YOLO Review

Cho Sung Man

YOLO ver.1 ?
(You Only Look Once)

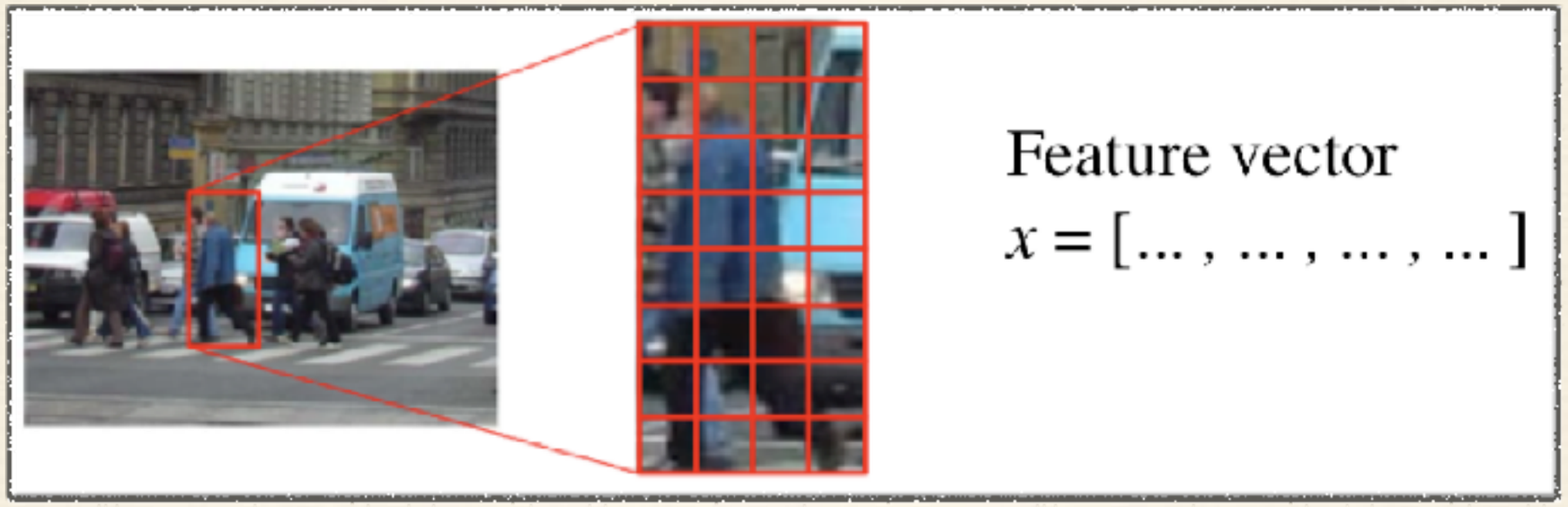
- Frame object detection as a regression problem
- Single neural network predict
- FPS : 45
- More localization error
- Less false positive on background

Introduction

- DPM
- R-CNN
- Unified model . YOLO.

DPM
Deformable Parts Model

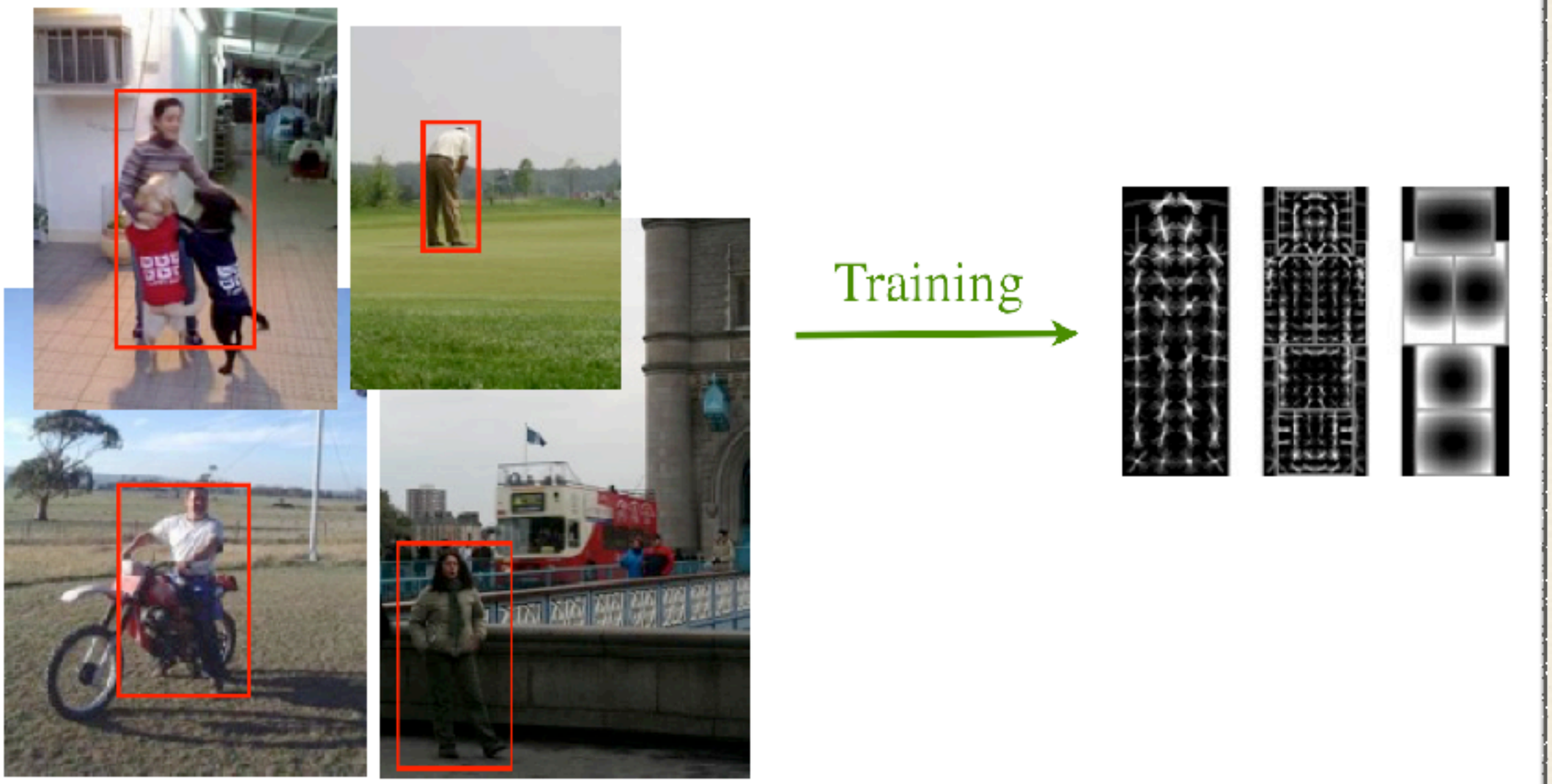
- Starting point : sliding window classifiers



- Detect objects by testing each sub window
  - Reduces object detection to binary classification
  - Dalal & Triggs : HOG features + linear SVM

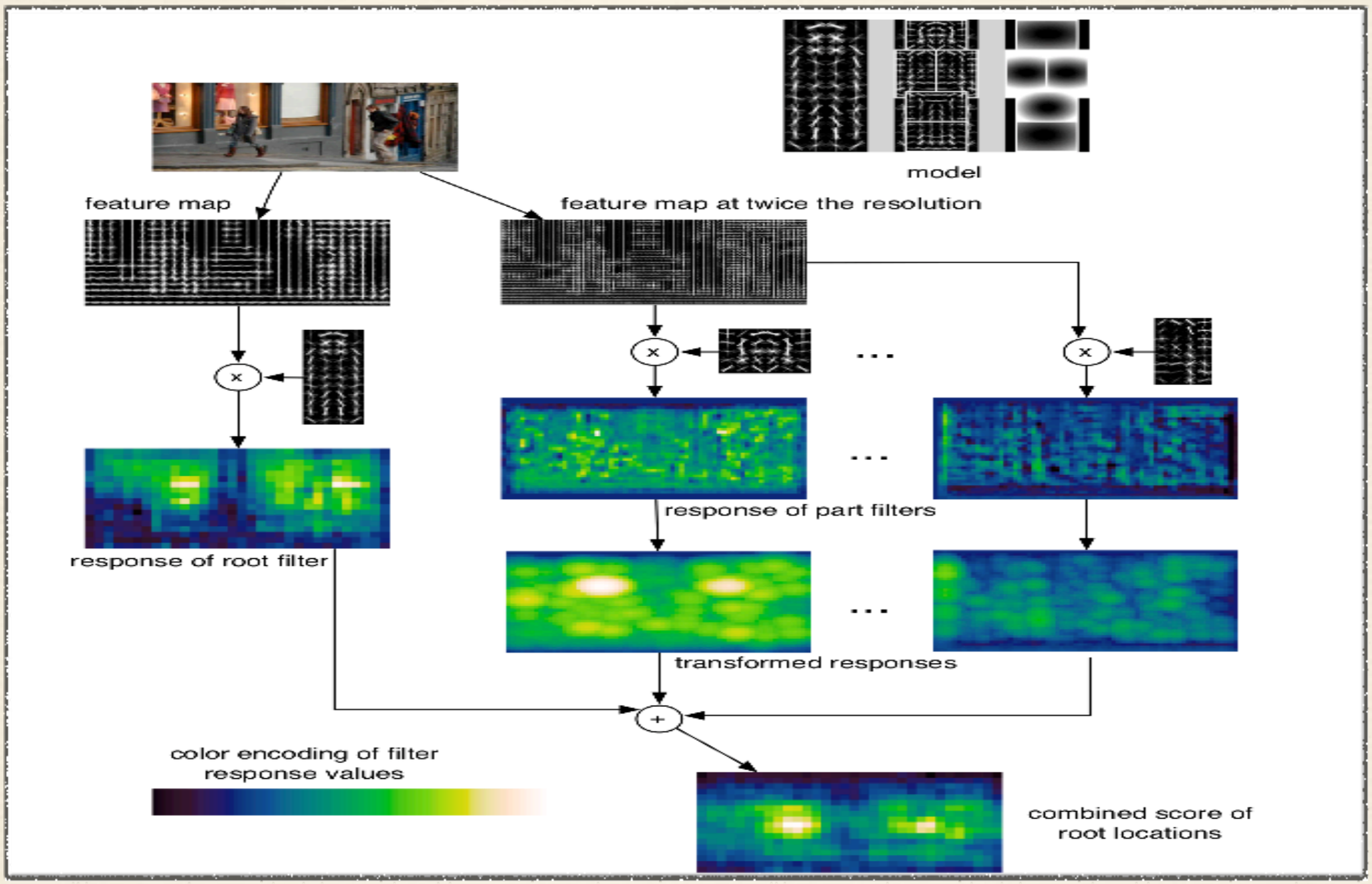
DPM

Deformable Parts Model



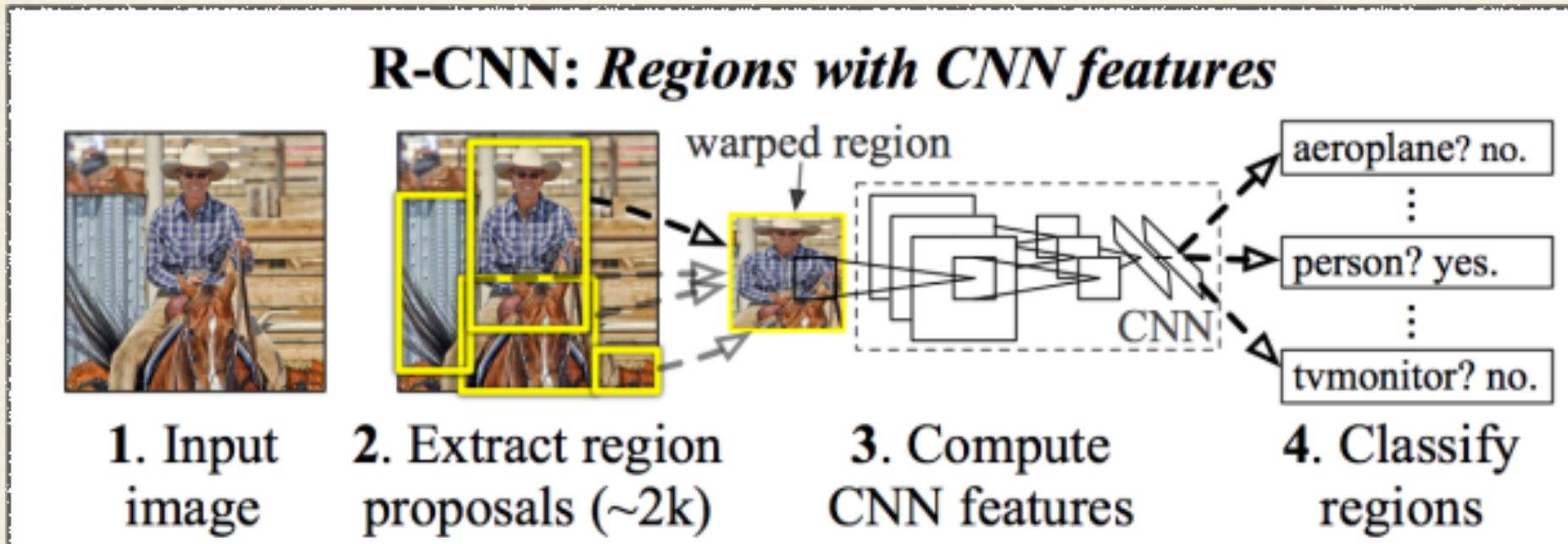
DPM

Deformable Parts Model



R-CNN

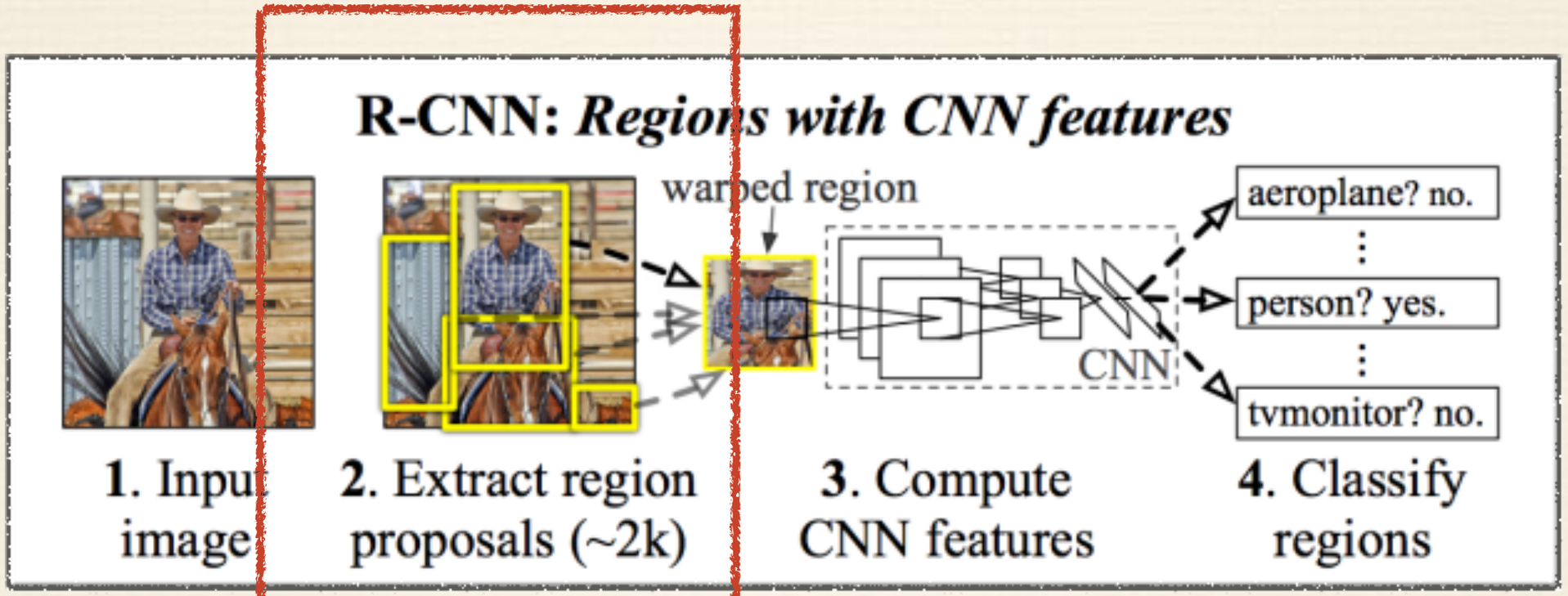
Region Proposals + CNN



	localization	feature extraction	classification
this paper:	selective search	deep learning CNN	binary linear SVM
alternatives:	objectness, constrained parametric min-cuts, sliding window ...	HOG, SIFT, LBP, BoW, DPM ...	SVM, Neural networks, Logistic regression ...

R-CNN

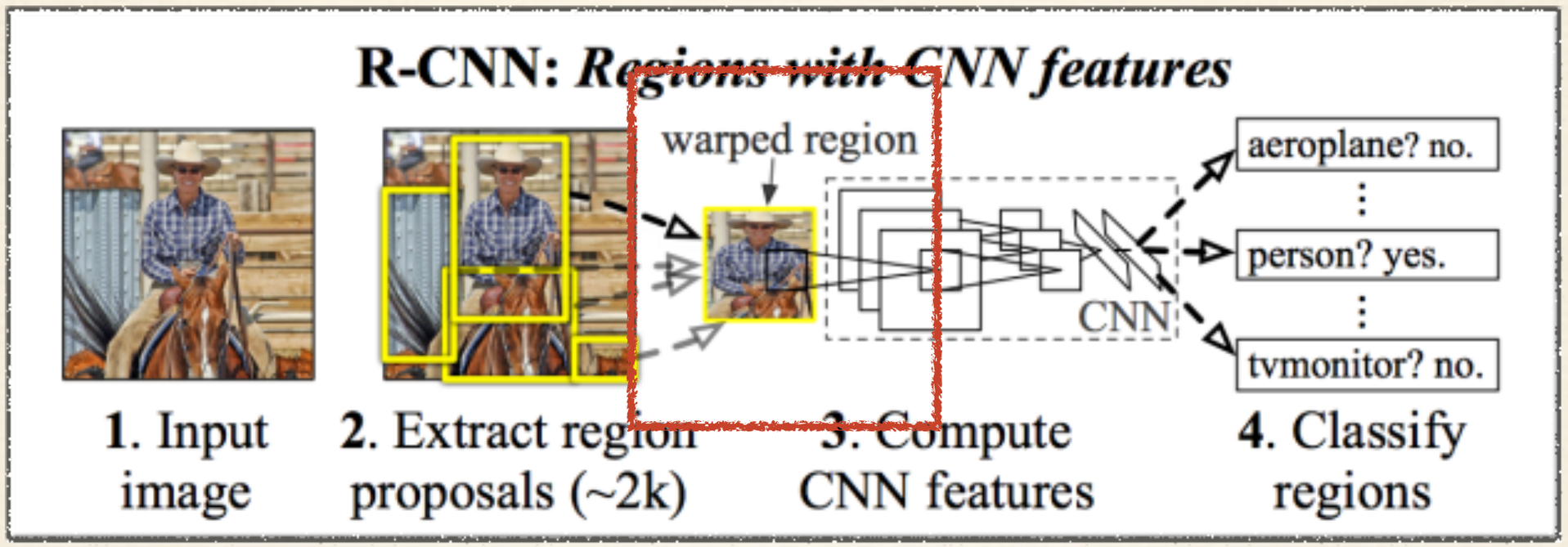
Region Proposal



- Sliding window + CNN = High computational cost
- Selective search !

R-CNN

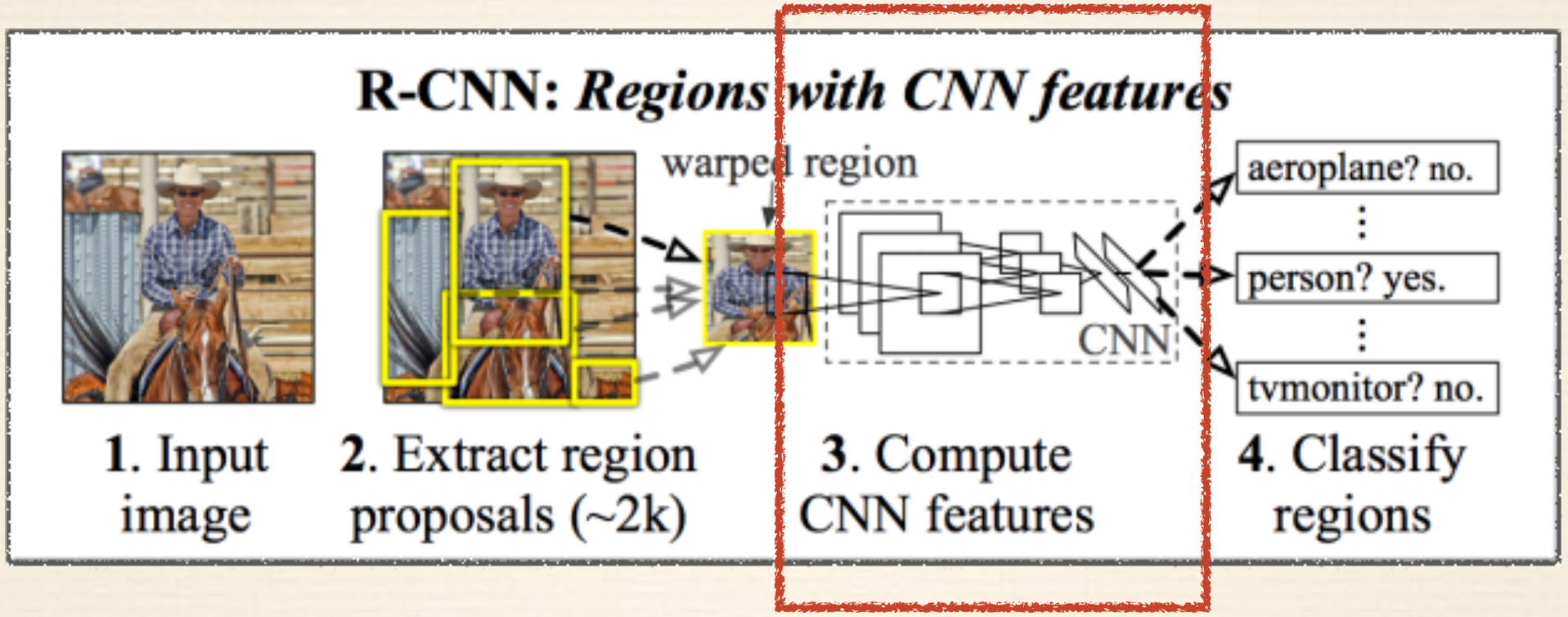
Region Warping



- Regardless of Size and aspect ratio
- Warp to 224x224 patch

R-CNN

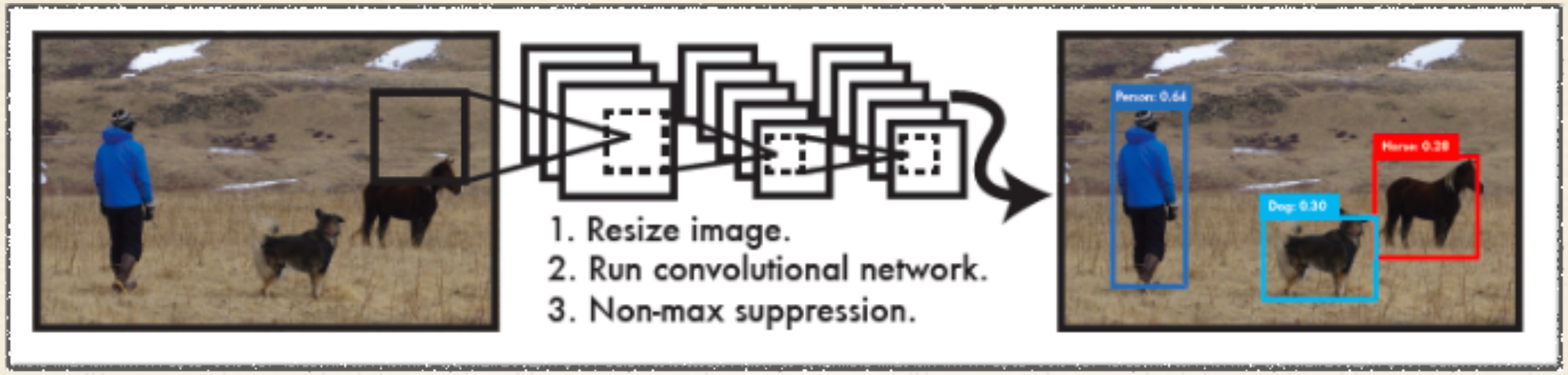
Feature Extraction



- 4096-dimensional feature vector
- Their own implementation of the CNN

YOLO

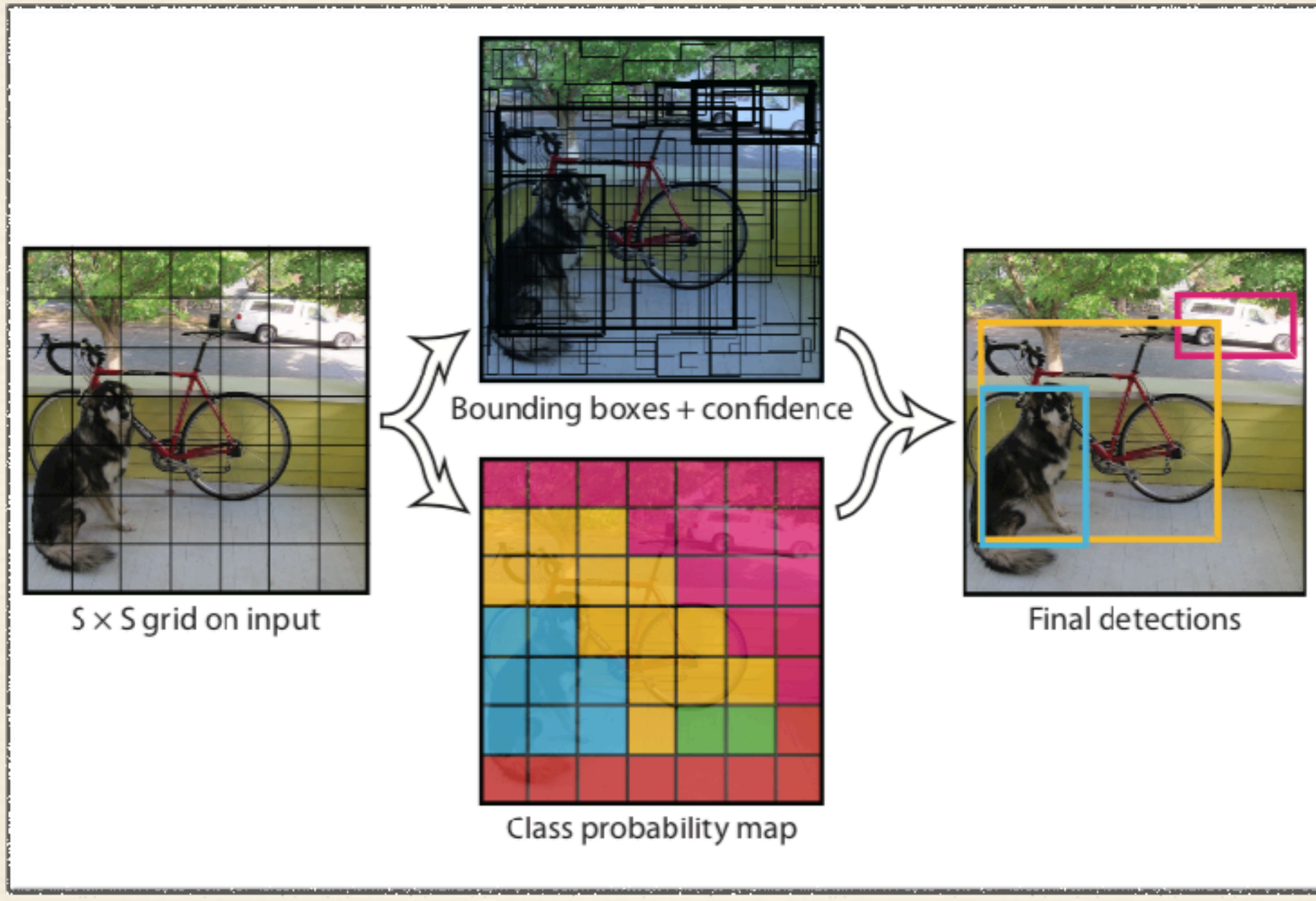
Pipe line



- Resize the input image to 448x448
- Run a single convolutional network on the image
- Threshold the resulting detections

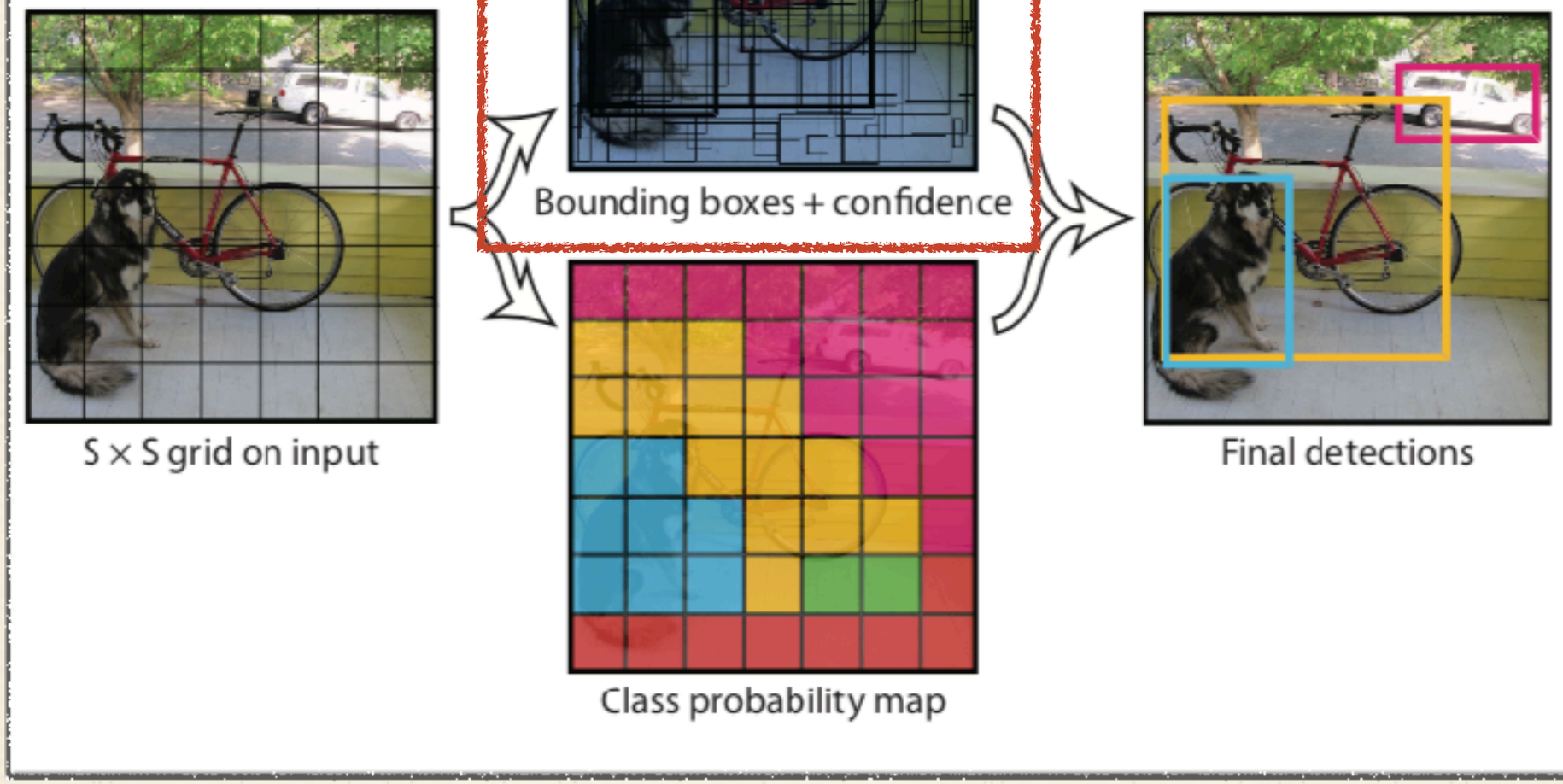
YOLO

Unified Detection



YOLO

Unified Detection



YOLO

Bounding Box + Confidence

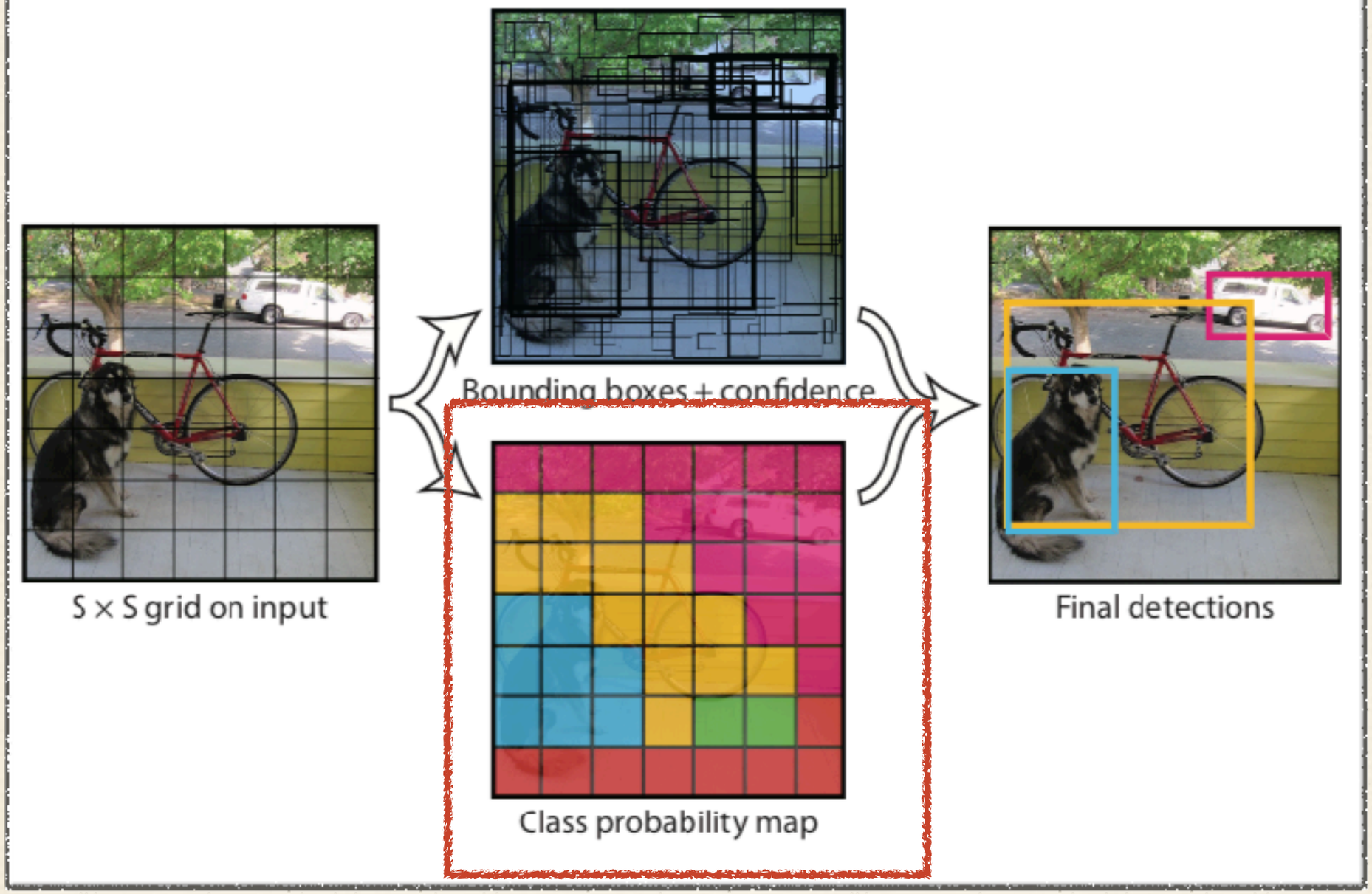
- Confidence = Pr(Object) * IOU
- Each bounding box consists of 5 Predictions

$$x,\ y,\ w,\ h,\ conf$$

x, y *: coordinates represent the center of the box relative to the bounds of the grid cell.*
width, height *: predicted relative to the whole image*
confidence *: prediction represents the IOU between the predicted box and any ground truth box*

YOLO

Unified Detection



YOLO

Class Probability map

- Each grid cell predicts C conditional class prob.

$$\Pr(\text{Class} \mid \text{Object})$$

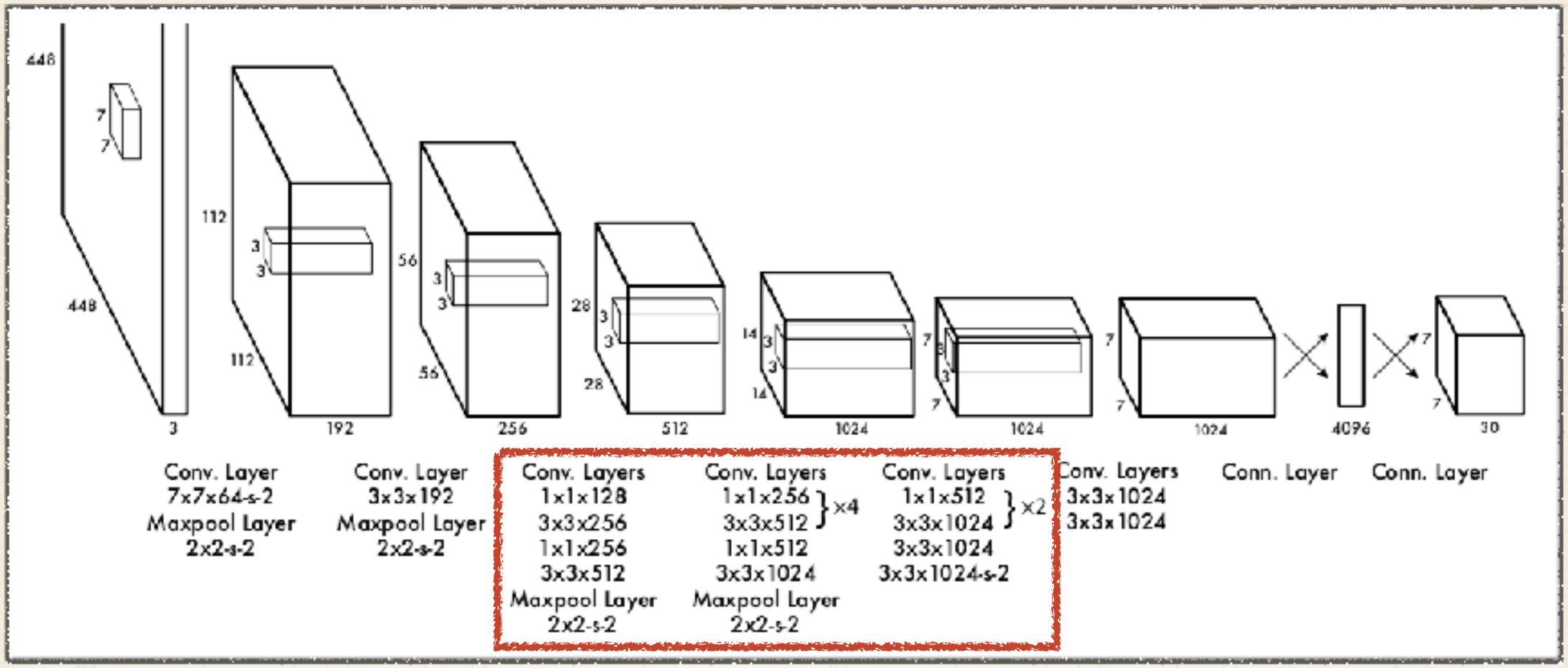
YOLO

The Model

- Image is divided into $S \times S$ grid.
- Each grid predicts B bounding boxes.
  - C class probabilities
- $S \times S \times (B \times 5 + C)$ tensor !

YOLO

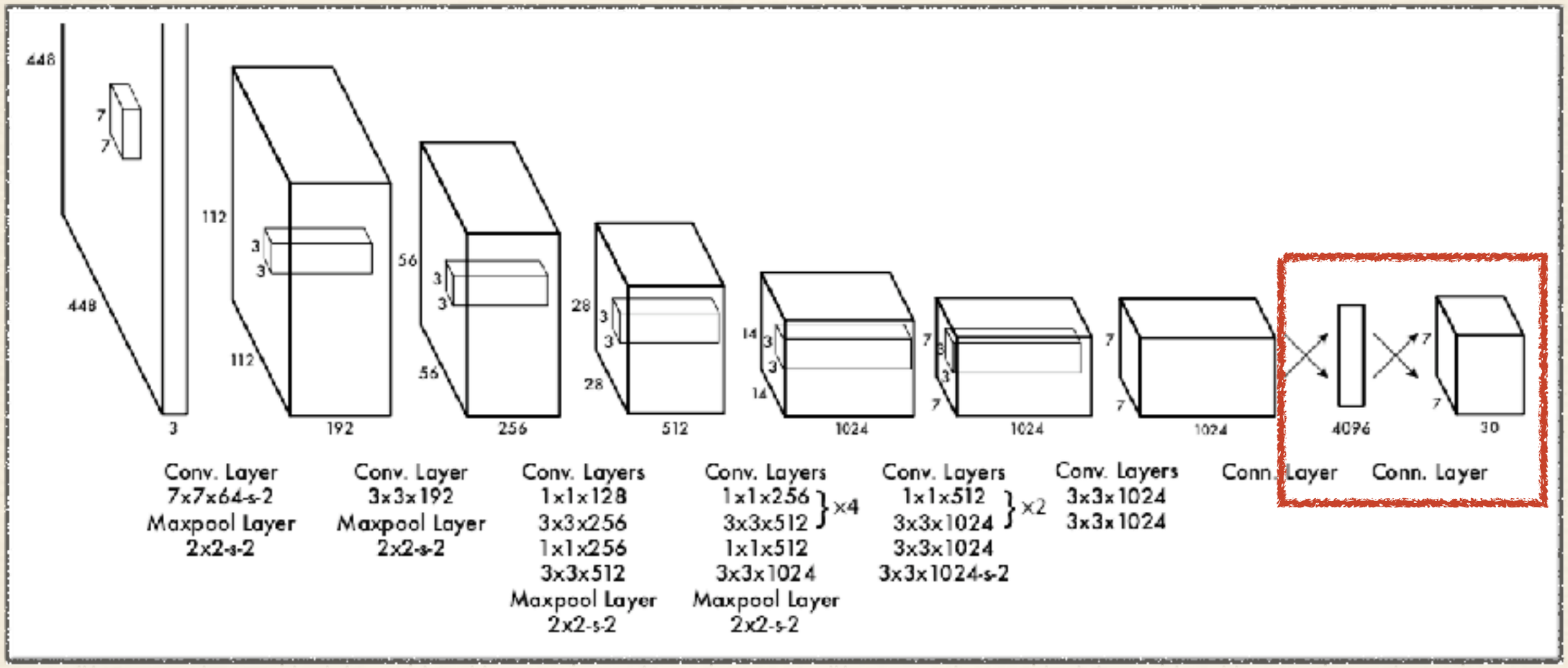
Network Architecture



- Inspired by the GoogLeNet model
- Instead of inception module, use **1x1 reduction layers**

YOLO

Network Architecture



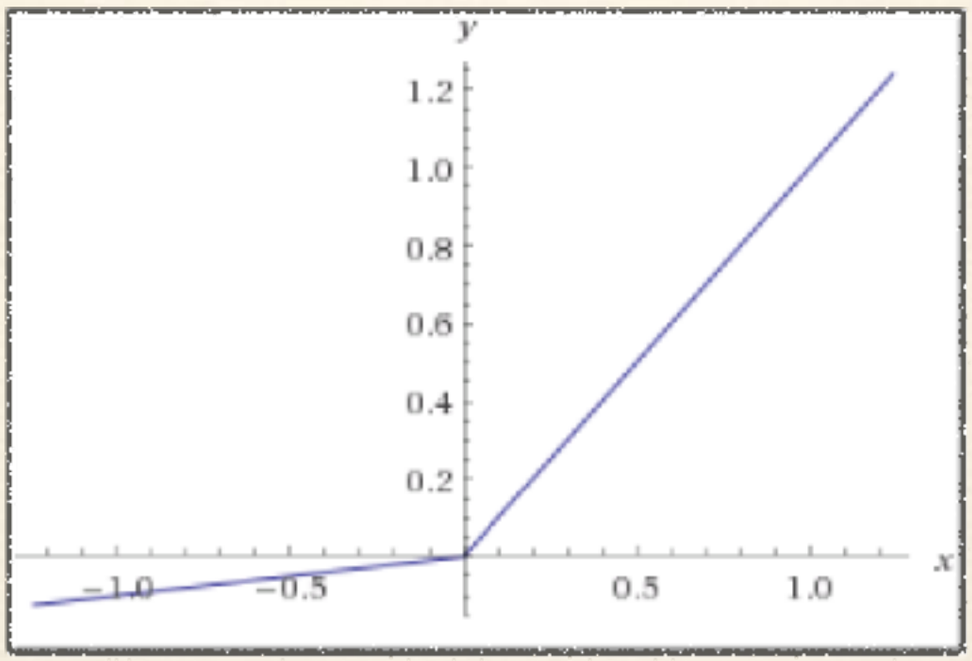
- 24 convolutional layers, 2 fully connected layers
- Final output 7 x 7 x 30 tensor

YOLO
Training

$$\sum_{i=0}^{48}\left(\lambda \mathbb{1}_i^{\text{obj}}\left((x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2+(\sqrt{w_i}-\sqrt{\hat{w}_i})^2+(\sqrt{h_i}-\sqrt{\hat{h}_i})^2\right)+\sum_{c\in\text{classes}}(p_i(c)-\hat{p}_i(c))^2\right)$$

- Loss function : sum squared error.

- Activation function : leaky ReLU



YOLO

Training - Loss function

$$\sum_{i=0}^{48}\left(\lambda\mathbb{1}_i^{\text{obj}}\left((x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2+(\sqrt{w_i}-\sqrt{\hat{w}_i})^2+(\sqrt{h_i}-\sqrt{\hat{h}_i})^2\right)+\sum_{c\in\text{classes}}(p_i(c)-\hat{p}_i(c))^2\right)$$

Grid

Object

Class Probability

- 0~ 49 Grid / Object $\{x, y, w, h\}$ / Class Prob.
- w, h - square root
- Class Prob - L2 norm.
- λ - weight between two terms.

YOLO

Training - Loss function

$$\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right]$$

$$+ \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} \left[(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 \right]$$

$$+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} (C_i - \hat{C}_i)^2$$

$$+ \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{noobj} (C_i - \hat{C}_i)^2$$

$$+ \sum_{i=0}^{S^2} \mathbb{1}_{i}^{obj} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2 \qquad (3)$$

$\lambda_{coord} = 5$ $\lambda_{noobj} = 0.5$

YOLO

Parameters

- *batch size : 64*
- *momentum : 0.9*
- *decay : 0.0005*
- *learning rate :* *slowly raise from 10^{-3} to 10^{-2}*
- *drop out layer rate : 0.5*
- *data augmentation : scaling and translating 20% of images*
- *adjust exposure and saturation of images*

YOLO

Result

VOC 2012 test	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
MR_CNN_MORE_DATA [11]	**73.9**	**85.5**	**82.9**	**76.6**	**57.8**	**62.7**	**79.4**	77.2	86.6	**55.0**	**79.1**	**62.2**	87.0	**83.4**	**84.7**	78.9	45.3	73.4	65.8	80.3	74.0
HyperNet_VGG	71.4	84.2	78.5	73.6	55.6	53.7	78.7	**79.8**	87.7	49.6	74.9	52.1	86.0	81.7	83.3	**81.8**	**48.6**	**73.5**	59.4	79.9	65.7
HyperNet_SP	71.3	84.1	78.3	73.3	55.5	53.6	78.6	79.6	87.5	49.5	74.9	52.1	85.6	81.6	83.2	81.6	48.4	73.2	59.3	79.7	65.6
Fast R-CNN + YOLO	70.7	83.4	78.5	73.5	55.8	43.4	79.1	73.1	**89.4**	49.4	75.5	57.0	**87.5**	80.9	81.0	74.7	41.8	71.5	68.5	**82.1**	67.2
MR_CNN_S_CNN [11]	70.7	85.0	79.6	71.5	55.3	57.7	76.0	73.9	84.6	50.5	74.3	61.7	85.5	79.9	81.7	76.4	41.0	69.0	61.2	77.7	72.1
Faster R-CNN [28]	70.4	84.9	79.8	74.3	53.9	49.8	77.5	75.9	88.5	45.6	77.1	55.3	86.9	81.7	80.9	79.6	40.1	72.6	60.9	81.2	61.5
DEEP_ENS_COCO	70.1	84.0	79.4	71.6	51.9	51.1	74.1	72.1	88.6	48.3	73.4	57.8	86.1	80.0	80.7	70.4	46.6	69.6	**68.8**	75.9	71.4
NoC [29]	68.8	82.8	79.0	71.6	52.3	53.7	74.1	69.0	84.9	46.9	74.3	53.1	85.0	81.3	79.5	72.2	38.9	72.4	59.5	76.7	68.1
Fast R-CNN [14]	68.4	82.3	78.4	70.8	52.3	38.7	77.8	71.6	89.3	44.2	73.0	55.0	**87.5**	80.5	80.8	72.0	35.1	68.3	65.7	80.4	64.2
UMICH_FGS_STRUCT	66.4	82.9	76.1	64.1	44.6	49.4	70.3	71.2	84.6	42.7	68.6	55.8	82.7	77.1	79.9	68.7	41.4	69.0	60.0	72.0	66.2
NUS_NIN_C2000 [7]	63.8	80.2	73.8	61.9	43.7	43.0	70.3	67.6	80.7	41.9	69.7	51.7	78.2	75.2	76.9	65.1	38.6	68.3	58.0	68.7	63.3
BabyLearning [7]	63.2	78.0	74.2	61.3	45.7	42.7	68.2	66.8	80.2	40.6	70.0	49.8	79.0	74.5	77.9	64.0	35.3	67.9	55.7	68.7	62.6
NUS_NIN	62.4	77.9	73.1	62.6	39.5	43.3	69.1	66.4	78.9	39.1	68.1	50.0	77.2	71.3	76.1	64.7	38.4	66.9	56.2	66.9	62.7
R-CNN VGG BB [13]	62.4	79.6	72.7	61.9	41.2	41.9	65.9	66.4	84.6	38.5	67.2	46.7	82.0	74.8	76.0	65.2	35.6	65.4	54.2	67.4	60.3
R-CNN VGG [13]	59.2	76.8	70.9	56.6	37.5	36.9	62.9	63.6	81.1	35.7	64.3	43.9	80.4	71.6	74.0	60.0	30.8	63.4	52.0	63.5	58.7
YOLO	57.9	77.0	67.2	57.7	38.3	22.7	68.3	55.9	81.4	36.2	60.8	48.5	77.2	72.3	71.3	63.5	28.9	52.2	54.8	73.9	50.8
Feature Edit [33]	56.3	74.6	69.1	54.4	39.1	33.1	65.2	62.7	69.7	30.8	56.0	44.6	70.0	64.4	71.1	60.2	33.3	61.3	46.4	61.7	57.8
R-CNN BB [13]	53.3	71.8	65.8	52.0	34.1	32.6	59.6	60.0	69.8	27.6	52.0	41.7	69.6	61.3	68.3	57.8	29.6	57.8	40.9	59.3	54.1
SDS [16]	50.7	69.7	58.4	48.5	28.3	28.8	61.3	57.5	70.8	24.1	50.7	35.9	64.9	59.1	65.8	57.1	26.0	58.8	38.6	58.9	50.7
R-CNN [13]	49.6	68.1	63.8	46.1	29.4	27.9	56.6	57.0	65.9	26.5	48.7	39.5	66.2	57.3	65.4	53.2	26.2	54.5	38.1	50.6	51.6

Thank You.

Q&A