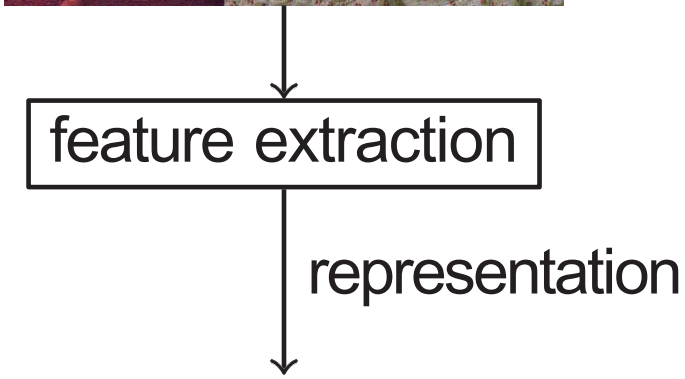
Computer Vision I:

Jingya Wang
Email: wangjingya@shanghaitech.edu.cn

data-driven approach

data-driven approach



data-driven approach

feature extraction

hand-craft feature

representation

parameters → classifier

"dog"

"cat"

data-driven approach

parameters → model

“dog”

“cat”

data-driven approach

parameters → model → representation

model → “dog”

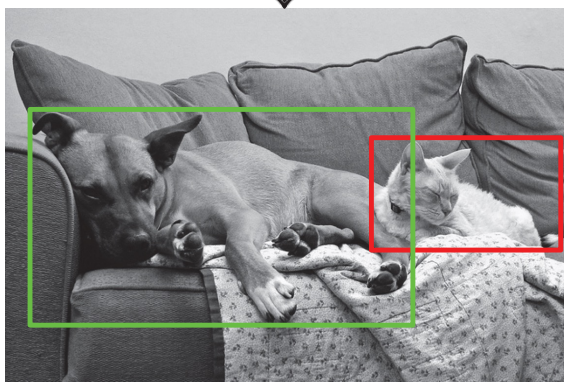
model → “cat”

data-driven approach

parameters → model → representation

model → "dog"

model → "cat"

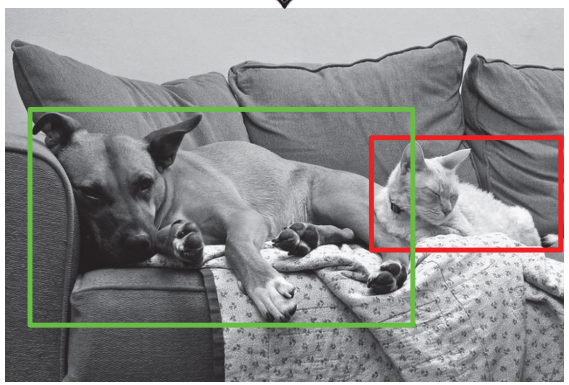


detection

data-driven approach

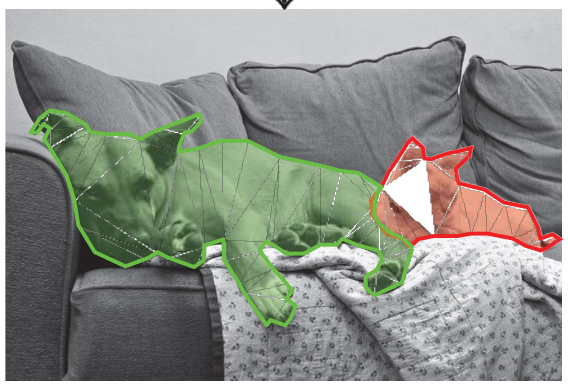
parameters → model → representation

更细粒度 pixel level



"dog"

"cat"



beyond classification

object localization

classify + regress

bounding box (x, y, w, h)

beyond classification

object localization

classify + regress

bounding box (x, y, w, h)

semantic segmentation

pixel-wise classify

beyond classification

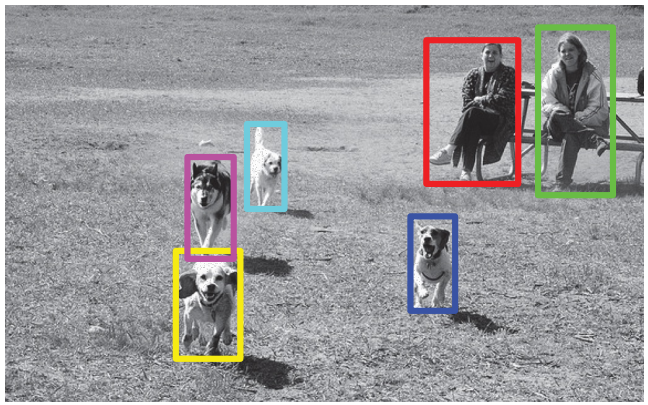
object localization

classify + regress

bounding box (x, y, w, h)

semantic segmentation

pixel-wise classify



object detection

per region: classify + regress

bounding box (x, y, w, h)

beyond classification

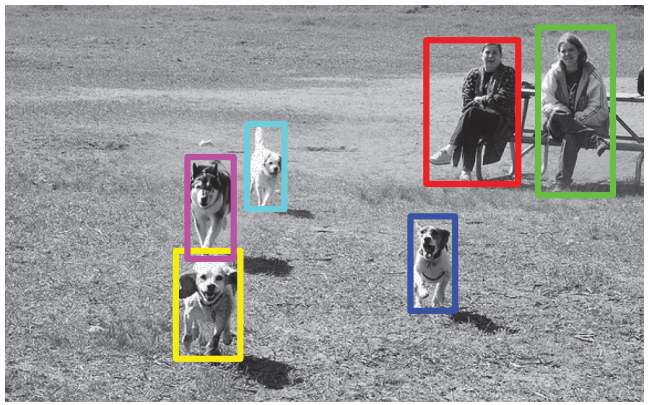
object localization

classify + regress

bounding box (x, y, w, h)

semantic segmentation

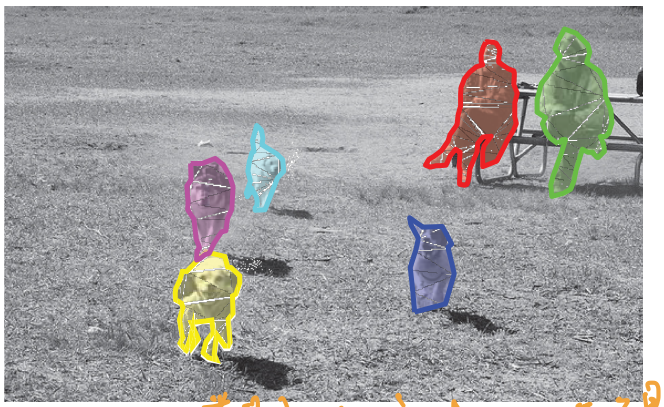
pixel-wise classify



object detection

per region: classify + regress

bounding box (x, y, w, h)



instance segmentation

per region: pixel-wise classify

template matching, or sliding window

- slide template over image at multiple positions
- positions can be overlapping, or even dense (every pixel)
- seek maximum similarity score

template matching, or sliding window

- slide template over image at multiple positions
- positions can be overlapping, or even dense (every pixel)
- seek maximum similarity score

template matching, or sliding window

- slide template over image at multiple positions
- positions can be overlapping, or even dense (every pixel)
- seek maximum similarity score

template matching, or sliding window

- slide template over image at multiple positions
- positions can be overlapping, or even dense (every pixel)
- seek maximum similarity score

template matching, or sliding window

- slide template over image at multiple positions
- positions can be overlapping, or even dense (every pixel)
- seek maximum similarity score

template matching, or sliding window

- slide template over image at multiple positions
- positions can be overlapping, or even dense (every pixel)
- seek maximum similarity score

template matching, or sliding window

- slide template over image at multiple positions
- positions can be overlapping, or even dense (every pixel)
- seek maximum similarity score

template matching, or sliding window

- slide template over image at multiple positions
- positions can be overlapping, or even dense (every pixel)
- seek maximum similarity score

template matching, or sliding window

- slide template over image at multiple positions
- positions can be overlapping, or even dense (every pixel)
- seek maximum similarity score

template matching, or sliding window

- slide template over image at multiple positions
- positions can be overlapping, or even dense (every pixel)
- seek maximum similarity score

template matching, or sliding window

- slide template over image at multiple positions
- positions can be overlapping, or even dense (every pixel)
- seek maximum similarity score

template matching, or sliding window

- slide template over image at multiple positions
- positions can be overlapping, or even dense (every pixel)
- seek maximum similarity score

template matching, or sliding window

- slide template over image at multiple positions
- positions can be overlapping, or even dense (every pixel)
- seek maximum similarity score

template matching, or sliding window

- slide template over image at multiple positions
- positions can be overlapping, or even dense (every pixel)
- seek maximum similarity score

template matching, or sliding window

- slide template over image at multiple positions
- positions can be overlapping, or even dense (every pixel)
- seek maximum similarity score

template matching, or sliding window

- slide template over image at multiple positions
- positions can be overlapping, or even dense (every pixel)
- seek maximum similarity score

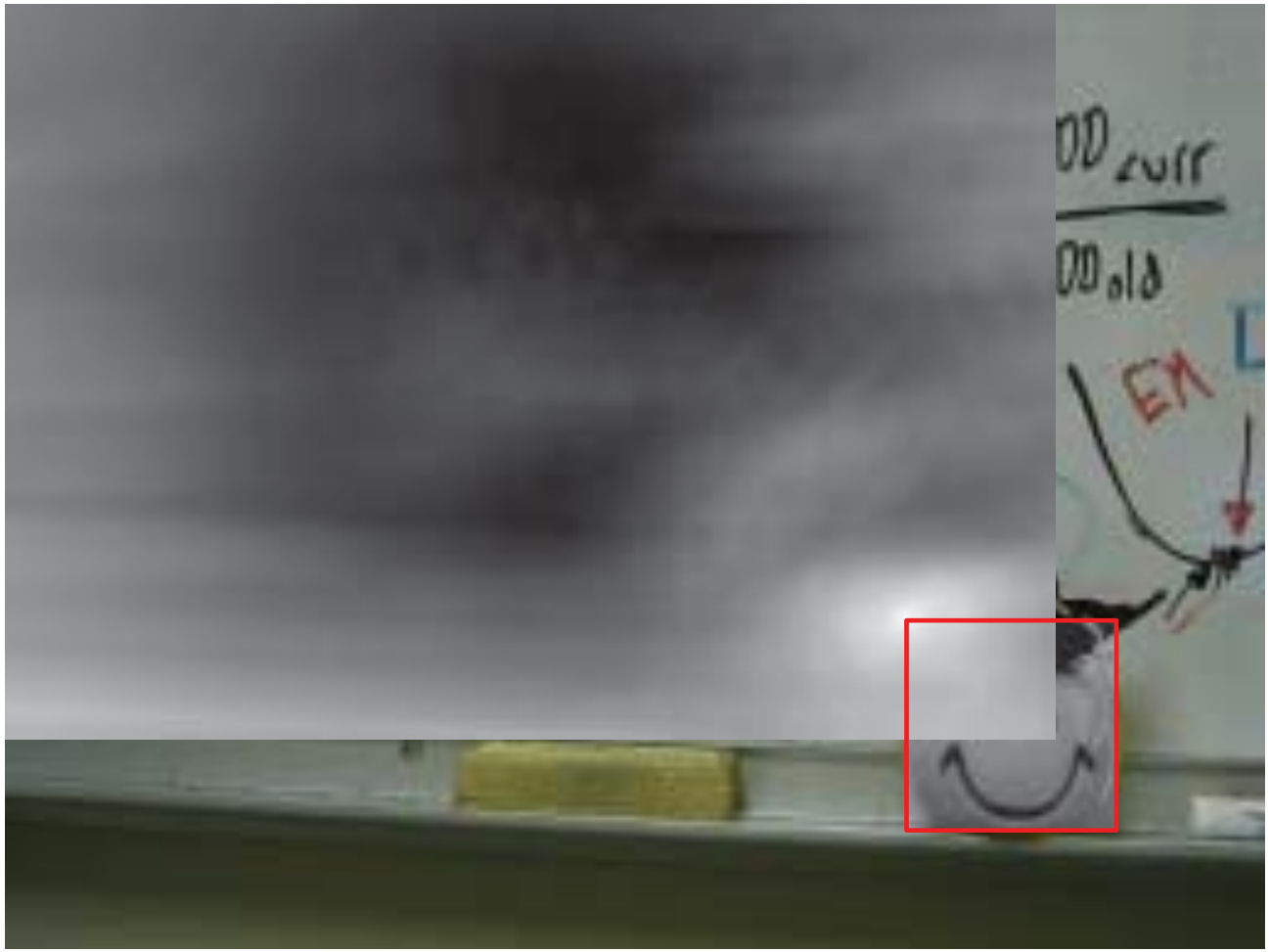
template matching, or sliding window

- slide template over image at multiple positions
- positions can be overlapping, or even dense (every pixel)
- seek maximum similarity score

template matching, or sliding window

- slide template over image at multiple positions
- positions can be overlapping, or even dense (every pixel)
- seek maximum similarity score

template matching, or sliding window



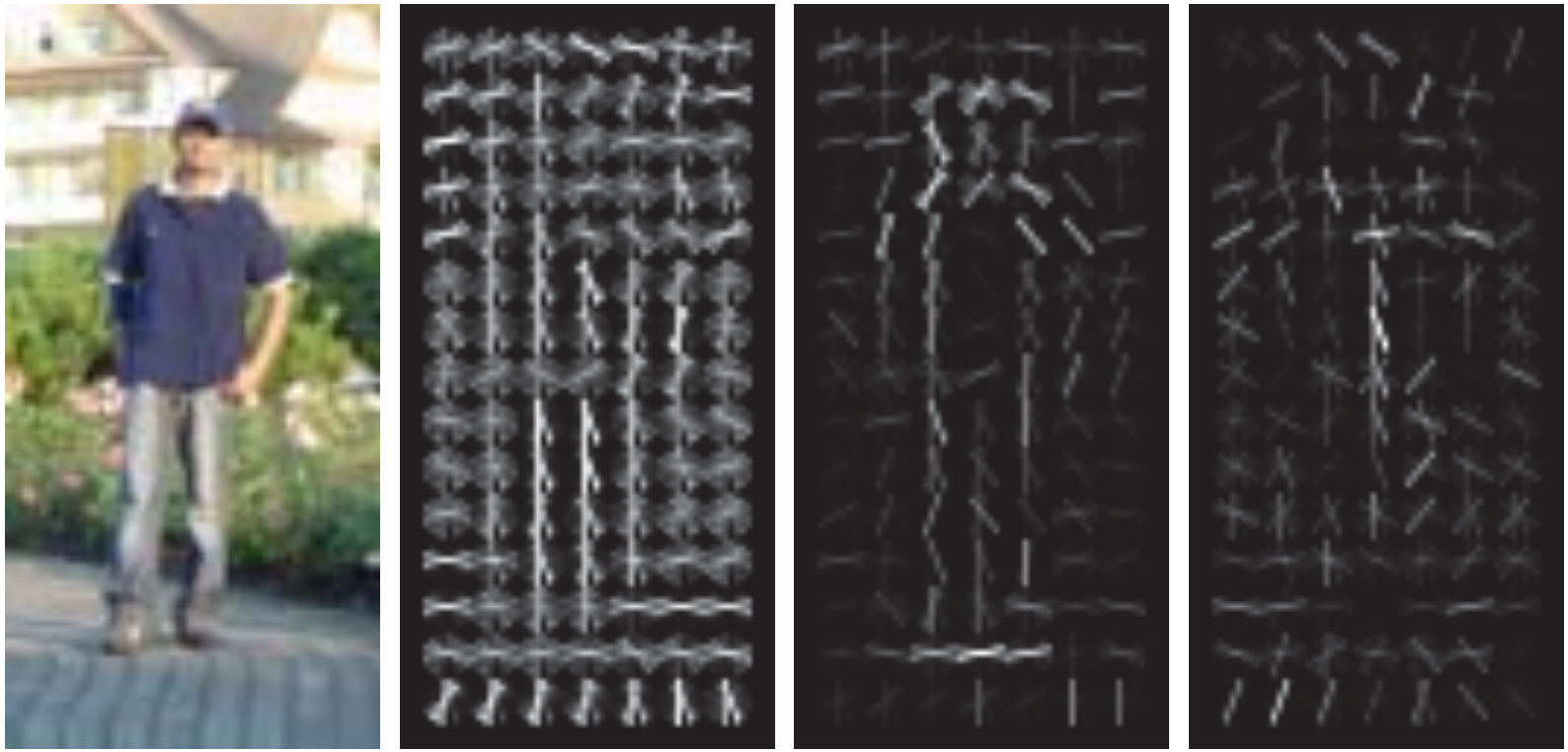
- slide template over image at multiple positions
- positions can be overlapping, or even dense (every pixel)
- seek maximum similarity score (*e.g.* cross-correlation)

two problems

- to detect a given instance (template), a similarity score may be enough; but to detect an object of a given class, we need strong features and a good classifier
- with unknown position, scale and aspect ratio, the search space is 4-dimensional: to search efficiently, we need something better than exhaustive search

histogram of oriented gradients (HOG)

[Dalal and Triggs 2005]

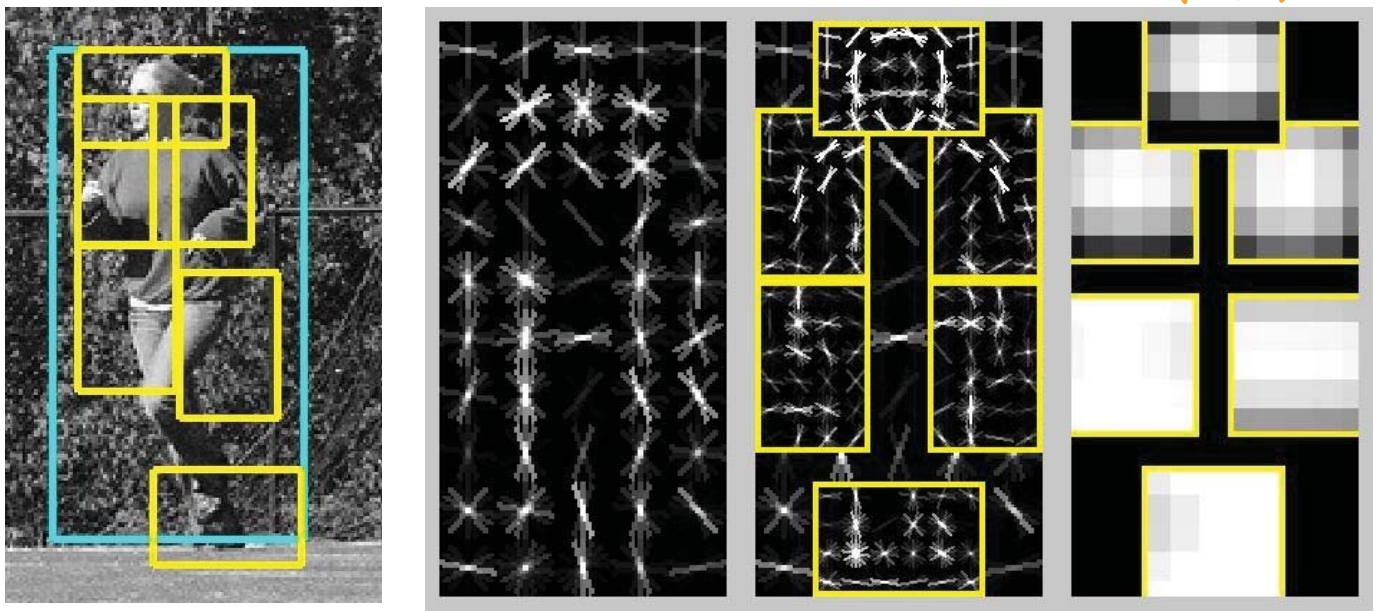


- dense, SIFT-like descriptors
- SVM classifier
- sliding window detection at all positions and scales

Dalal and Triggs. CVPR 2005. Histograms of Oriented Gradients for Human Detection.

deformable part model (DPM)

[Felzenszwalb et al. 2008]



- appearance represented by HOG
- spatial configuration inspired by “pictorial structures”
- part locations treated as latent variables: latent SVM

Felzenszwalb, Mcallester and Ramanan. CVPR 2008. A Discriminatively Trained, Multiscale, Deformable Part Model.

selective search (SS)

[van de Sande et al. 2011]

input image

ground truth

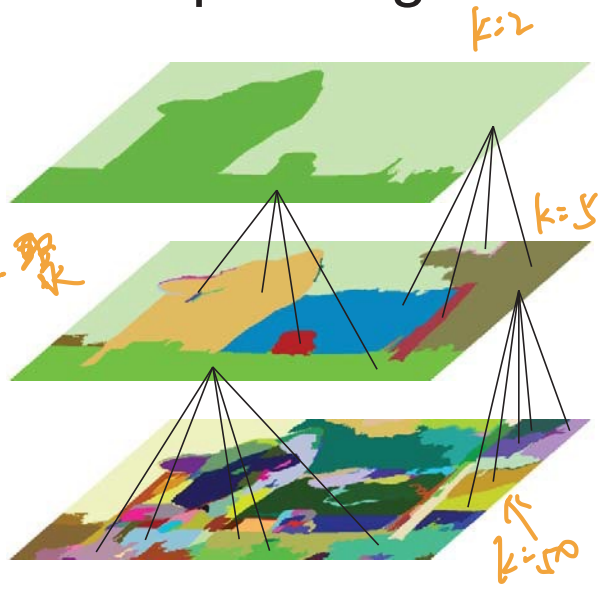
van de Sande, Uijlings, Gevers and Smeulders. ICCV 2011. Segmentation As Selective Search for Object Recognition.

selective search (SS)

[van de Sande et al. 2011]

input image

ground truth

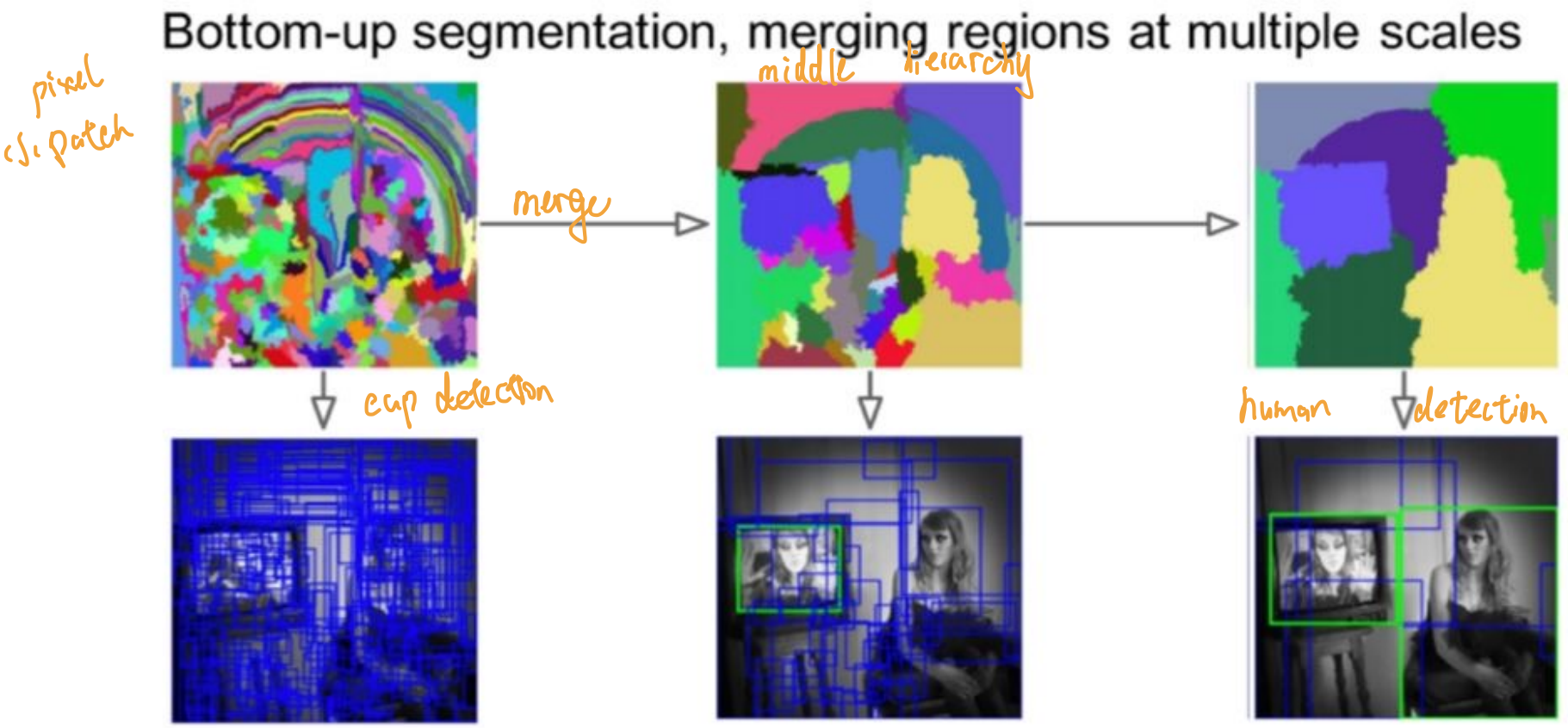


hierarchical grouping

object proposals

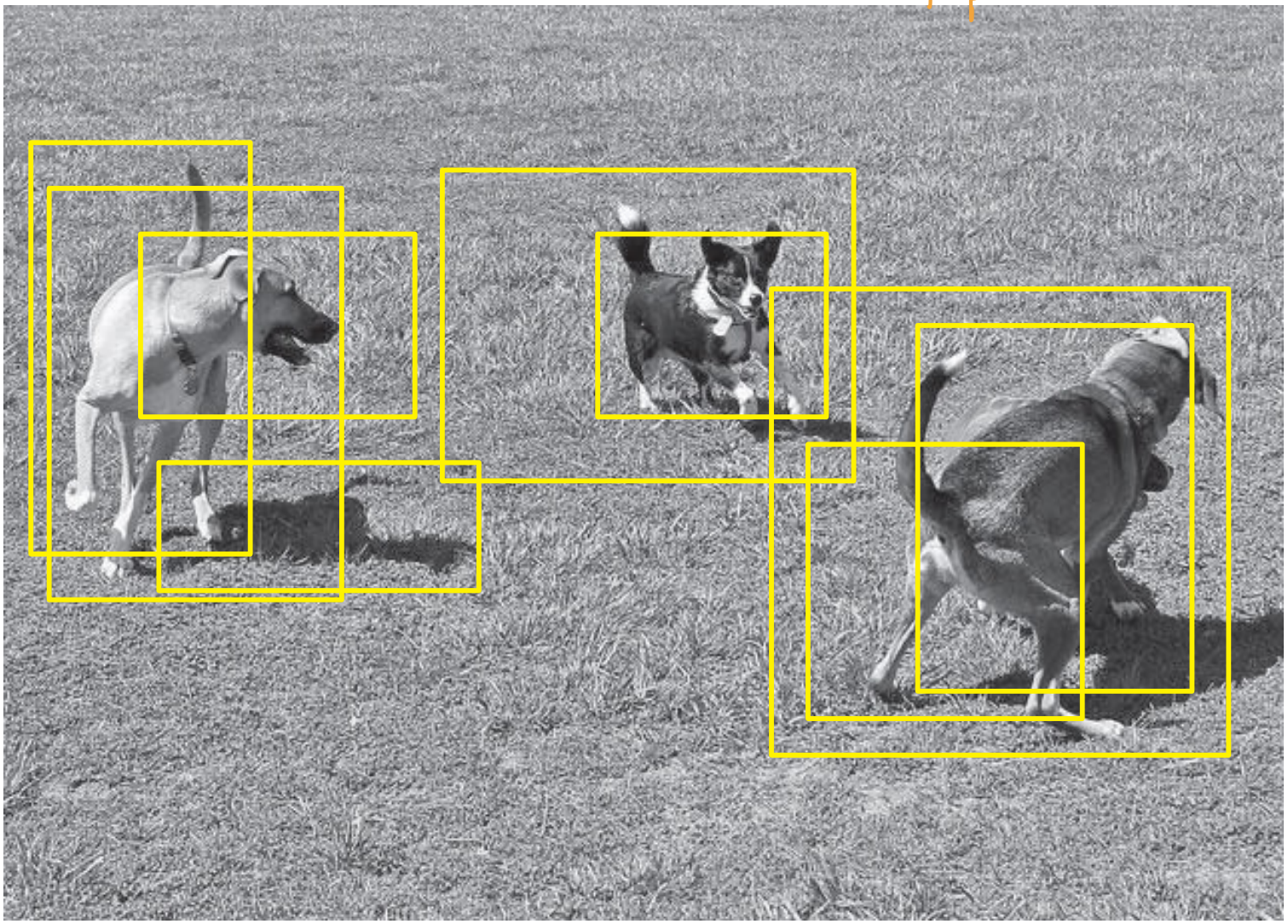
selective search (SS)

- hierarchical segmentation at all scales
- simple geometric and appearance features (*e.g.* size, texture)
- high recall: ~ 97% of ground truth objects found with ~ 1000 – 2000 proposals/image at ~ 2-5s/image

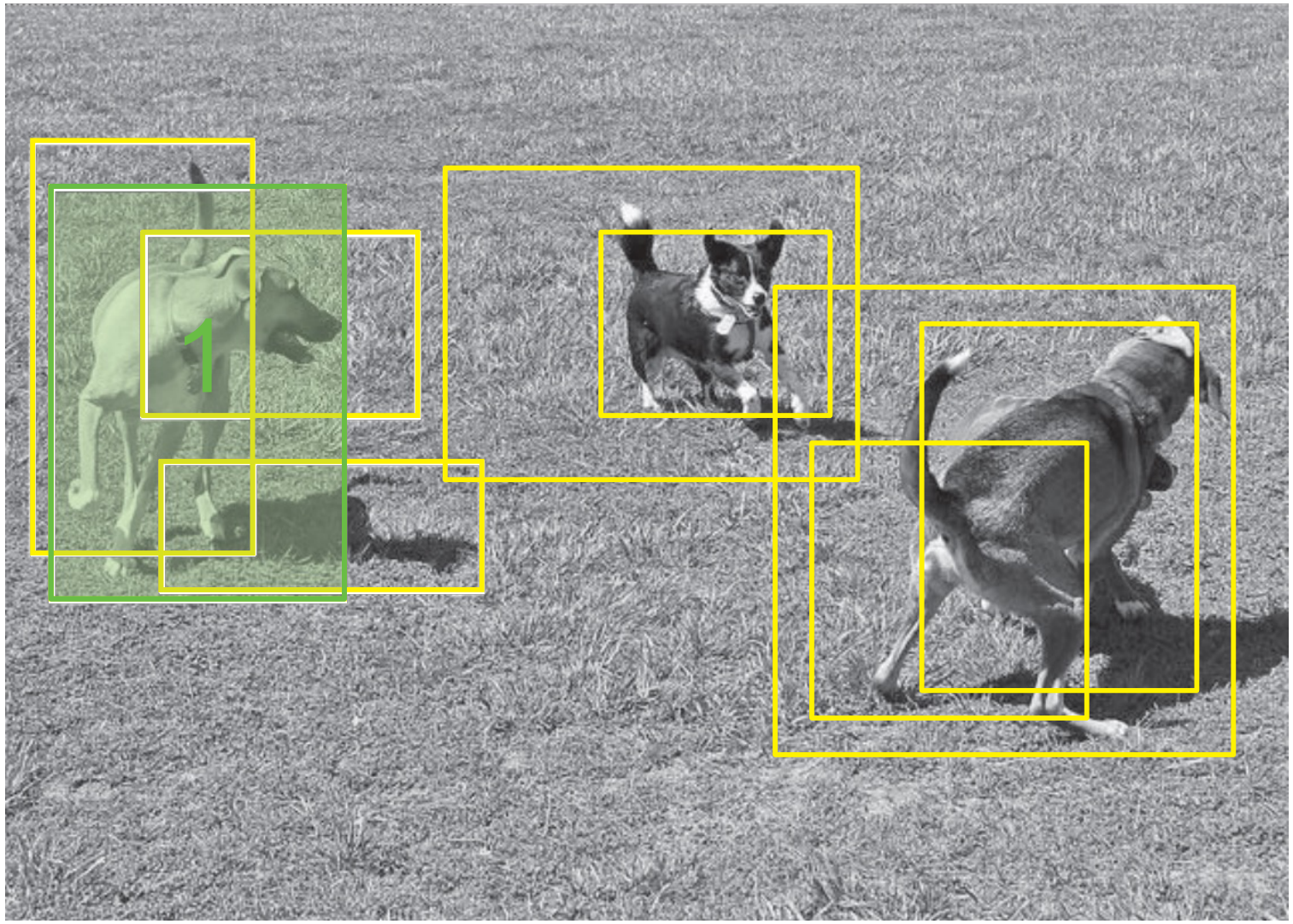


van de Sande, Uijlings, Gevers and Smeulders. ICCV 2011. Segmentation As Selective Search for Object Recognition.

non-maximum suppression (NMS)

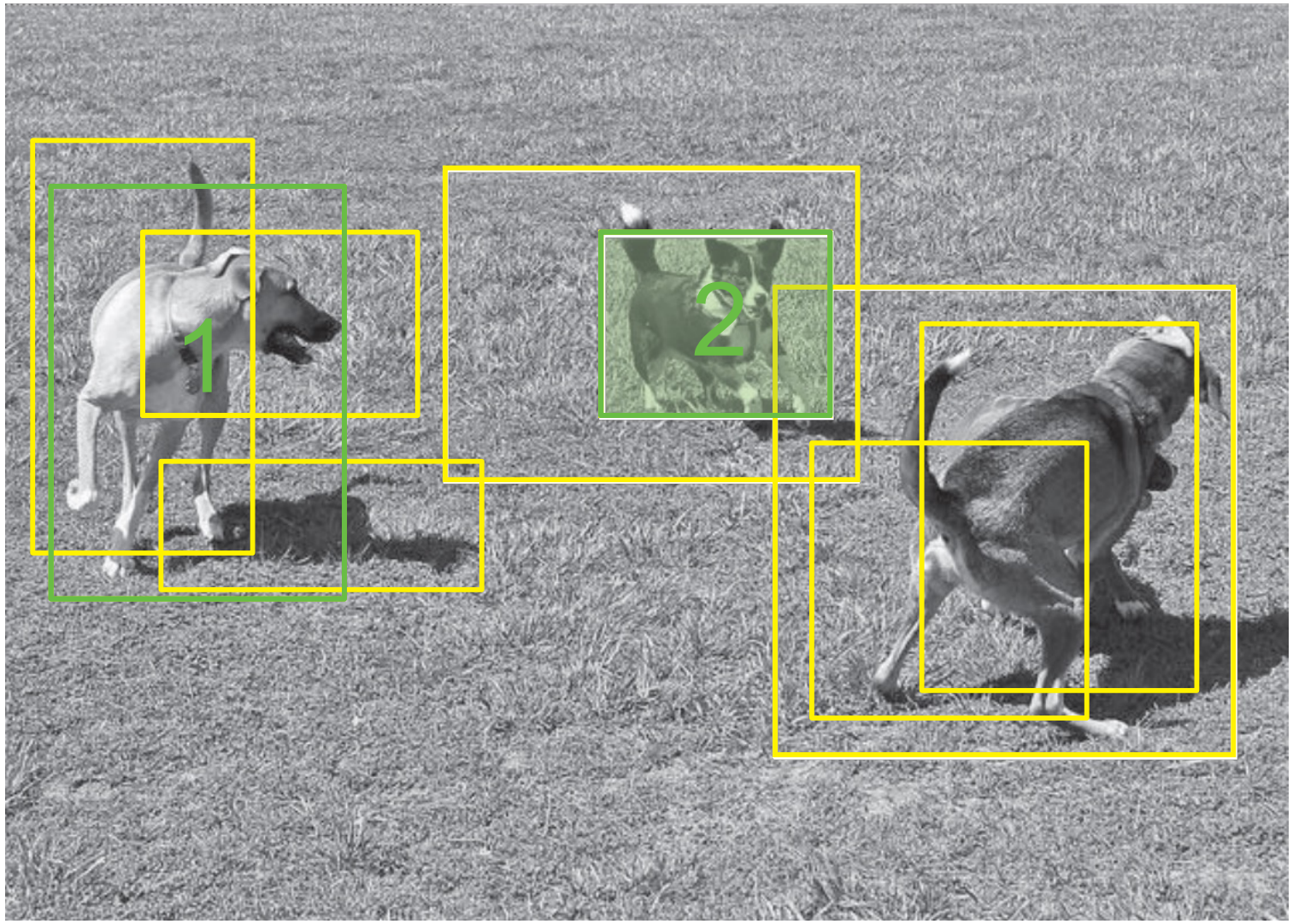


non-maximum suppression (NMS)



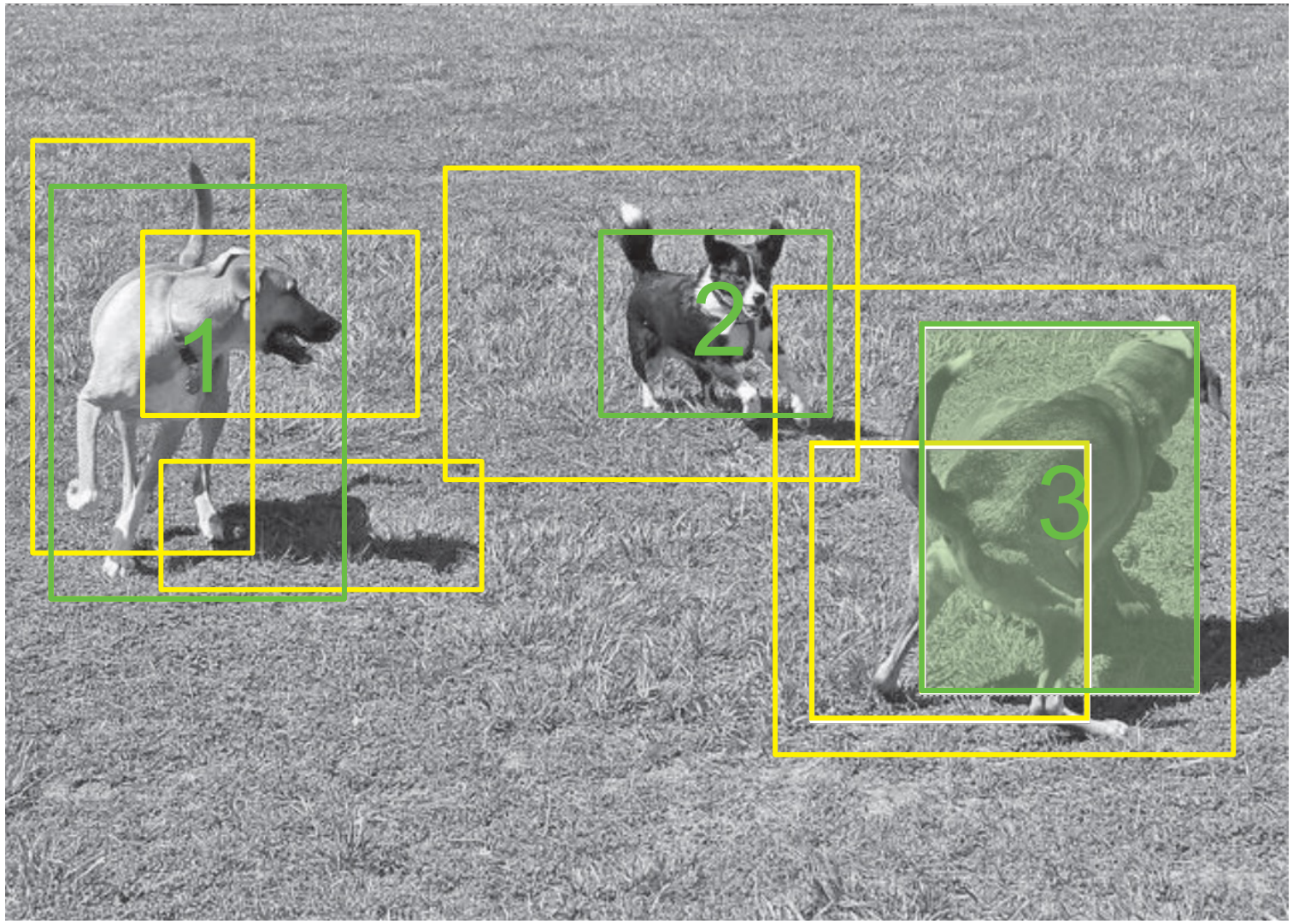
region 1 remains

non-maximum suppression (NMS)



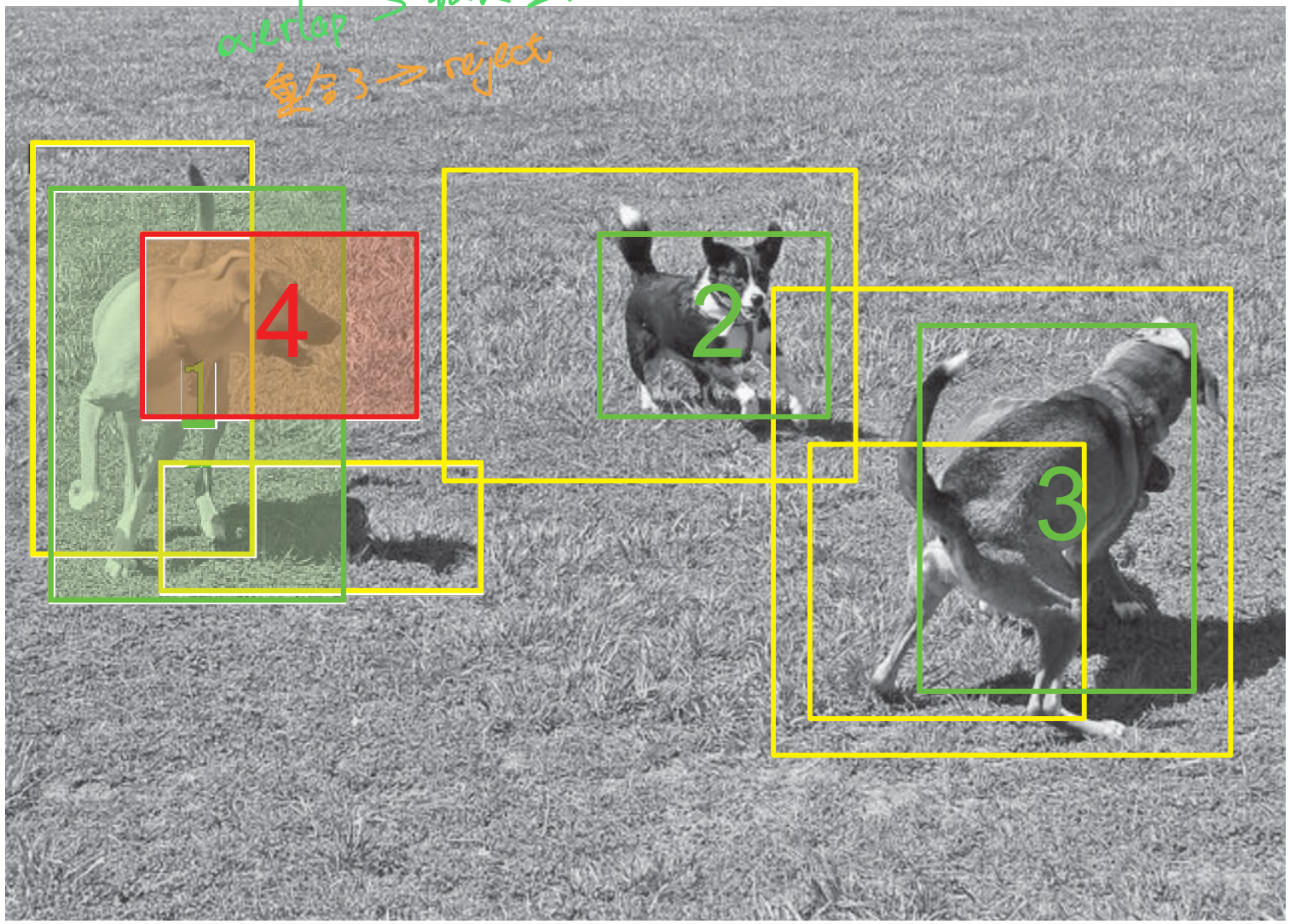
region 2 remains

non-maximum suppression (NMS)



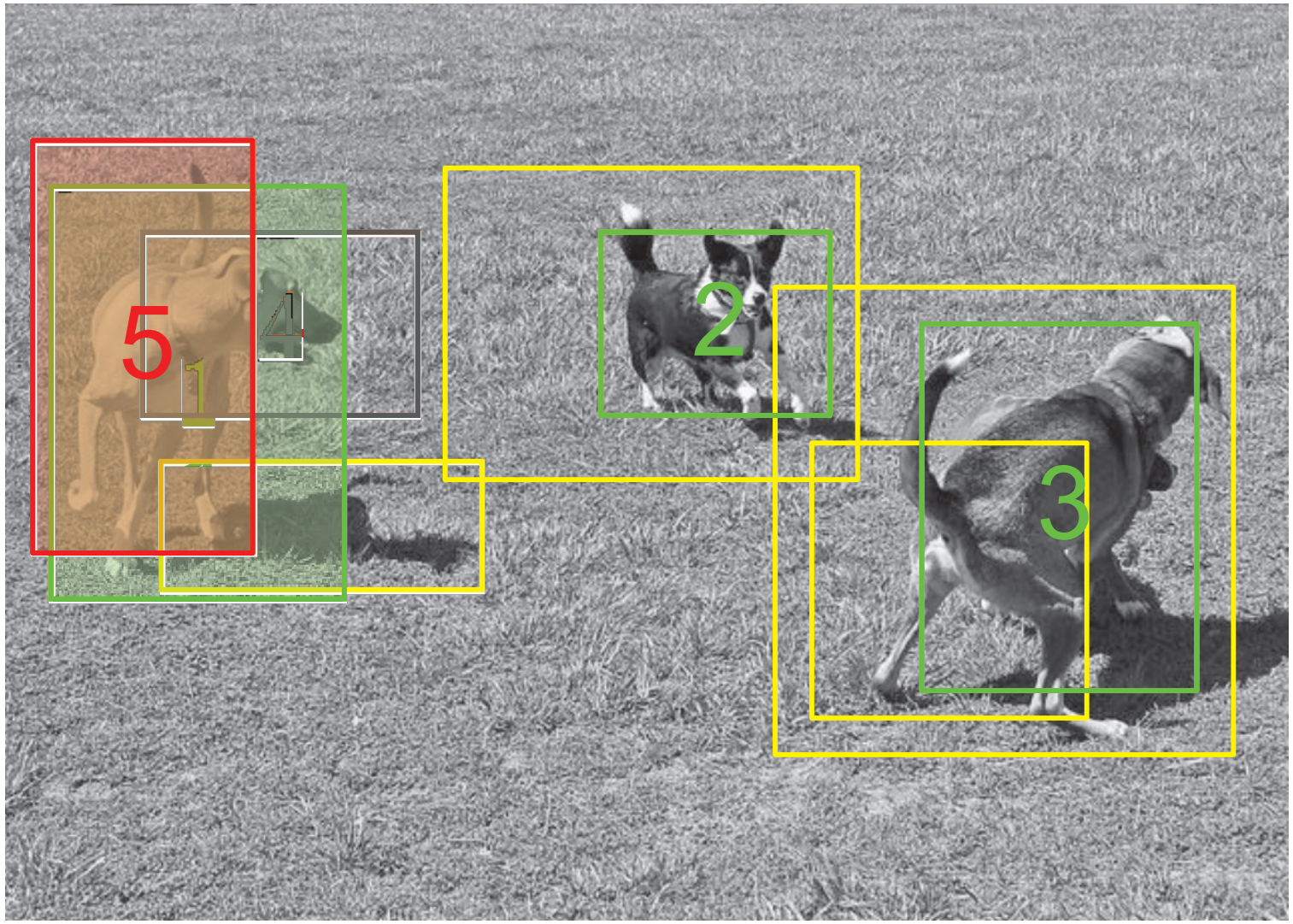
region 2 remains

non-maximum suppression (NMS)



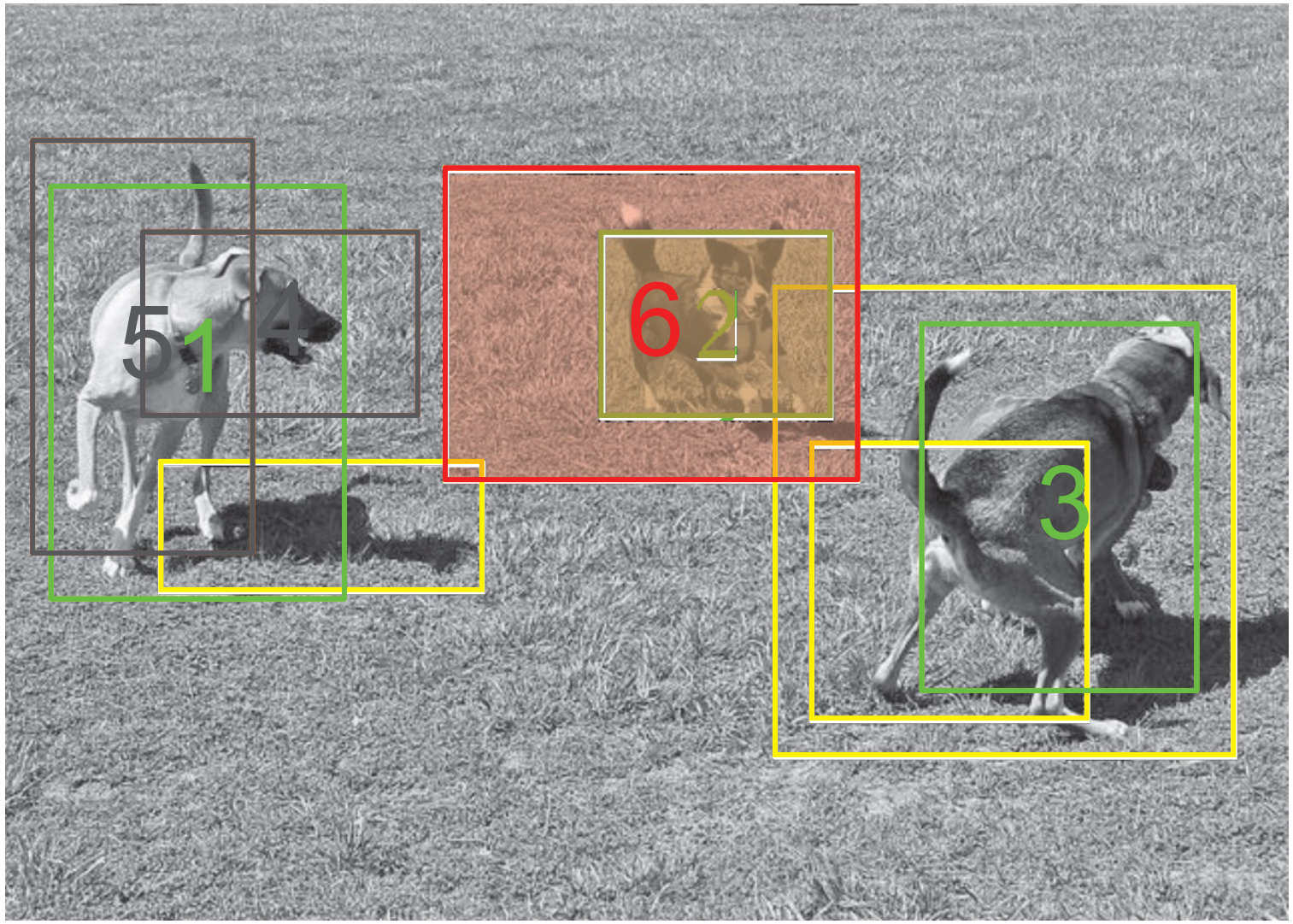
region 4 is rejected because $J(r_4, r_1) = 0.2750 > 0.25$

non-maximum suppression (NMS)



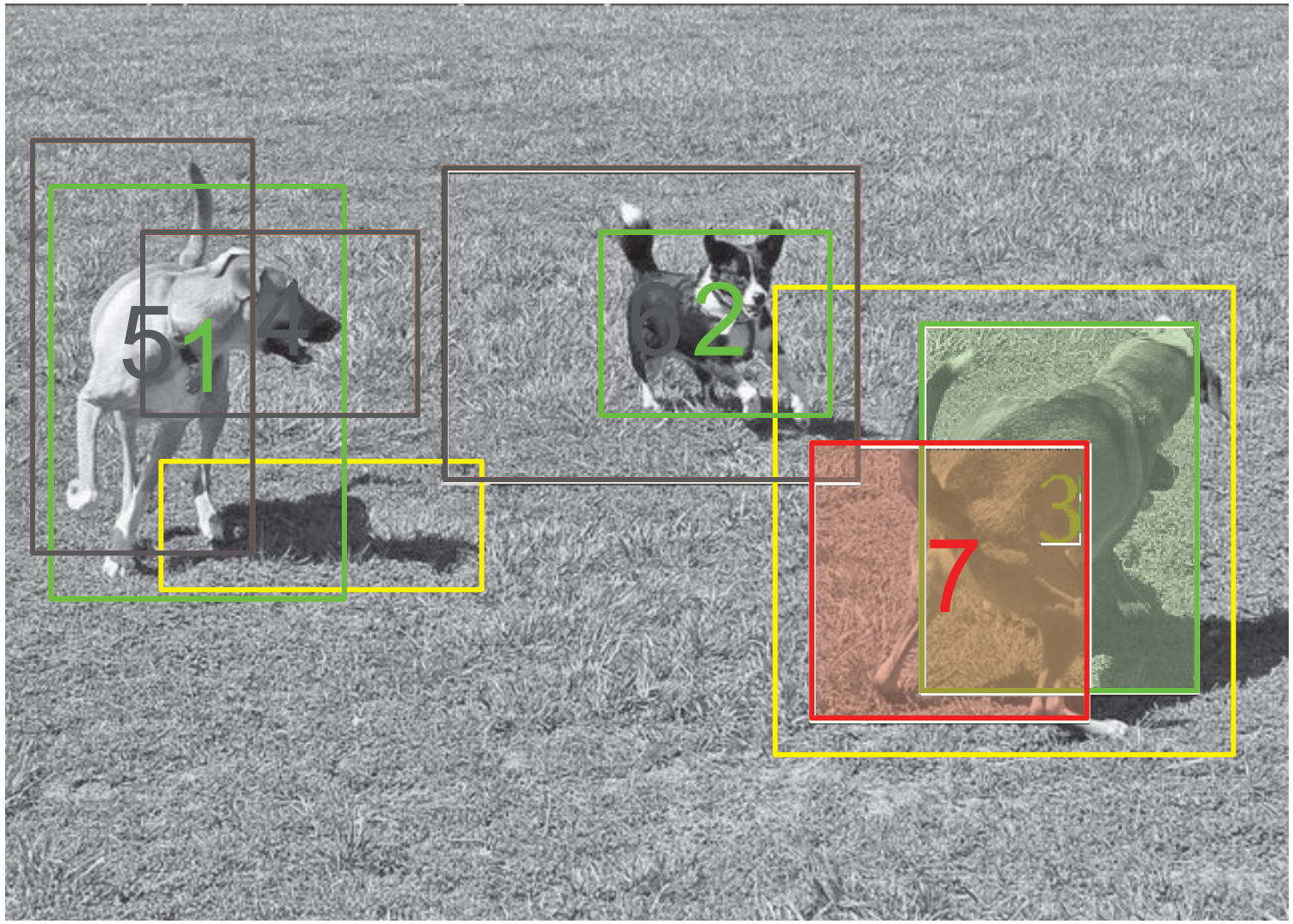
region 5 is rejected because $J(r_4, r_1) = 0.2750 > 0.25$

non-maximum suppression (NMS)



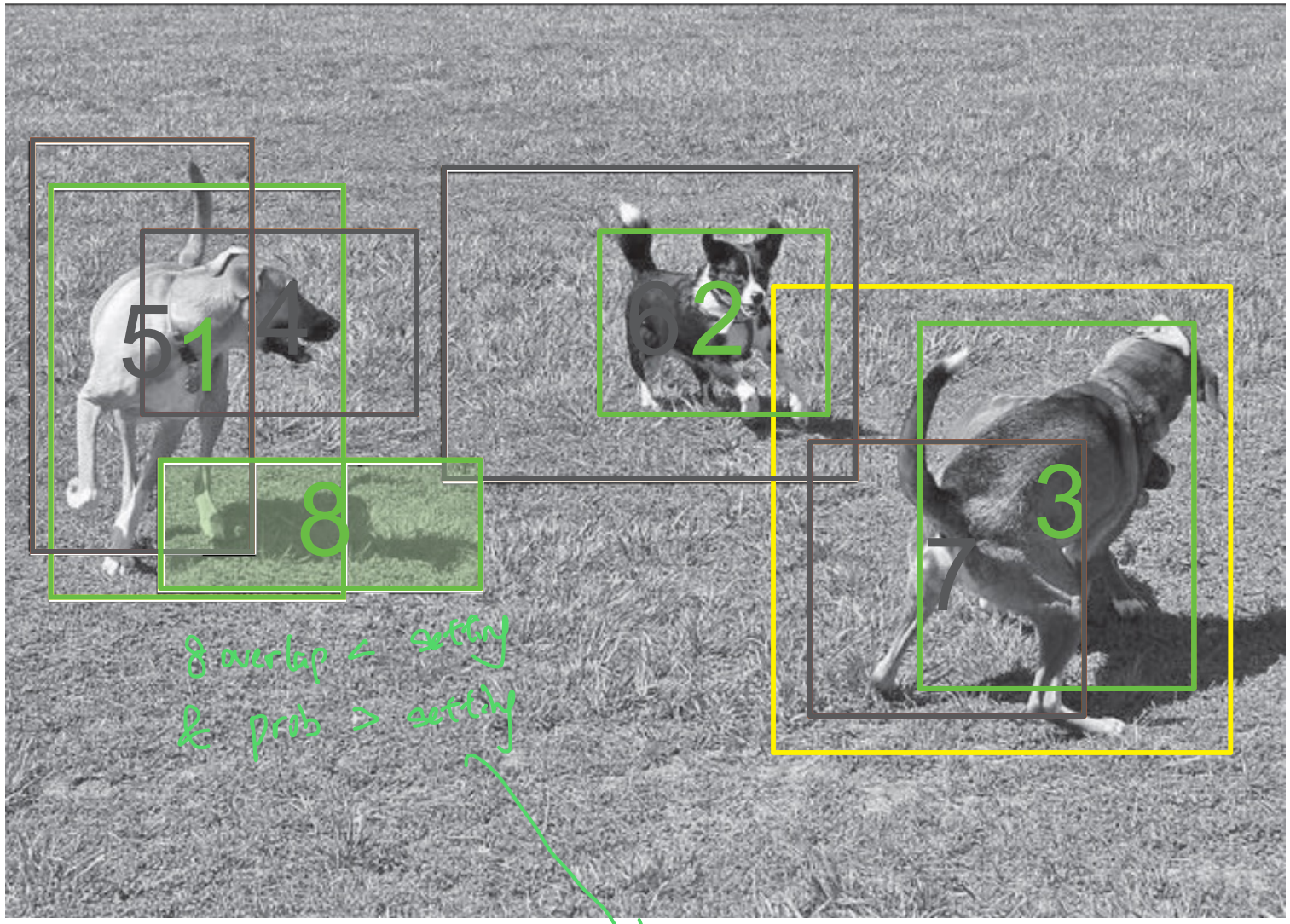
region 6 is rejected because $J(r_4, r_1) = 0.2750 > 0.25$

non-maximum suppression (NMS)

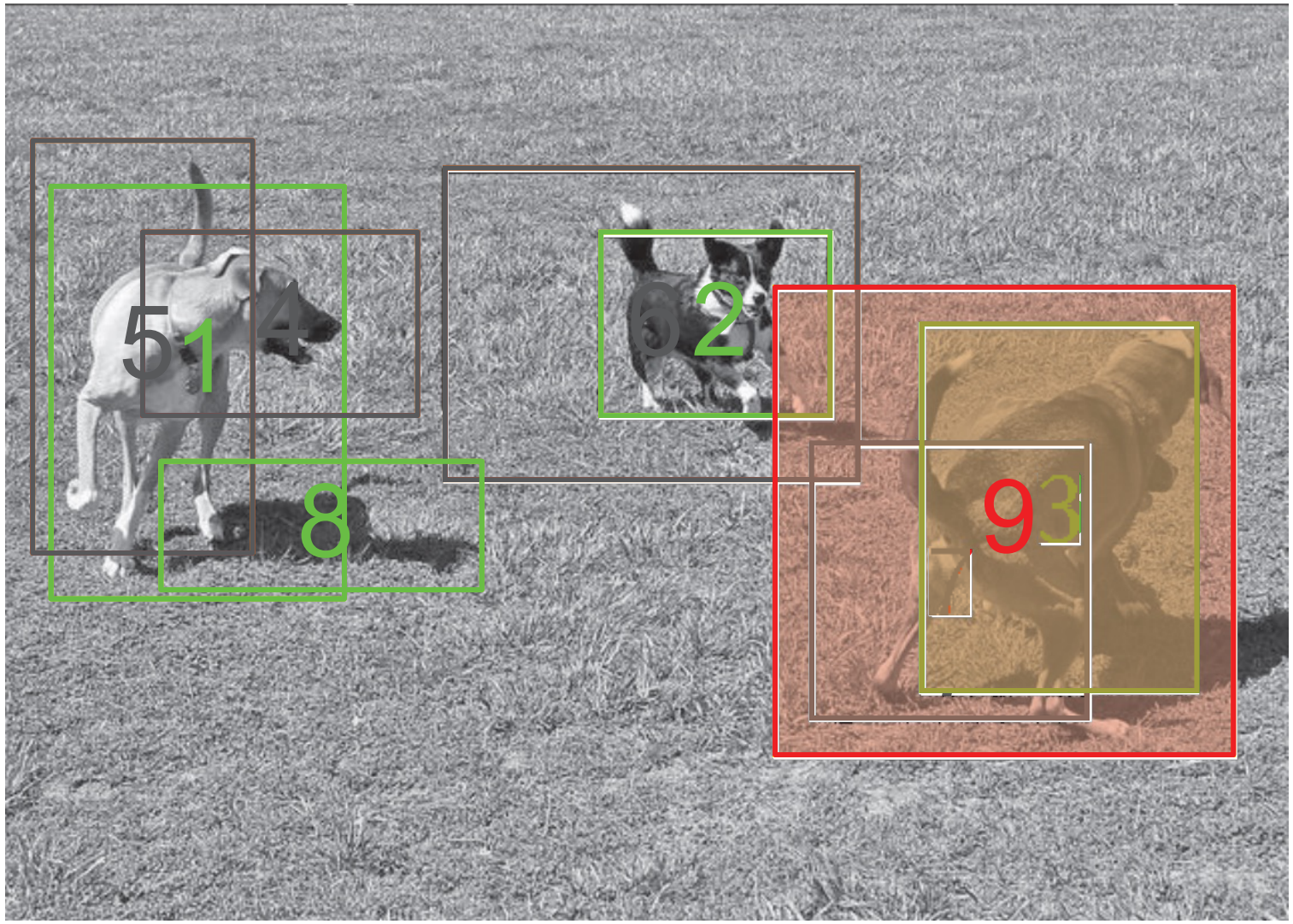


region 7 is rejected because $J(r_4, r_1) = 0.2750 > 0.25$

non-maximum suppression (NMS)

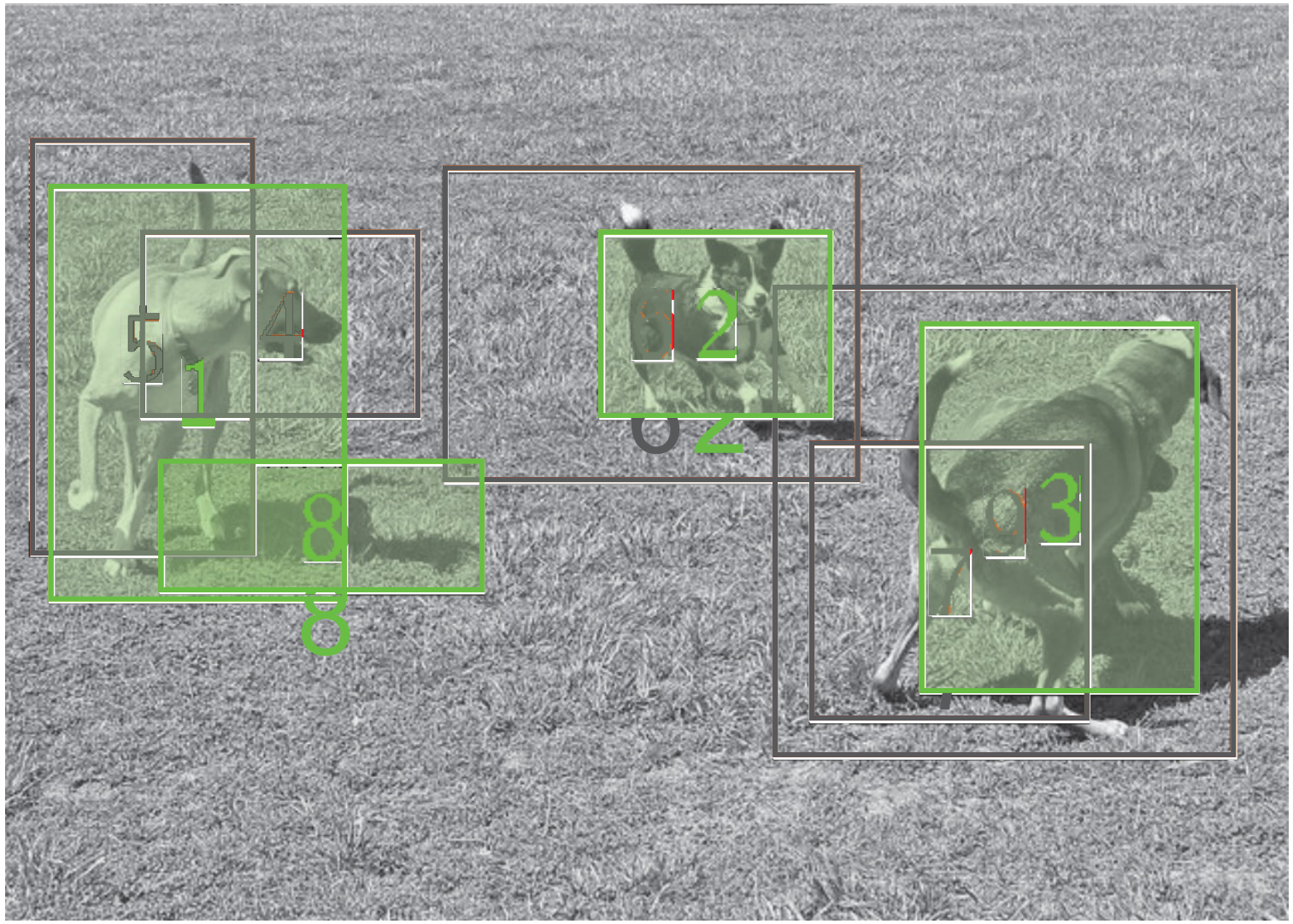


non-maximum suppression (NMS)



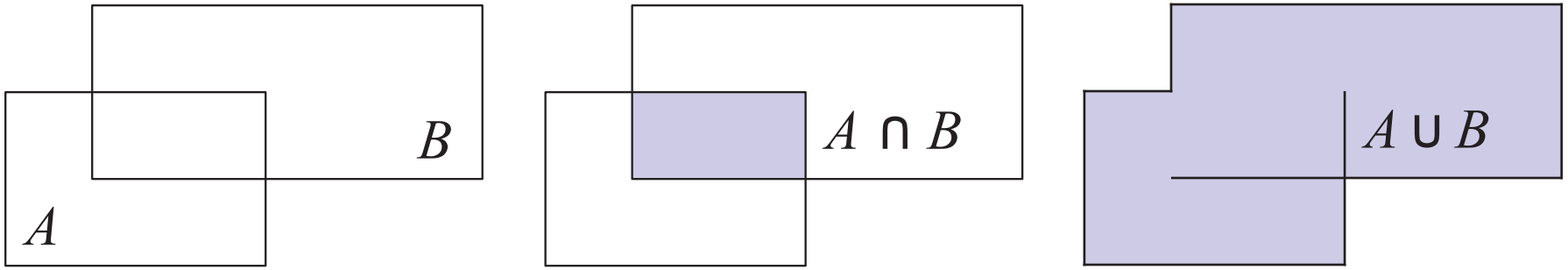
region 9 is rejected because $J(r_9, r_3) = 0.4706 > 0.25$

non-maximum suppression (NMS)



in the end, regions 1, 2, 3, 8 remain

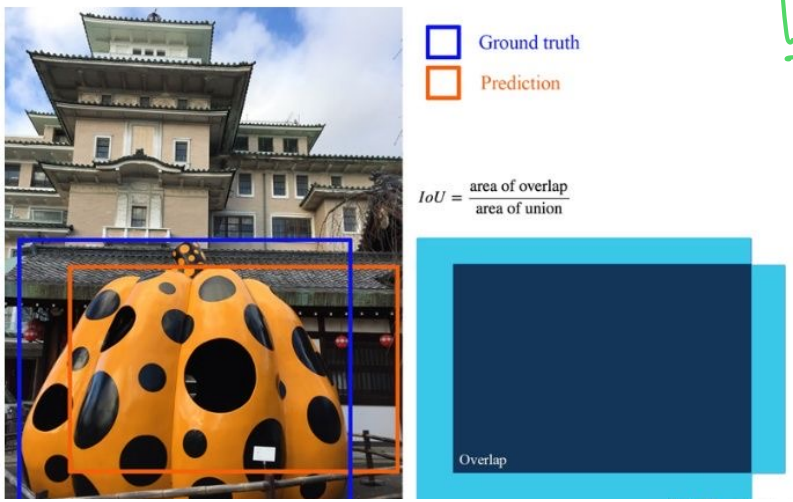
region overlap



- given regions $A, B \subset \mathrm{R}^2$ represented as planar point sets (including interior)
- their intersection over union (IoU) or Jaccard index is

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

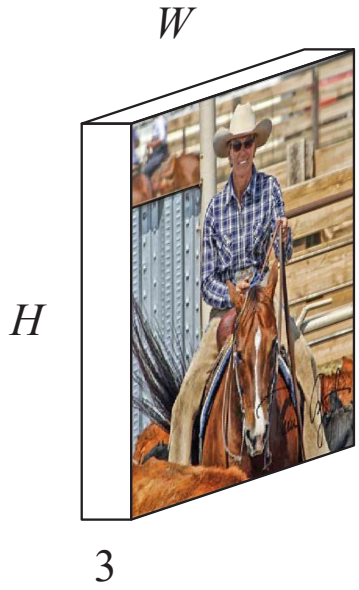
> setting → reject



two-stage detection

regions with CNN features (R-CNN)

[Girshick et al. 2014]

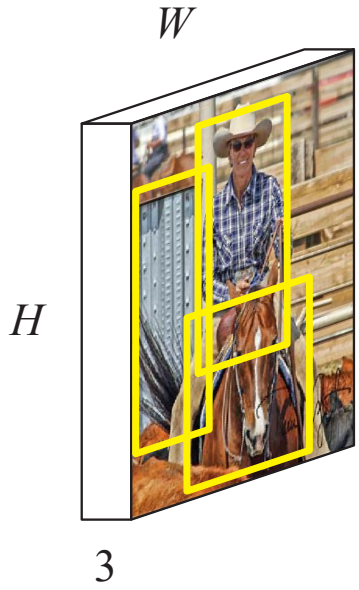


- 3-channel RGB input, fixed width $W = 500$ pixels
- $\sim$ 2000 SS region proposals warped into fixed $w \times h = 227 \times 227$
- each proposal yields a $k = 4096$ dimensional feature by CaffeNet
- each feature is classified into c classes by c one-*vs*. -rest SVMs and localized by bounding box regression

Girshick, Donahue, Darrell and Malik. CVPR 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.

regions with CNN features (R-CNN)

[Girshick et al. 2014]

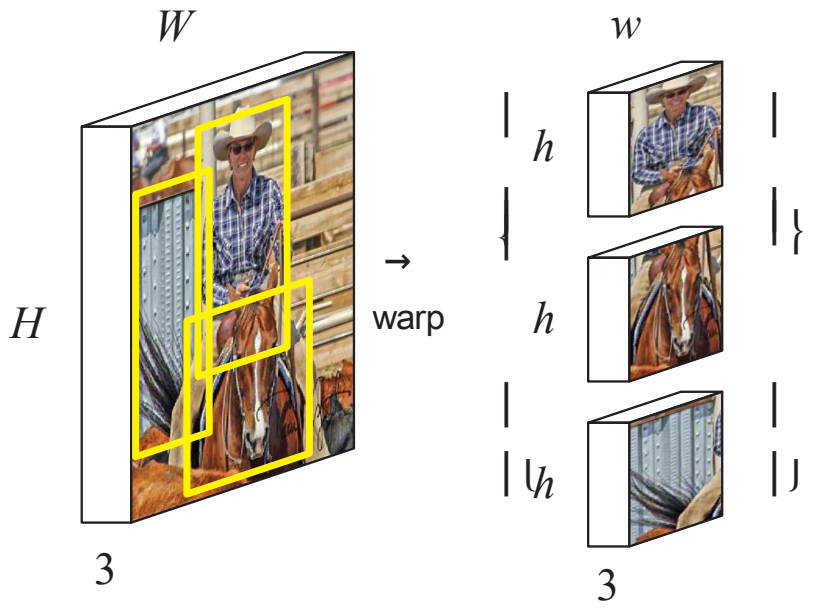


- 3-channel RGB input, fixed width $W = 500$ pixels
- ~ 2000 SS region proposals warped into fixed $w \times h = 227 \times 227$
- each proposal yields a $k = 4096$ dimensional feature by CaffeNet
- each feature is classified into c classes by c one-*vs*. -rest SVMs and localized by bounding box regression

Girshick, Donahue, Darrell and Malik. CVPR 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.

regions with CNN features (R-CNN)

[Girshick et al. 2014]

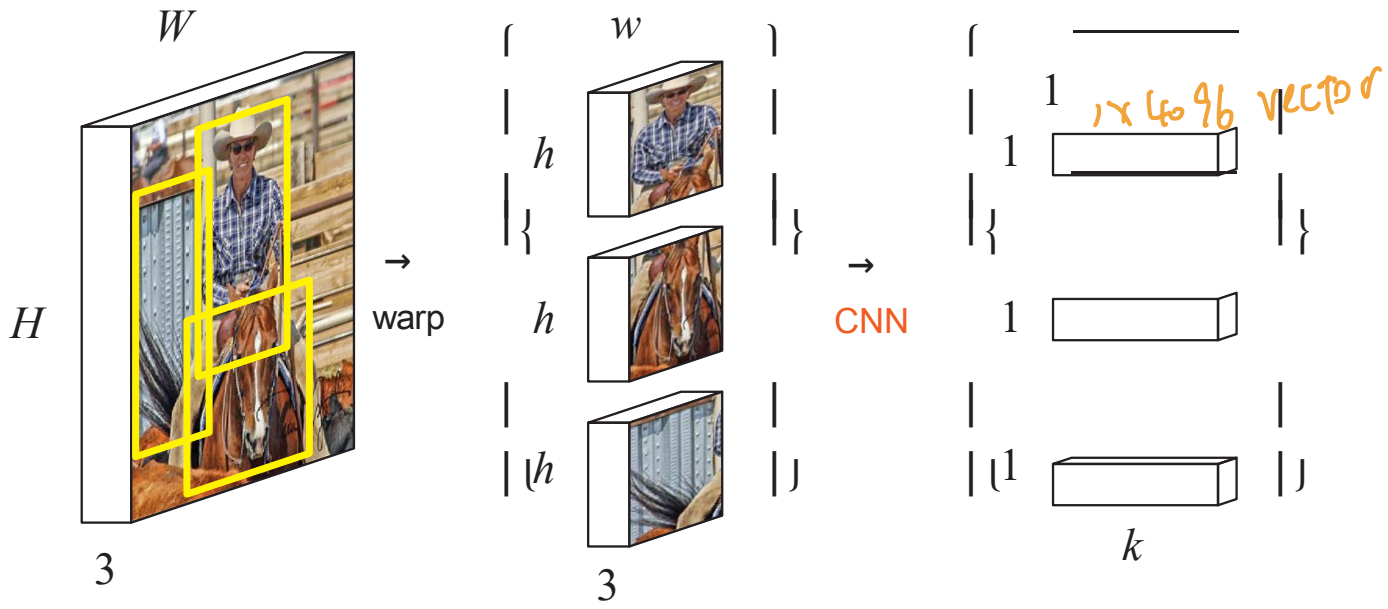


- 3-channel RGB input, fixed width $W = 500$ pixels
- ~ 2000 SS region proposals warped into fixed $w \times h = 227 \times 227$
- each proposal yields a $k = 4096$ dimensional feature by CaffeNet
- each feature is classified into c classes by c one-*vs*. -rest SVMs and localized by bounding box regression

Girshick, Donahue, Darrell and Malik. CVPR 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.

regions with CNN features (R-CNN)

[Girshick et al. 2014]

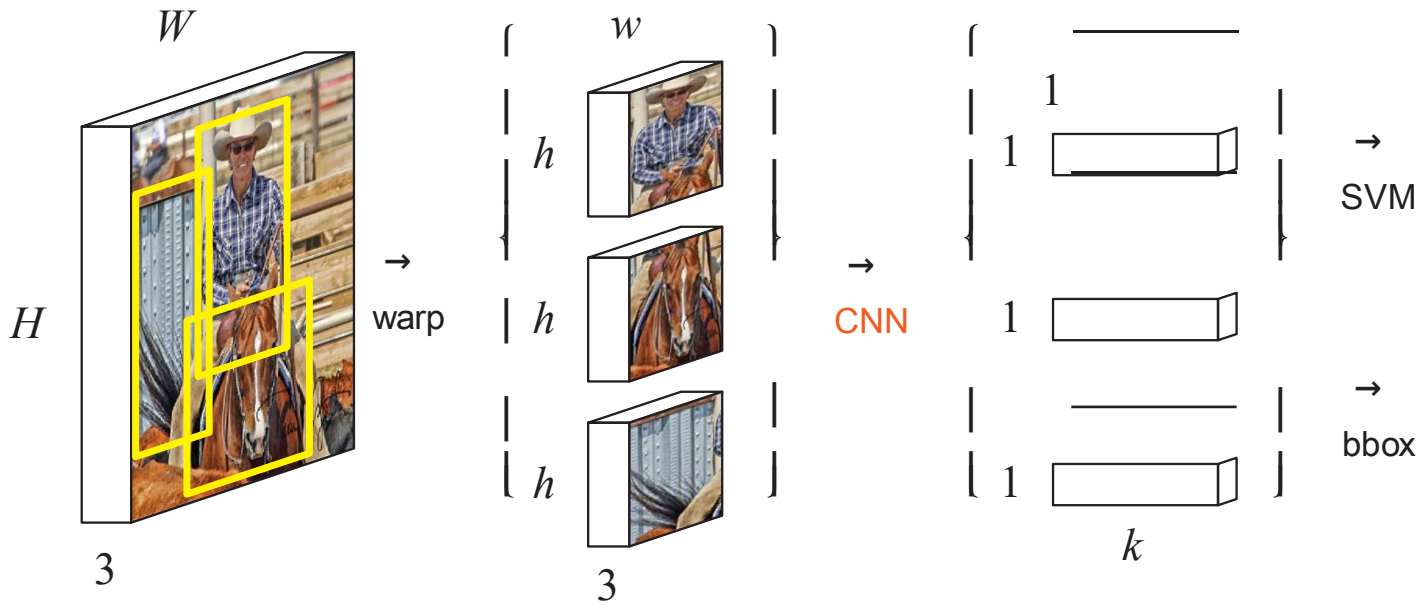


- 3-channel RGB input, fixed width $W = 500$ pixels
- ~ 2000 SS region proposals warped into fixed $w \times h = 227 \times 227$
- each proposal yields a $k = 4096$ dimensional feature by CaffeNet
- each feature is classified into c classes by c one-*vs*. -rest SVMs and localized by bounding box regression

Girshick, Donahue, Darrell and Malik. CVPR 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.

regions with CNN features (R-CNN)

[Girshick et al. 2014]

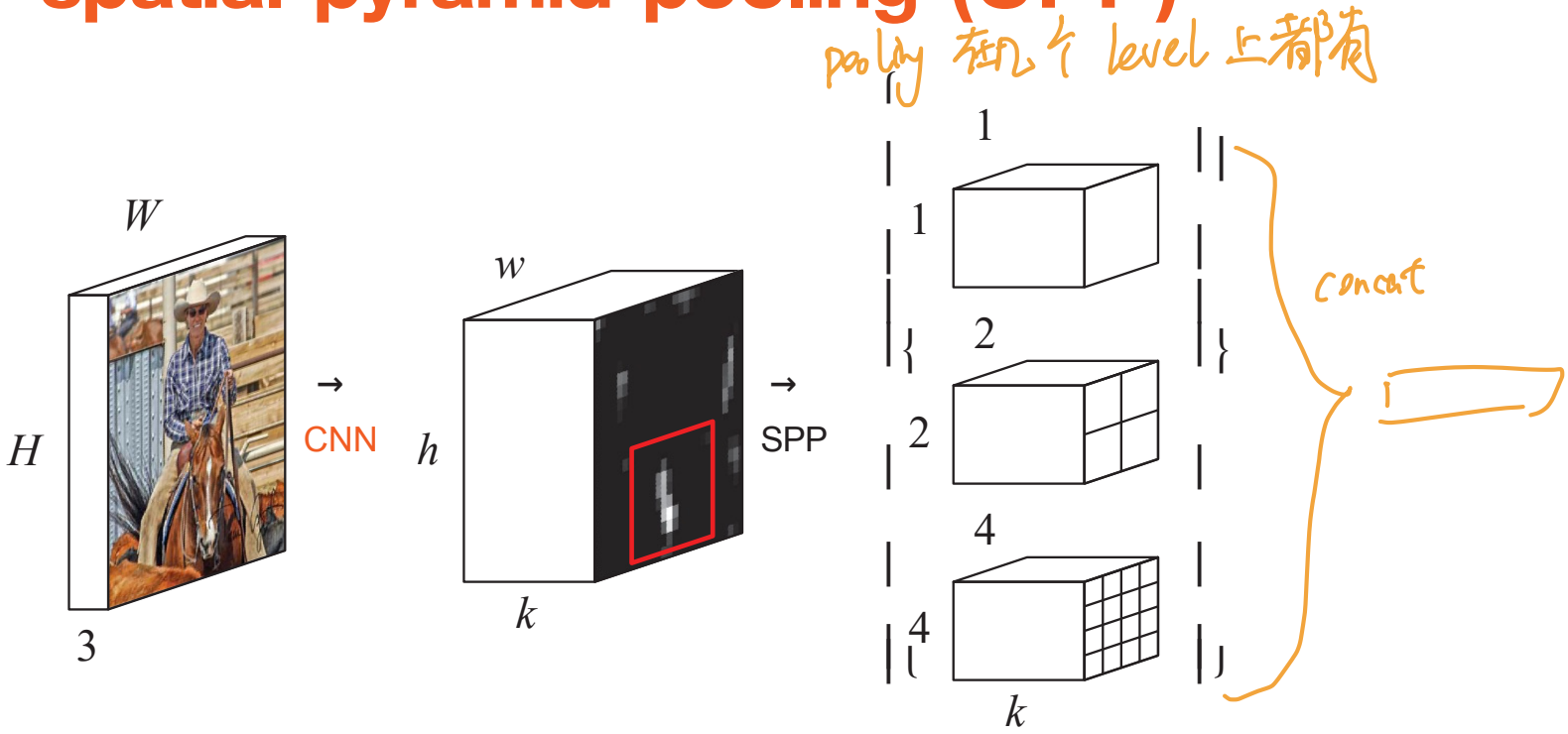


- 3-channel RGB input, fixed width $W = 500$ pixels
- ~ 2000 SS region proposals warped into fixed $w \times h = 227 \times 227$
- each proposal yields a $k = 4096$ dimensional feature by CaffeNet
- each feature is classified into c classes by c one-*vs*. -rest SVMs and localized by bounding box regression

binary

Girshick, Donahue, Darrell and Malik. CVPR 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.

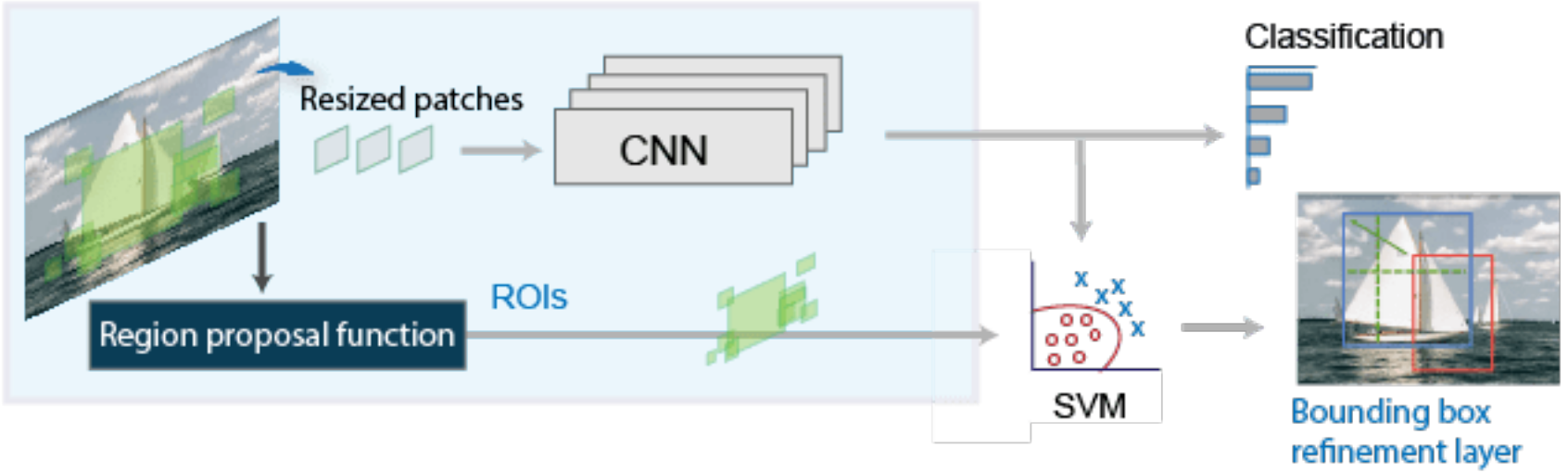
spatial pyramid pooling (SPP)



- 3-channel RGB input, arbitrary size
- input yields a single k dimensional feature map
- each region proposal projected onto feature maps
- then max-pooled into a number of fixed sizes $1 \times 1, 2 \times 2, 4 \times 4$ *etc.* and concatenated into fixed-length representation
- when the pyramid has only one level, we call this RoI pooling

He, Zhang, Ren and Sun. ECCV 2014. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition.

fast R-CNN

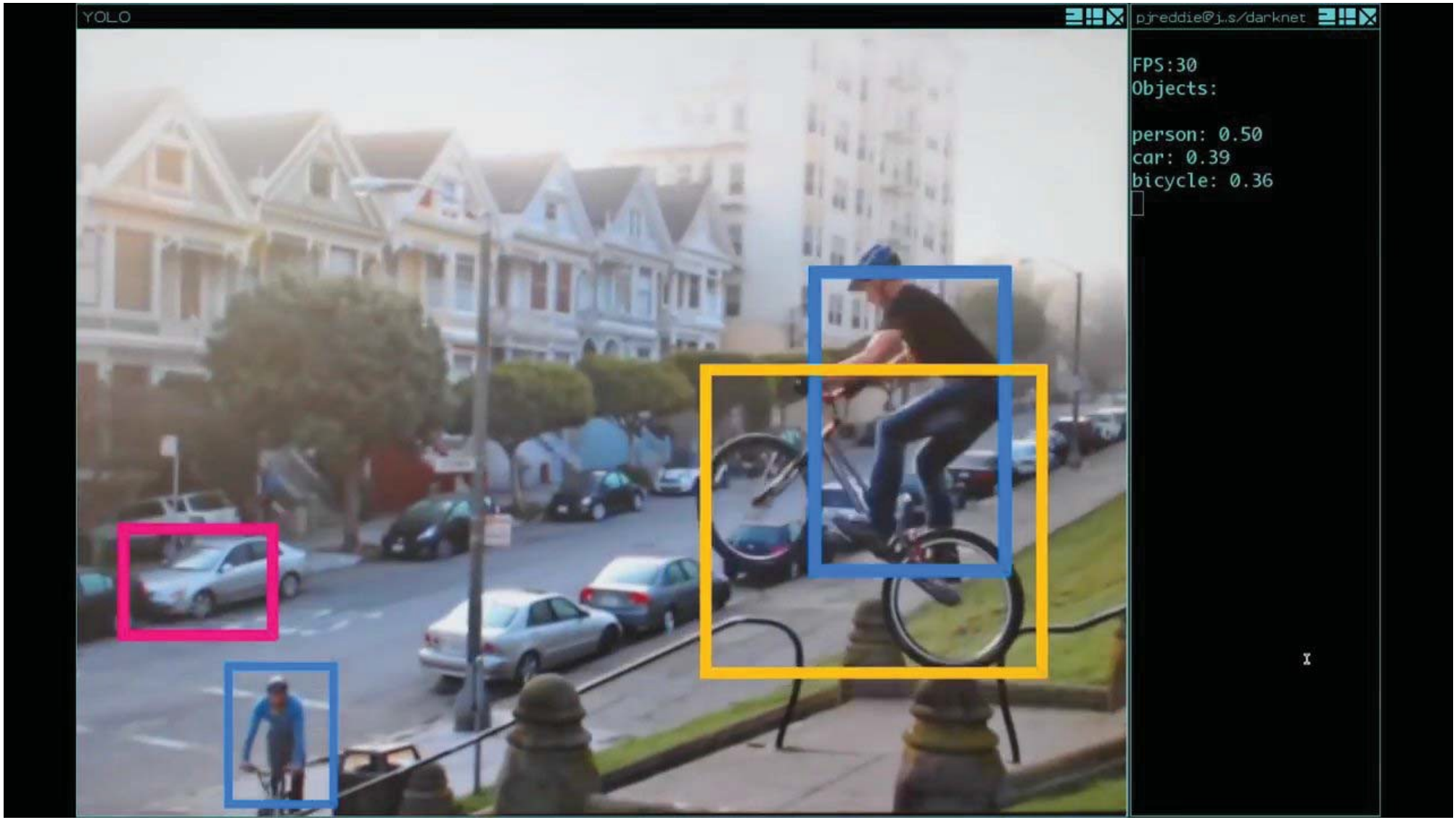

Classification
Resized patches
CNN
Region proposal function
ROIs
SVM
Bounding box
refinement layer

比不上 two-stage → region proposal + 物体分类与位置精修

one-stage detection

“you only look once” (YOLO)

[Redmon et al. 2016]



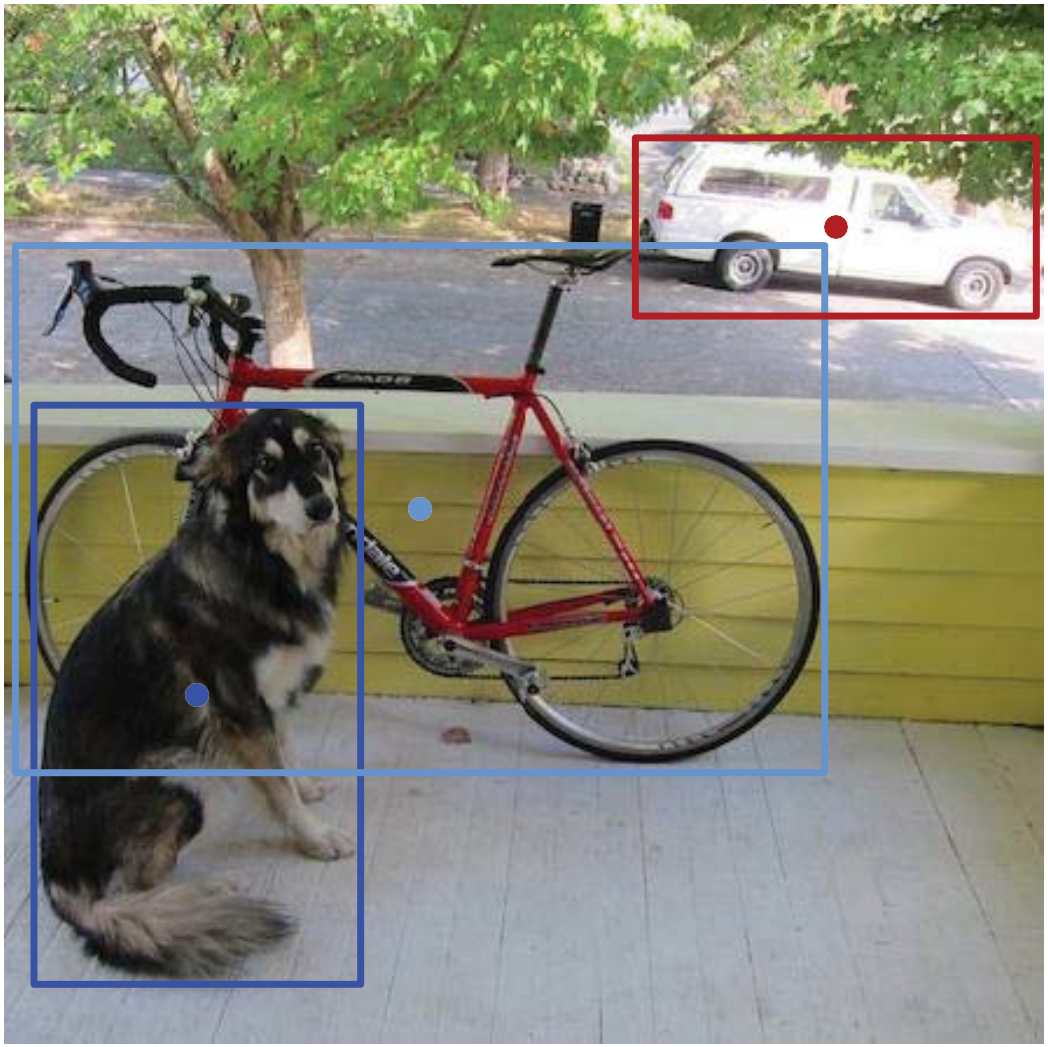
Redmon, Divvala, Girshick and Farhadi. CVPR 2016. You Only Look Once: Unified, Real-Time Object Detection.

“you only look once” (YOLO)

- input image

Redmon, Divvala, Girshick and Farhadi. CVPR 2016. You Only Look Once: Unified, Real-Time Object Detection.

“you only look once” (YOLO)



- groung truth bounding boxes and their centers

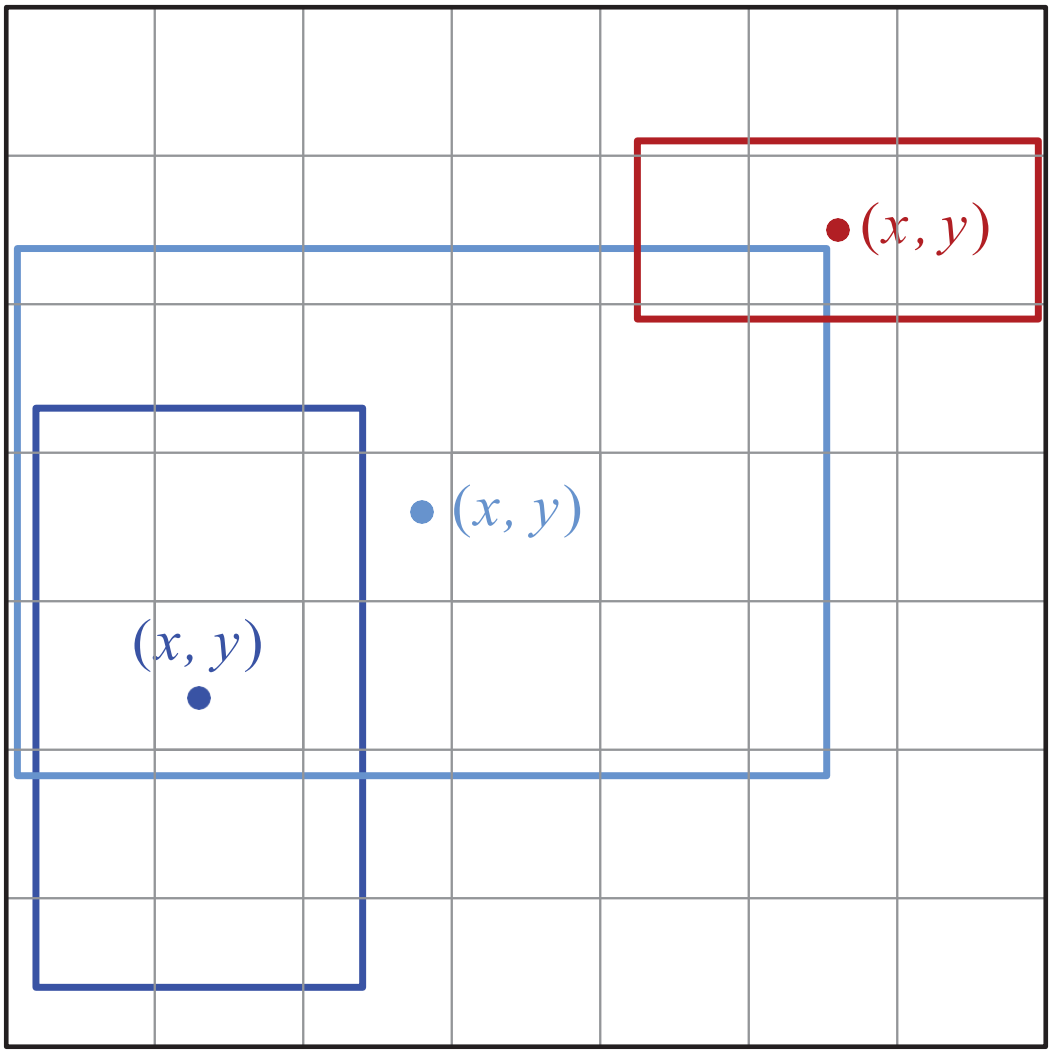
Redmon, Divvala, Girshick and Farhadi. CVPR 2016. You Only Look Once: Unified, Real-Time Object Detection.

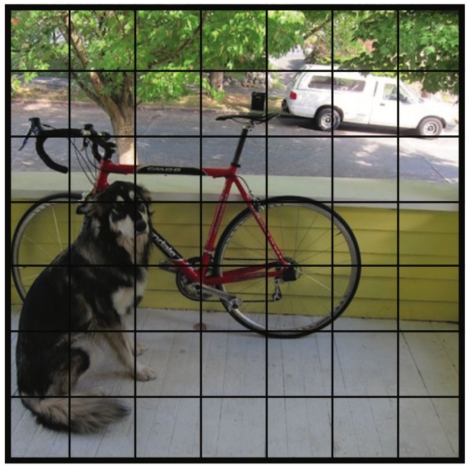
“you only look once” (YOLO)

1. Yolo is fast because it uses a regression approach and does not use complex frameworks.
2. Yolo makes predictions based on the information of the entire image, while other sliding-window detection frameworks can only make inferences based on local image information.

Redmon, Divvala, Girshick and Farhadi. CVPR 2016. You Only Look Once: Unified, Real-Time Object Detection.

“you only look once” (YOLO)



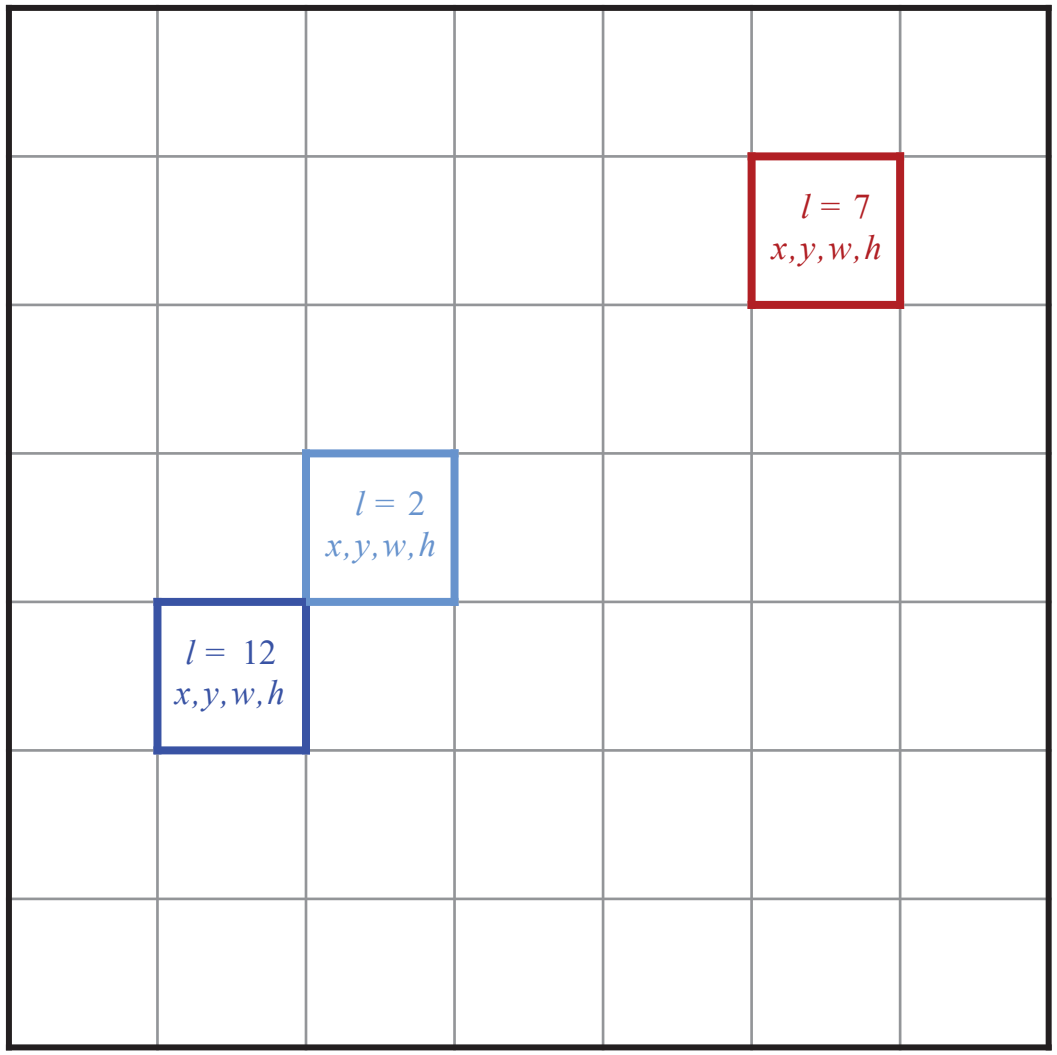


S × S grid on input

- image partitioned into 7×7 grid and center coordinates assigned to cells

Redmon, Divvala, Girshick and Farhadi. CVPR 2016. You Only Look Once: Unified, Real-Time Object Detection.

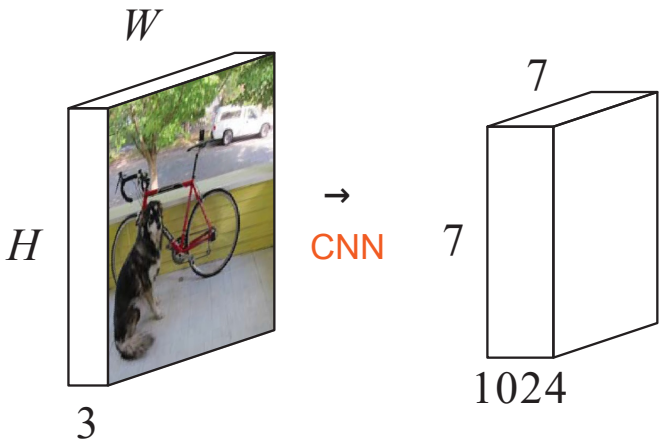
“you only look once” (YOLO)



- network learns to predict up to one object per cell, including class label l, center coordinates x, y and bounding box size w, h

Redmon, Divvala, Girshick and Farhadi. CVPR 2016. You Only Look Once: Unified, Real-Time Object Detection.

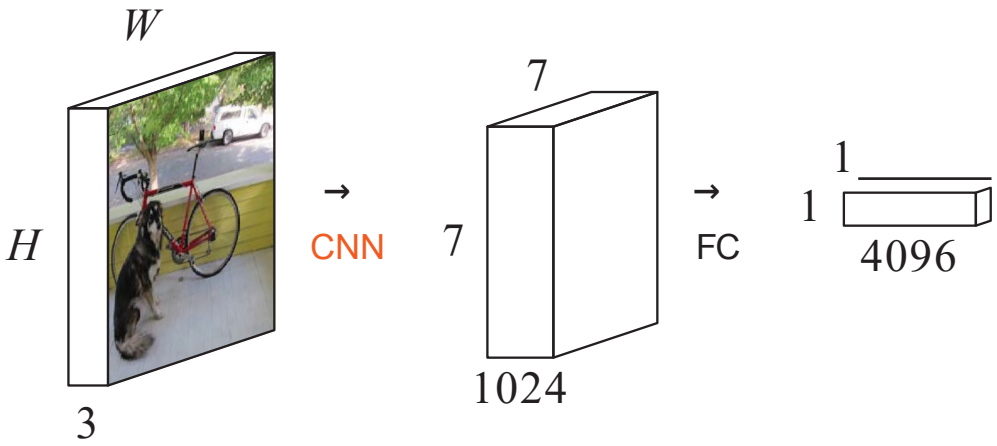
“you only look once” (YOLO)



- 3-channel input $W = H = 448$, 24-layer NiN-like network
- fully connected layer, increasing to 4096 features
- $c = 20$ class scores and 4 bounding box coordinates at each position
- in a single stage, network performs regression from the image to a $7 \times 7 \times 24$ tensor encoding detected classes and positions
- regression (f_2) loss on both class scores and coordinates
- “objectness” score makes it look like two-stage

Redmon, Divvala, Girshick and Farhadi. CVPR 2016. You Only Look Once: Unified, Real-Time Object Detection.

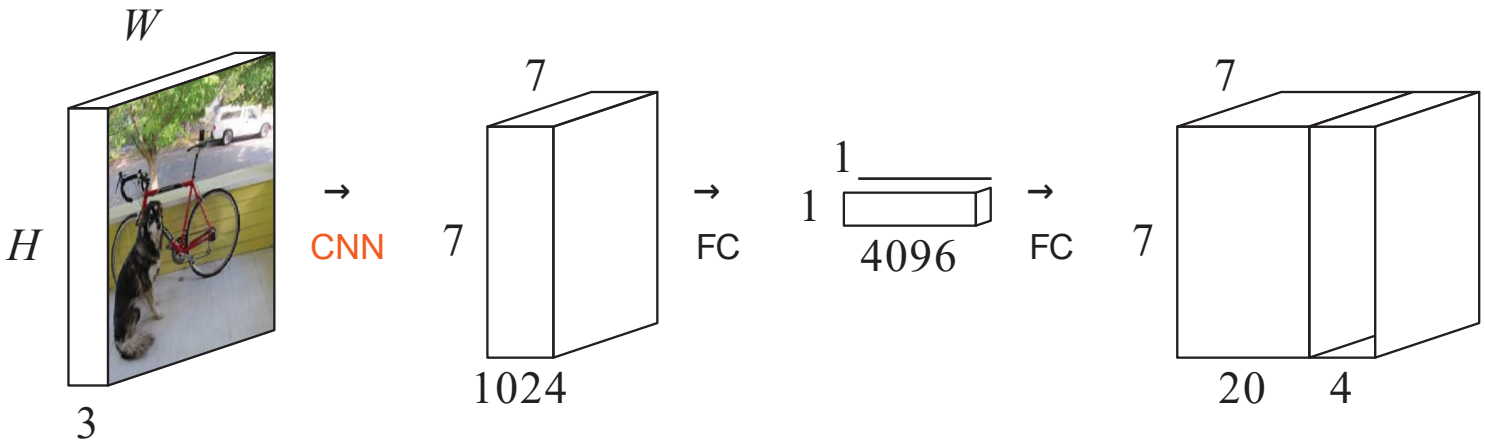
“you only look once” (YOLO)



- 3-channel input $W = H = 448$, 24-layer NiN-like network
- fully connected layer, increasing to 4096 features
- $c = 20$ class scores and 4 bounding box coordinates at each position
- in a single stage, network performs regression from the image to a $7 \times 7 \times 24$ tensor encoding detected classes and positions
- regression (f_2) loss on both class scores and coordinates
- “objectness” score makes it look like two-stage

Redmon, Divvala, Girshick and Farhadi. CVPR 2016. You Only Look Once: Unified, Real-Time Object Detection.

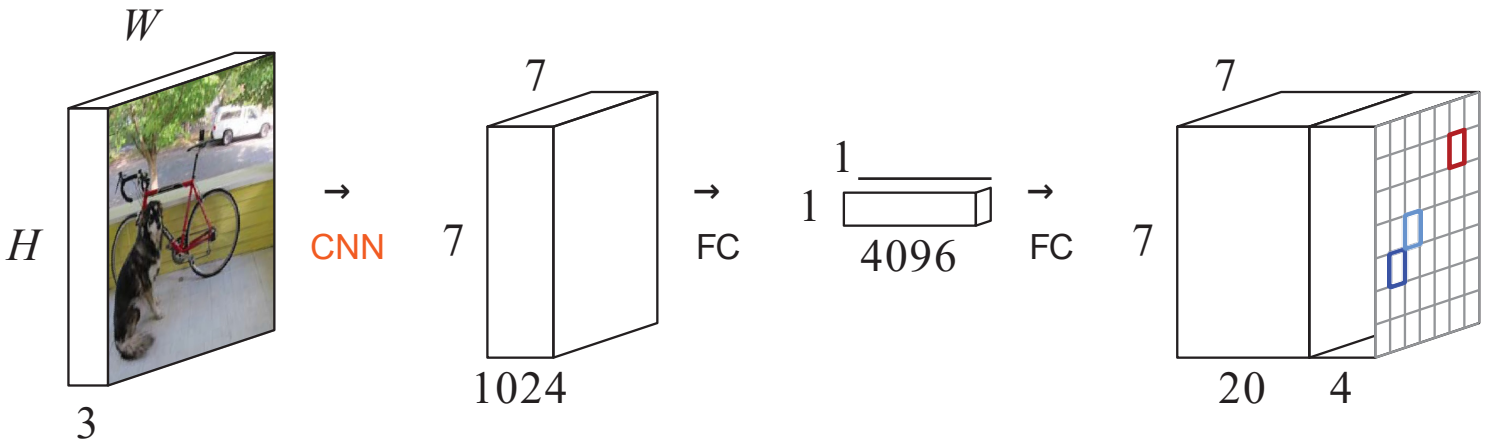
“you only look once” (YOLO)



- 3-channel input $W = H = 448$, 24-layer NiN-like network
- fully connected layer, increasing to 4096 features
- $c = 20$ class scores and 4 bounding box coordinates at each position
- in a single stage, network performs regression from the image to a $7 \times 7 \times 24$ tensor encoding detected classes and positions
- regression (f_2) loss on both class scores and coordinates
- “objectness” score makes it look like two-stage

Redmon, Divvala, Girshick and Farhadi. CVPR 2016. You Only Look Once: Unified, Real-Time Object Detection.

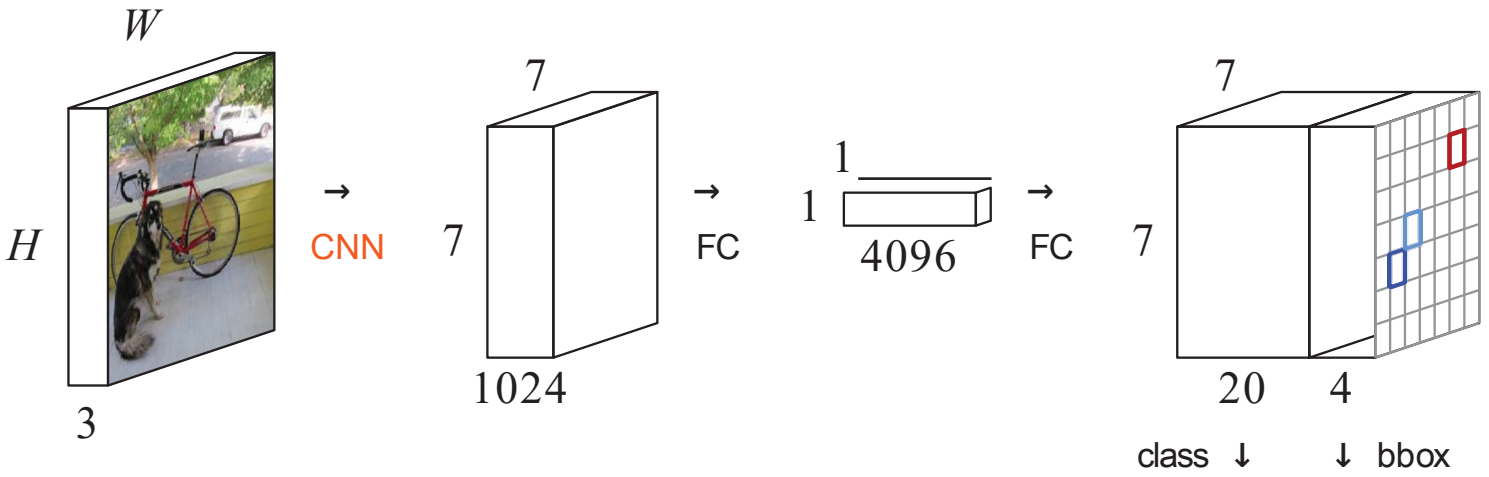
“you only look once” (YOLO)



- 3-channel input $W = H = 448$, 24-layer NiN-like network
- fully connected layer, increasing to 4096 features
- $c = 20$ class scores and 4 bounding box coordinates at each position
- in a single stage, network performs regression from the image to a $7 \times 7 \times 24$ tensor encoding detected classes and positions
- regression (f_2) loss on both class scores and coordinates
- “objectness” score makes it look like two-stage

Redmon, Divvala, Girshick and Farhadi. CVPR 2016. You Only Look Once: Unified, Real-Time Object Detection.

“you only look once” (YOLO)



- 3-channel input $W = H = 448$, 24-layer NiN-like network
- fully connected layer, increasing to 4096 features
- $c = 20$ class scores and 4 bounding box coordinates at each position
- in a single stage, network performs regression from the image to a $7 \times 7 \times 24$ tensor encoding detected classes and positions
- regression (f_2) loss on both class scores and coordinates
- “objectness” score makes it look like two-stage

Redmon, Divvala, Girshick and Farhadi. CVPR 2016. You Only Look Once: Unified, Real-Time Object Detection.

“you only look once” (YOLO)

pros

- extremely fast: 45fps; 93× to 500× test speedup *vs.* R-CNN on AlexNet, with similar performance
- end-to-end trainable, fully convolutional, one-stage detection

cons

- only up to one prediction per cell (fixed in later versions)
- trouble localizing small objects
- low-performance compared to two-stage detectors on strong networks

Redmon, Divvala, Girshick and Farhadi. CVPR 2016. You Only Look Once: Unified, Real-Time Object Detection.

“you only look once” (YOLO)

pros

- extremely fast: 45fps; 93× to 500× test speedup *vs.* R-CNN on AlexNet, with similar performance
- end-to-end trainable, fully convolutional, one-stage detection

cons

- only up to one prediction per cell (fixed in later versions)
- trouble localizing small objects
- low-performance compared to two-stage detectors on strong networks

Redmon, Divvala, Girshick and Farhadi. CVPR 2016. You Only Look Once: Unified, Real-Time Object Detection.