlapping patches scores (heat maps)**

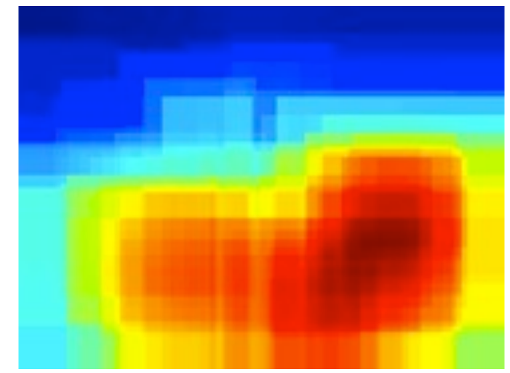


CNN
person
classifier

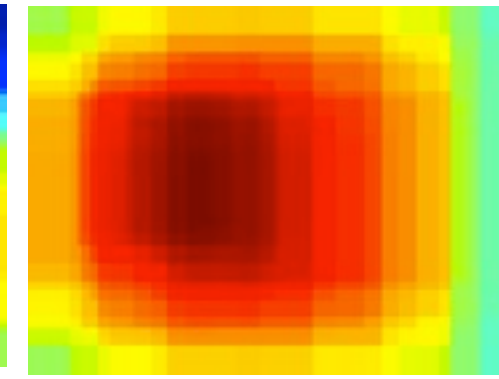


Qualitative results

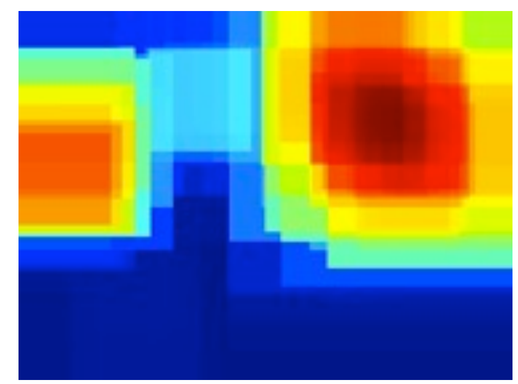
Chair



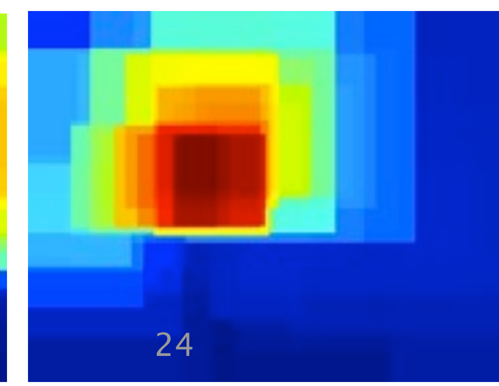
Dining table



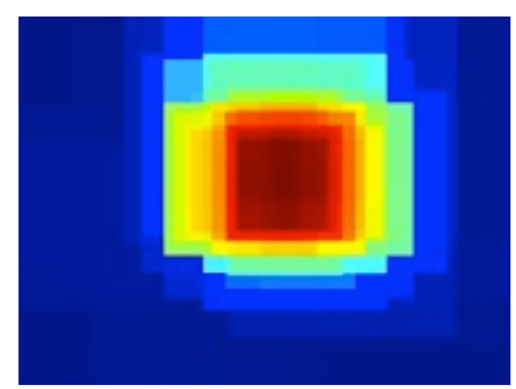
Potted plant



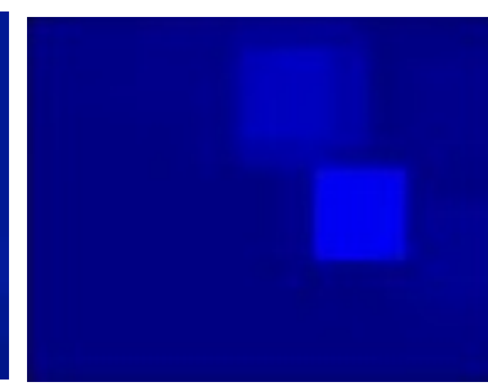
Sofa



Person

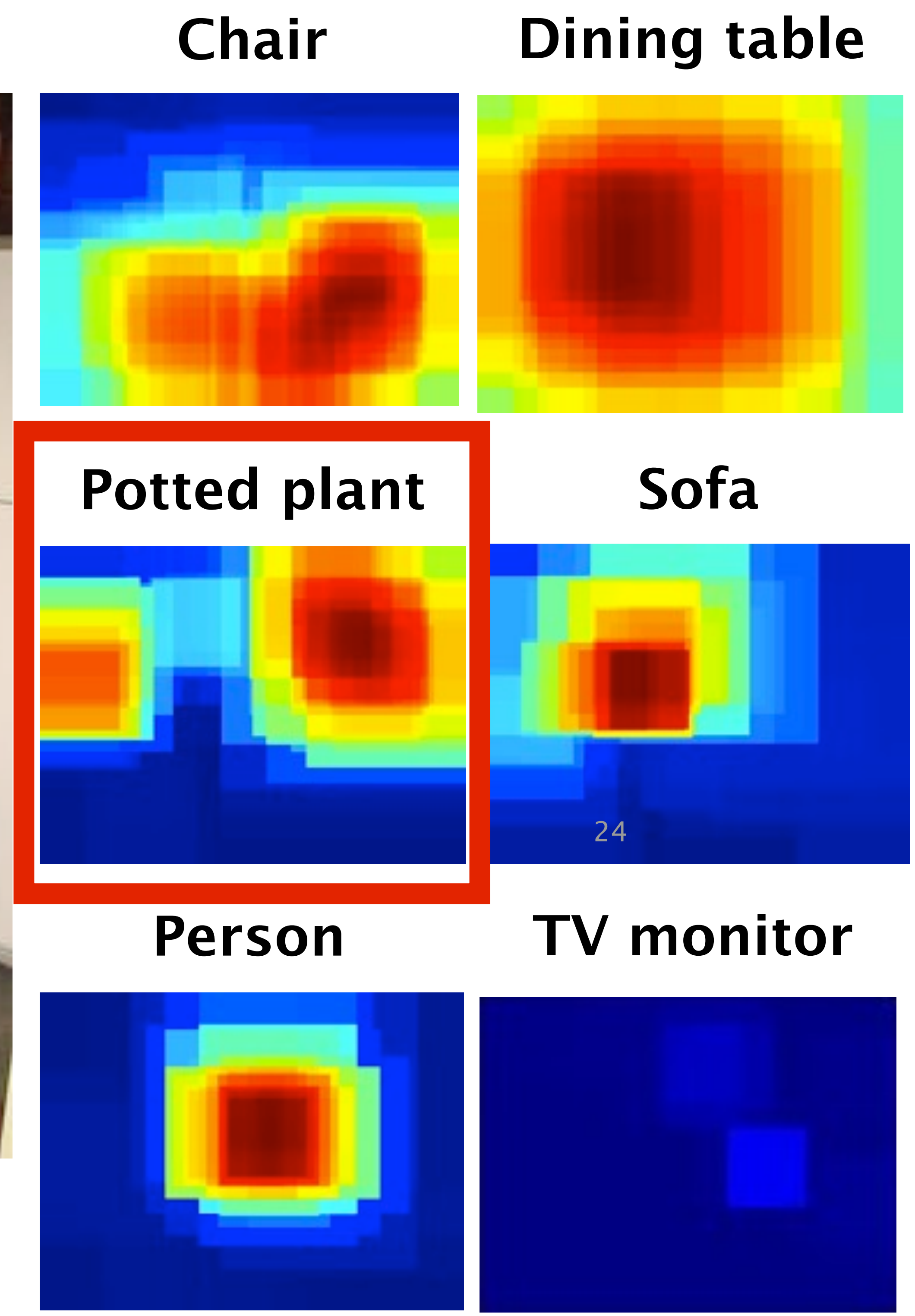


TV monitor



Source : Pascal VOC'12 test set

Qualitative results

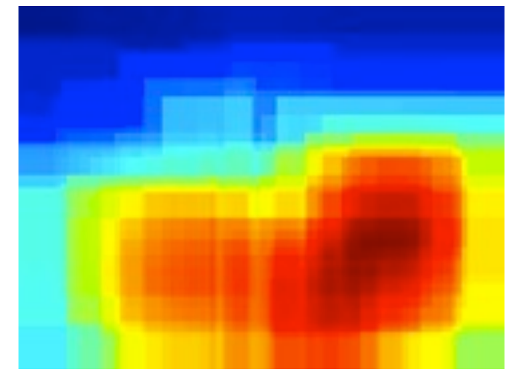


Source : Pascal VOC'12 test set

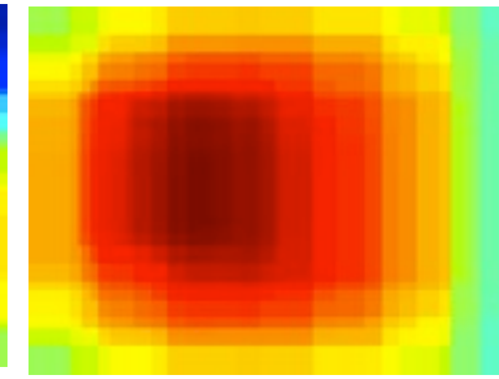
Qualitative results

Source : Pascal VOC'12 test set

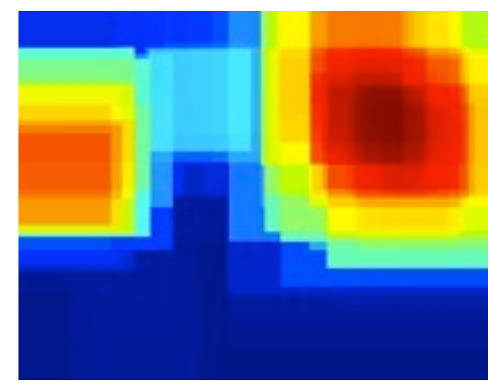
Chair



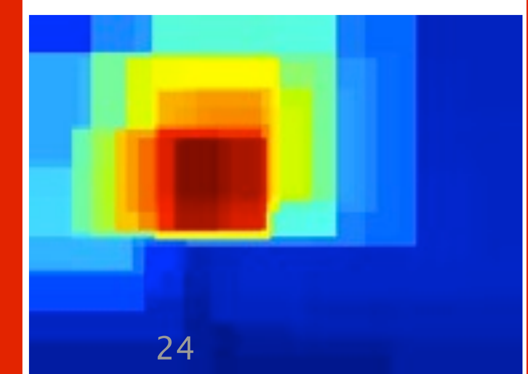
Dining table



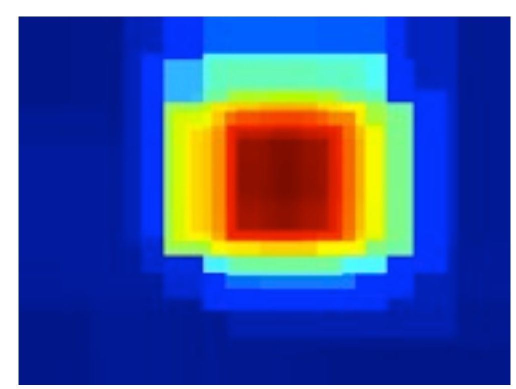
Potted plant



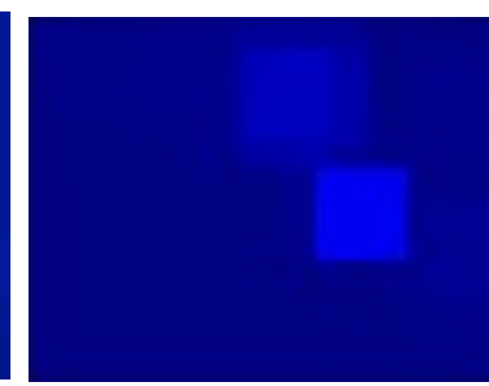
Sofa



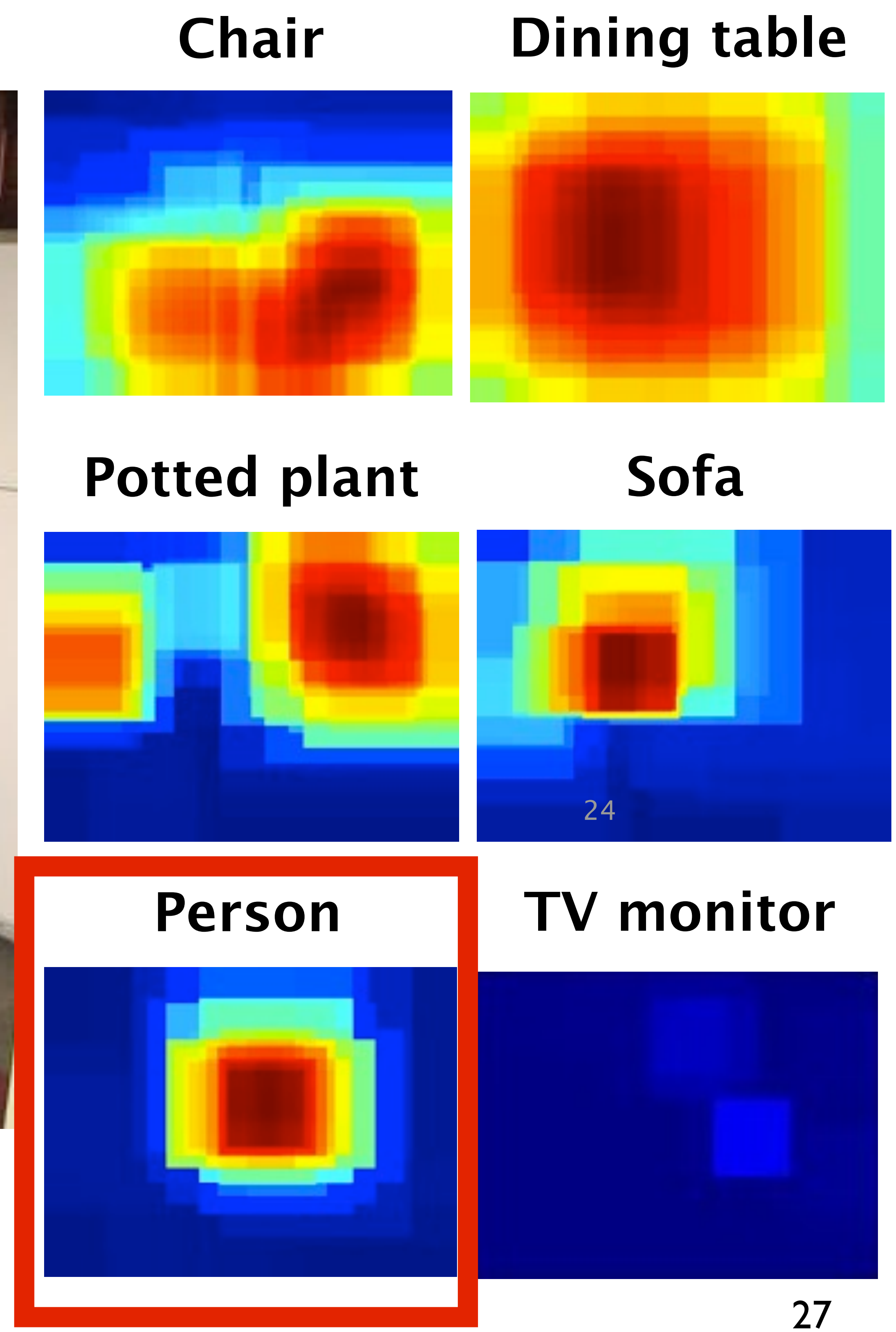
Person



TV monitor

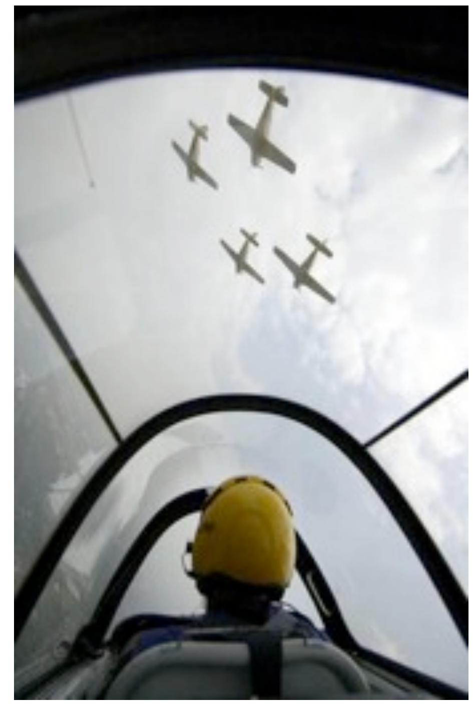


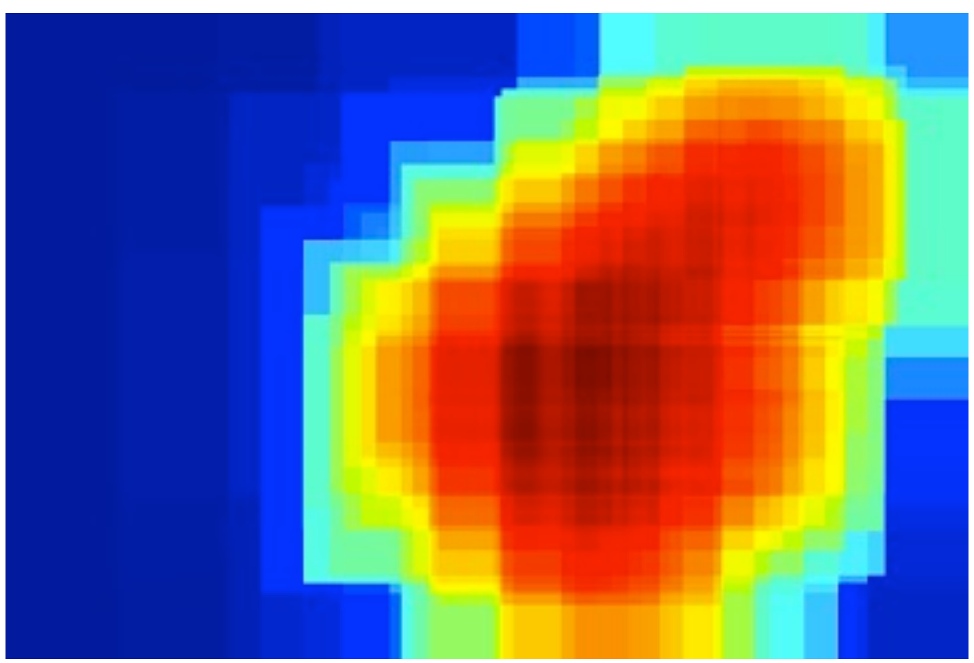
Qualitative results

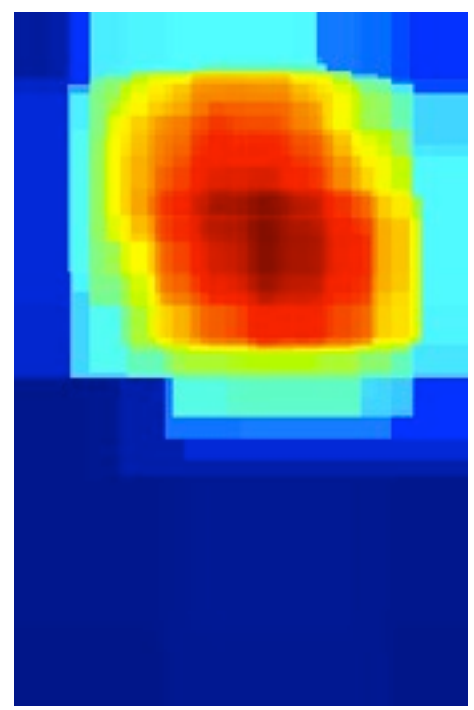


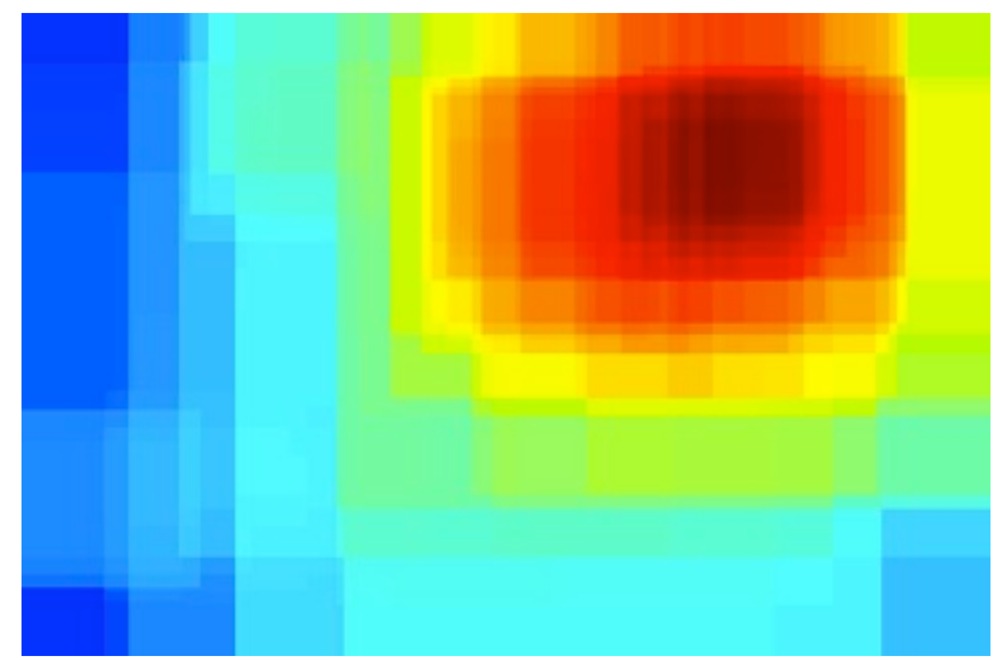
Source : Pascal VOC'12 test set

Visualizations (aeroplane)





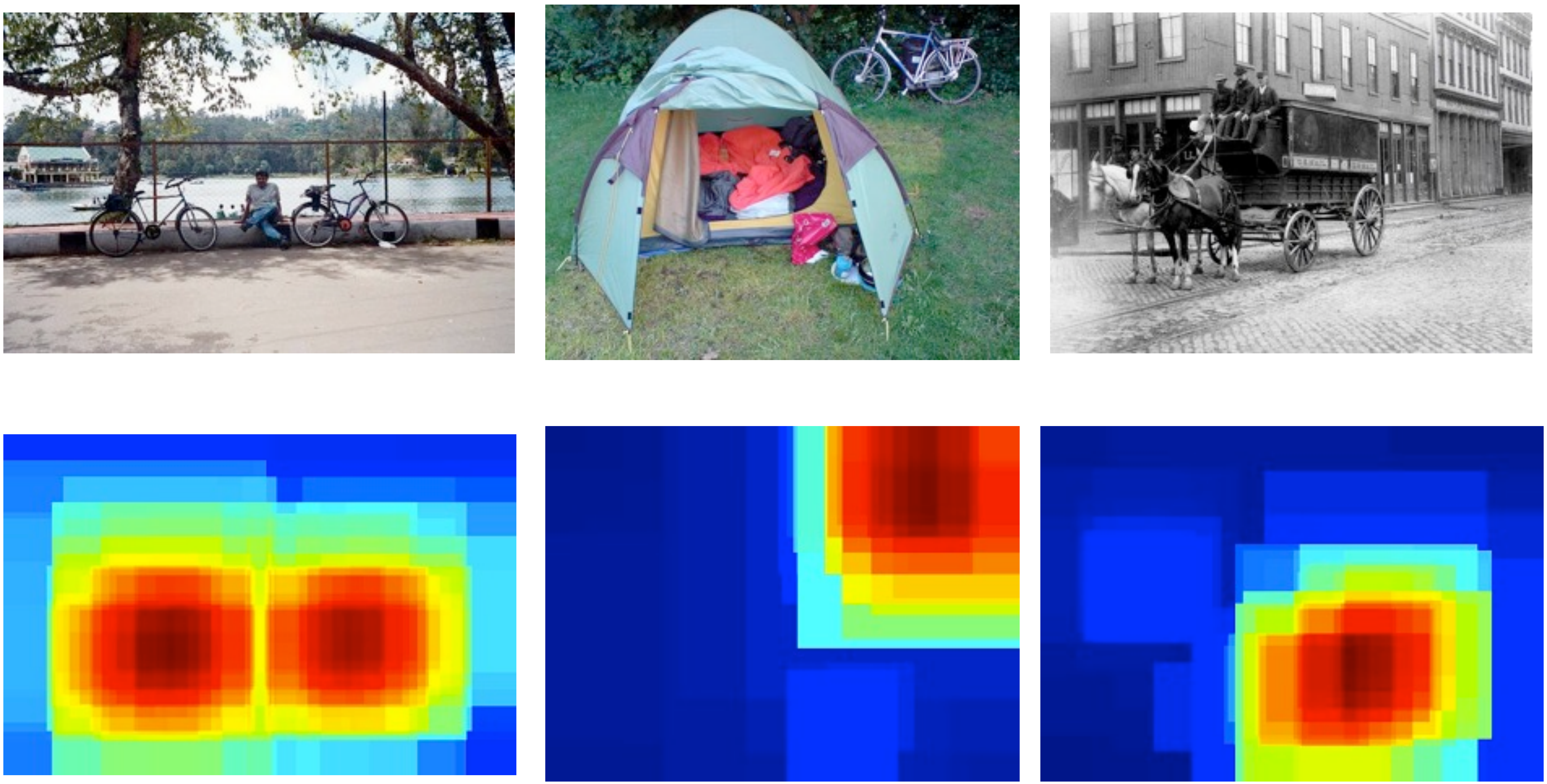




First false positive

Source : Pascal VOC'12 test set

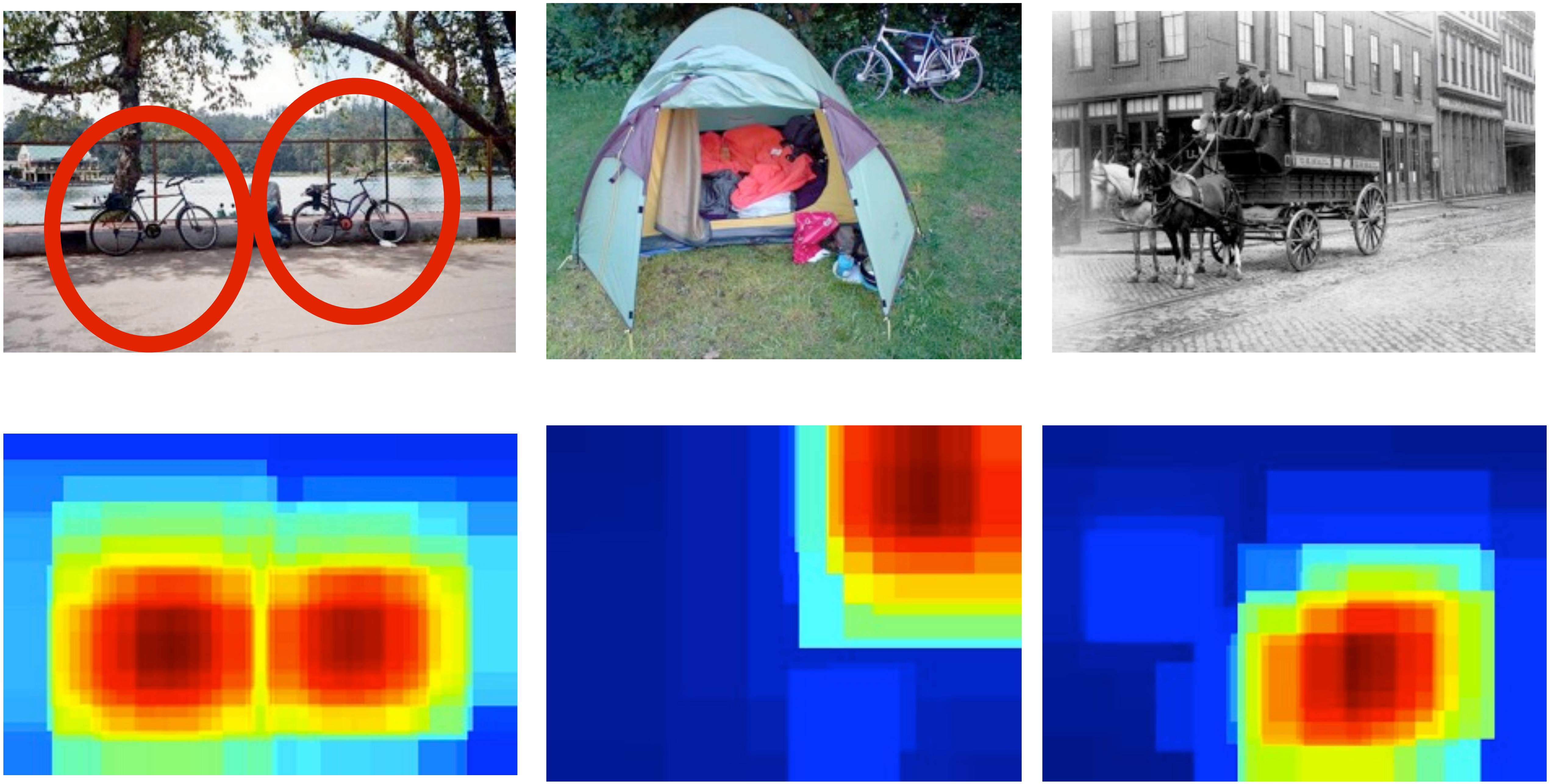
Visualizations (bicycle)



First false positive

Source : Pascal VOC'12 test set

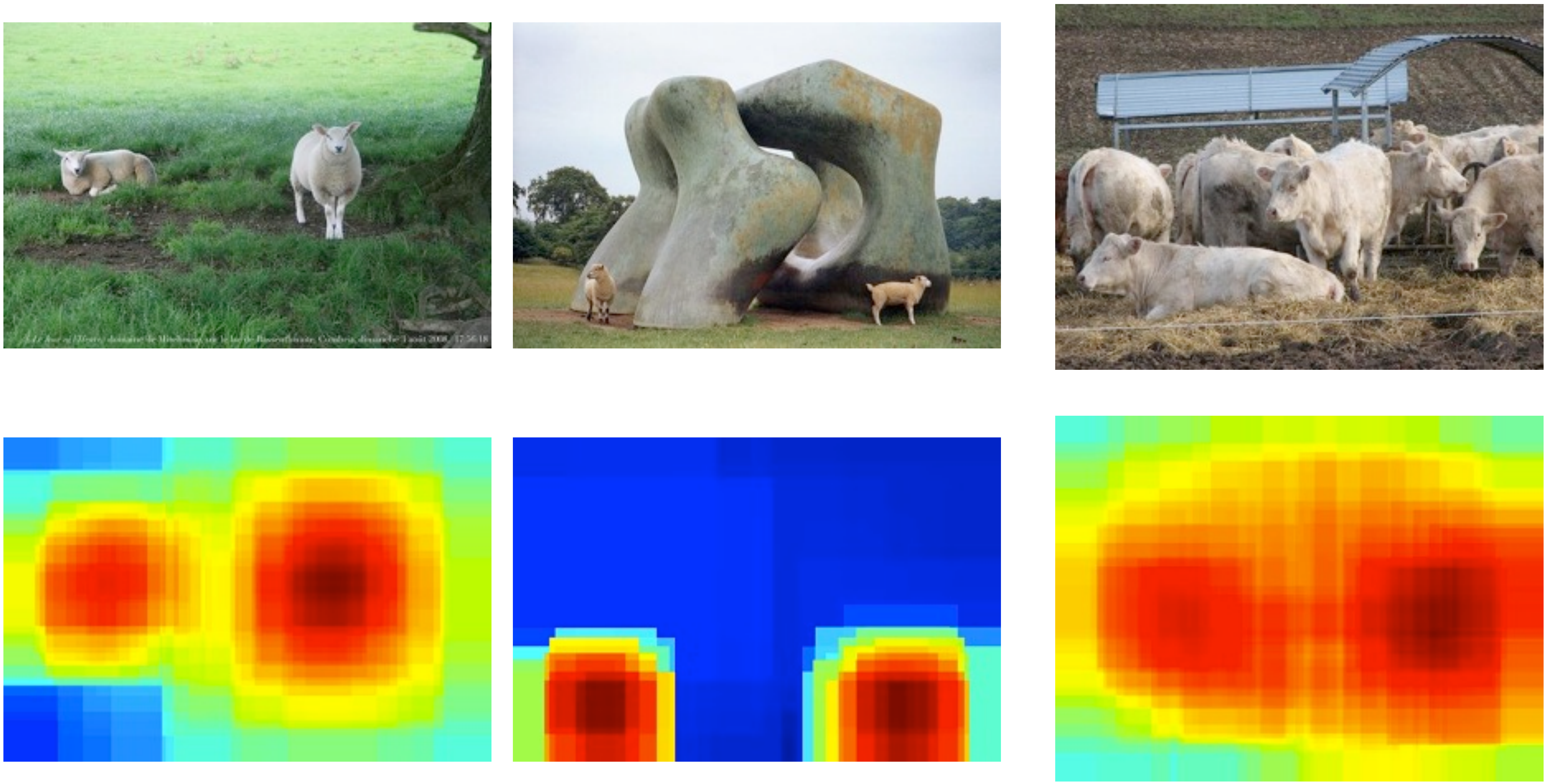
Visualizations (bicycle)



First false positive

Source : Pascal VOC'12 test set

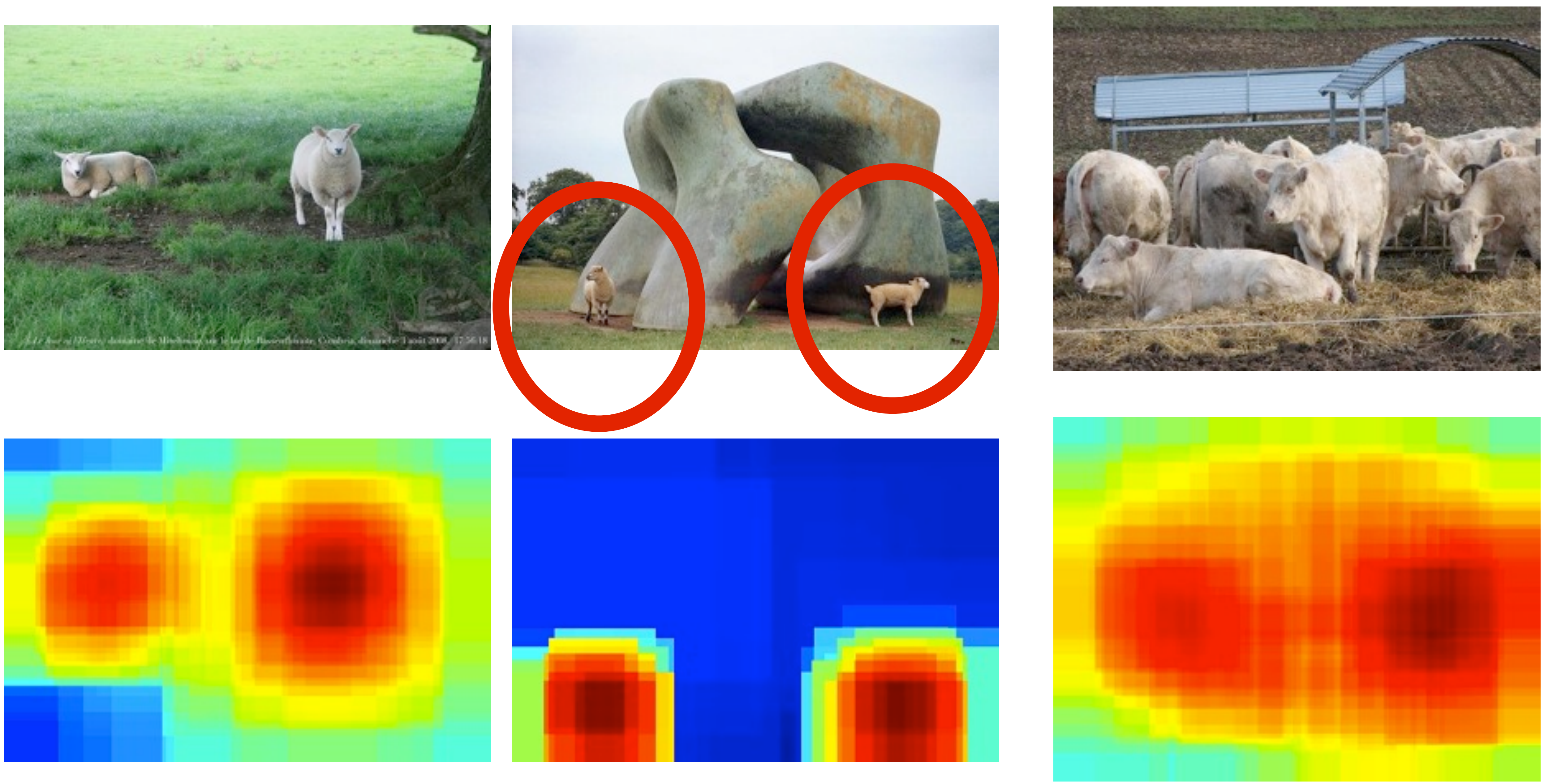
Visualizations (sheep)



First false positive

Source : Pascal VOC'12 test set

Visualizations (sheep)



First false positive

Source : Pascal VOC'12 test set

Quantitative results

Pascal VOC'12 object classification :

	plane	bike	bird	boat	btl	bus	car	cat	chair	cow	table	dog	horse	moto	pers	plant	sheep	sofa	train	tv	**mAP**
NUS-PSL [48]	**97.3**	**84.2**	80.8	**85.3**	**60.8**	**89.9**	**86.8**	89.3	**75.4**	77.8	**75.1**	83.0	87.5	**90.1**	95.0	57.8	79.2	**73.4**	**94.5**	**80.7**	82.2
NO PRETRAIN	85.2	75.0	69.4	66.2	48.8	82.1	79.5	79.8	62.4	61.9	49.8	75.9	71.4	82.7	93.1	59.1	69.7	49.3	80.0	76.7	70.9
PRE-1000C	93.5	78.4	*87.7*	80.9	57.3	85.0	81.6	*89.4*	66.9	73.8	62.0	*89.5*	83.2	87.6	*95.8*	*61.4*	79.0	54.3	88.0	78.3	78.7
PRE-1000R	93.2	77.9	83.8	80.0	55.8	82.7	79.0	84.3	66.2	71.7	59.5	83.4	81.4	84.8	95.2	59.8	74.9	52.9	83.8	75.7	76.3
PRE-1512	94.6	82.9	**88.2**	84.1	60.3	89.0	84.4	**90.7**	72.1	**86.8**	69.0	**92.1**	**93.4**	88.6	**96.1**	**64.3**	**86.6**	62.3	91.1	79.8	**82.8**

State of the art : $\boxed{82.2}$

Quantitative results

Pascal VOC'12 object classification :

	plane	bike	bird	boat	btl	bus	car	cat	chair	cow	table	dog	horse	moto	pers	plant	sheep	sofa	train	tv	**mAP**
NUS-PSL [48]	**97.3**	**84.2**	80.8	**85.3**	**60.8**	**89.9**	**86.8**	89.3	**75.4**	77.8	**75.1**	83.0	87.5	**90.1**	95.0	57.8	79.2	**73.4**	**94.5**	**80.7**	82.2
NO PRETRAIN	85.2	75.0	69.4	66.2	48.8	82.1	79.5	79.8	62.4	61.9	49.8	75.9	71.4	82.7	93.1	59.1	69.7	49.3	80.0	76.7	70.9
PRE-1000C	93.5	78.4	*87.7*	80.9	57.3	85.0	81.6	*89.4*	66.9	73.8	62.0	*89.5*	83.2	87.6	*95.8*	*61.4*	79.0	54.3	88.0	78.3	78.7
PRE-1000R	93.2	77.9	83.8	80.0	55.8	82.7	79.0	84.3	66.2	71.7	59.5	83.4	81.4	84.8	95.2	59.8	74.9	52.9	83.8	75.7	76.3
PRE-1512	94.6	82.9	**88.2**	84.1	60.3	89.0	84.4	**90.7**	72.1	**86.8**	69.0	**92.1**	**93.4**	88.6	**96.1**	**64.3**	**86.6**	62.3	91.1	79.8	**82.8**

State of the art : 82.2

No pre-training baseline : 70.9

Quantitative results

Pascal VOC'12 object classification :

	plane	bike	bird	boat	btl	bus	car	cat	chair	cow	table	dog	horse	moto	pers	plant	sheep	sofa	train	tv	mAP
NUS-PSL [48]	**97.3**	**84.2**	80.8	**85.3**	**60.8**	**89.9**	**86.8**	89.3	**75.4**	77.8	**75.1**	83.0	87.5	**90.1**	95.0	57.8	79.2	**73.4**	**94.5**	**80.7**	$\boxed{82.2}$
NO PRETRAIN	85.2	75.0	69.4	66.2	48.8	82.1	79.5	79.8	62.4	61.9	49.8	75.9	71.4	82.7	93.1	59.1	69.7	49.3	80.0	76.7	$\boxed{70.9}$
PRE-1000C	93.5	78.4	*87.7*	80.9	57.3	85.0	81.6	*89.4*	66.9	73.8	62.0	*89.5*	83.2	87.6	*95.8*	*61.4*	79.0	54.3	88.0	78.3	$\boxed{78.7}$
PRE-1000R	93.2	77.9	83.8	80.0	55.8	82.7	79.0	84.3	66.2	71.7	59.5	83.4	81.4	84.8	95.2	59.8	74.9	52.9	83.8	75.7	76.3
PRE-1512	94.6	82.9	**88.2**	84.1	60.3	89.0	84.4	**90.7**	72.1	**86.8**	69.0	**92.1**	**93.4**	88.6	**96.1**	**64.3**	**86.6**	62.3	91.1	79.8	**82.8**

State of the art : $\boxed{82.2}$

No pre-training baseline : $\boxed{70.9}$

1000 ILSVRC classes : $\boxed{78.7}$

Quantitative results

Pascal VOC'12 object classification :

	plane	bike	bird	boat	btl	bus	car	cat	chair	cow	table	dog	horse	moto	pers	plant	sheep	sofa	train	tv	**mAP**
NUS-PSL [48]	**97.3**	**84.2**	80.8	**85.3**	**60.8**	**89.9**	**86.8**	89.3	**75.4**	77.8	**75.1**	83.0	87.5	**90.1**	95.0	57.8	79.2	**73.4**	**94.5**	**80.7**	82.2
NO PRETRAIN	85.2	75.0	69.4	66.2	48.8	82.1	79.5	79.8	62.4	61.9	49.8	75.9	71.4	82.7	93.1	59.1	69.7	49.3	80.0	76.7	70.9
PRE-1000C	93.5	78.4	*87.7*	80.9	57.3	85.0	81.6	*89.4*	66.9	73.8	62.0	*89.5*	83.2	87.6	*95.8*	*61.4*	79.0	54.3	88.0	78.3	78.7
PRE-1000R	93.2	77.9	83.8	80.0	55.8	82.7	79.0	84.3	66.2	71.7	59.5	83.4	81.4	84.8	95.2	59.8	74.9	52.9	83.8	75.7	76.3
PRE-1512	94.6	82.9	**88.2**	84.1	60.3	89.0	84.4	**90.7**	72.1	**86.8**	69.0	**92.1**	**93.4**	88.6	**96.1**	**64.3**	**86.6**	62.3	91.1	79.8	**82.8**

State of the art : $\boxed{82.2}$

No pre-training baseline : $\boxed{70.9}$

1000 ILSVRC classes : $\boxed{78.7}$

1512 classes (our best) : $\boxed{\mathbf{82.8}}$

Quantitative results

Pascal VOC'12 object classification :

	plane	bike	bird	boat	btl	bus	car	cat	chair	cow	table	dog	horse	moto	pers	plant	sheep	sofa	train	tv	**mAP**
NUS-PSL [48]	**97.3**	**84.2**	80.8	**85.3**	**60.8**	**89.9**	**86.8**	89.3	**75.4**	77.8	**75.1**	83.0	87.5	**90.1**	95.0	57.8	79.2	**73.4**	**94.5**	**80.7**	82.2
NO PRETRAIN	85.2	75.0	69.4	66.2	48.8	82.1	79.5	79.8	62.4	61.9	49.8	75.9	71.4	82.7	93.1	59.1	69.7	49.3	80.0	76.7	70.9
PRE-1000C	93.5	78.4	*87.7*	80.9	57.3	85.0	81.6	*89.4*	66.9	73.8	62.0	*89.5*	83.2	87.6	*95.8*	*61.4*	79.0	54.3	88.0	78.3	78.7
PRE-1000R	93.2	77.9	83.8	80.0	55.8	82.7	79.0	84.3	66.2	71.7	59.5	83.4	81.4	84.8	95.2	59.8	74.9	52.9	83.8	75.7	76.3
PRE-1512	94.6	82.9	**88.2**	84.1	60.3	89.0	84.4	**90.7**	72.1	**86.8**	69.0	**92.1**	**93.4**	88.6	**96.1**	**64.3**	**86.6**	62.3	91.1	79.8	**82.8**

State of the art : 82.2

No pre-training baseline : 70.9

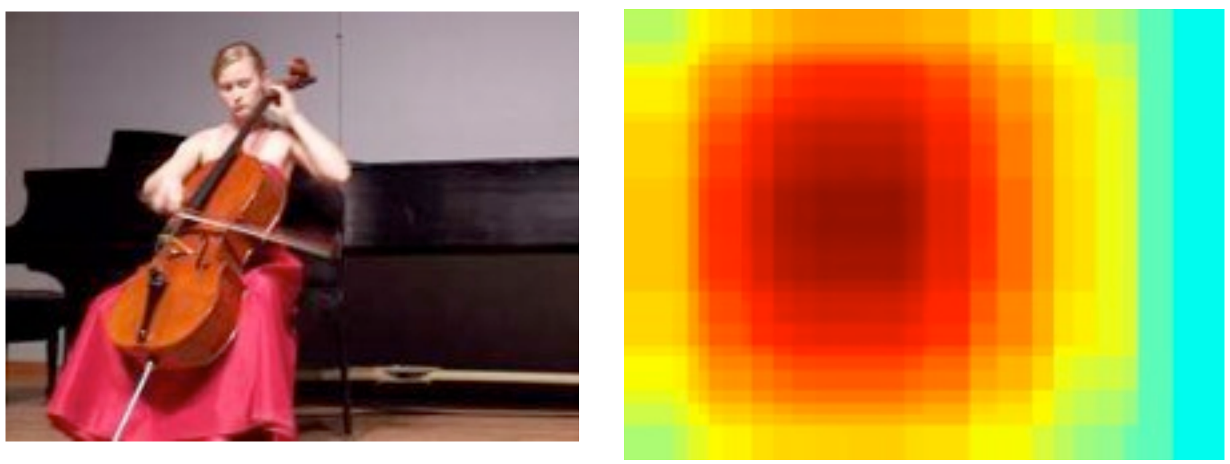
1000 ILSVRC classes : 78.7

Random 1000 classes : 76.3

1512 classes (our best) : **82.8**

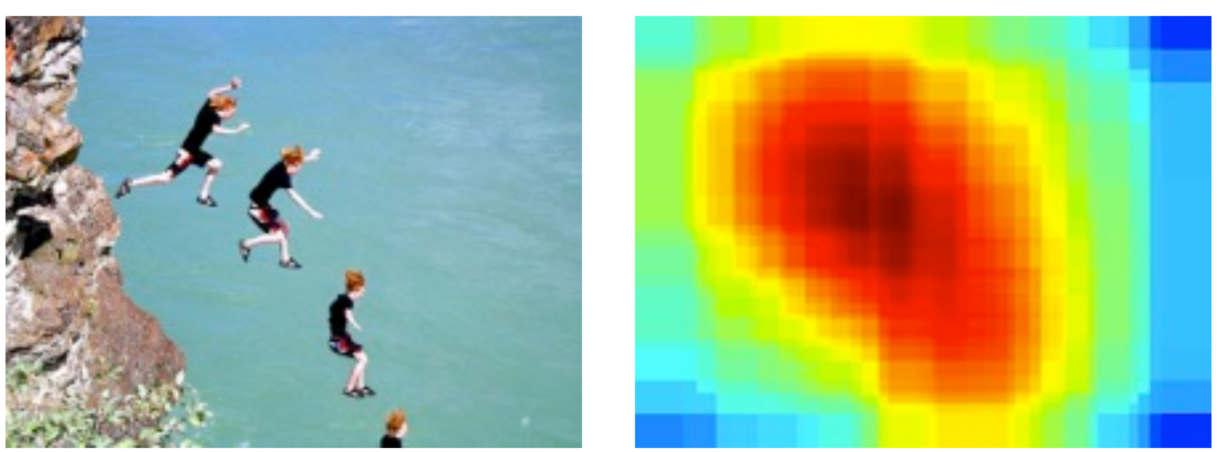
Different task : action classification (still images)

playing instrument

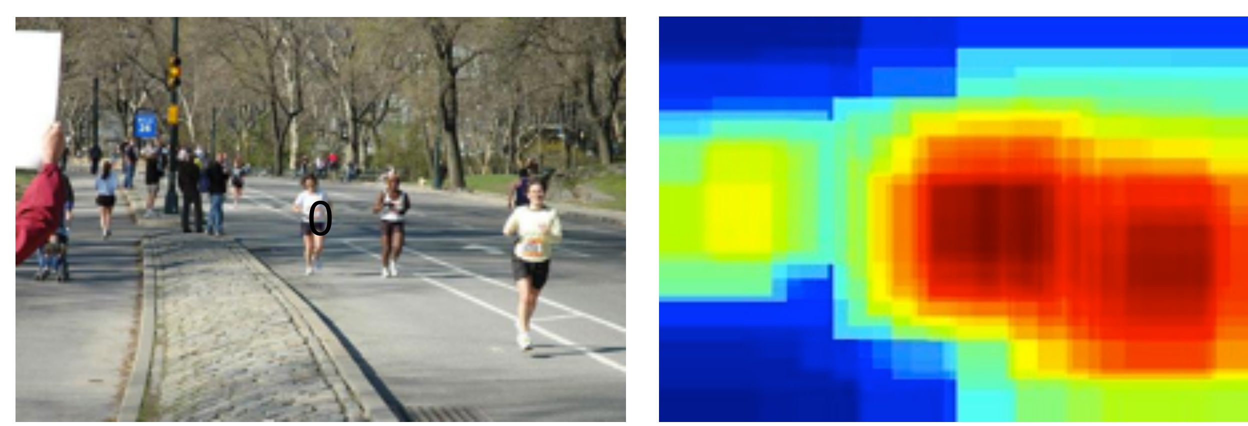


playing instrument

jumping



running

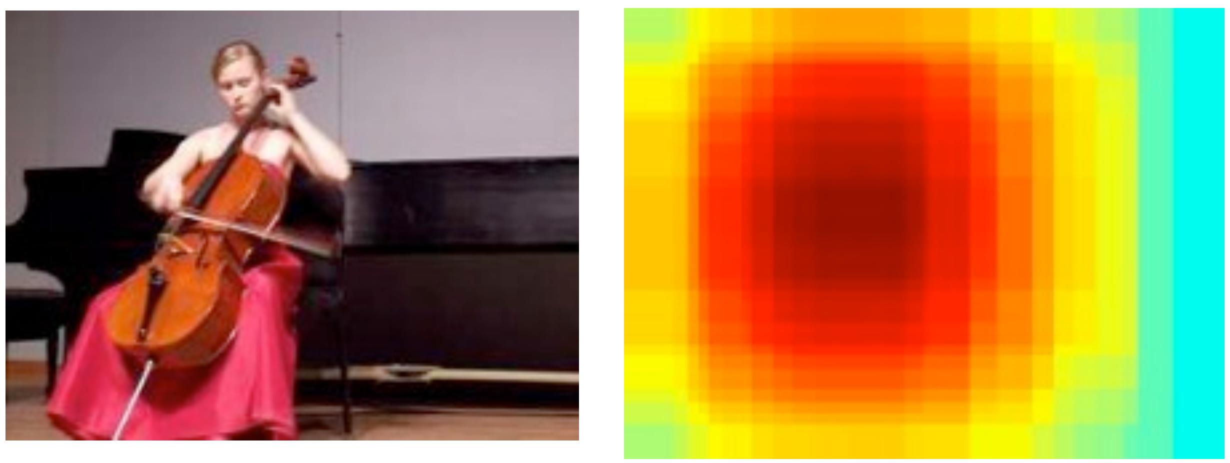


Source : Pascal VOC'12 Action classification test set

State-of-the-art 70.2% mAP result

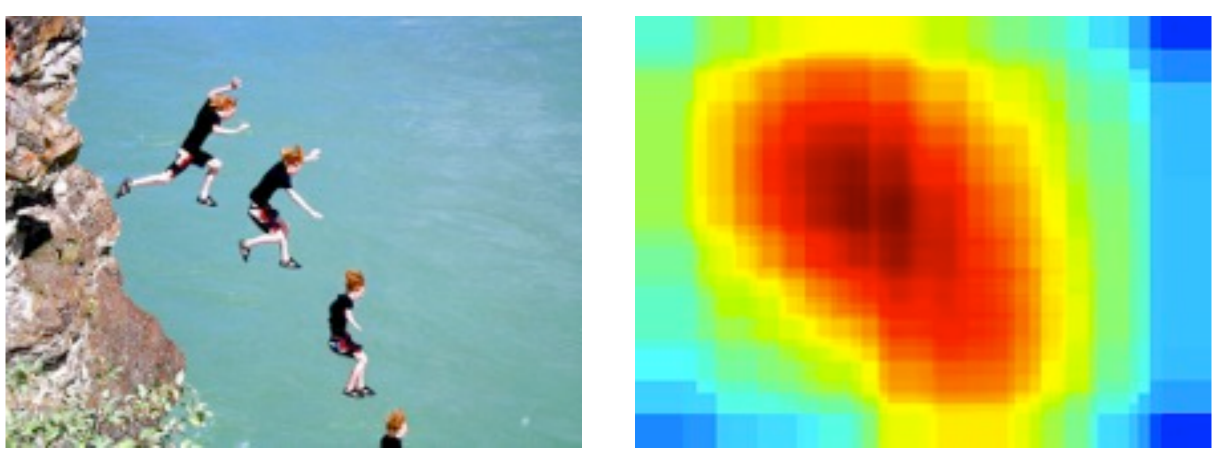
Different task : action classification (still images)

playing instrument

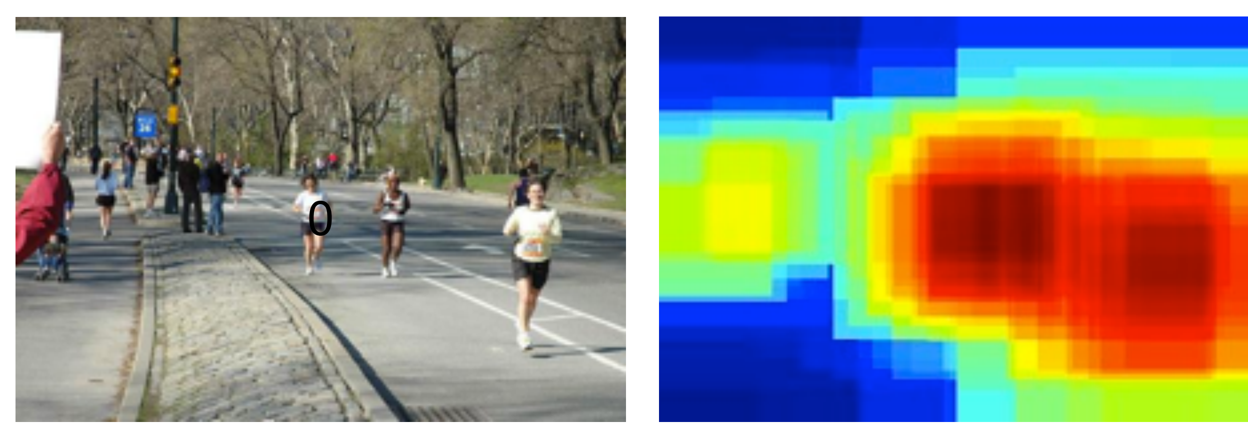


playing instrument

jumping



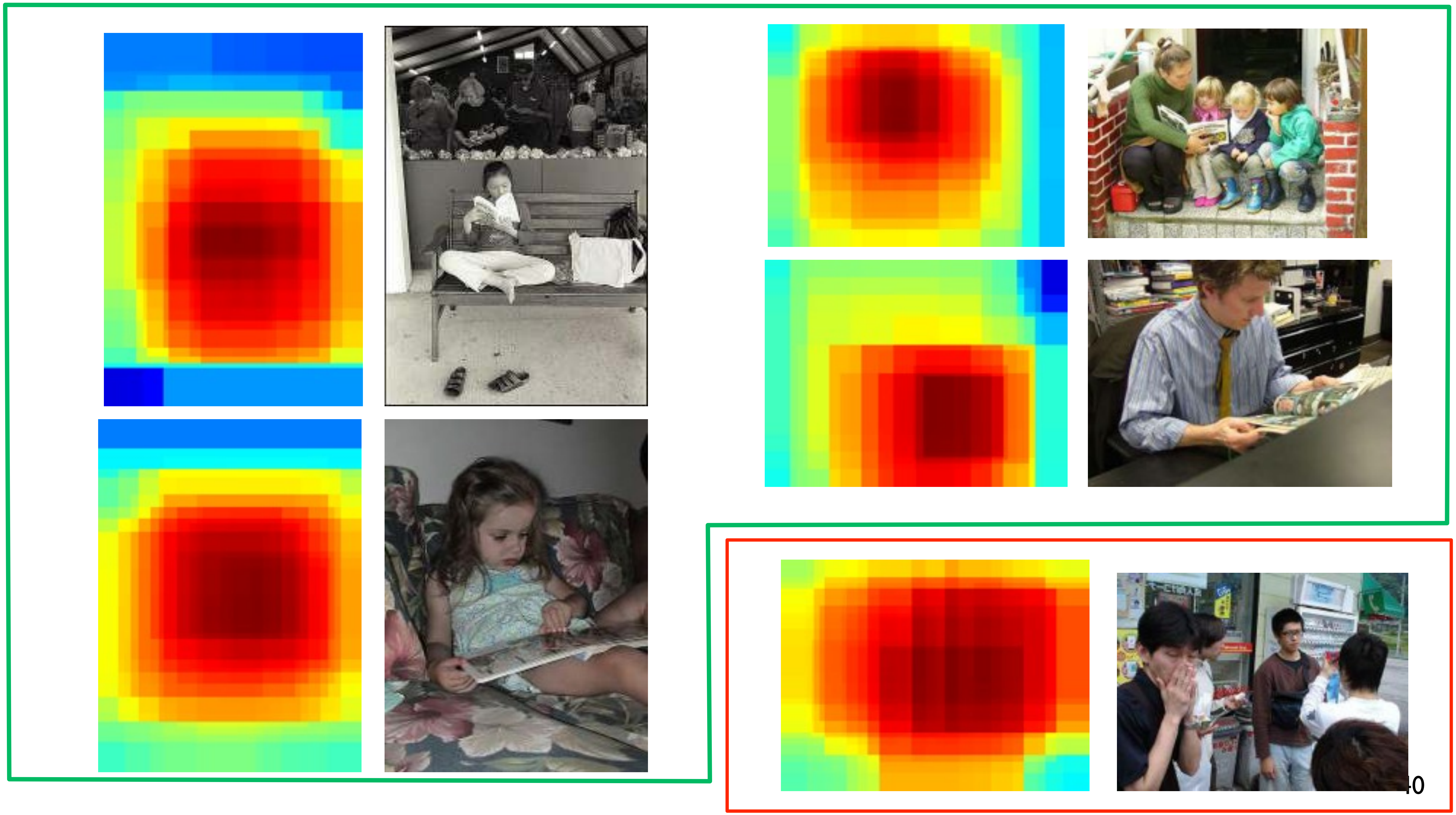
running



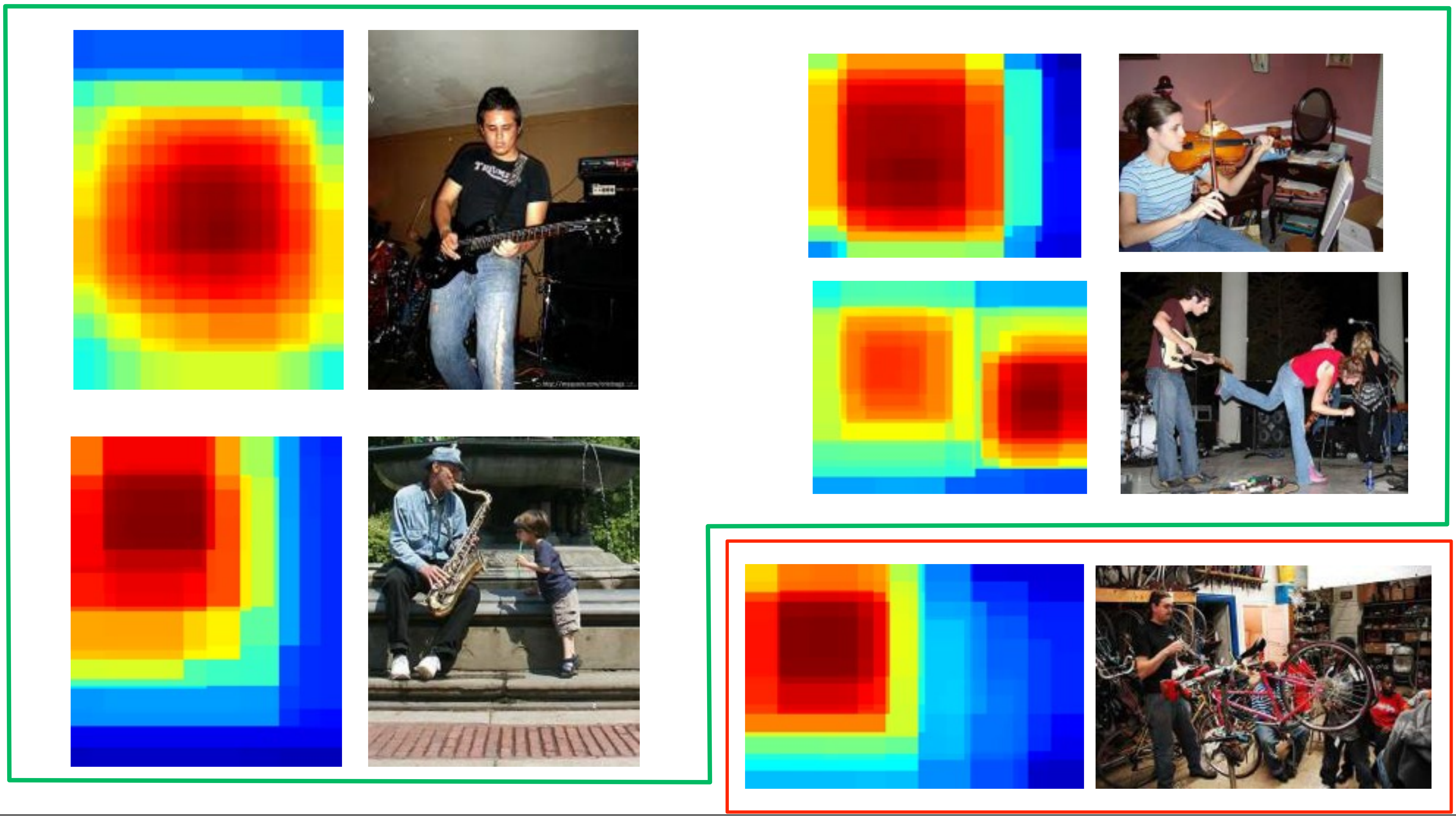
Source : Pascal VOC'12 Action classification test set

State-of-the-art 70.2% mAP result

Qualitative results (reading)



Qualitative results (playing instrument)

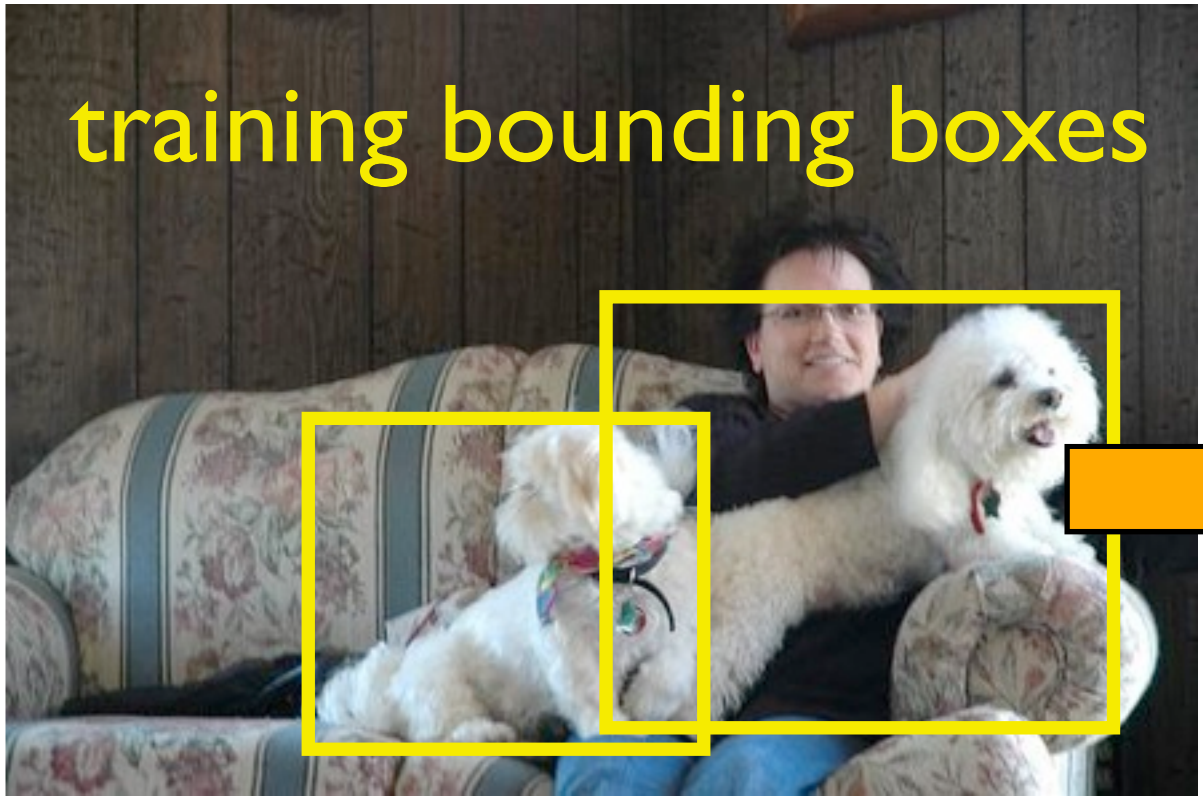


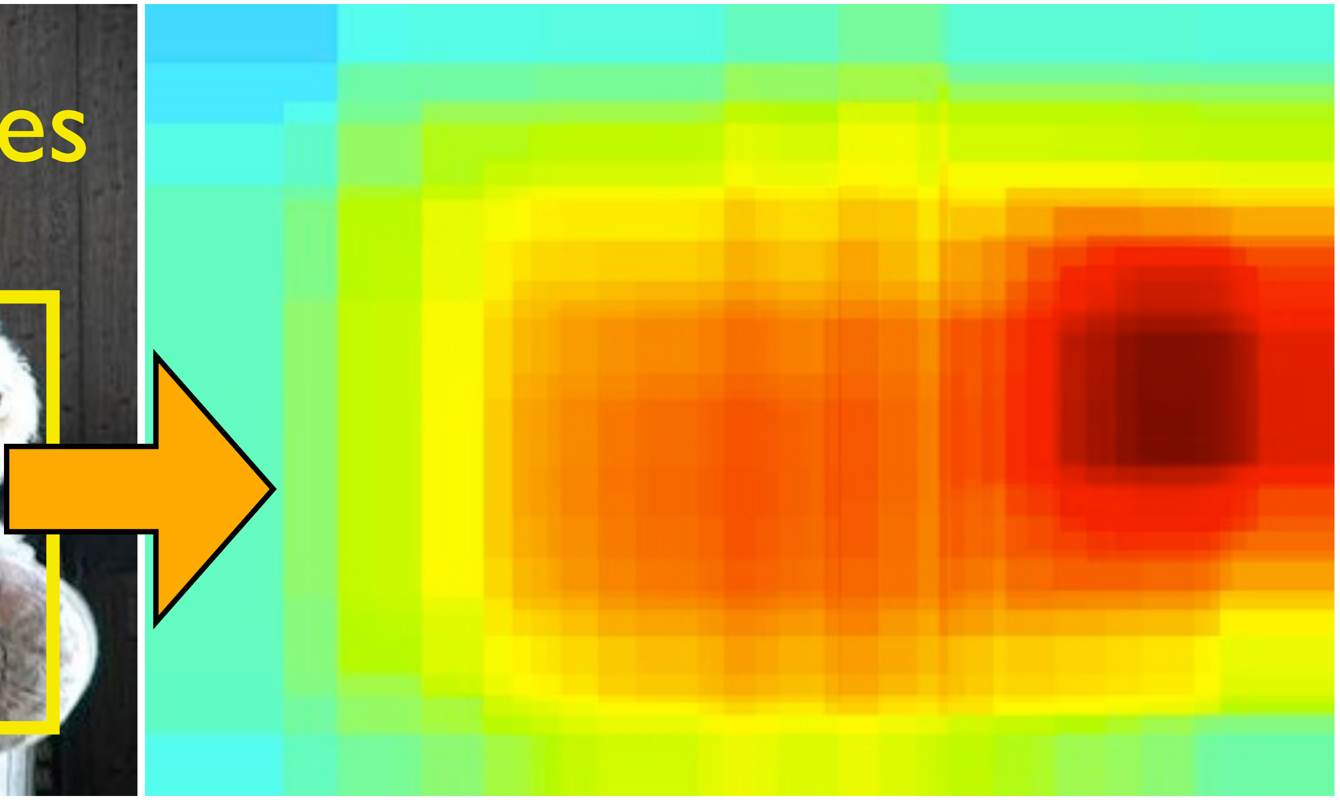
Qualitative results (phoning)

Take-home messages

- **Transfer learning with CNNs avoids overfitting**
  - See also : [Girshick et al.'14], [Sermanet et al.'13], [Donahue et al. '13], [Zeiler & Fergus '13], [Razavian et al. '14], [Chatfield et al. '14]
- **We study the effect of pre-training data :**
  - More pre-training data => better
  - Related pre-training data => even better
- **Transfer to action classification.**
- **http://www.di.ens.fr/willow/research/cnn/**
  - Implementation (Torch7 modules) available soon
  - Includes efficient and flexible GPU training code

This work





«dog» heatmap

- Bounding box annotation is expensive.
  Can we avoid it?
- YES WE CAN !

Follow-up work

«dog» heatmap

- Weakly supervised, no bounding boxes required
- 82.8 => **86.3%** mean AP on VOC classification
- Appearing on Arxiv soon (check our webpage)
  - http://www.di.ens.fr/willow/research/weakcnn/

Willow project-team

Weakly supervised object recognition with convolutional neural networks

Maxime Oquab,
Léon Bottou, Ivan Laptev, Josef Sivic

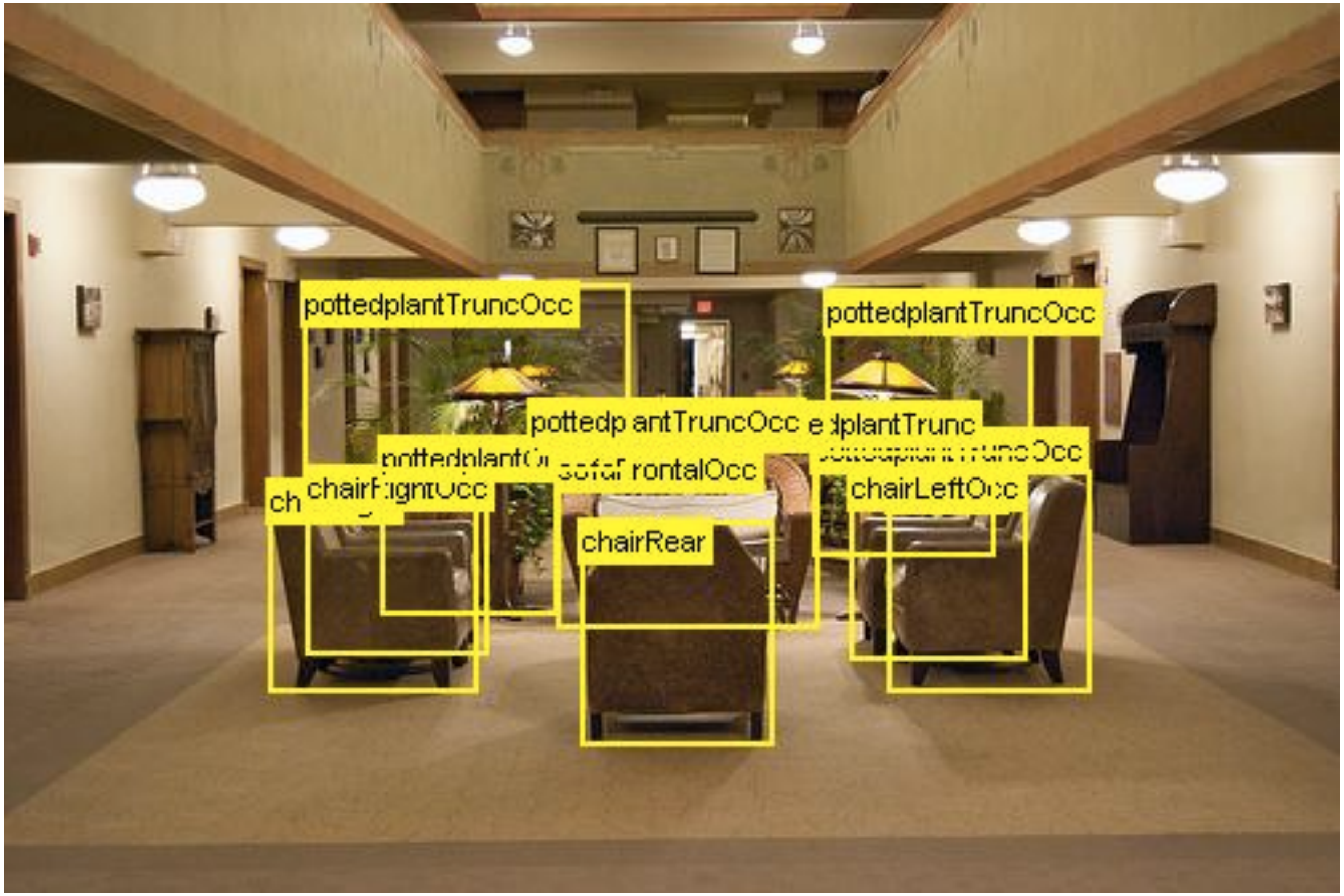
(All following slides stolen from Josef Sivic)

Are bounding boxes needed for training CNNs?

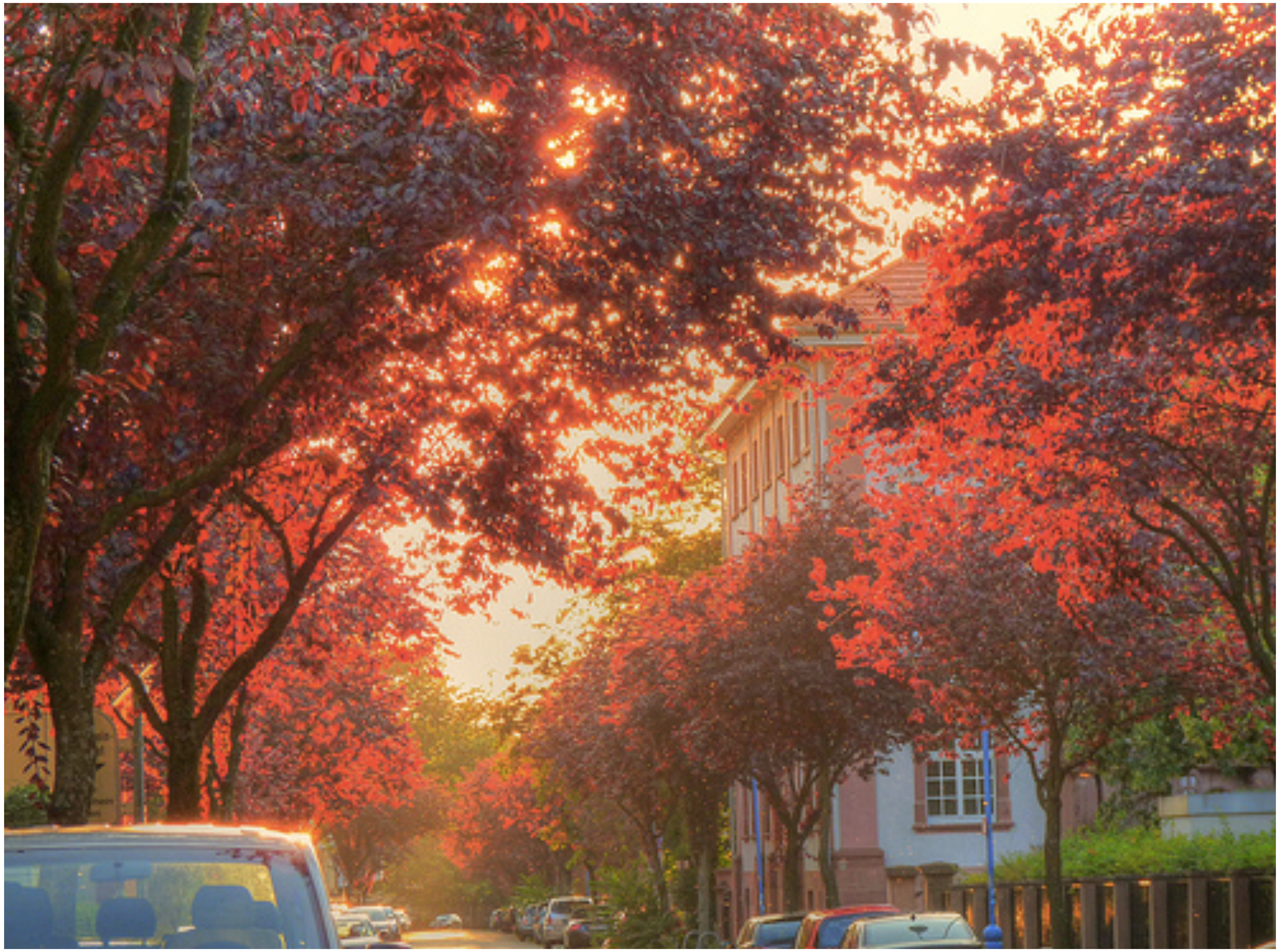
Image-level labels: Bicycle, Person

[Oquab, Bottou, Laptev, Sivic, In submission, 2014]

Motivation: labeling bounding boxes is tedious



Motivation: image-level labels are plentiful

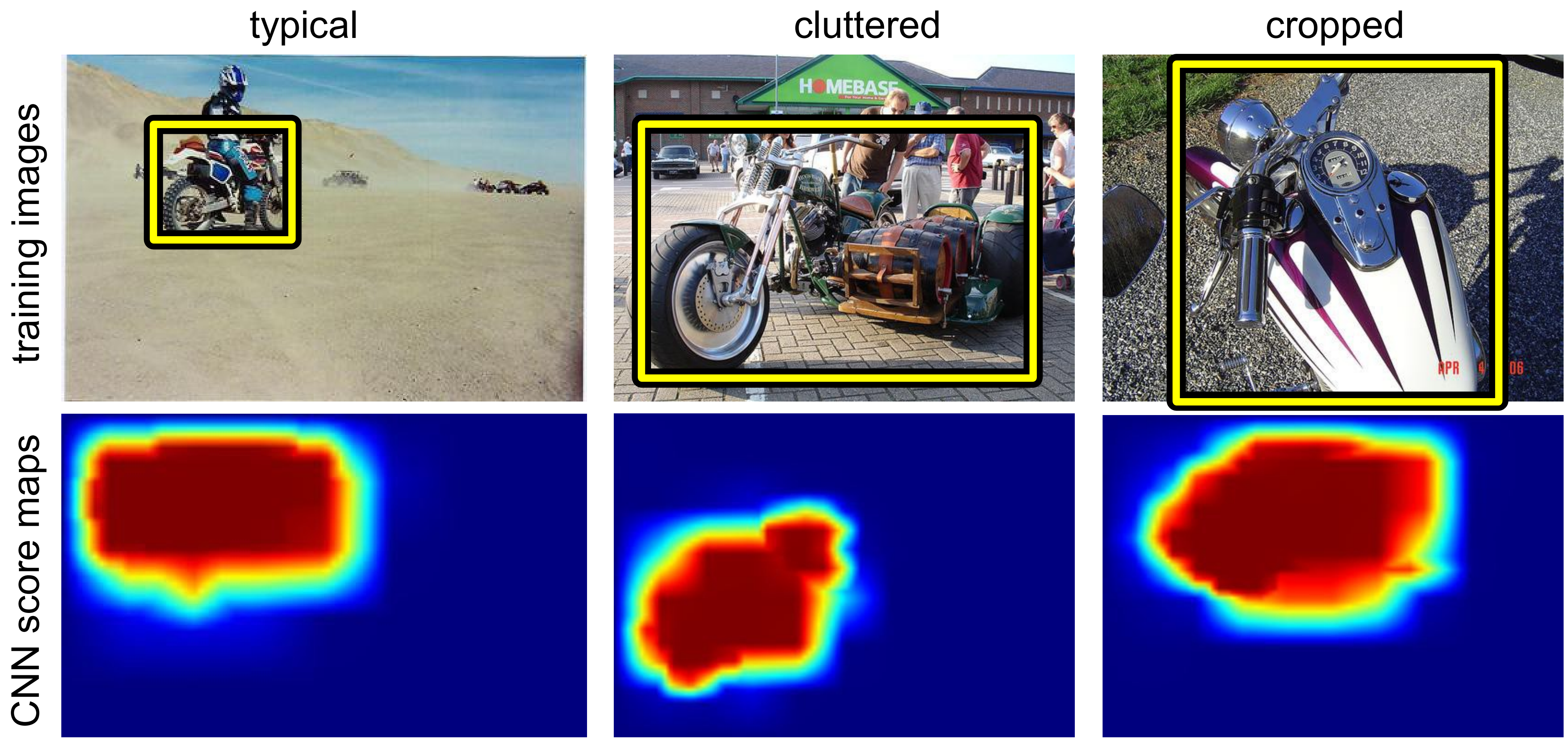


“Beautiful red leaves in a back street of Freiburg”

[Kuznetsova et al., ACL 2013]
http://www.cs.stonybrook.edu/~pkuznetsova/imgcaption/captions1K.html

Let the algorithm localize the object in the image

Example training images with bounding boxes

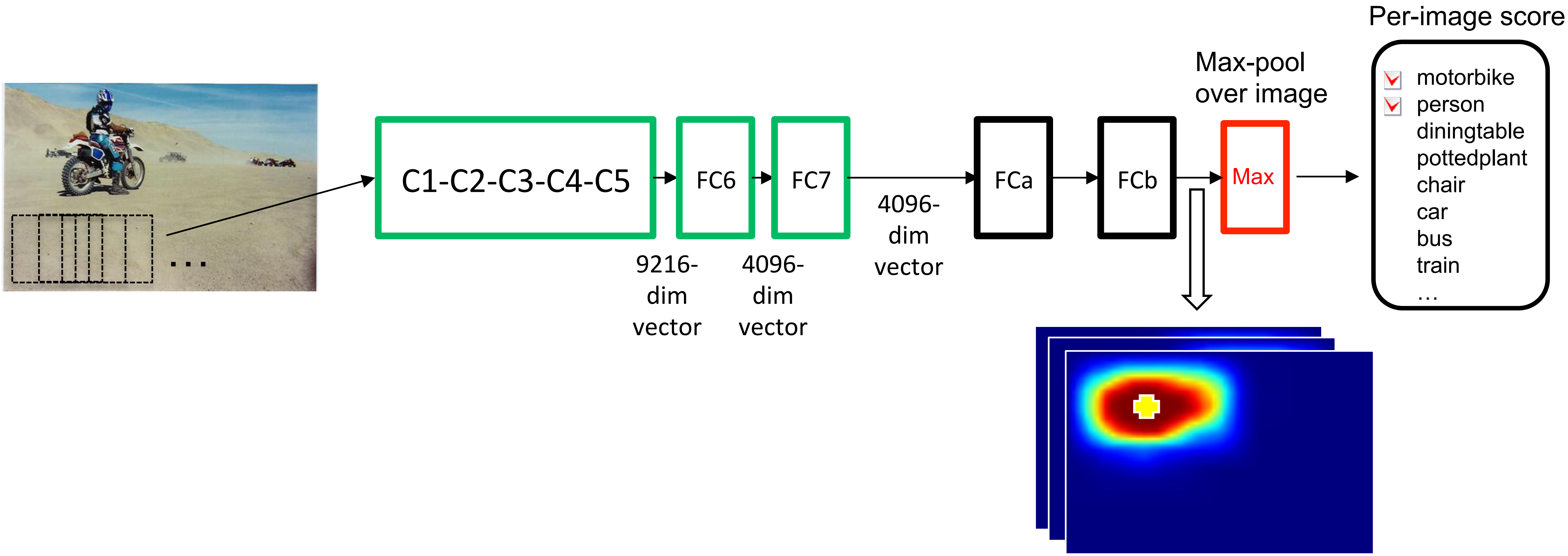


The locations of objects learnt by the CNN

NB: Related to multiple instance learning, e.g. [Viola et al.'05] and weakly supervised object localization, e.g. [Pandy and Lazebnik'11], [Prest et al.'12], …

[Oquab, Bottou, Laptev, Sivic, In submission, 2014]

Approach: search over object's location

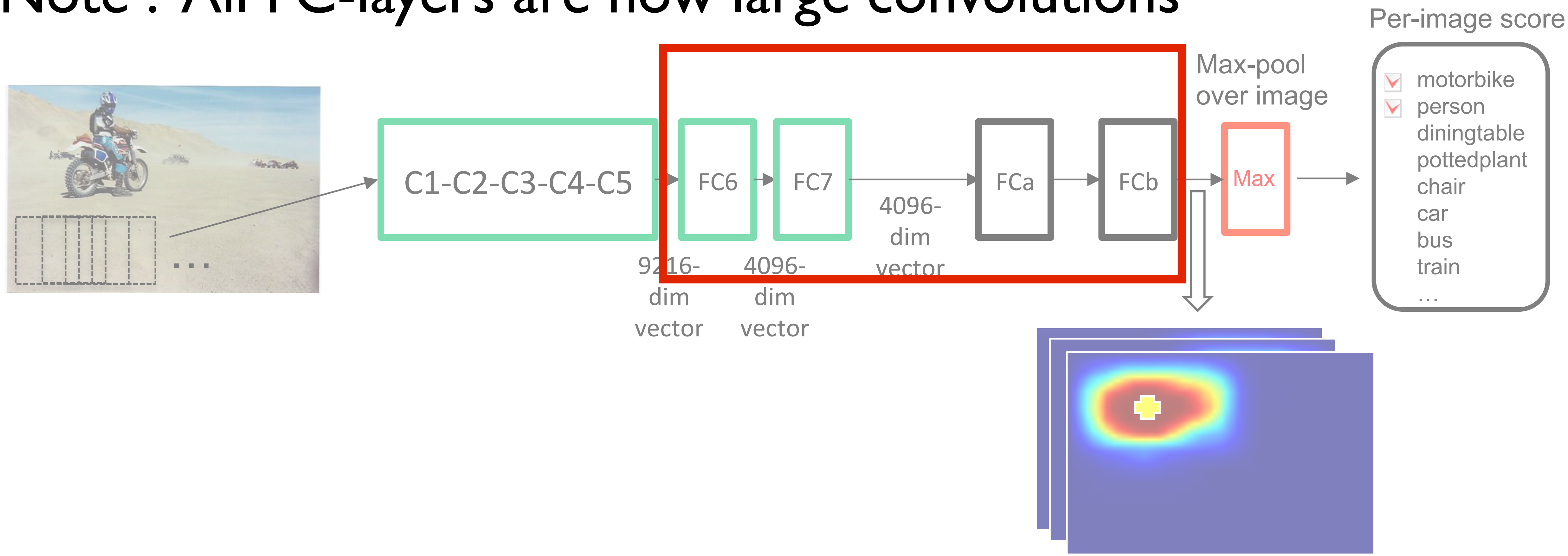


1. Efficient window sliding to find object location hypothesis
2. Image-level aggregation (max-pool)
3. Multi-label loss function (allow multiple objects in image)

See also [Sermanet et al. '14] and [Chaftield et al.'14]

Approach: search over object's location

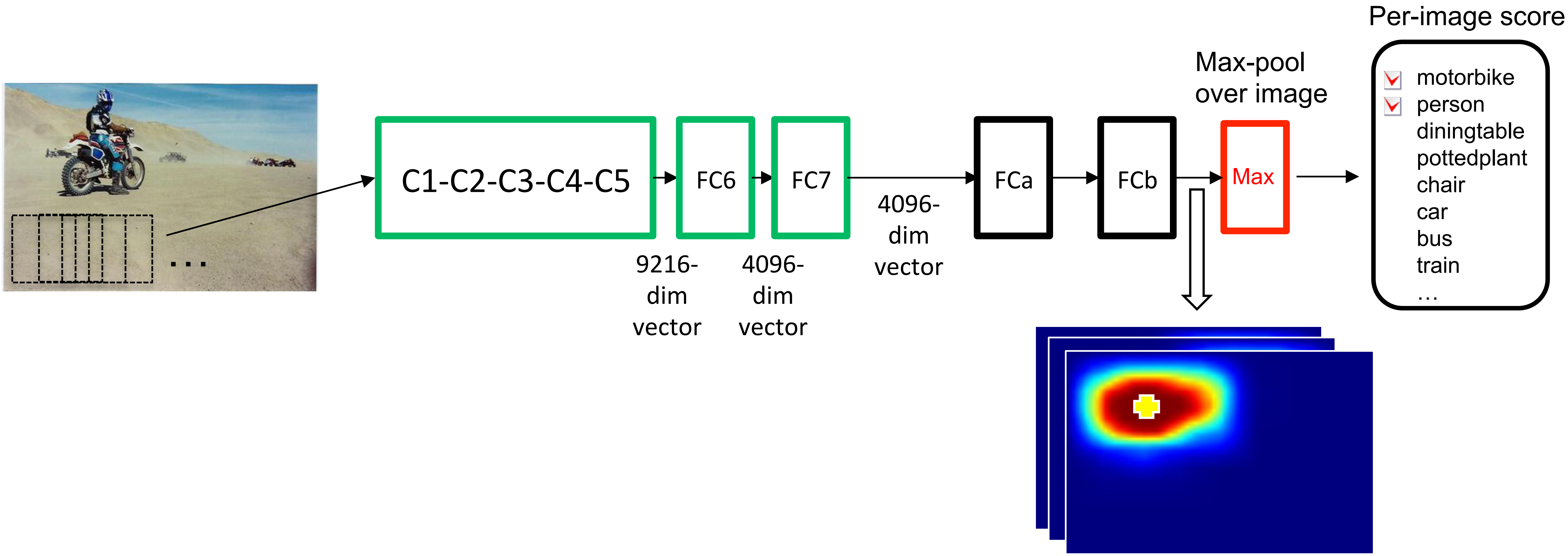
Note : All FC-layers are now large convolutions



1. Efficient window sliding to find object location hypothesis
2. Image-level aggregation (max-pool)
3. Multi-label loss function (allow multiple objects in image)

See also [Sermanet et al. '14] and [Chaftield et al.'14]

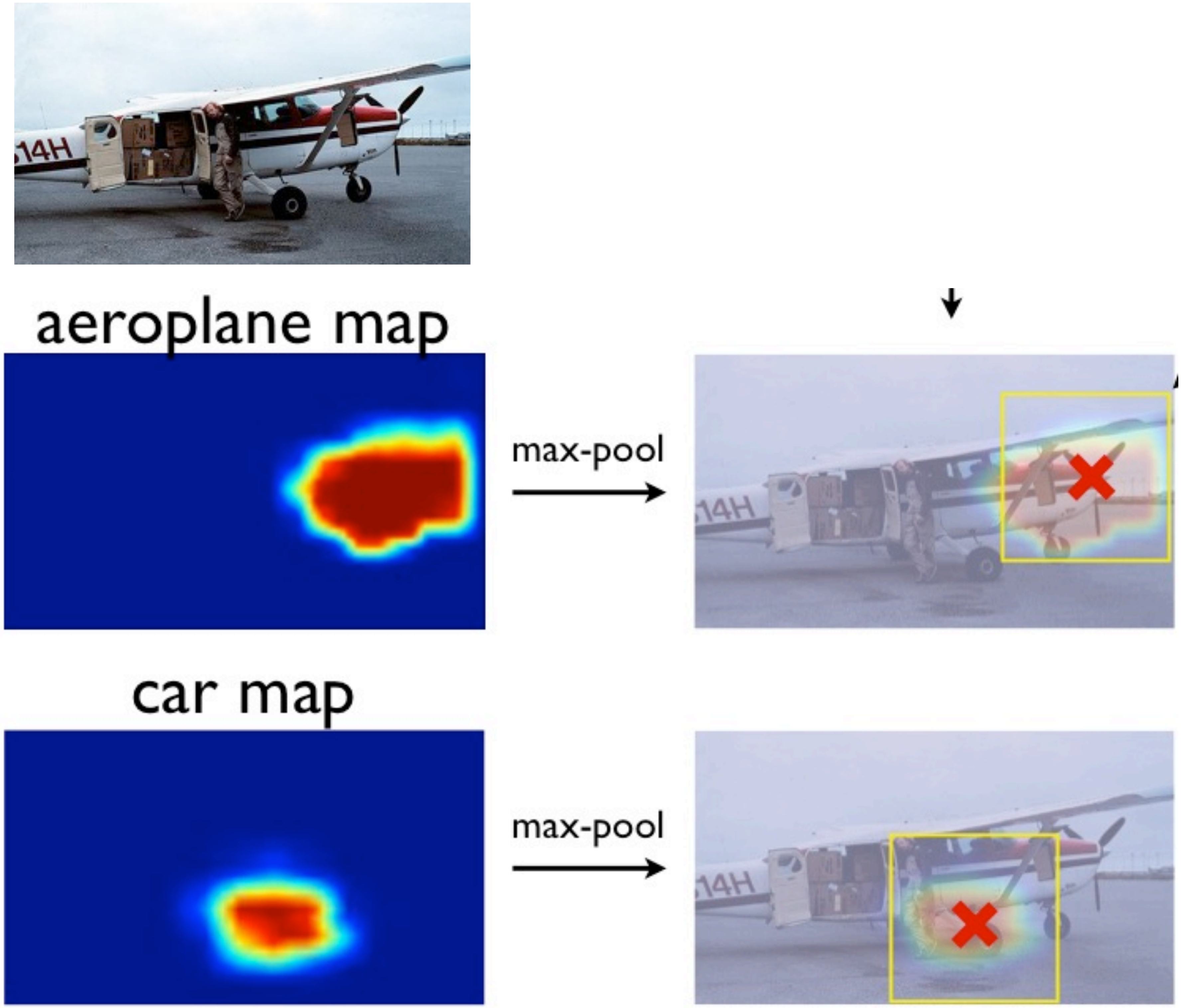
Approach: search over object's location



1. Efficient window sliding to find object location hypothesis
2. Image-level aggregation (max-pool)
3. Multi-label loss function (allow multiple objects in image)

See also [Sermanet et al. '14] and [Chaftield et al.'14]

Search for objects using max-pooling



Correct label:
increase score
for this class

Incorrect label:
decrease score
for this class

Search for objects using max-pooling

learn from :

at training time

<=>

learn from :

Most discriminative part

Hardest negative

What is the effect of errors?

Multi-scale training and testing

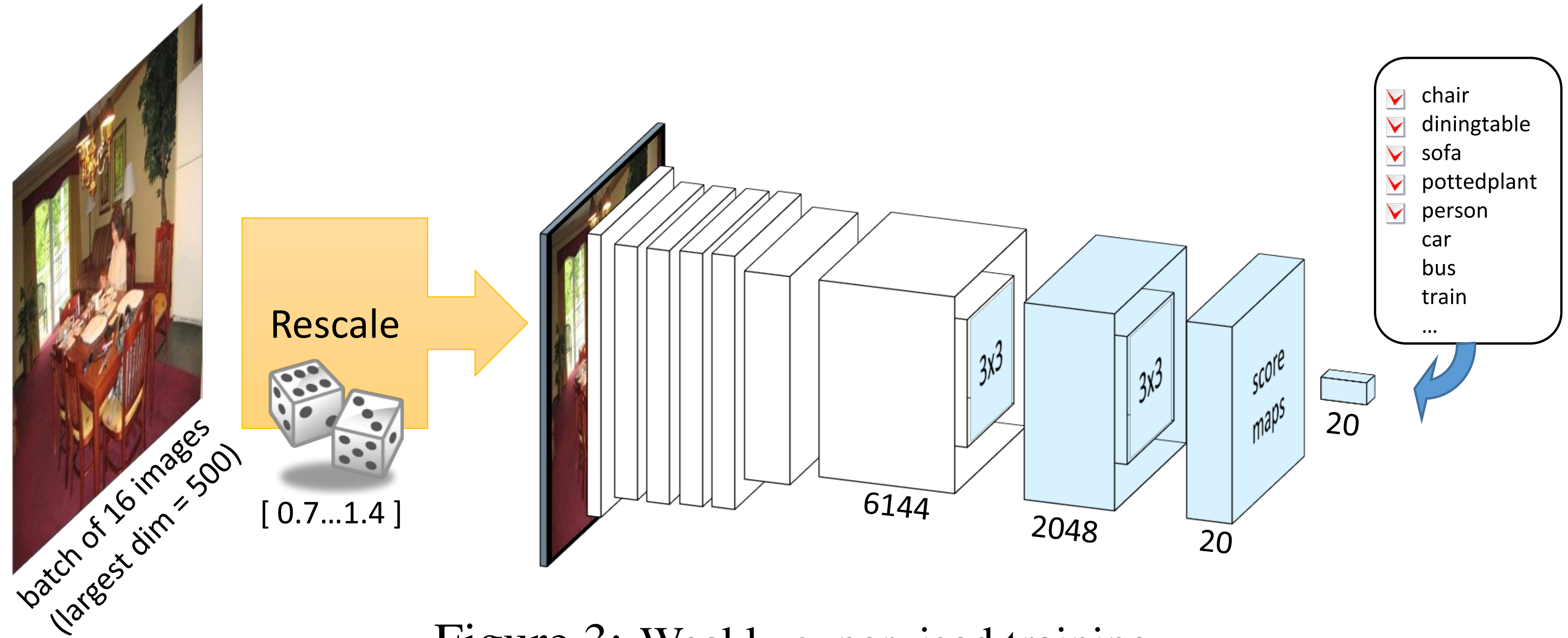


Figure 3: Weakly supervised training

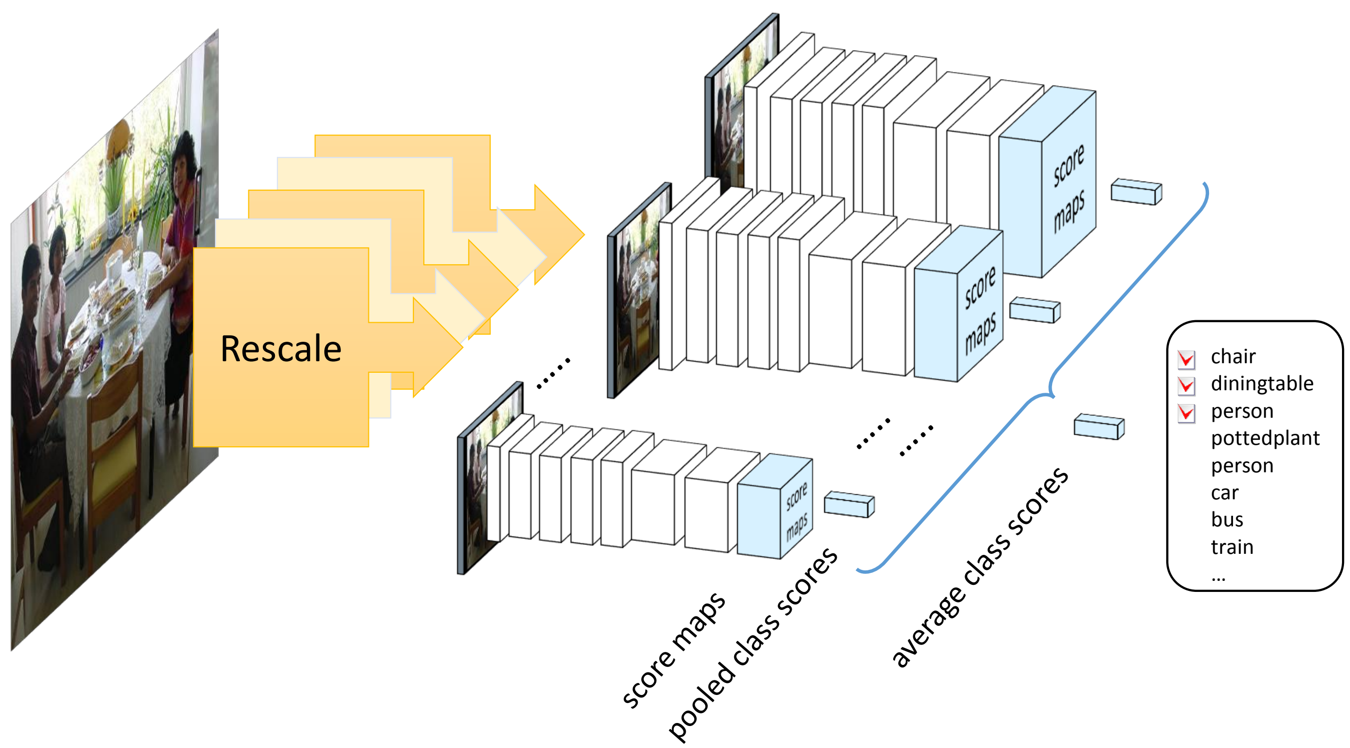


Figure 4: Multiscale object recognition

Evolution of maps during training

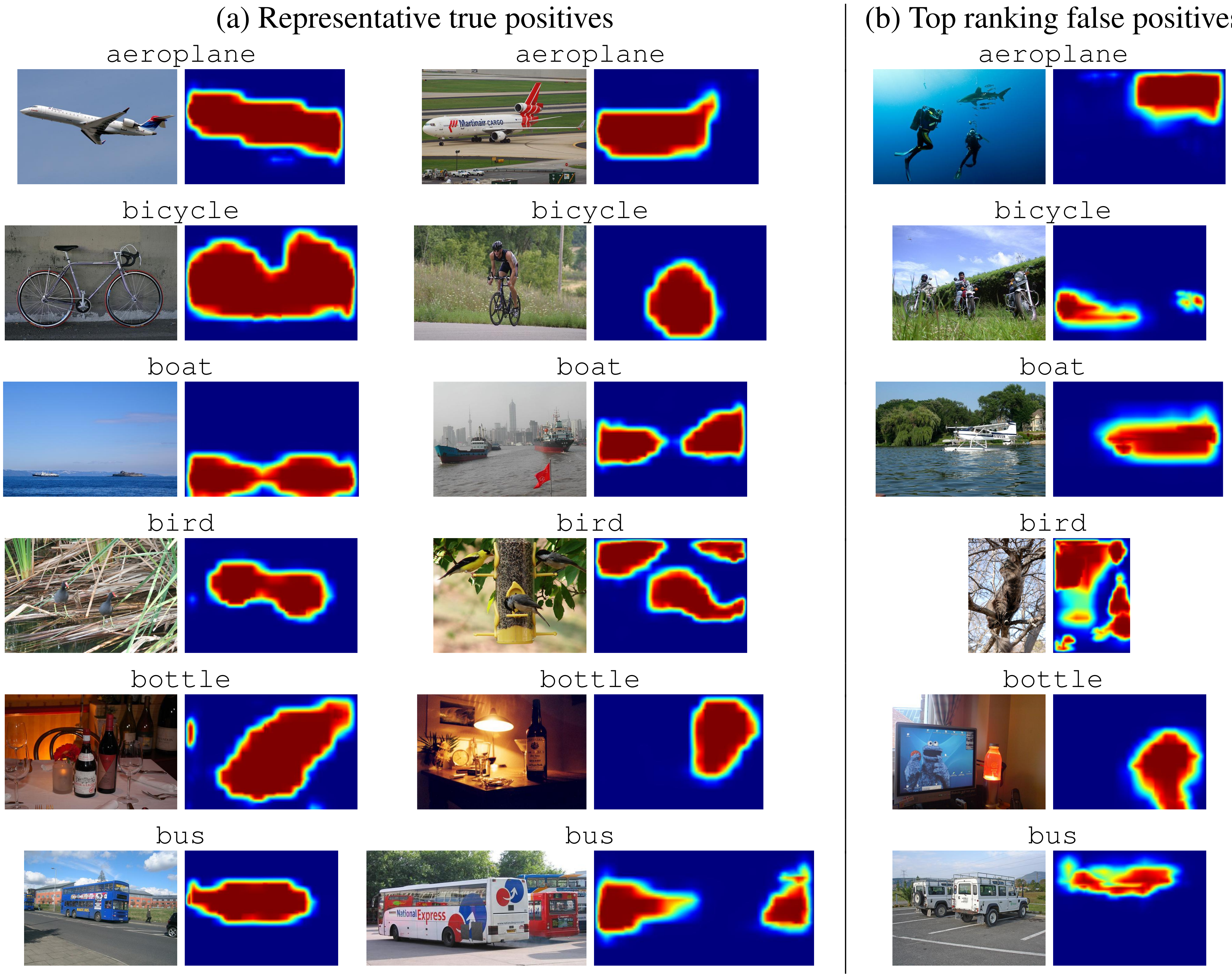
aeroplane - training iteration 0030

Results

	mAP	plane	bike	bird	boat	btl	bus	car	cat	chair	cow	table	dog	horse	moto	pers	plant	sheep	sofa	train	tv
A. Zeiler and Fergus [40]	79.0	96.0	77.1	88.4	85.5	55.8	85.8	78.6	91.2	65.0	74.4	67.7	87.8	86.0	85.1	90.9	52.2	83.6	61.1	91.8	76.1
B. Oquab et al. [26]	82.8	94.6	82.9	88.2	84.1	60.3	89.0	84.4	90.7	72.1	86.8	69.0	92.1	93.4	88.6	96.1	64.3	86.6	62.3	91.1	79.8
C. Chatfield et al. [4]	83.2	**96.8**	82.5	91.5	**88.1**	62.1	88.3	81.9	**94.8**	70.3	80.2	76.2	92.9	90.3	89.3	95.2	57.4	83.6	66.4	93.5	81.9
D. Full images (our)	78.7	95.3	77.4	85.6	83.1	49.9	86.7	77.7	87.2	67.1	79.4	73.5	85.3	90.3	85.6	92.7	47.8	81.5	63.4	91.4	74.1
E. Strong+weak (our)	86.0	96.5	88.3	91.9	87.7	64.0	90.3	86.8	93.7	74.0	**89.8**	76.3	93.4	94.9	91.2	97.3	66.0	90.9	69.9	93.9	83.2
F. Weak supervision (our)	**86.3**	96.7	**88.8**	**92.0**	87.4	**64.7**	**91.1**	**87.4**	94.4	**74.9**	89.2	**76.3**	**93.7**	**95.2**	**91.1**	**97.6**	**66.2**	**91.2**	**70.0**	**94.5**	**83.7**

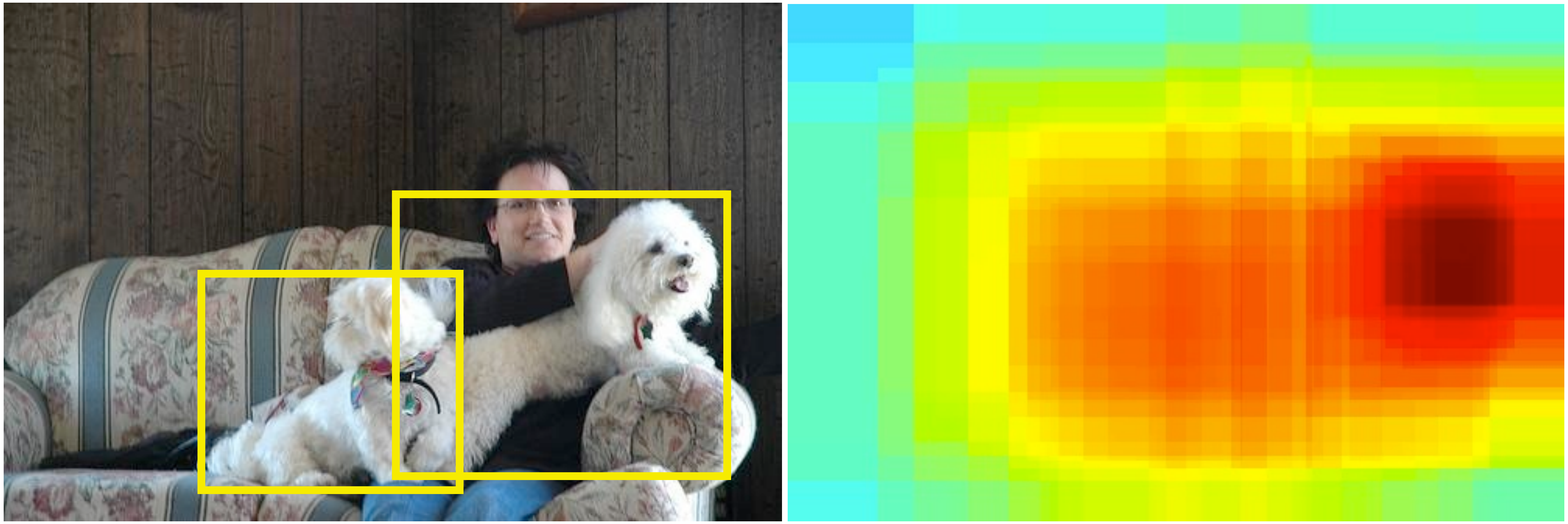
- Localizing objects by sliding helps
- Full supervision does not improve over weak supervision
- New state-of-the-art on Pascal VOC 2012 object classification

Object localization examples in testing data



Are bounding boxes harmful?

Output of the fully supervised CVPR’14 network:



- Why a higher score on the dog’s head?
- Responses are inconsistent with the annotations.
- Maybe we are doing it wrong.

Are bounding boxes harmful?

Bounding boxes are NOT alignment.

Should be treated as guidance not supervision
(at least for object classification)