Object Recognition with Neuronal Network (Deep Learning)

Daniel Saavedra Morales
Ingeniería Civil Matemática UC

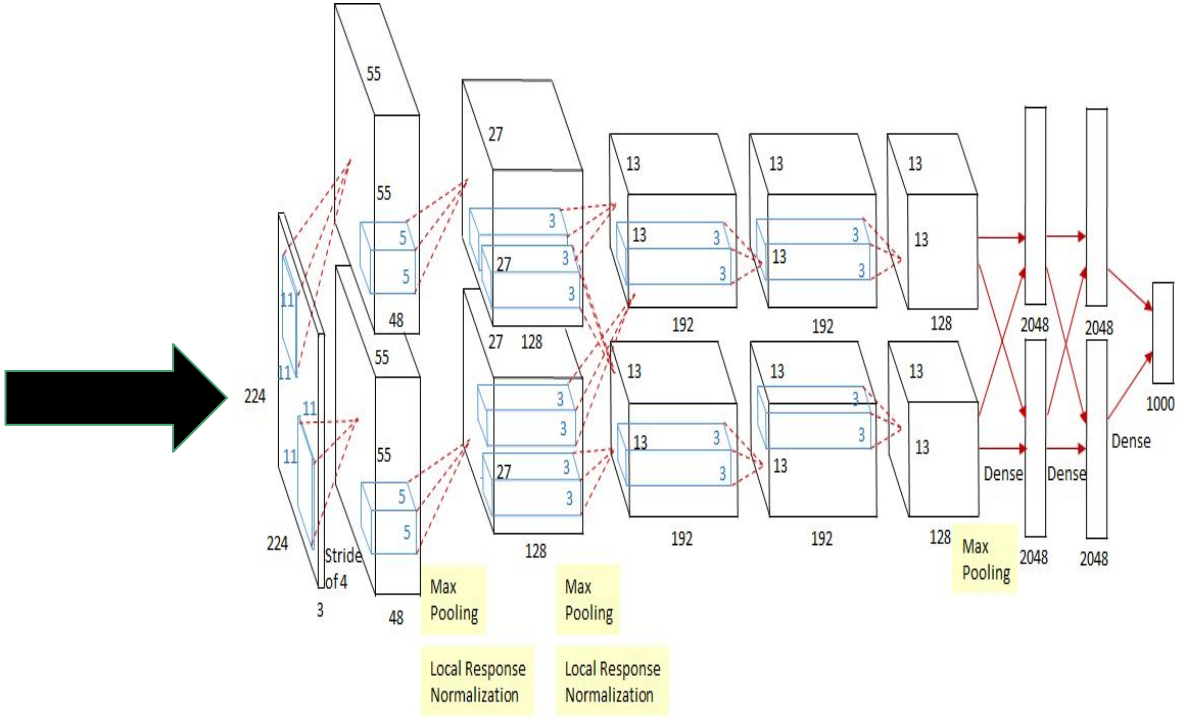
Agenda

1. Introducción
2. RCNN (Girshick, 2013)
3. ★Faster-RCNN (Ren, 2015)
4. SWCNN - Overfeat (Sermanet, 2013)
5. ★YOLO (Redmon, 2015)
6. Variantes
    a. ★SSD (Liu, 2015)
    b. Retinanet (Lin, 2017)
    c. YOLOv3 (Redmon, 2018)
7. Trabajos Futuros

Introducción

Introducción

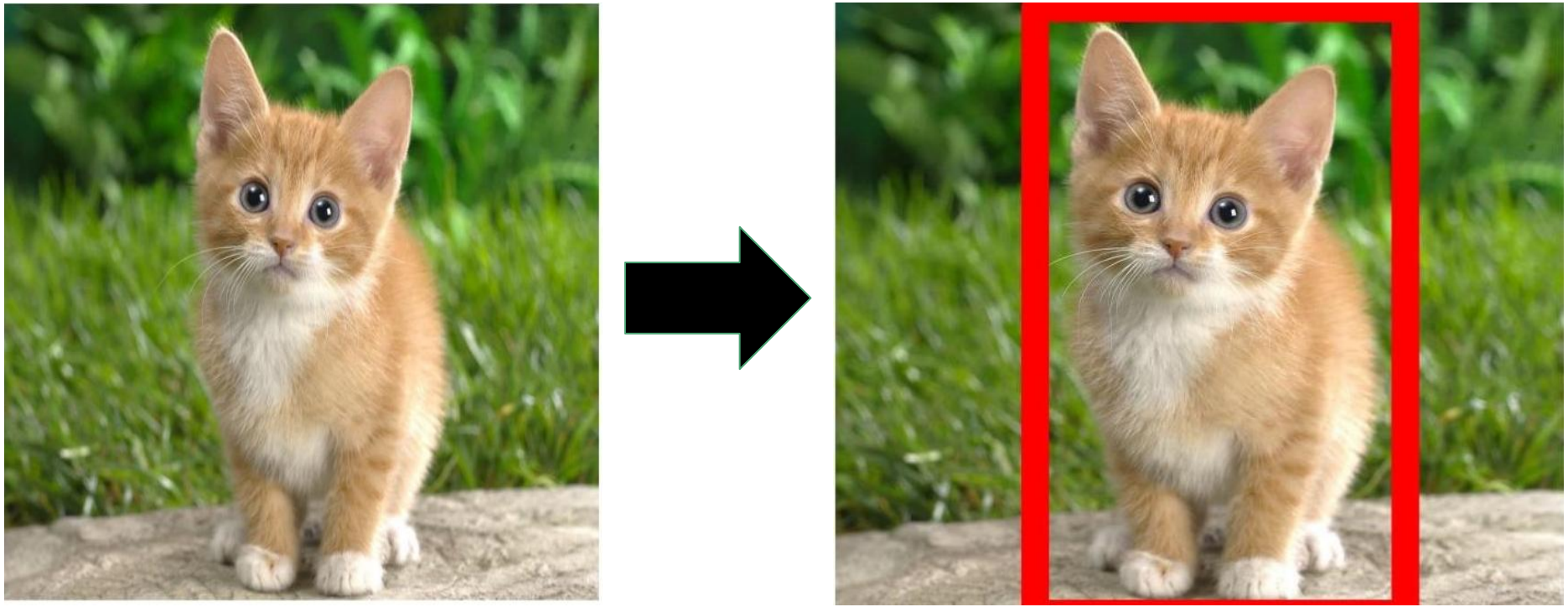
Clasificación



Alexnet (Krizhevsky,2012)

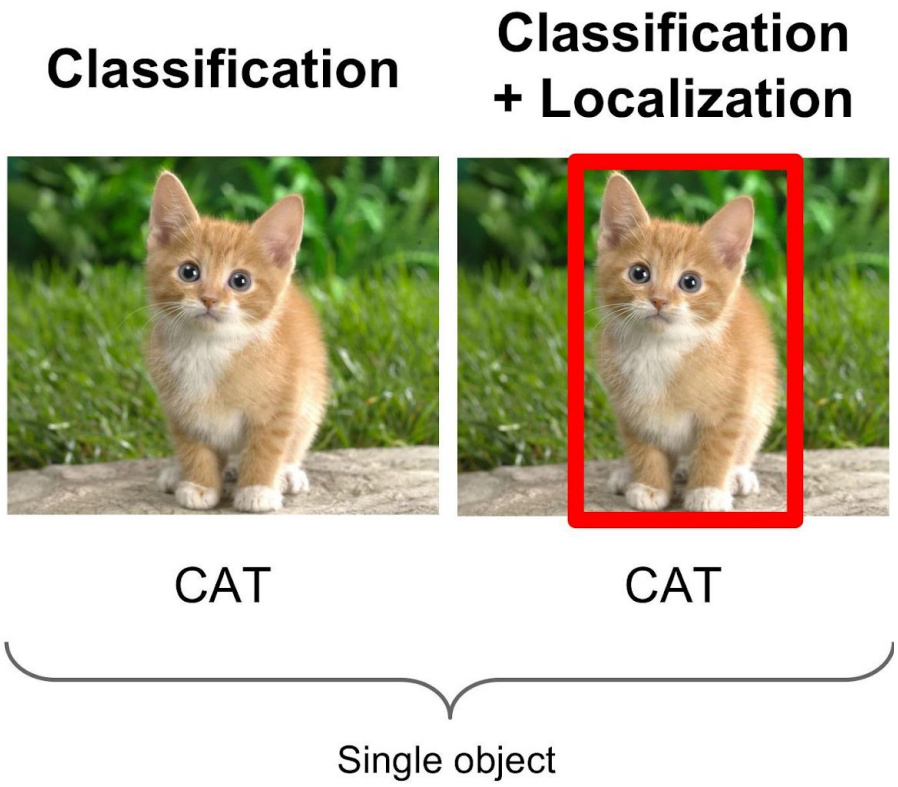
Introducción

Introducción



Localización

Introducción



Introducción

Classification

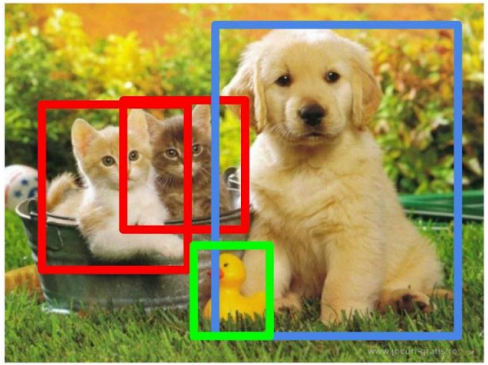
CAT

Classification + Localization

CAT

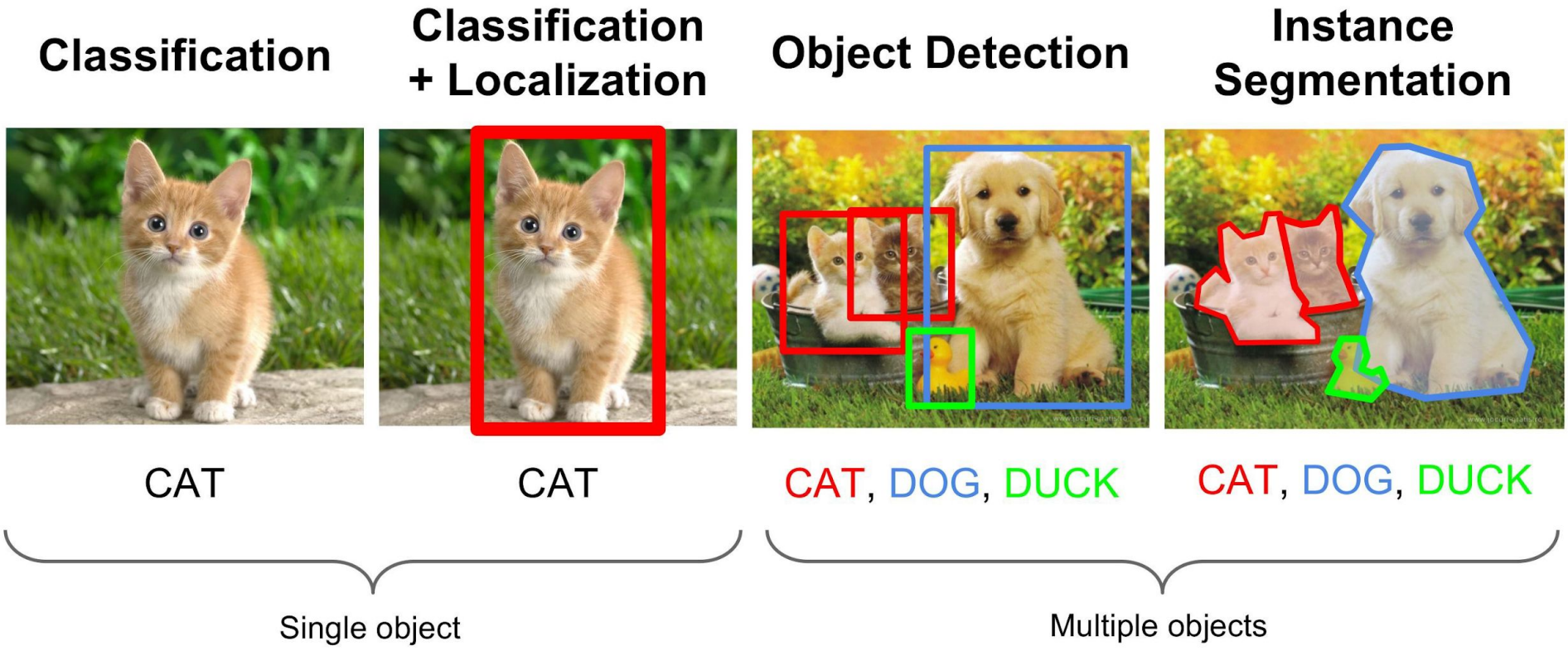
Single object

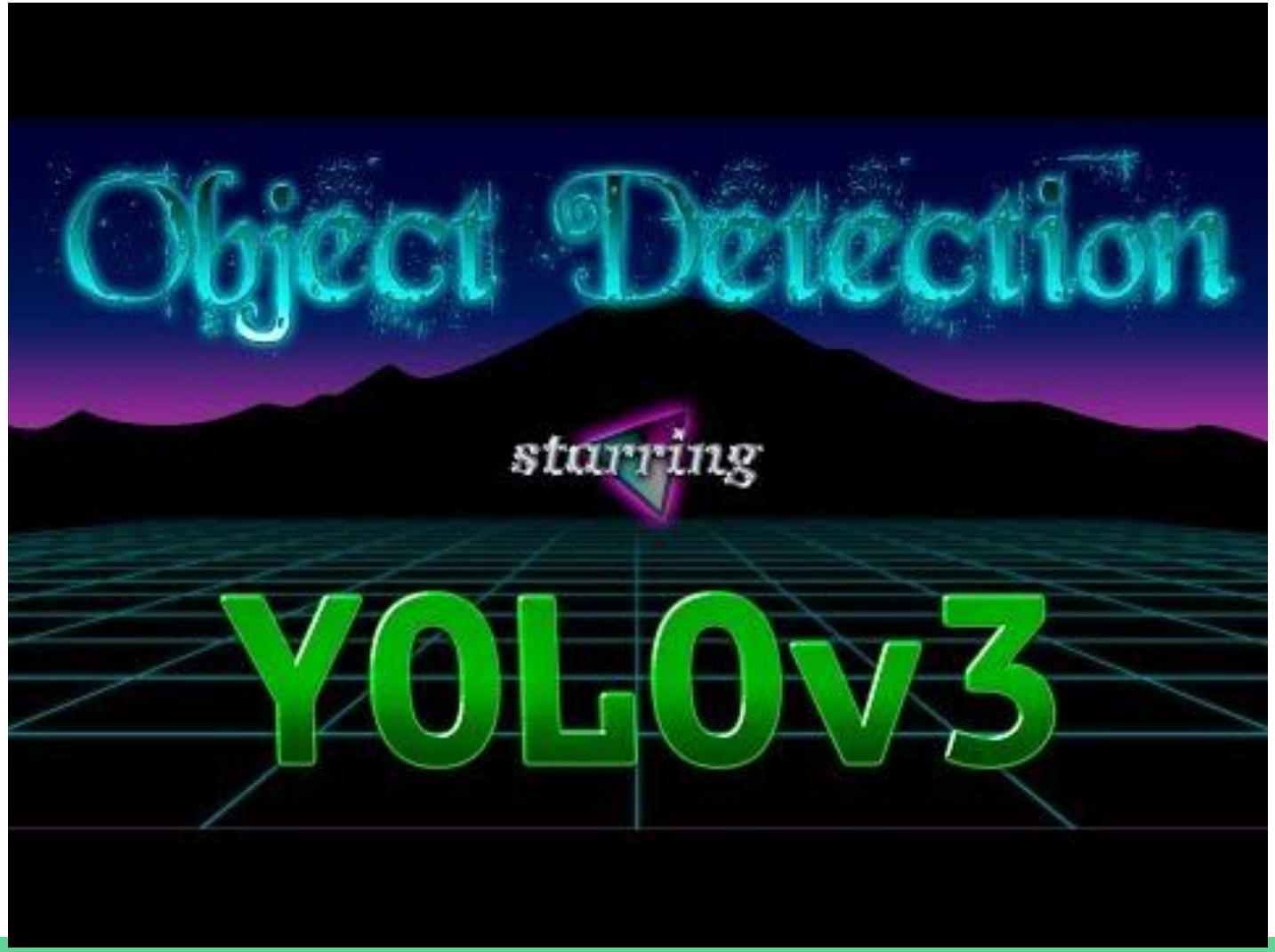
Object Detection

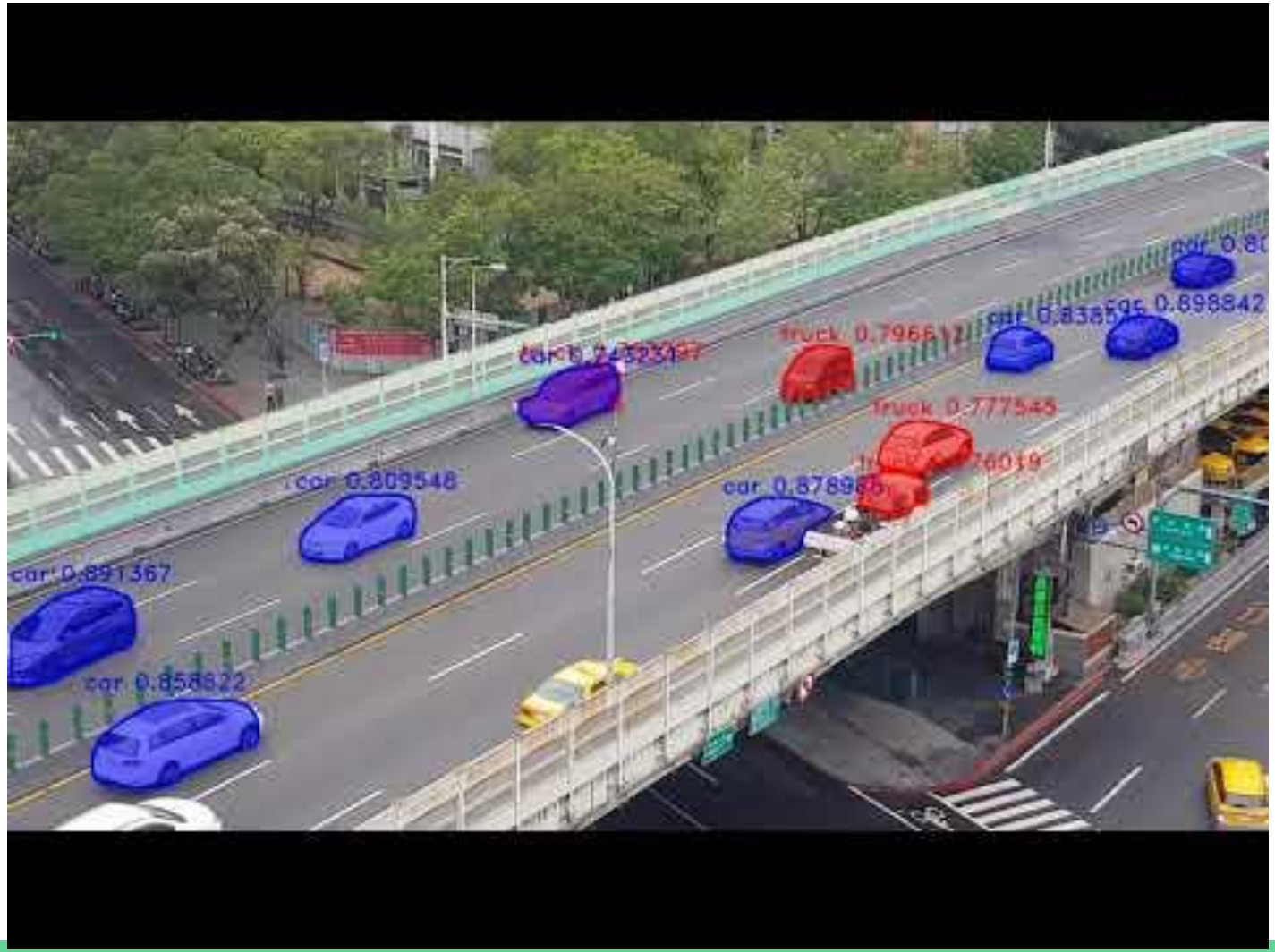


CAT, DOG, DUCK

Introducción

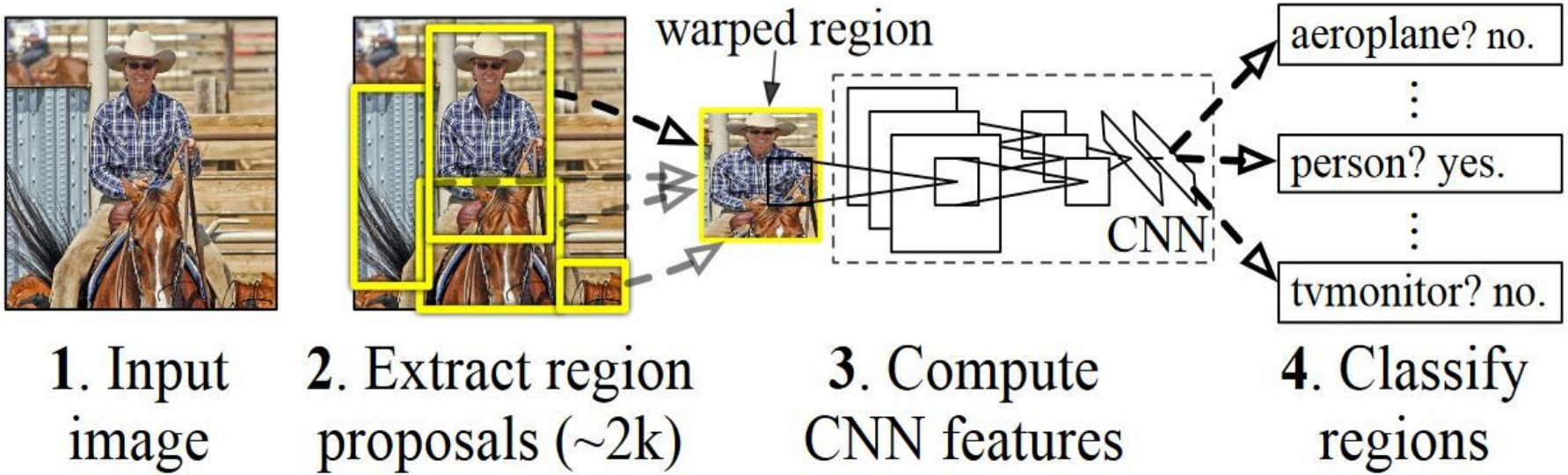



Object Detection
starring
YOLOv3

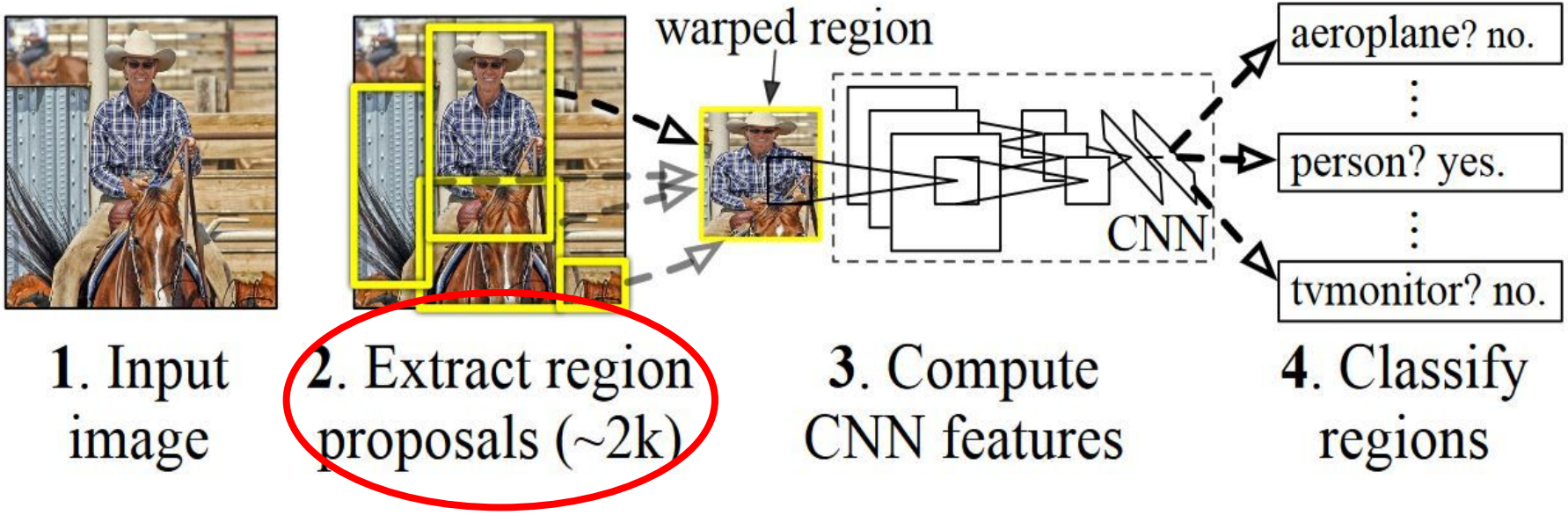

car 0.809548
car 0.891367
car 0.858822
truck 0.796612
truck 0.777545
car 0.898842

RCNN (Girshick, 2013)

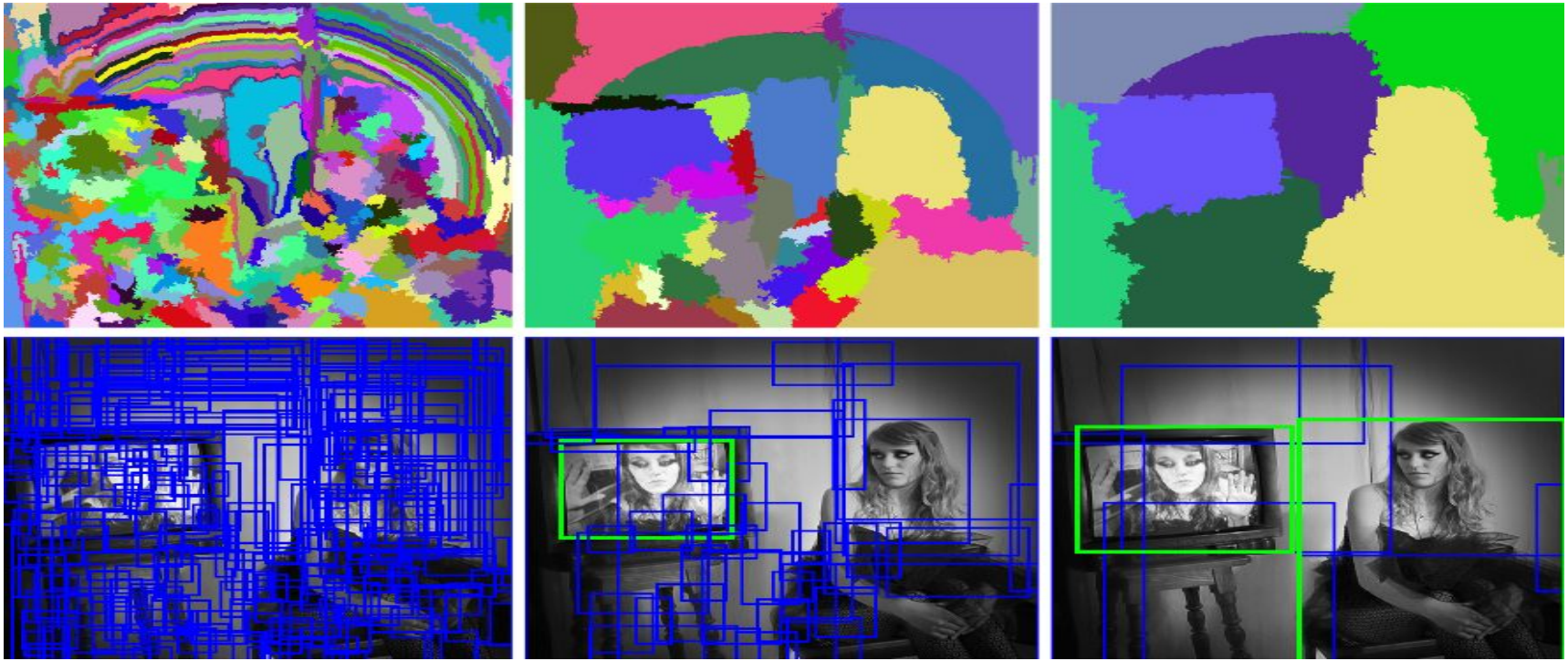
RCNN(2013)


warped region
CNN
aeroplane? no.
person? yes.
tvmonitor? no.
1. Input image
2. Extract region proposals (~2k)
3. Compute CNN features
4. Classify regions

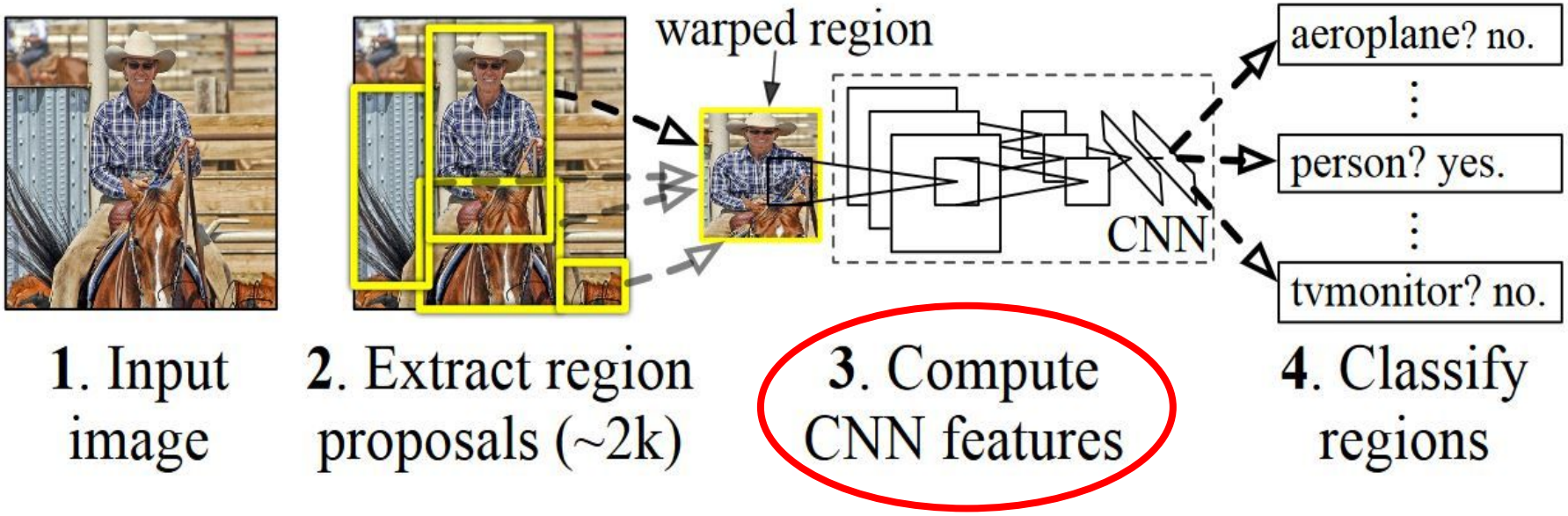
RCNN(2013)


warped region
CNN
aeroplane? no.
person? yes.
tvmonitor? no.
1. Input image
2. Extract region proposals (~2k)
3. Compute CNN features
4. Classify regions

Selective Search (Uijlings,2012)

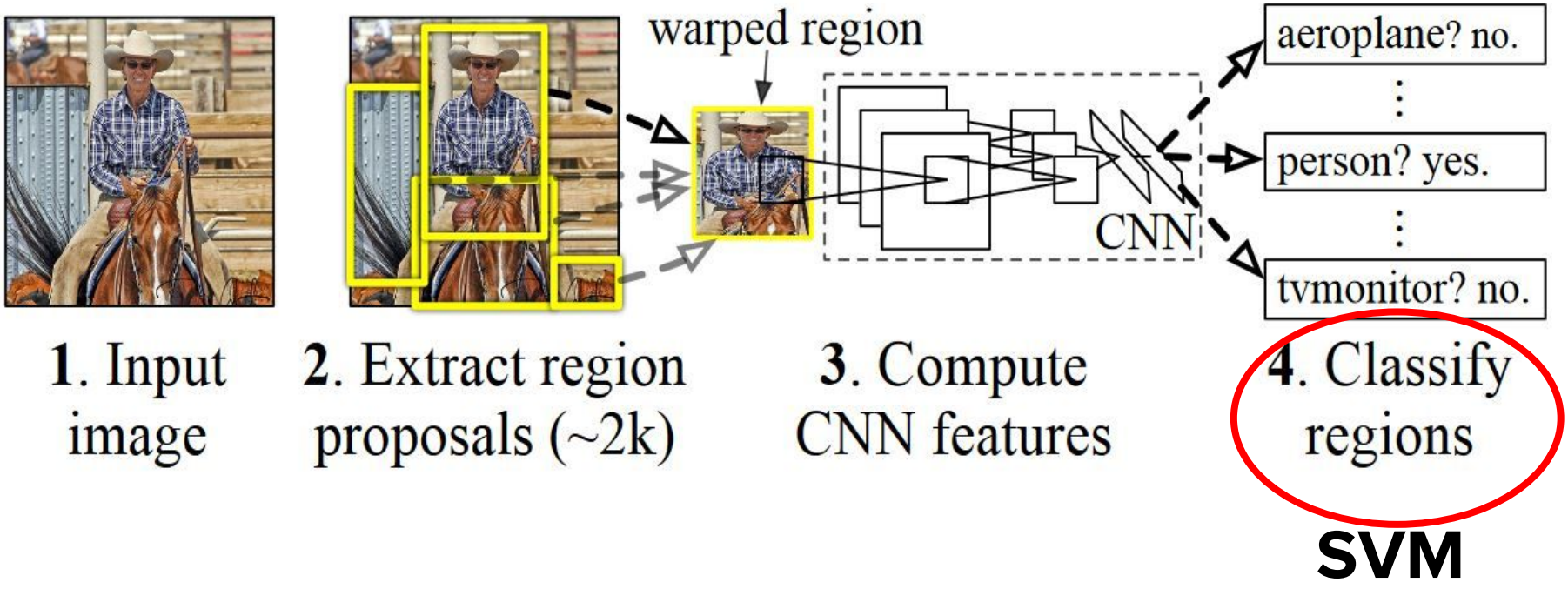


RCNN(2013)



VGG-16

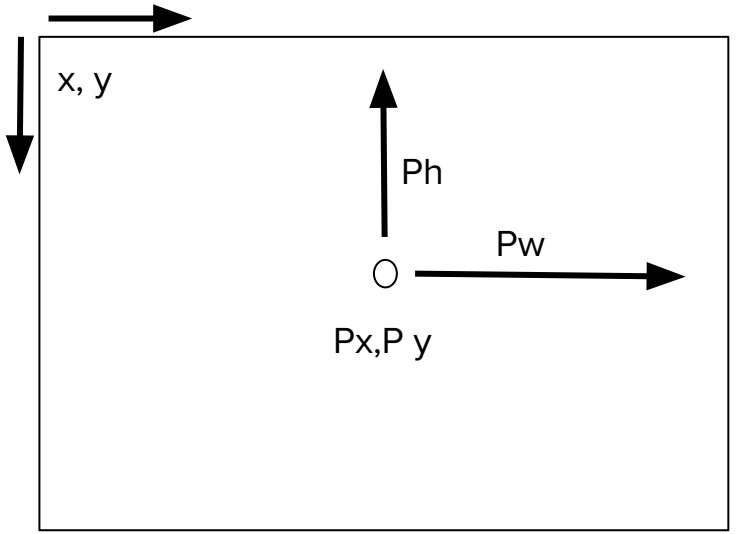
RCNN(2013)


warped region
aeroplane? no.
person? yes.
tvmonitor? no.
CNN
1. Input image
2. Extract region proposals (~2k)
3. Compute CNN features
4. Classify regions
SVM

RCNN(2013) Bounding-box Regression

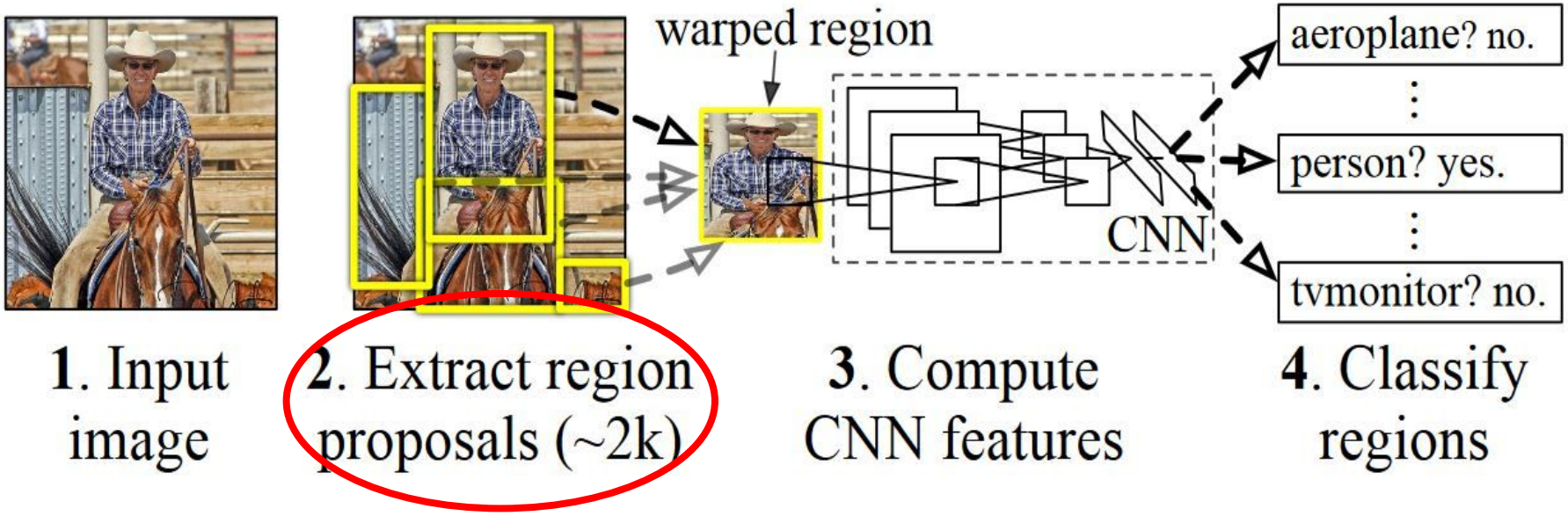
$$
\begin{aligned}
t_x &= (G_x - P_x)/P_w \\
t_y &= (G_y - P_y)/P_h \\
t_w &= \log(G_w/P_w) \\
t_h &= \log(G_h/P_h).
\end{aligned}
$$

t : Predicciones.
G: Bounding-box predicho.

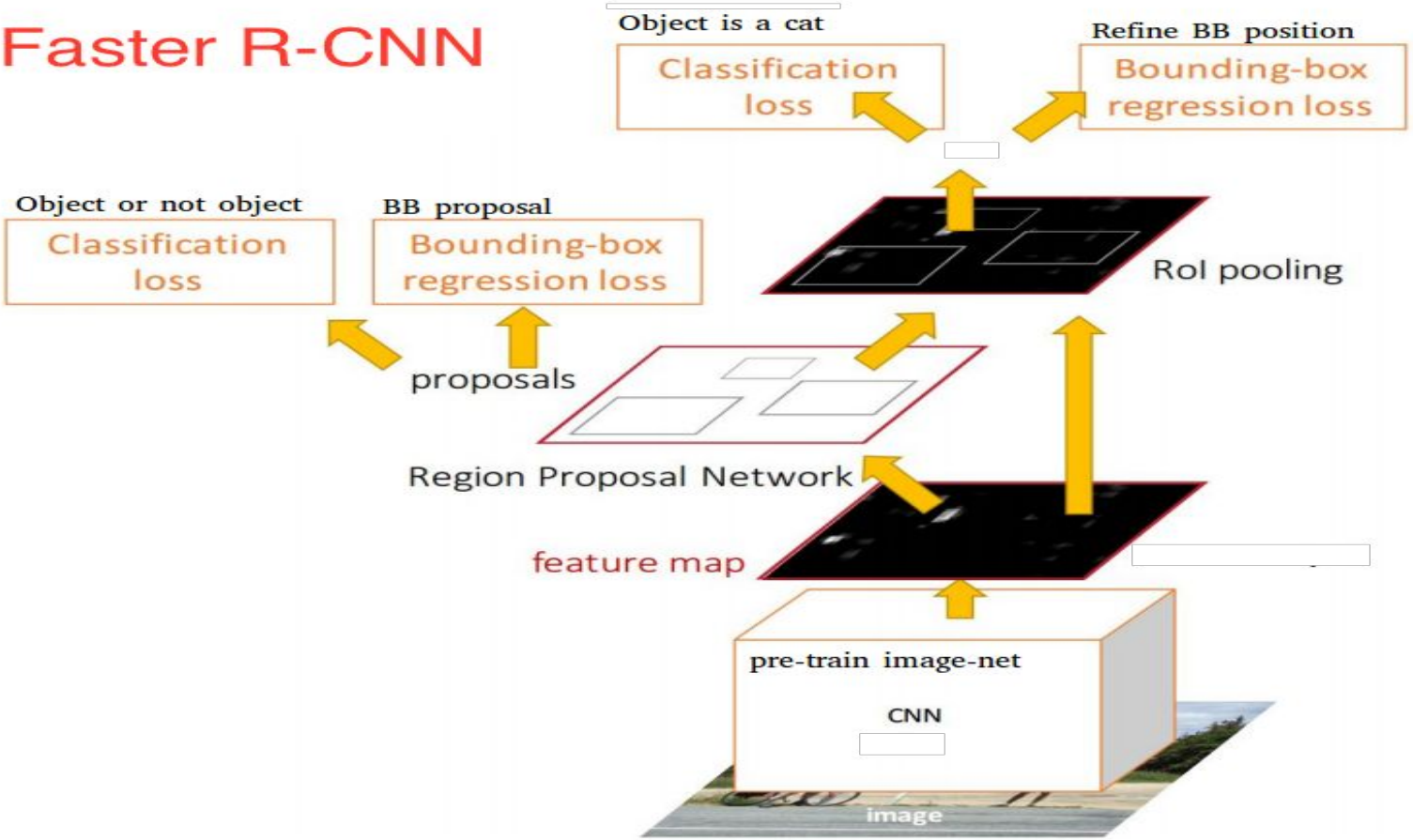


Faster-RCNN (Ren, 2015)

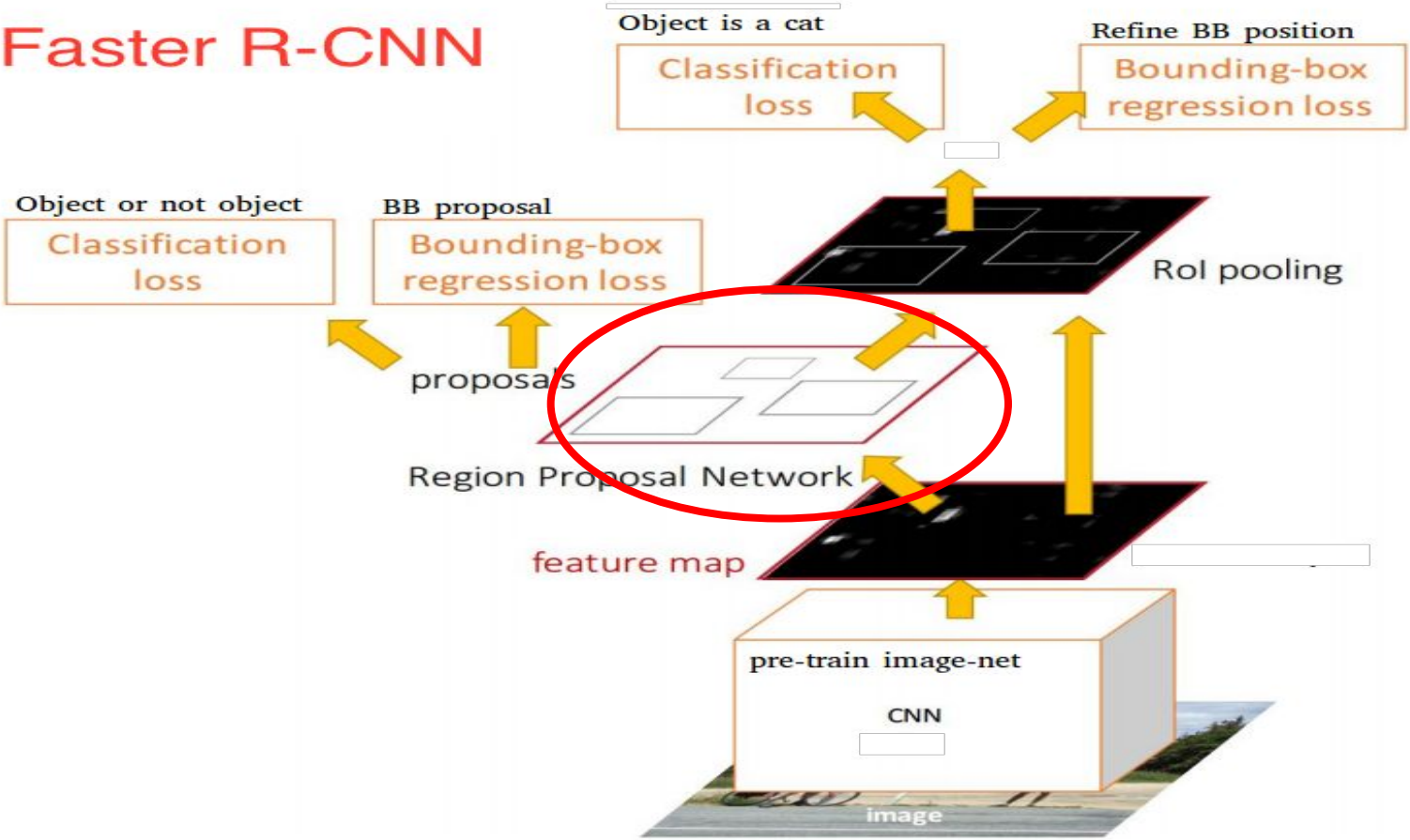
Faster-RCNN(2015)

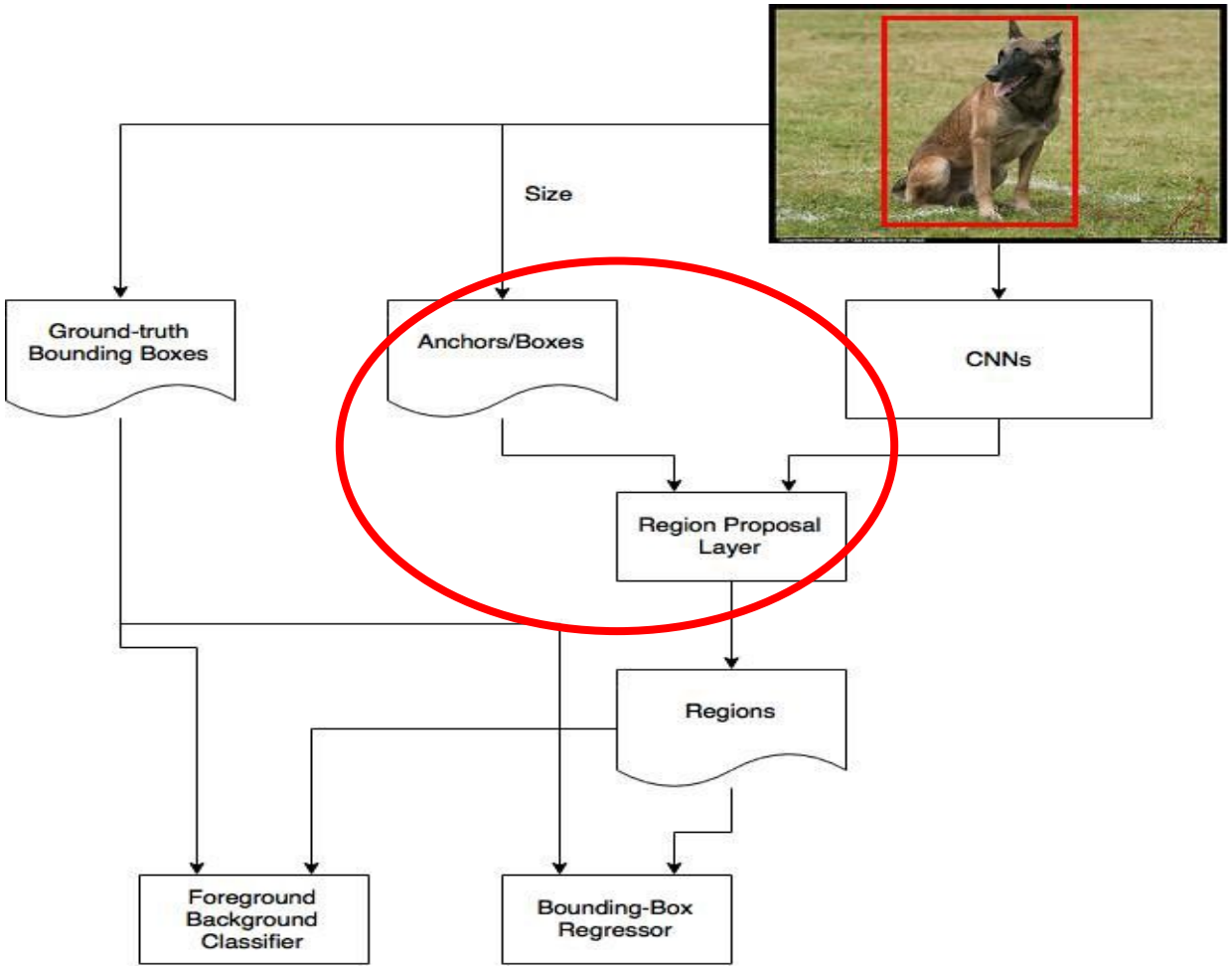


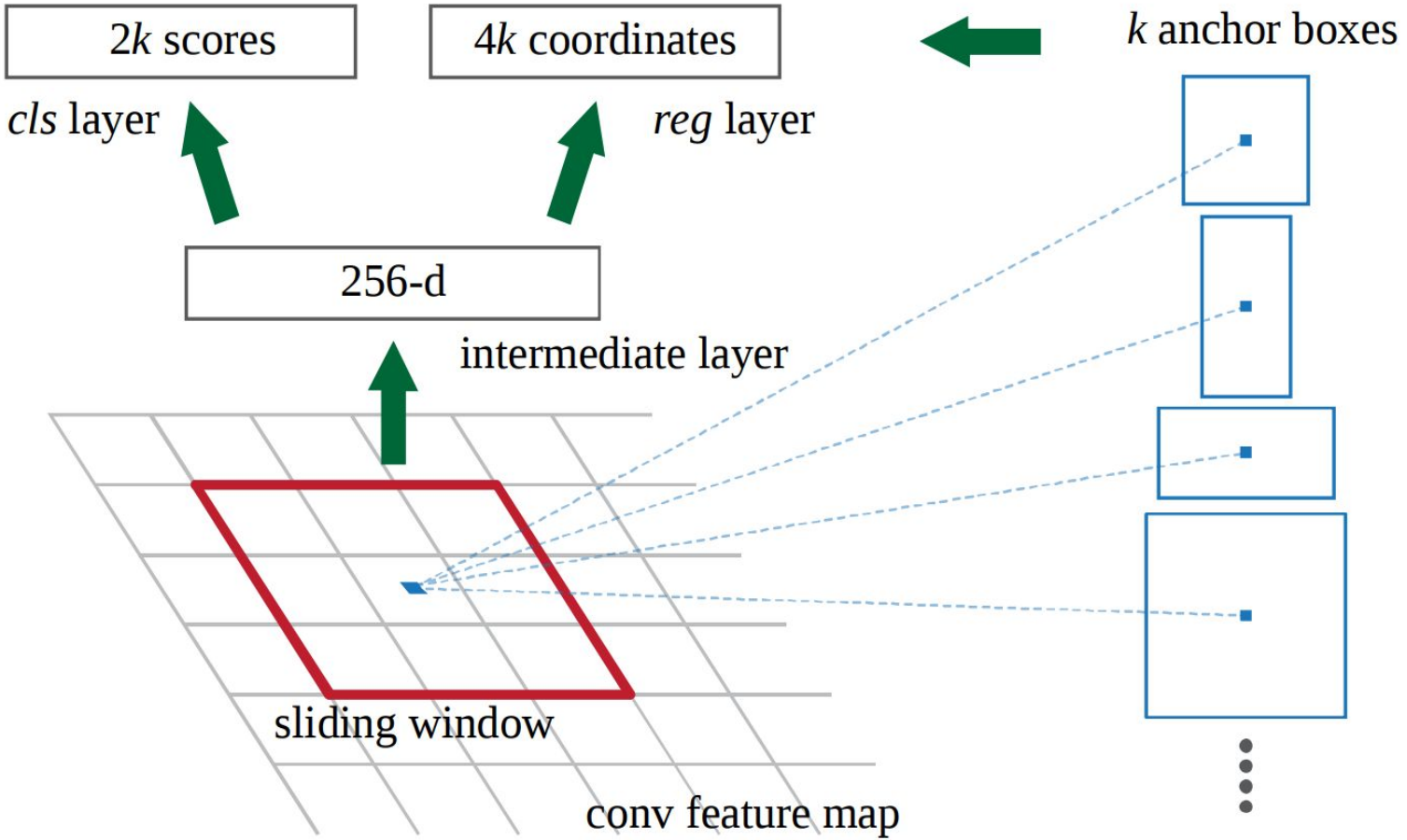
Faster R-CNN

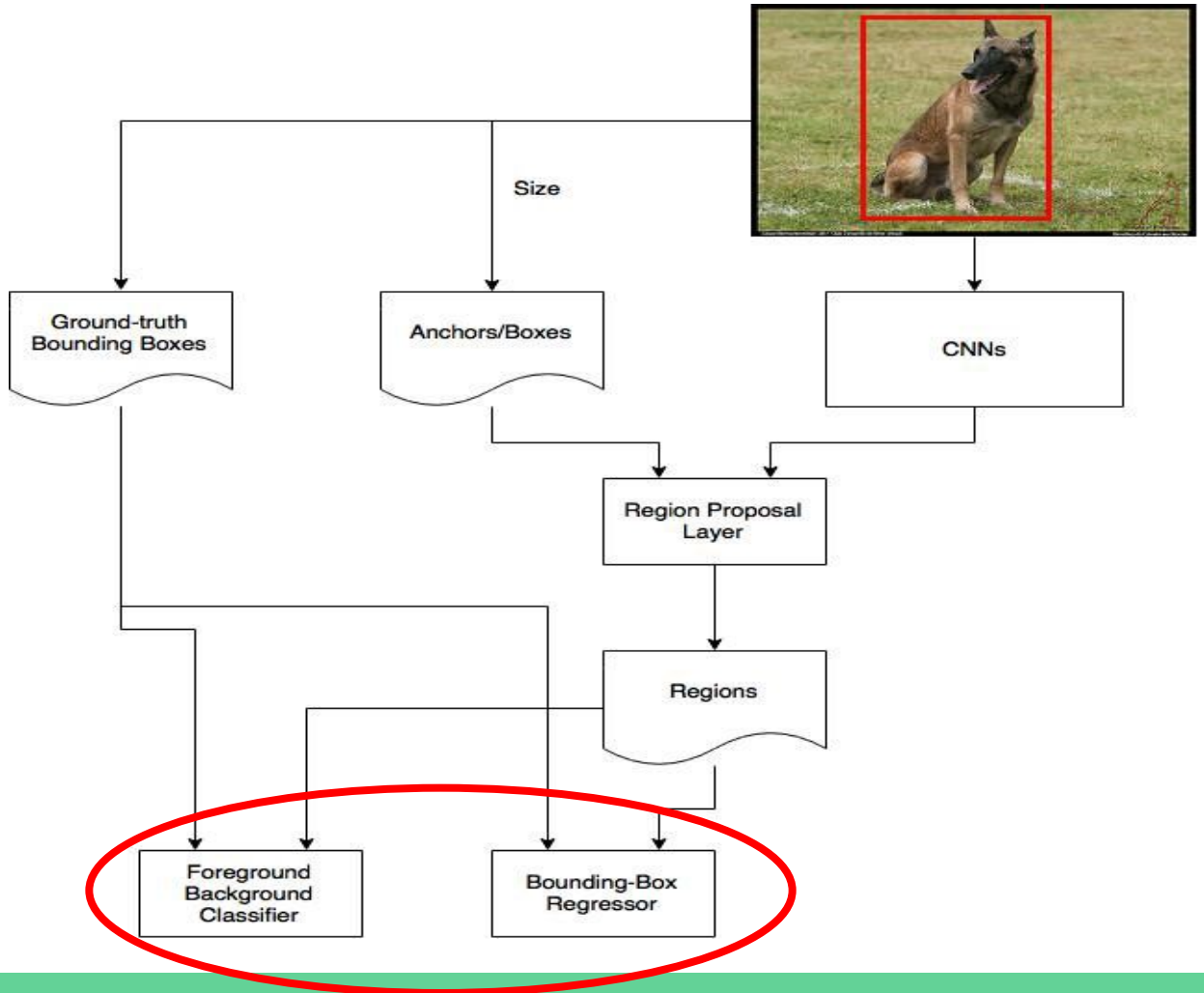

Object is a cat
Classification loss
Refine BB position
Bounding-box regression loss
Object or not object
Classification loss
BB proposal
Bounding-box regression loss
RoI pooling
proposals
Region Proposal Network
feature map
pre-train image-net
CNN
image

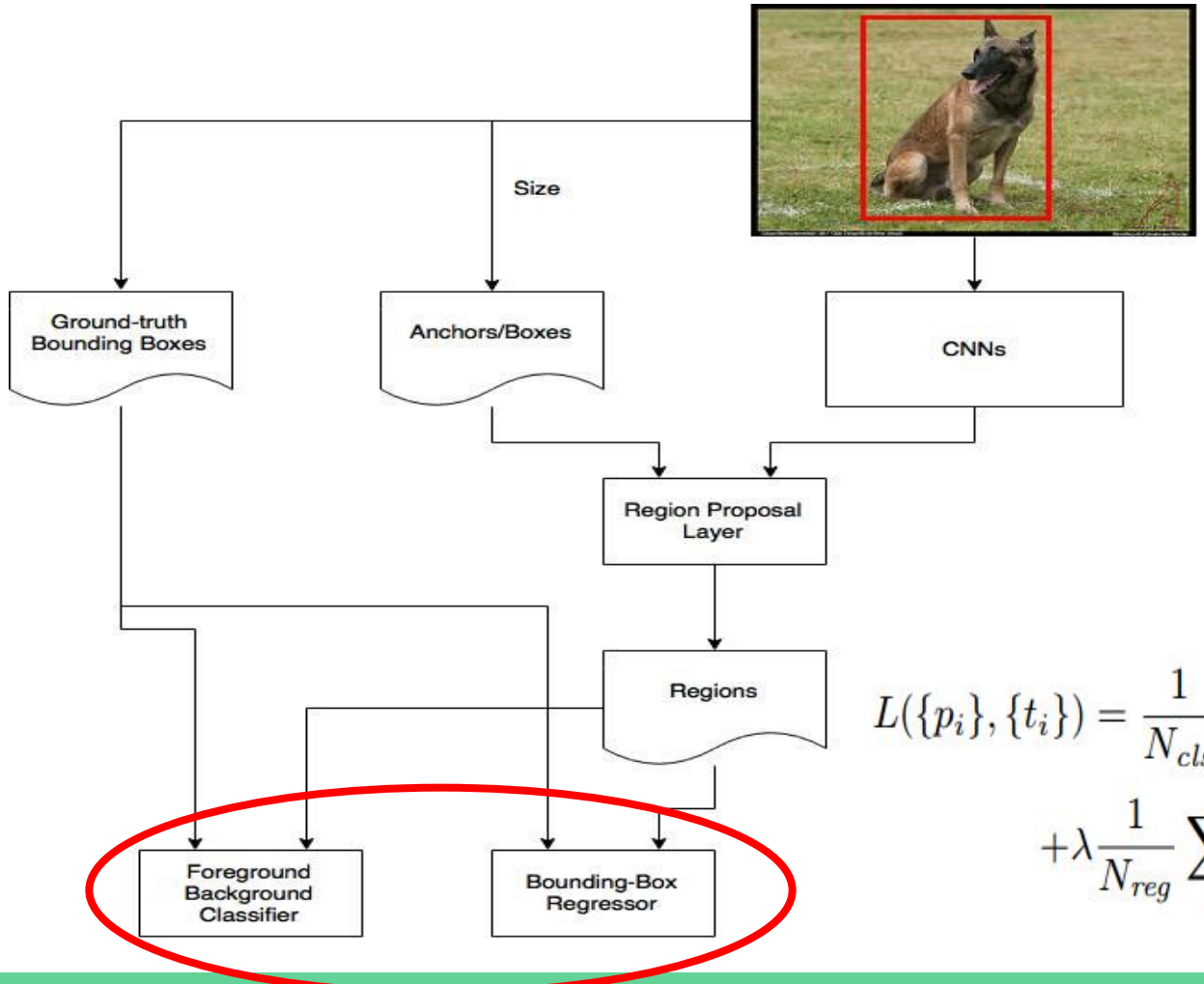
Faster R-CNN


Object is a cat
Classification loss
Refine BB position
Bounding-box regression loss
Object or not object
Classification loss
BB proposal
Bounding-box regression loss
RoI pooling
proposals
Region Proposal Network
feature map
pre-train image-net
CNN
image


Size
Ground-truth Bounding Boxes
Anchors/Boxes
CNNs
Region Proposal Layer
Regions
Foreground Background Classifier
Bounding-Box Regressor


2k scores
4k coordinates
k anchor boxes
cls layer
reg layer
256-d
intermediate layer
sliding window
conv feature map

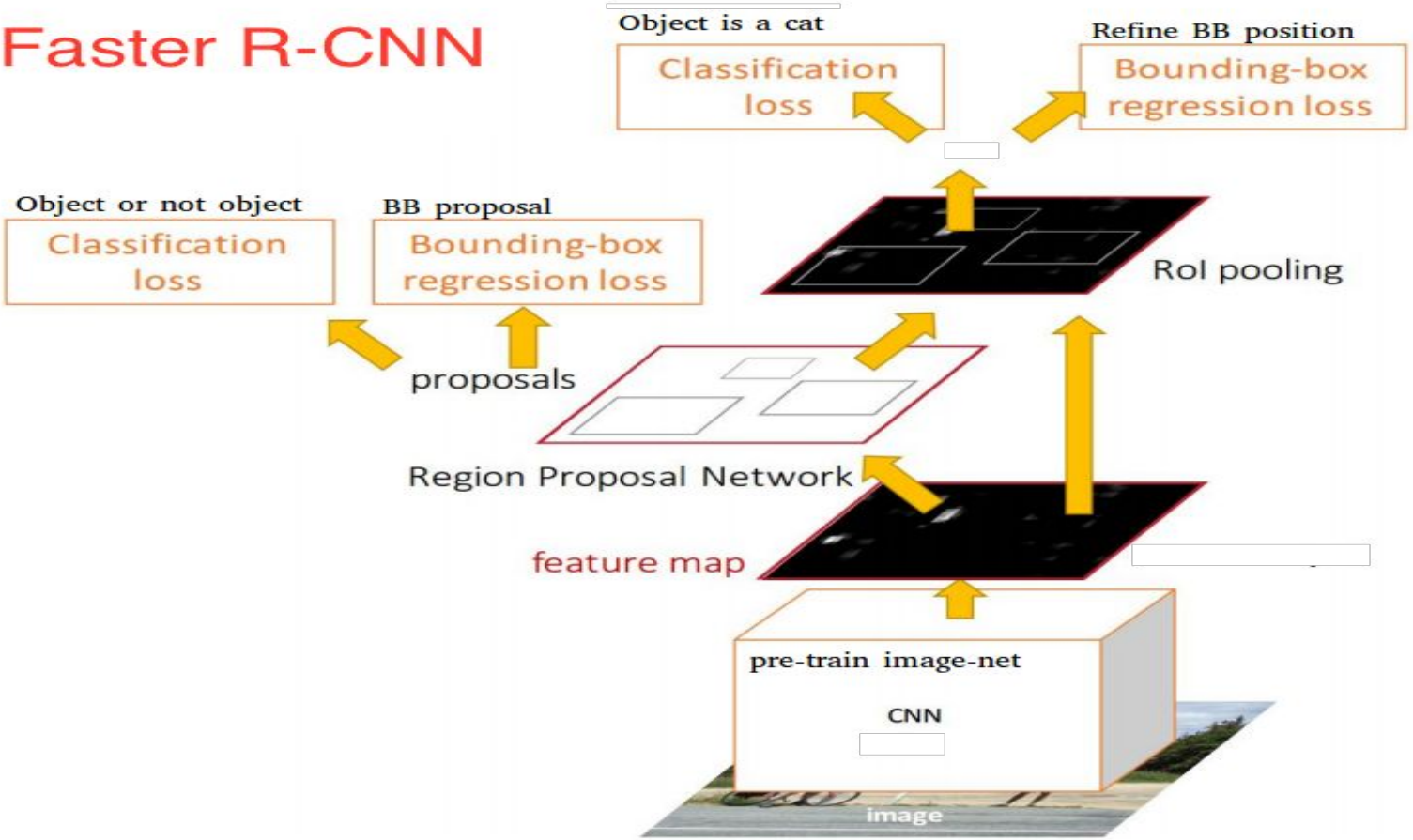

Size
Ground-truth Bounding Boxes
Anchors/Boxes
CNNs
Region Proposal Layer
Regions
Foreground Background Classifier
Bounding-Box Regressor



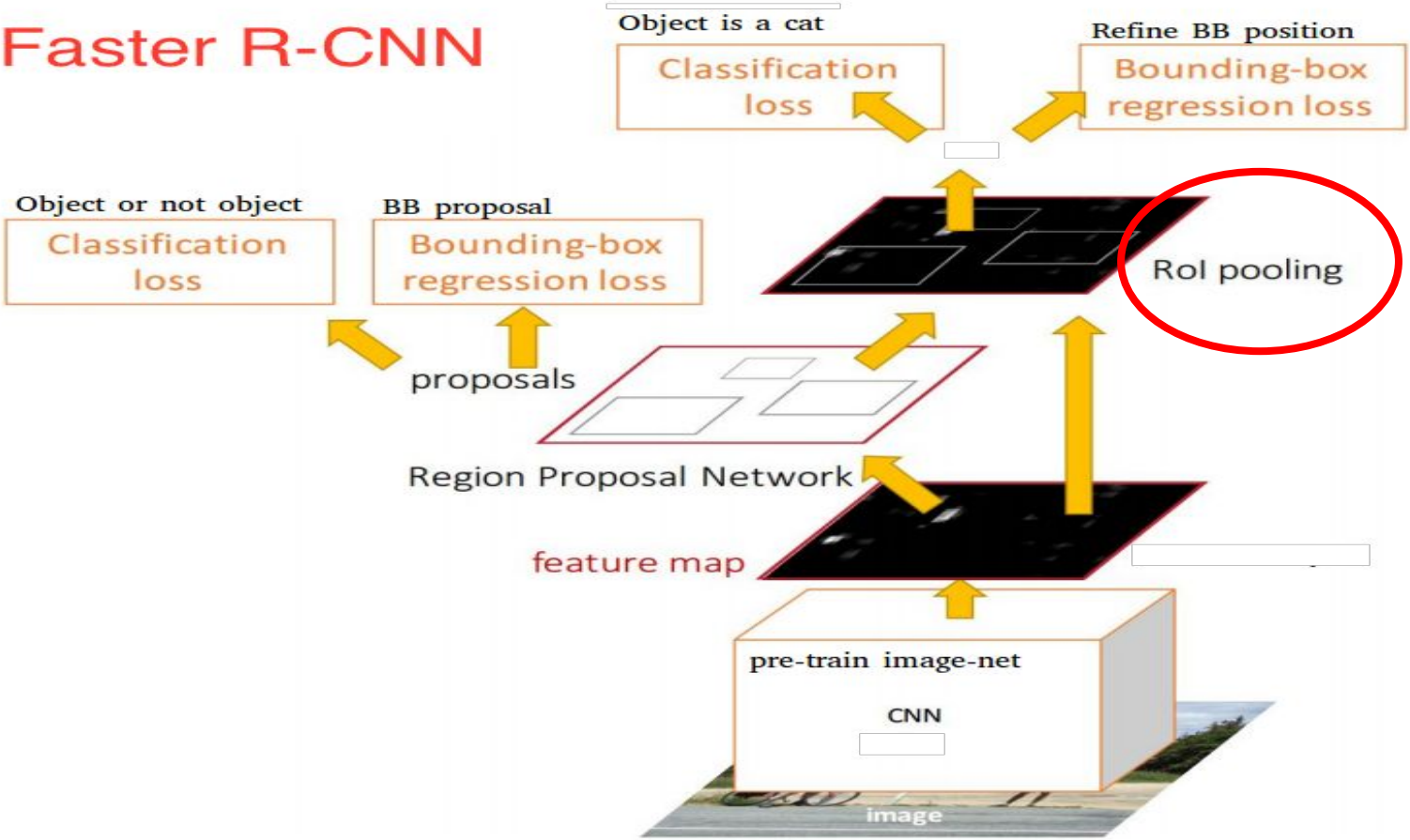
$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*)$$

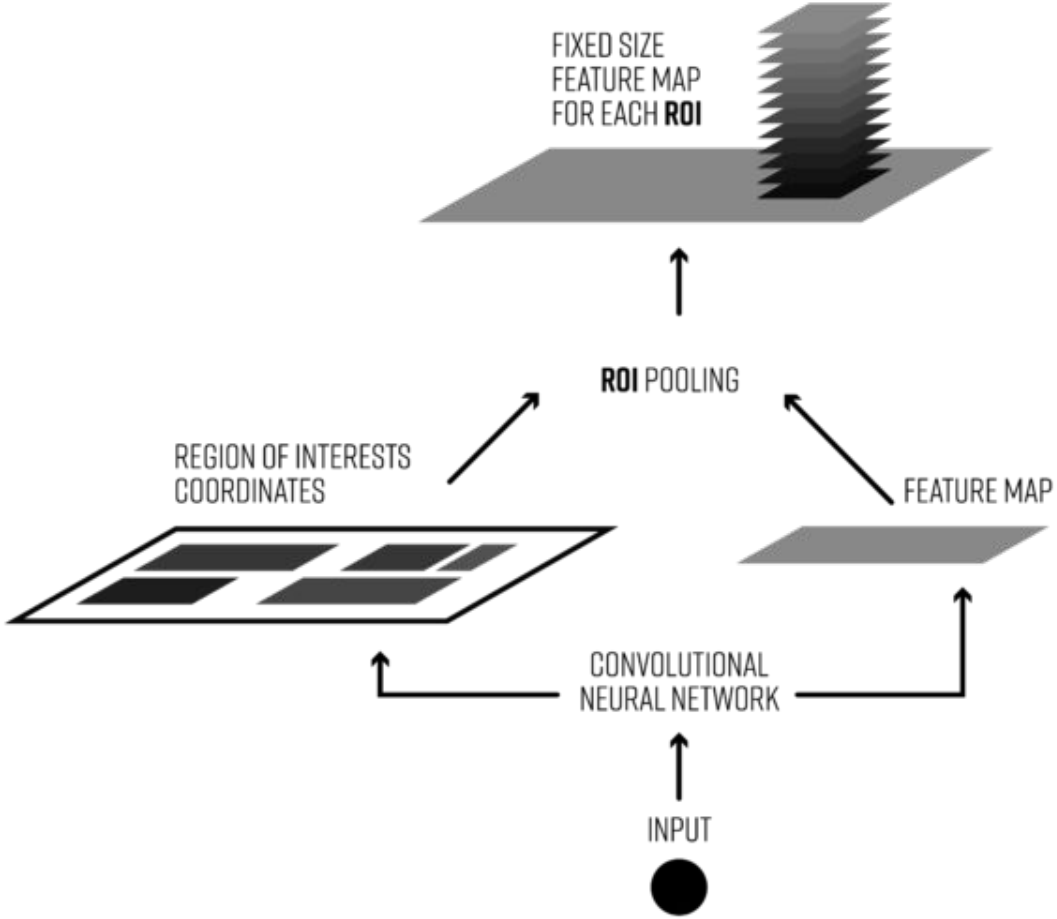
$$+\lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*).$$

Faster R-CNN

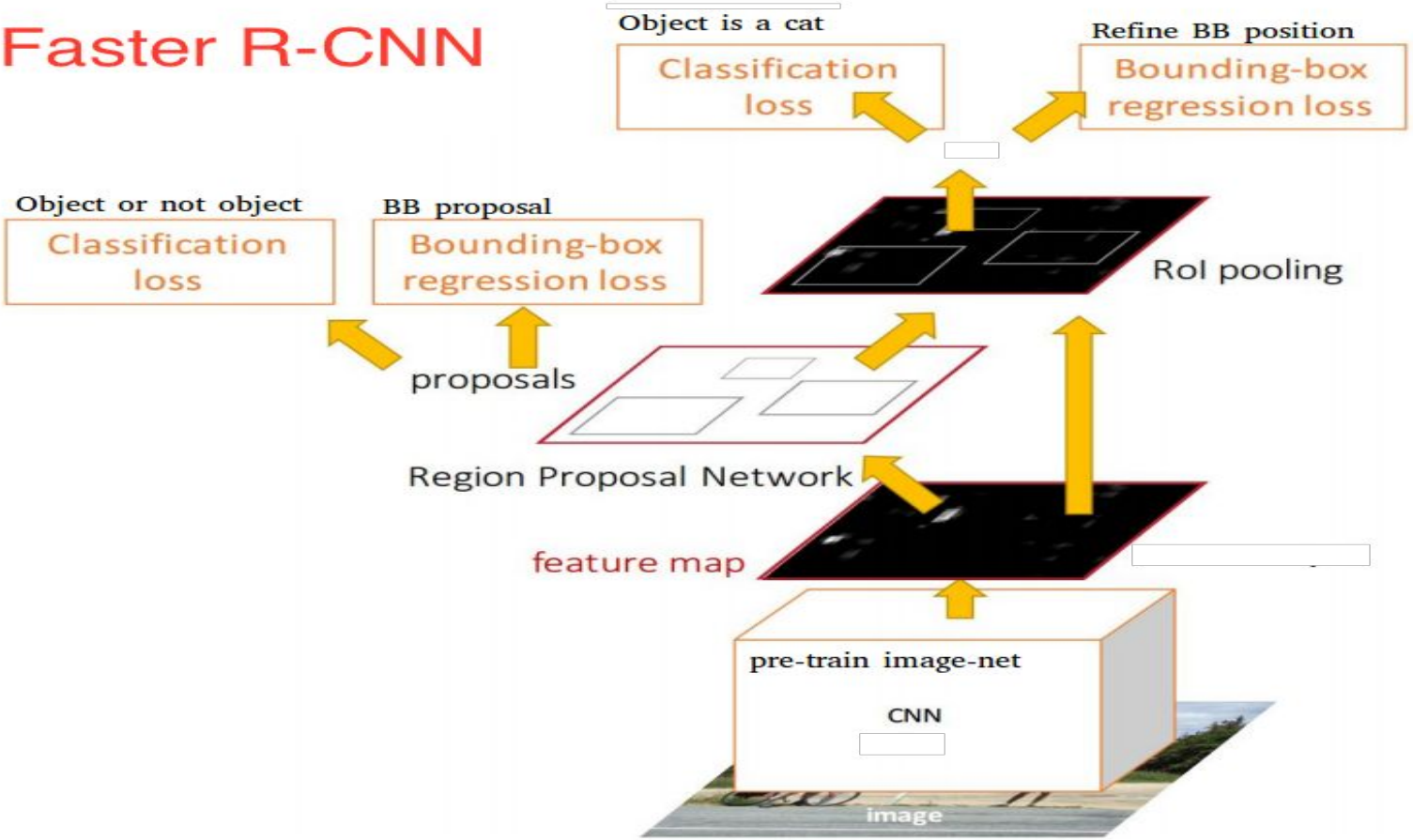

Object is a cat
Classification loss
Refine BB position
Bounding-box regression loss
Object or not object
Classification loss
BB proposal
Bounding-box regression loss
RoI pooling
proposals
Region Proposal Network
feature map
pre-train image-net
CNN
image

Faster R-CNN

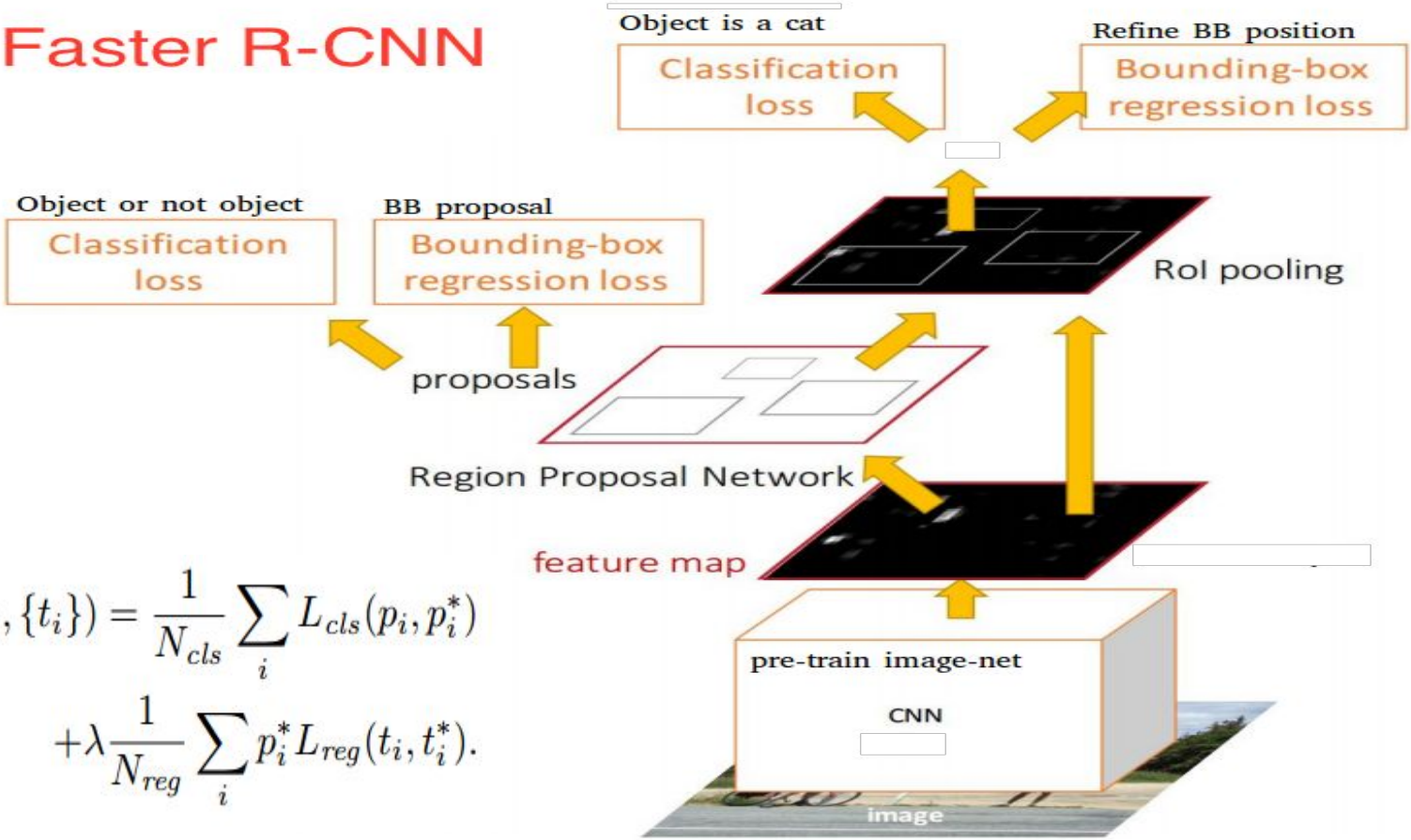

Object is a cat
Classification loss
Refine BB position
Bounding-box regression loss
Object or not object
Classification loss
BB proposal
Bounding-box regression loss
RoI pooling
proposals
Region Proposal Network
feature map
pre-train image-net
CNN
image


FIXED SIZE
FEATURE MAP
FOR EACH ROI
ROI POOLING
REGION OF INTERESTS
COORDINATES
FEATURE MAP
CONVOLUTIONAL
NEURAL NETWORK
INPUT

Faster R-CNN


Object is a cat
Classification loss
Refine BB position
Bounding-box regression loss
Object or not object
Classification loss
BB proposal
Bounding-box regression loss
RoI pooling
proposals
Region Proposal Network
feature map
pre-train image-net
CNN
image

Faster R-CNN

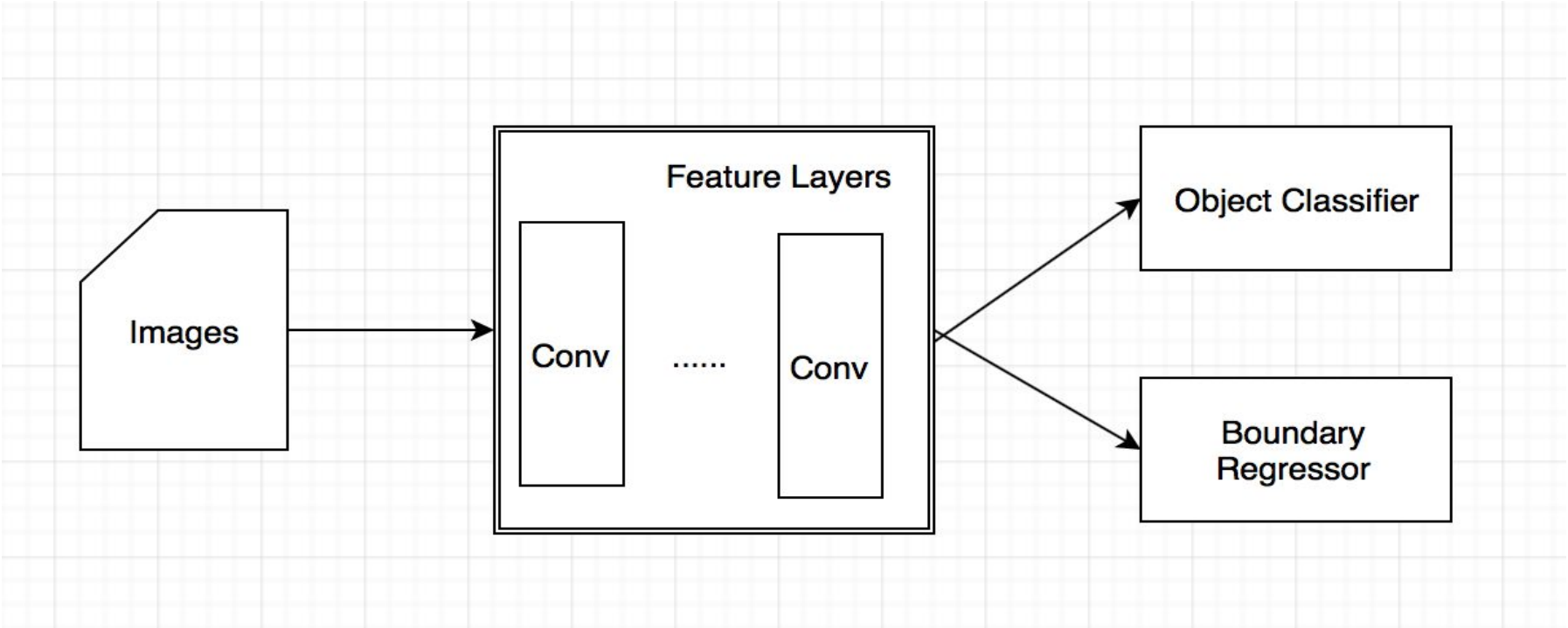


$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*)$$

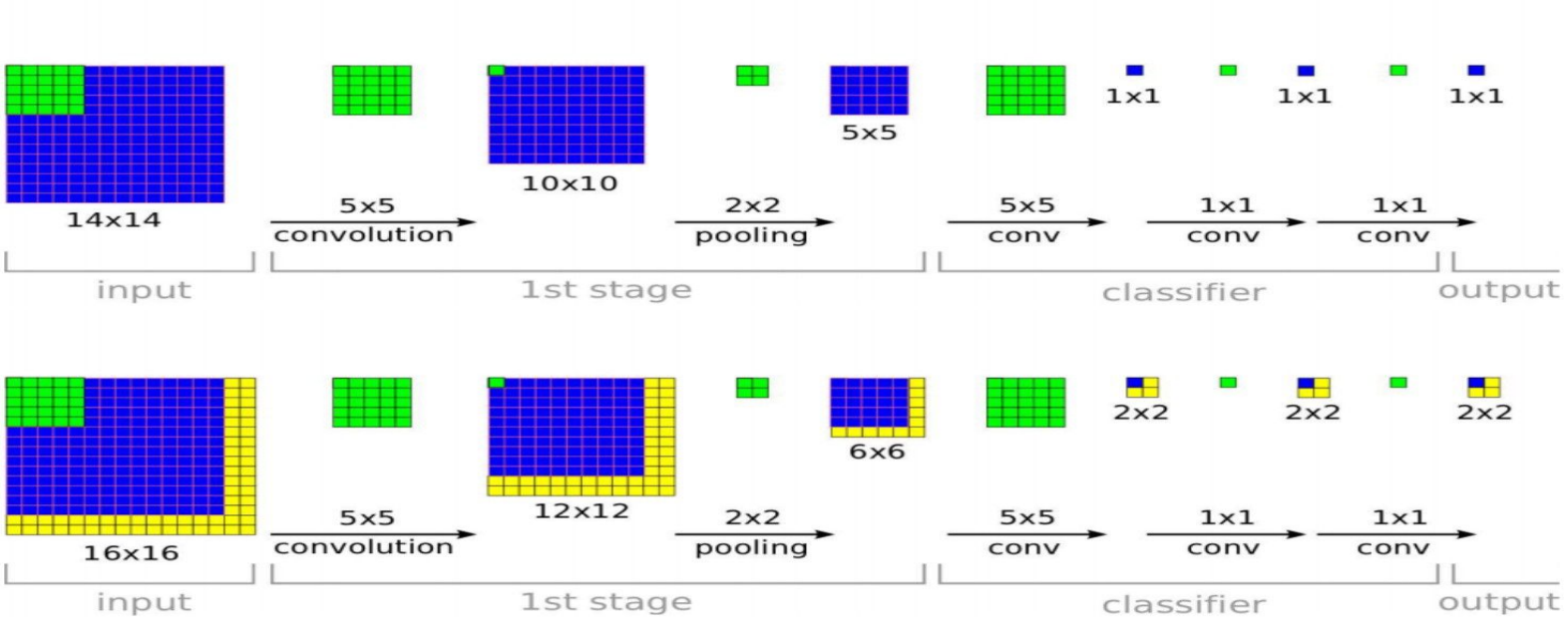
$$+\lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*).$$

SWCNN - Overfeat (Sermanet, 2013)

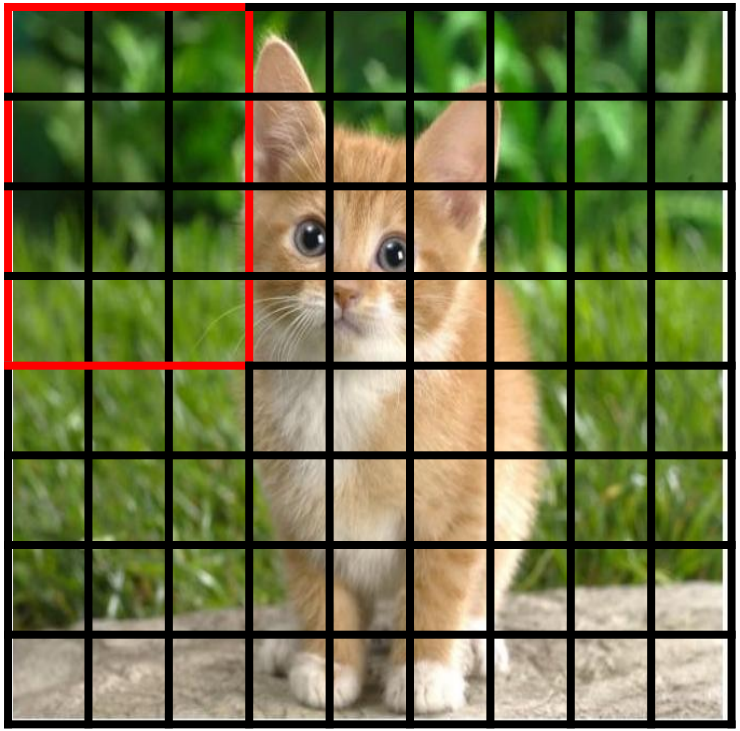
OverFeat(2013)

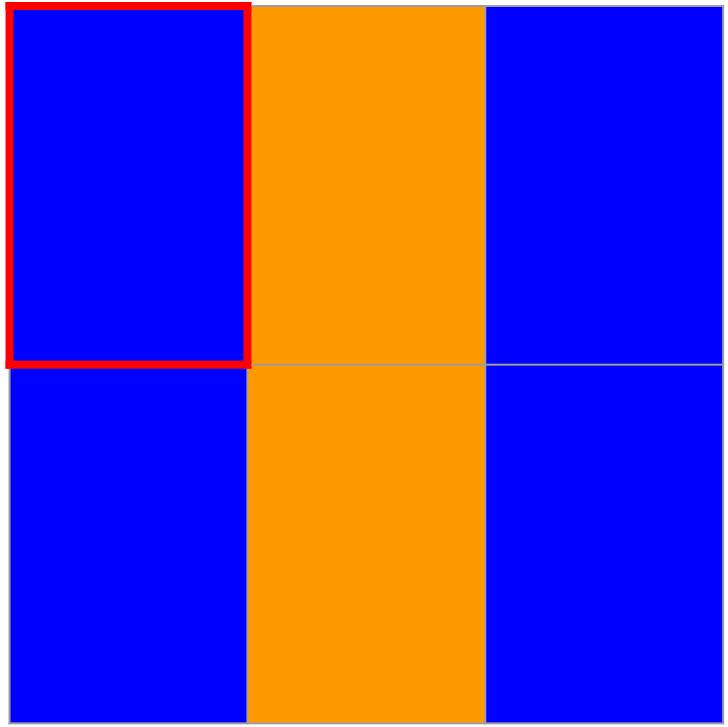


OverFeat(2013)



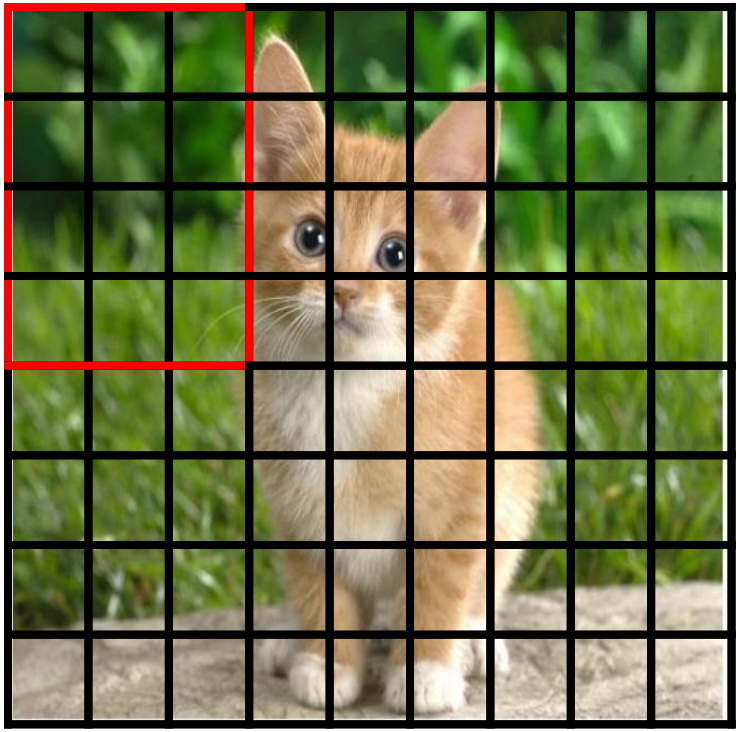
OverFeat(2013) Clasificador

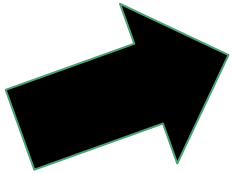


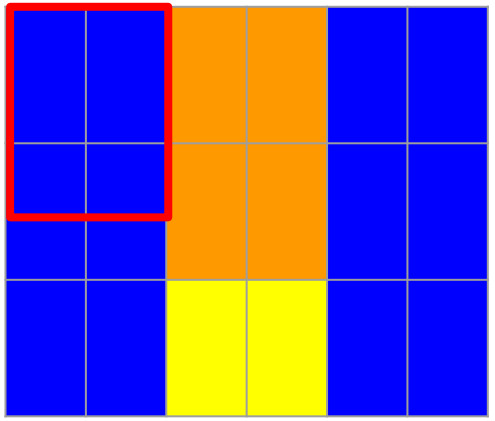


Mapa de Características (softmax)

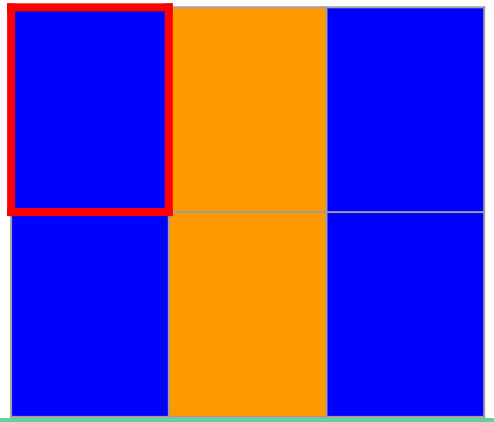
OverFeat(2013) Multi-scale





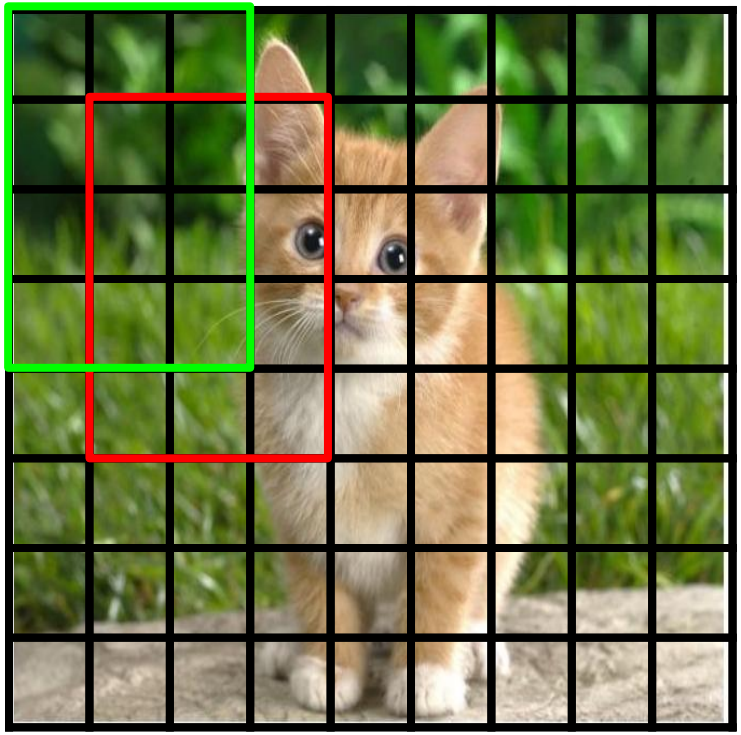


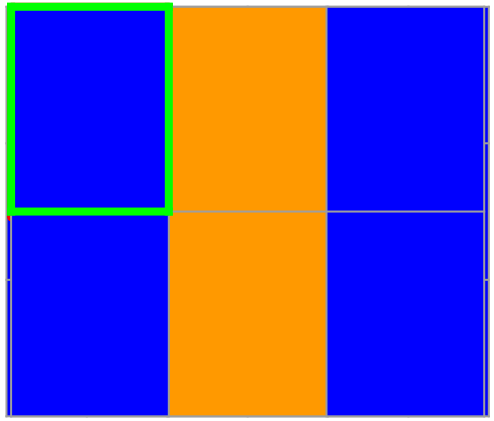
Salida 1



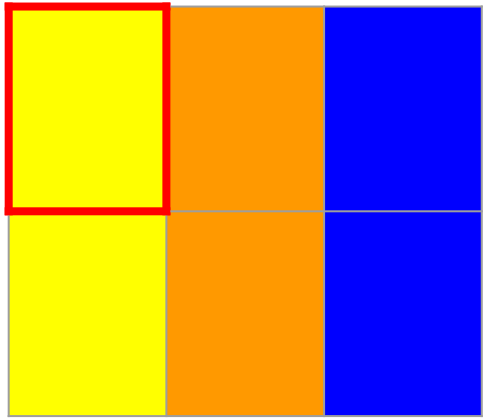
Salida 2

OverFeat(2013) (Δx, Δy)



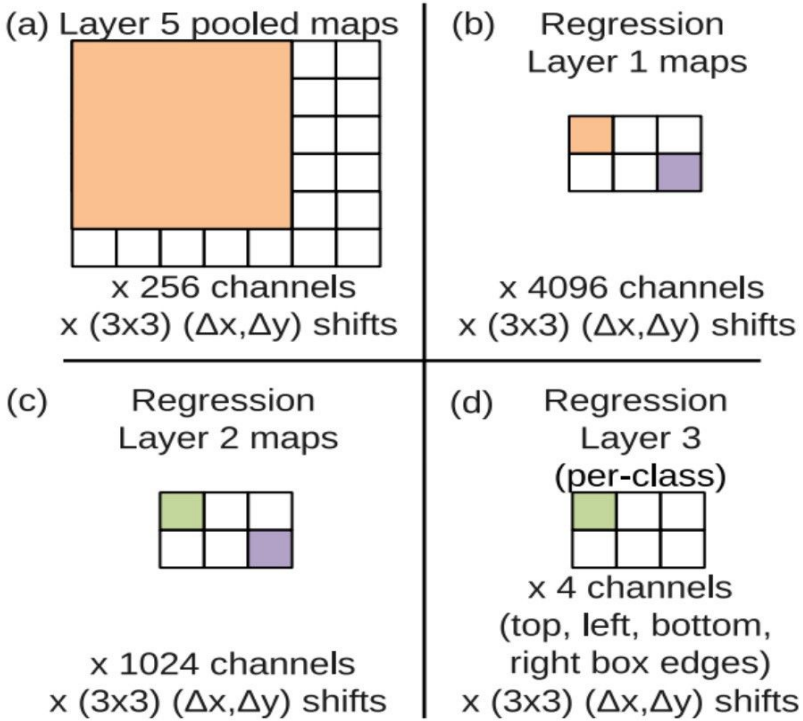


Salida 2

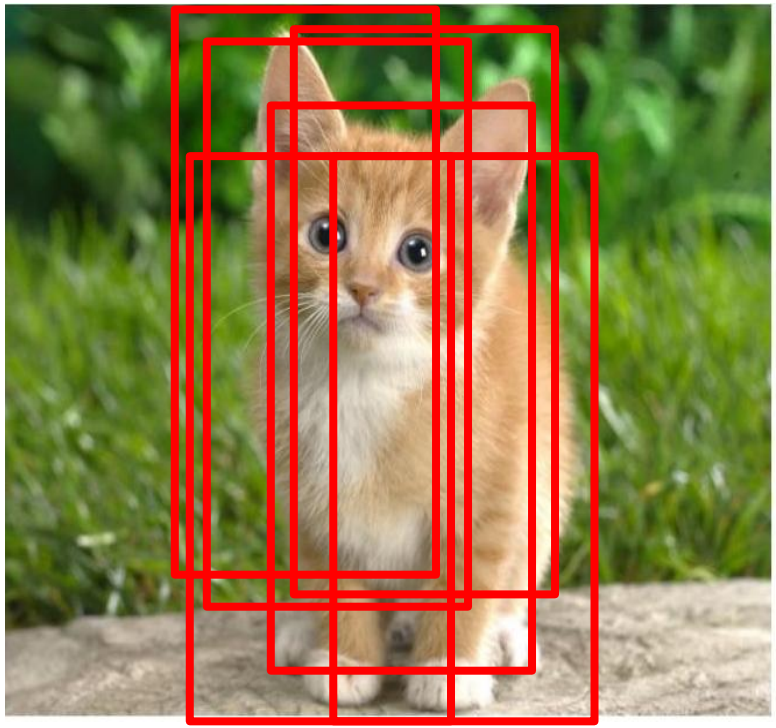


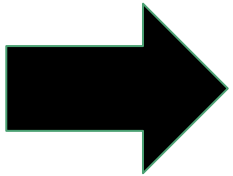
Salida 2

OverFeat(2013) Localización



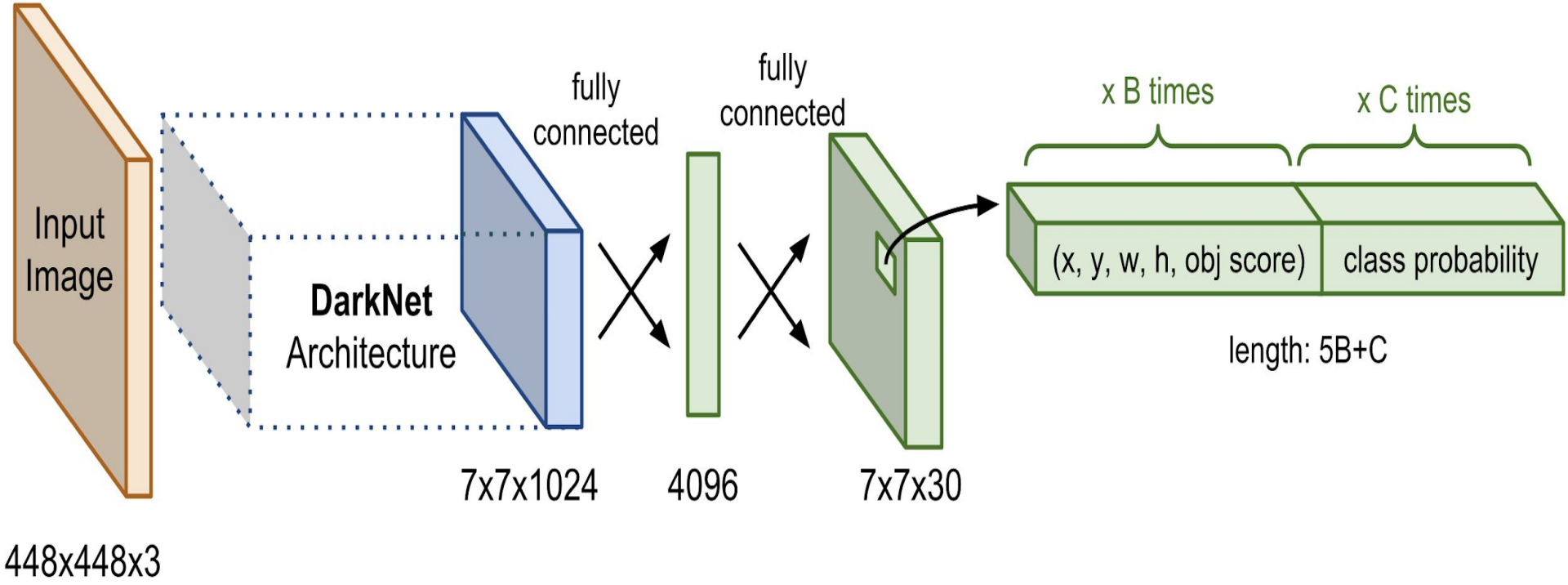
OverFeat(2013)



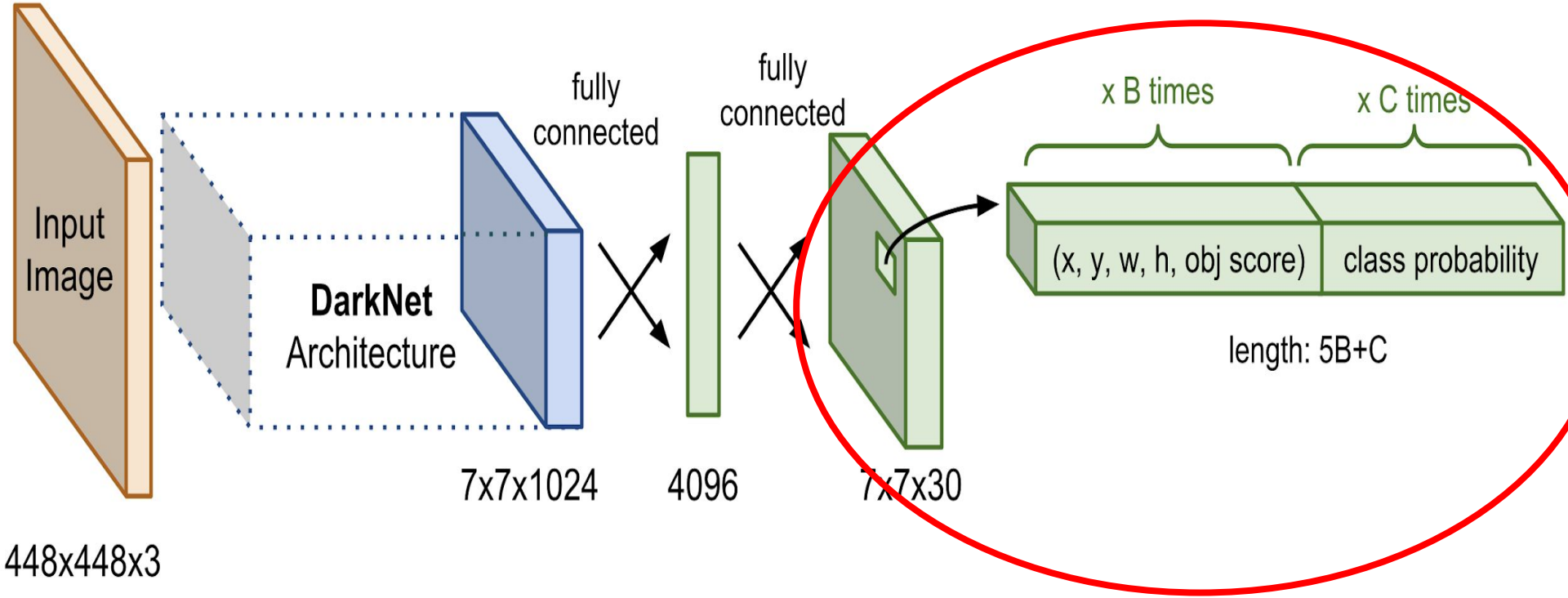


YOLO (Redmon, 2015)

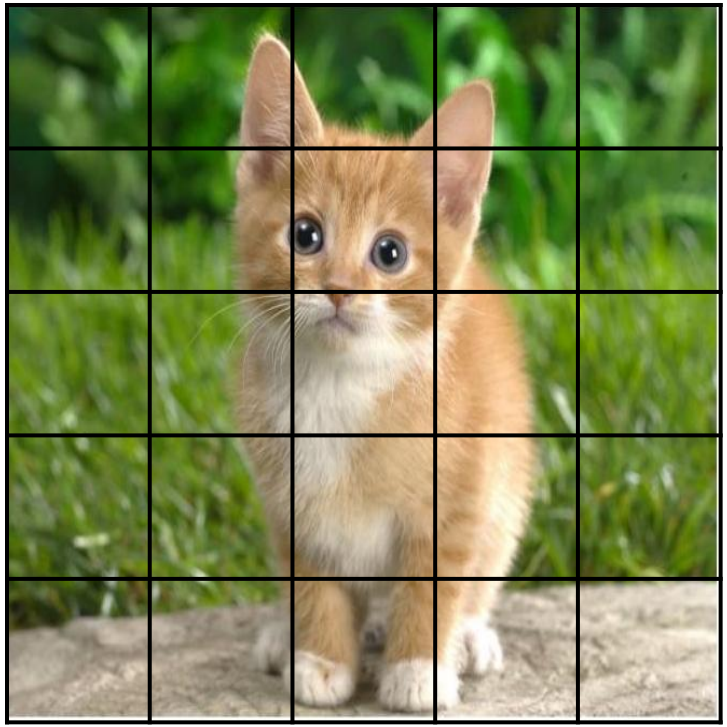
YOLO (2015)

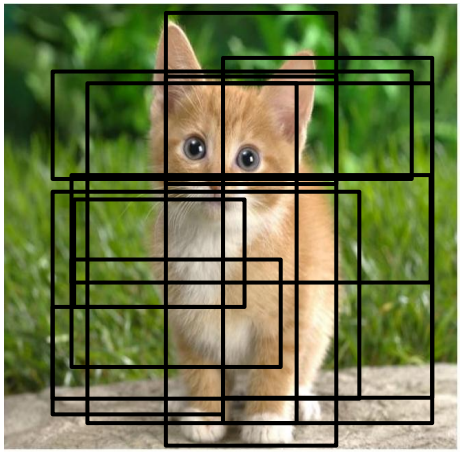

Input Image
448x448x3
DarkNet Architecture
7x7x1024
fully connected
4096
fully connected
7x7x30
x B times
x C times
(x, y, w, h, obj score)
class probability
length: 5B+C

YOLO (2015)

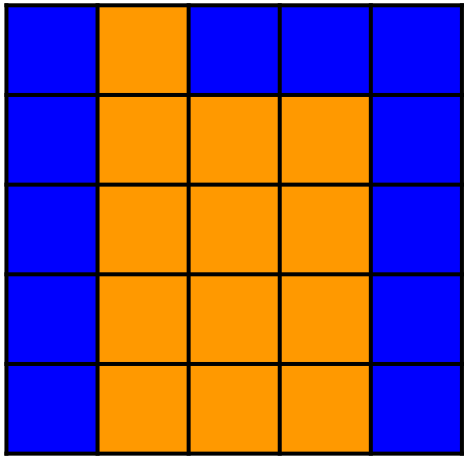

Input Image
448x448x3
DarkNet Architecture
7x7x1024
fully connected
4096
fully connected
7x7x30
x B times
x C times
(x, y, w, h, obj score)
class probability
length: 5B+C

YOLO (2015)





Bounding-box
+ Confidence



Class Probability
Map

Loss Function

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

Loss Function

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

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

Loss Function

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

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

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

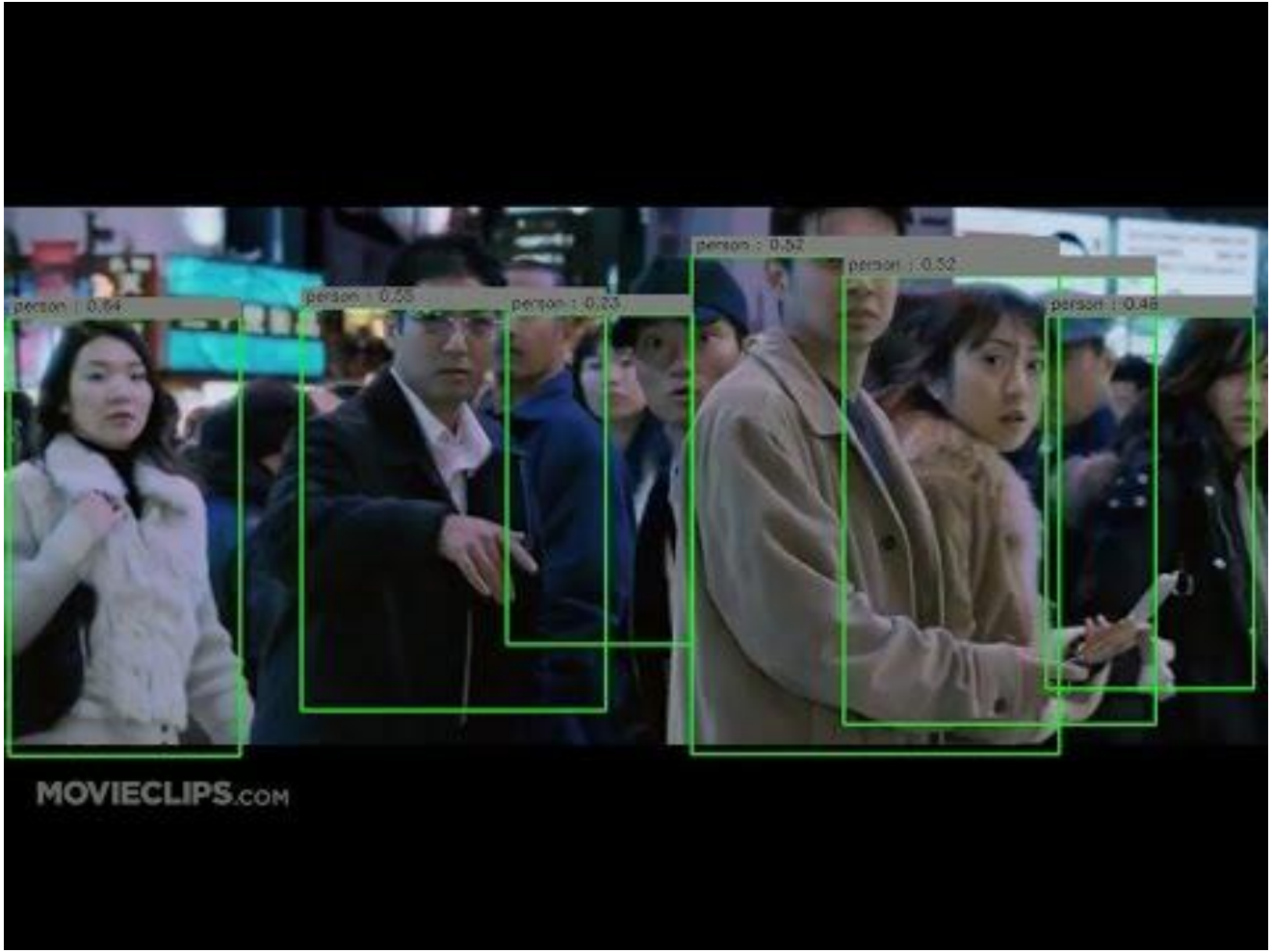
Loss Function

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

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

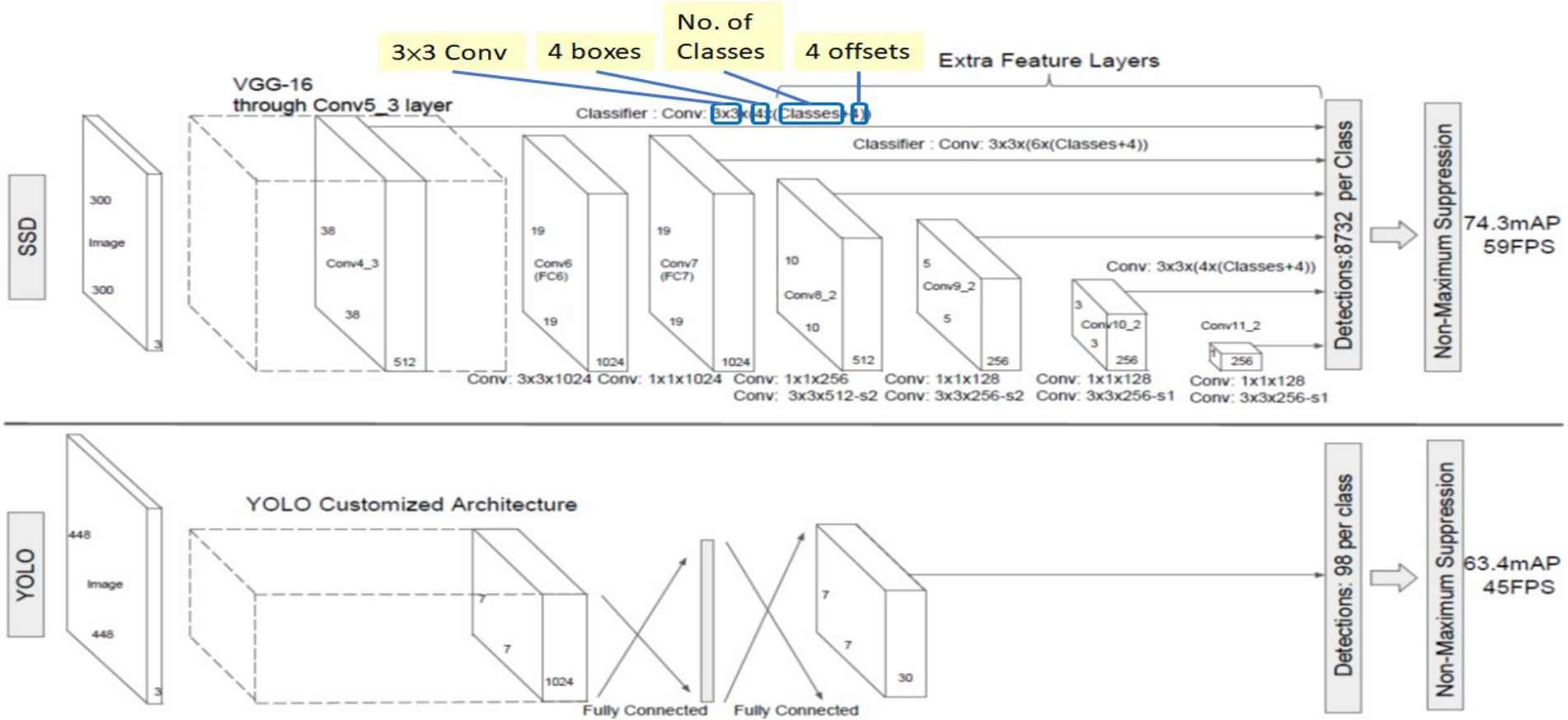
$$\sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} (C_i - \hat{C}_i)^2 + \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{noobj} (C_i - \hat{C}_i)^2$$

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

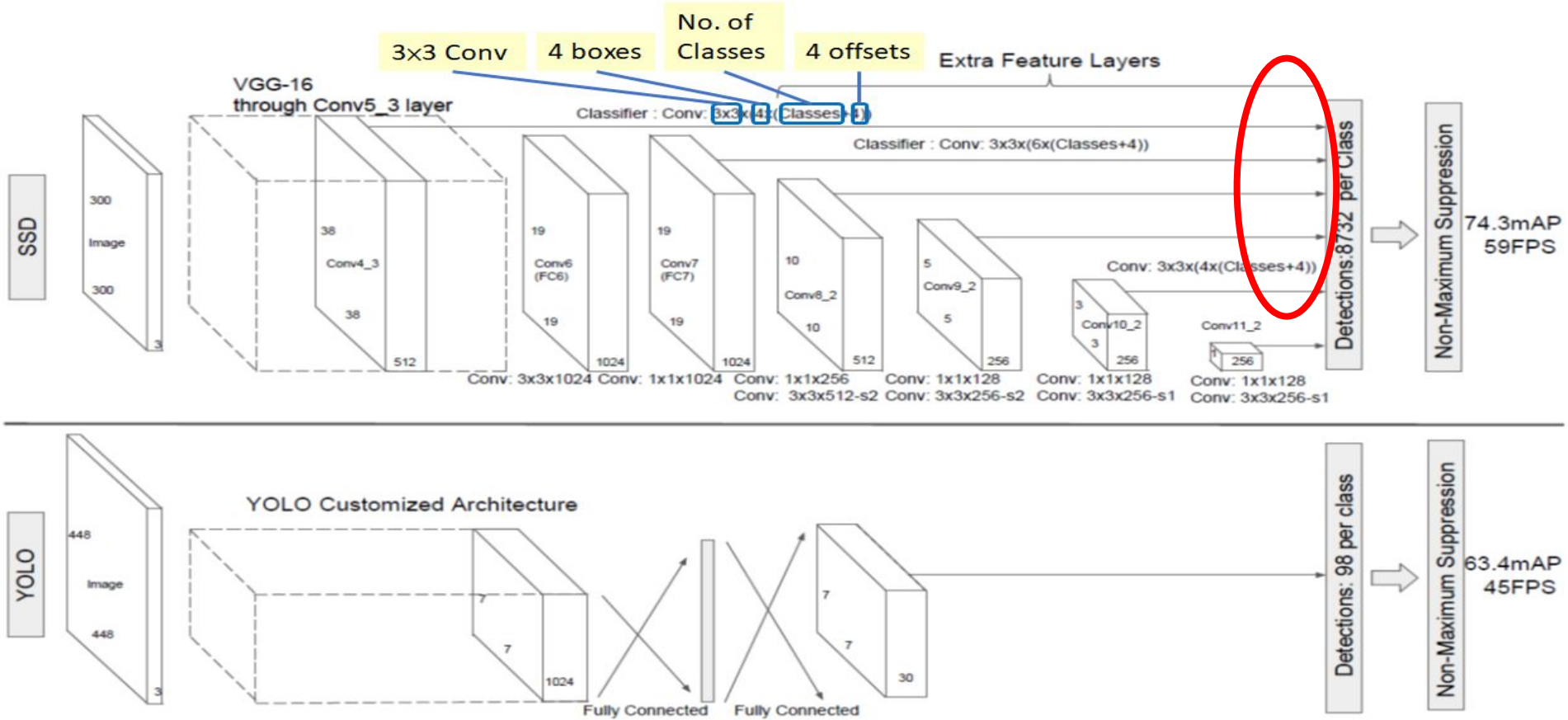

person : 0.64
person : 0.52
person : 0.52
person : 0.46
MOVIECLIPS.COM

SSD (Liu, 2015)

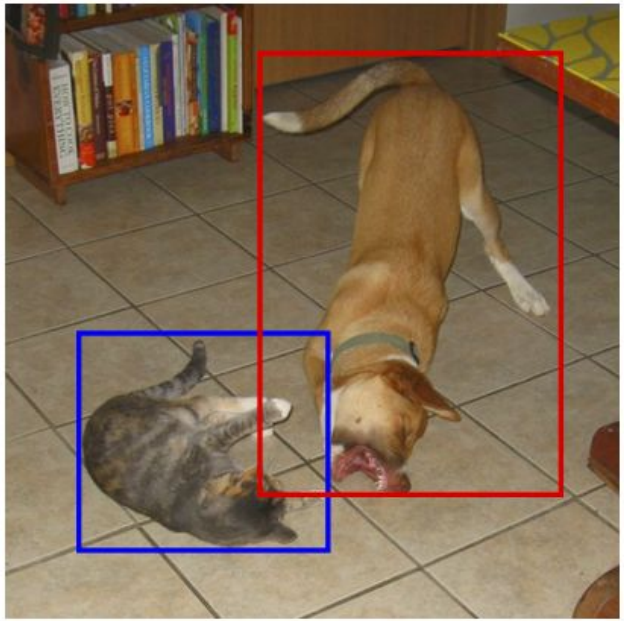
SSD (2015)


3x3 Conv
4 boxes
No. of Classes
4 offsets
Extra Feature Layers
VGG-16 through Conv5_3 layer
Classifier : Conv: 3x3x(4x(Classes+4))
Classifier : Conv: 3x3x(6x(Classes+4))
Conv: 3x3x(4x(Classes+4))
SSD
300
Image
300
3
38
Conv4_3
38
512
19
Conv6 (FC6)
19
1024
19
Conv7 (FC7)
19
1024
10
Conv8_2
10
512
5
Conv9_2
5
256
3
Conv10_2
3
256
Conv11_2
1
256
Detections:8732 per Class
Non-Maximum Suppression
74.3mAP
59FPS
Conv: 3x3x1024
Conv: 1x1x1024
Conv: 1x1x256
Conv: 3x3x512-s2
Conv: 1x1x128
Conv: 3x3x256-s2
Conv: 1x1x128
Conv: 3x3x256-s1
Conv: 1x1x128
Conv: 3x3x256-s1
YOLO
448
Image
448
3
YOLO Customized Architecture
7
7
1024
Fully Connected
Fully Connected
7
7
30
Detections: 98 per class
Non-Maximum Suppression
63.4mAP
45FPS

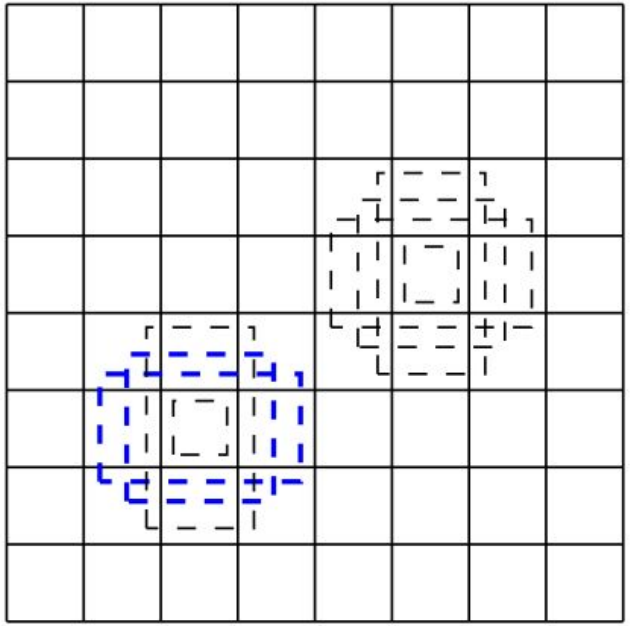
SSD (2015)


3x3 Conv
4 boxes
No. of Classes
4 offsets
Extra Feature Layers
VGG-16 through Conv5_3 layer
Classifier : Conv: 3x3x(4x(Classes+4))
Classifier : Conv: 3x3x(6x(Classes+4))
Conv: 3x3x(4x(Classes+4))
SSD
300
Image
300
3
38
Conv4_3
38
512
19
Conv6 (FC6)
19
1024
19
Conv7 (FC7)
19
1024
10
Conv8_2
10
512
5
Conv9_2
5
256
3
Conv10_2
3
256
Conv11_2
1
256
Detections:8732 per Class
Non-Maximum Suppression
74.3mAP
59FPS
Conv: 3x3x1024
Conv: 1x1x1024
Conv: 1x1x256
Conv: 3x3x512-s2
Conv: 1x1x128
Conv: 3x3x256-s2
Conv: 1x1x128
Conv: 3x3x256-s1
Conv: 1x1x128
Conv: 3x3x256-s1
YOLO
448
Image
448
3
YOLO Customized Architecture
7
7
1024
Fully Connected
Fully Connected
7
7
30
Detections: 98 per class
Non-Maximum Suppression
63.4mAP
45FPS

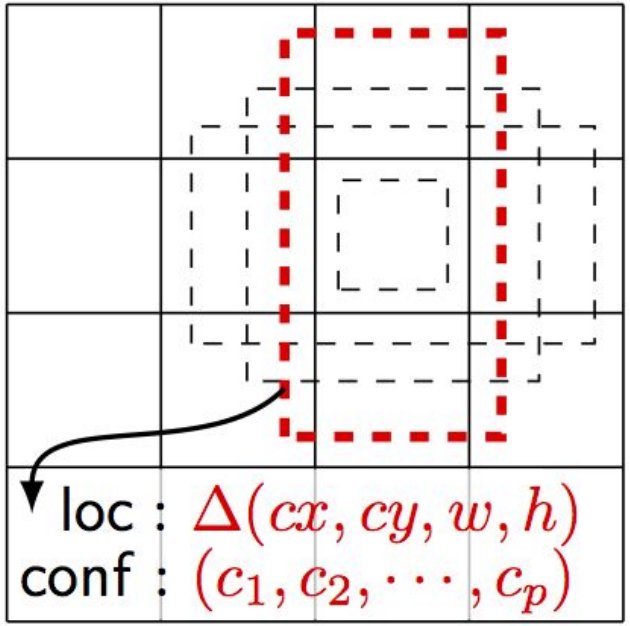
SSD (2015)



(a) Image with GT boxes



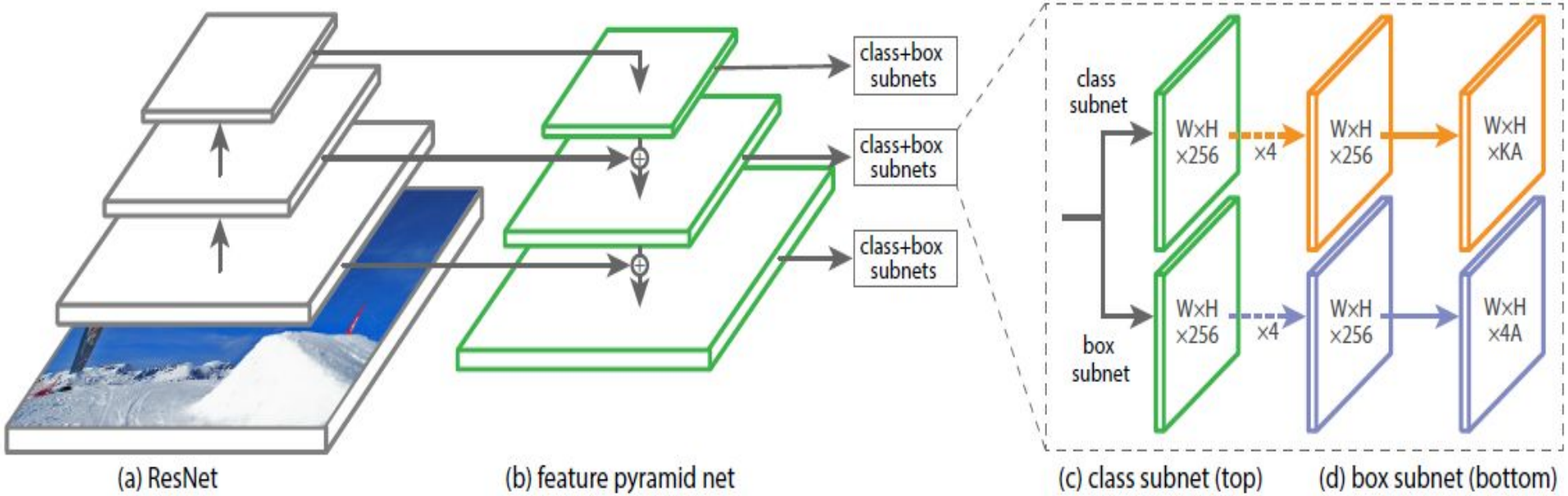
(b) 8×8 feature map



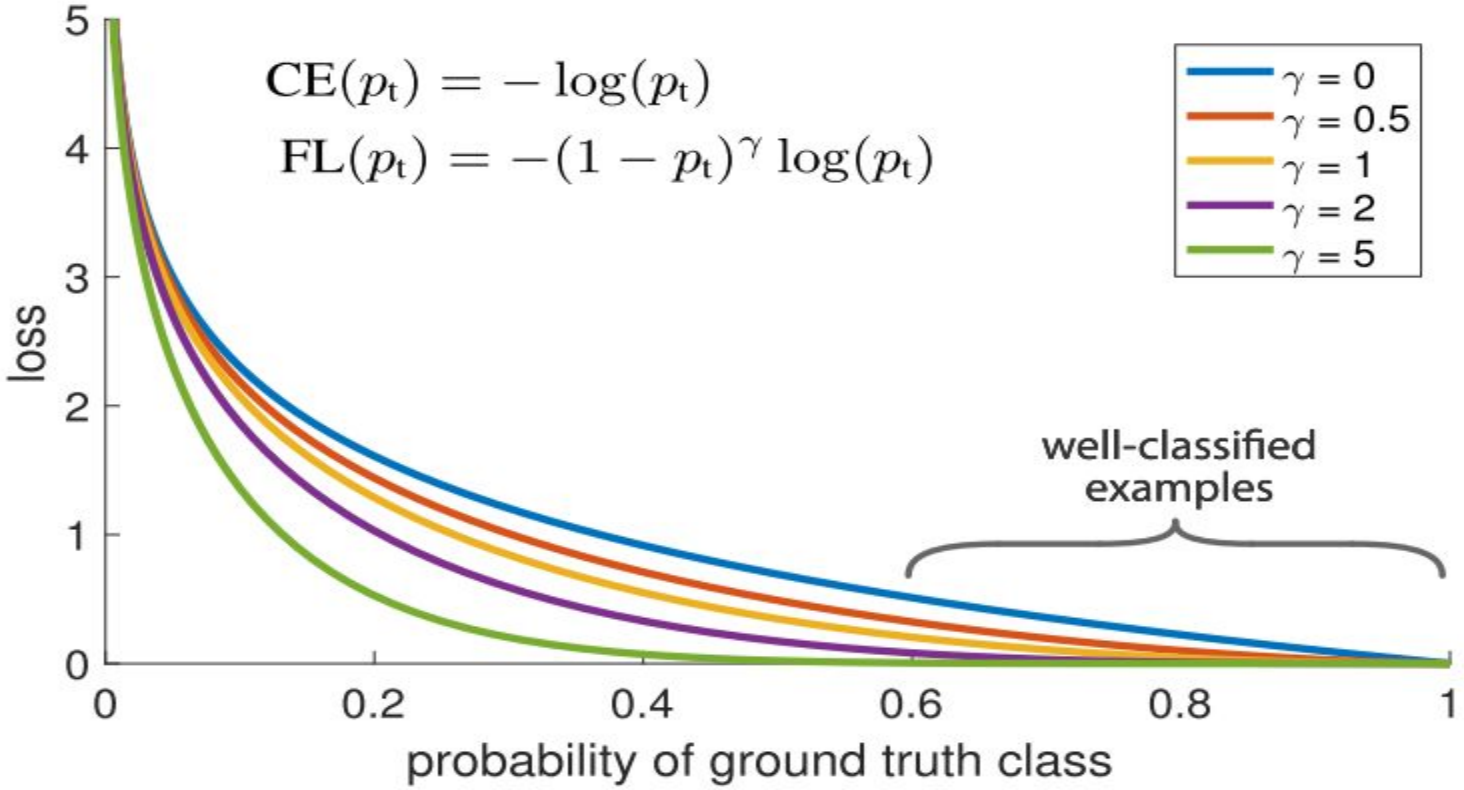
(c) 4×4 feature map

Retinanet (Lin, 2017)

RetinaNet (2017)

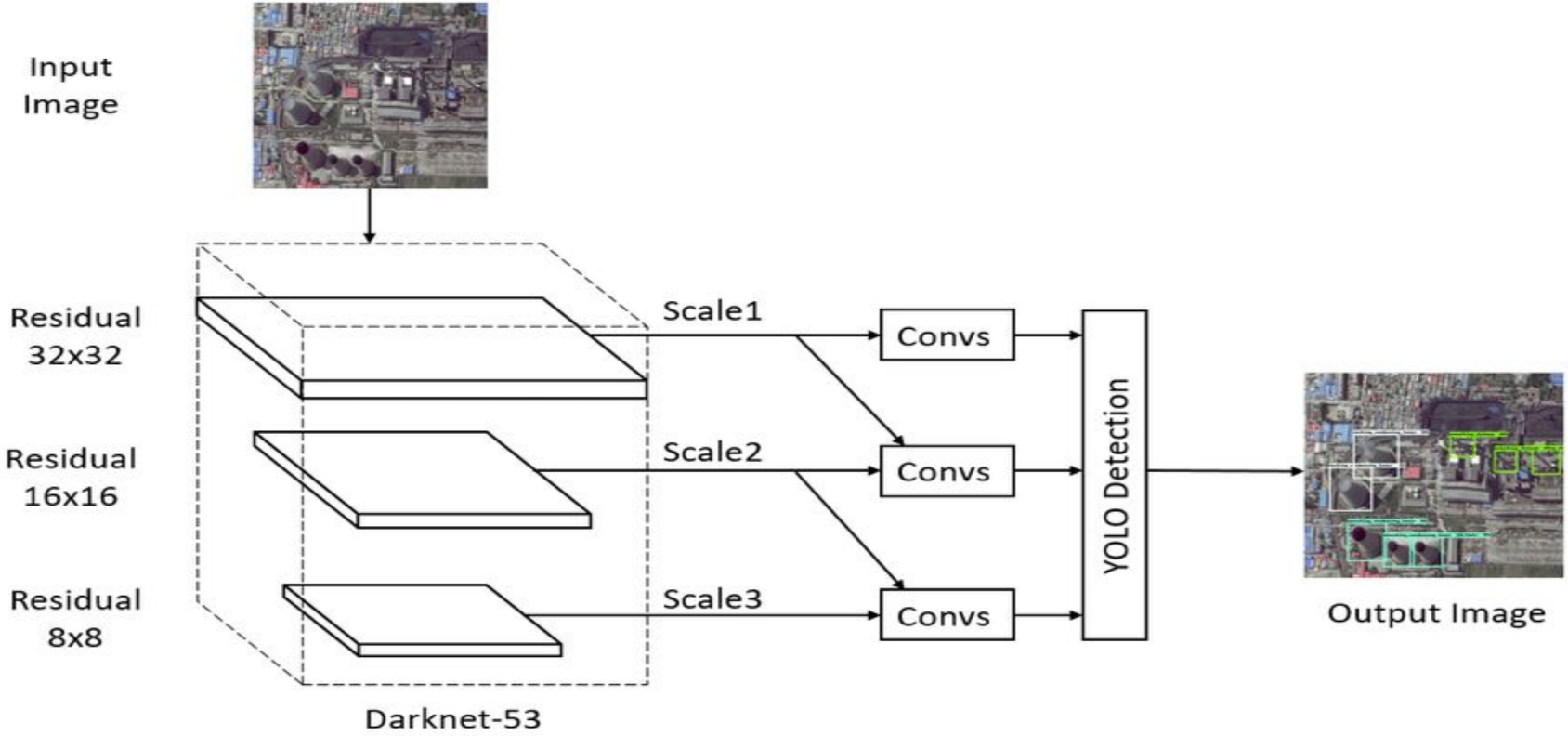


RetinaNet (2017)

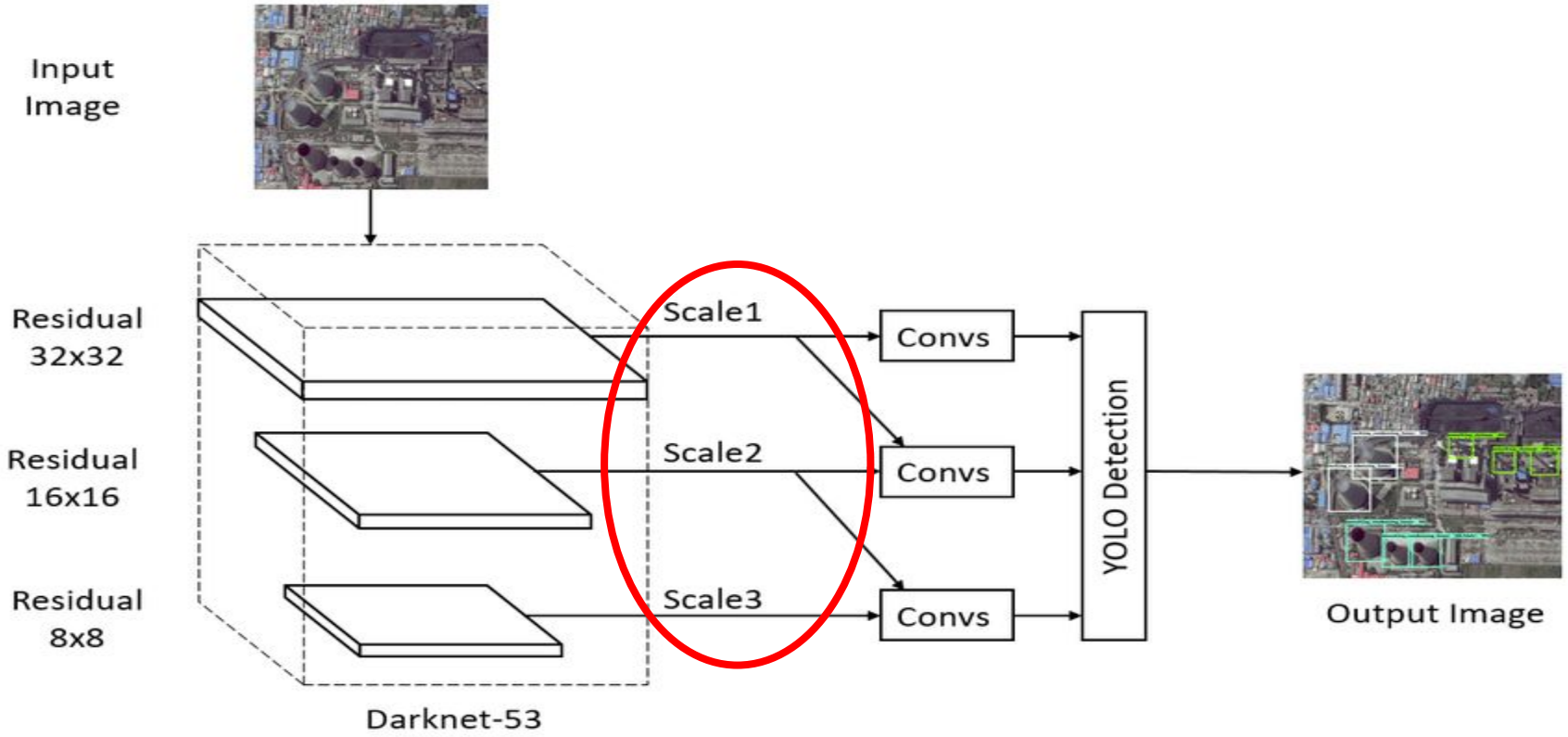


YOLOv3 (Redmon, 2018)

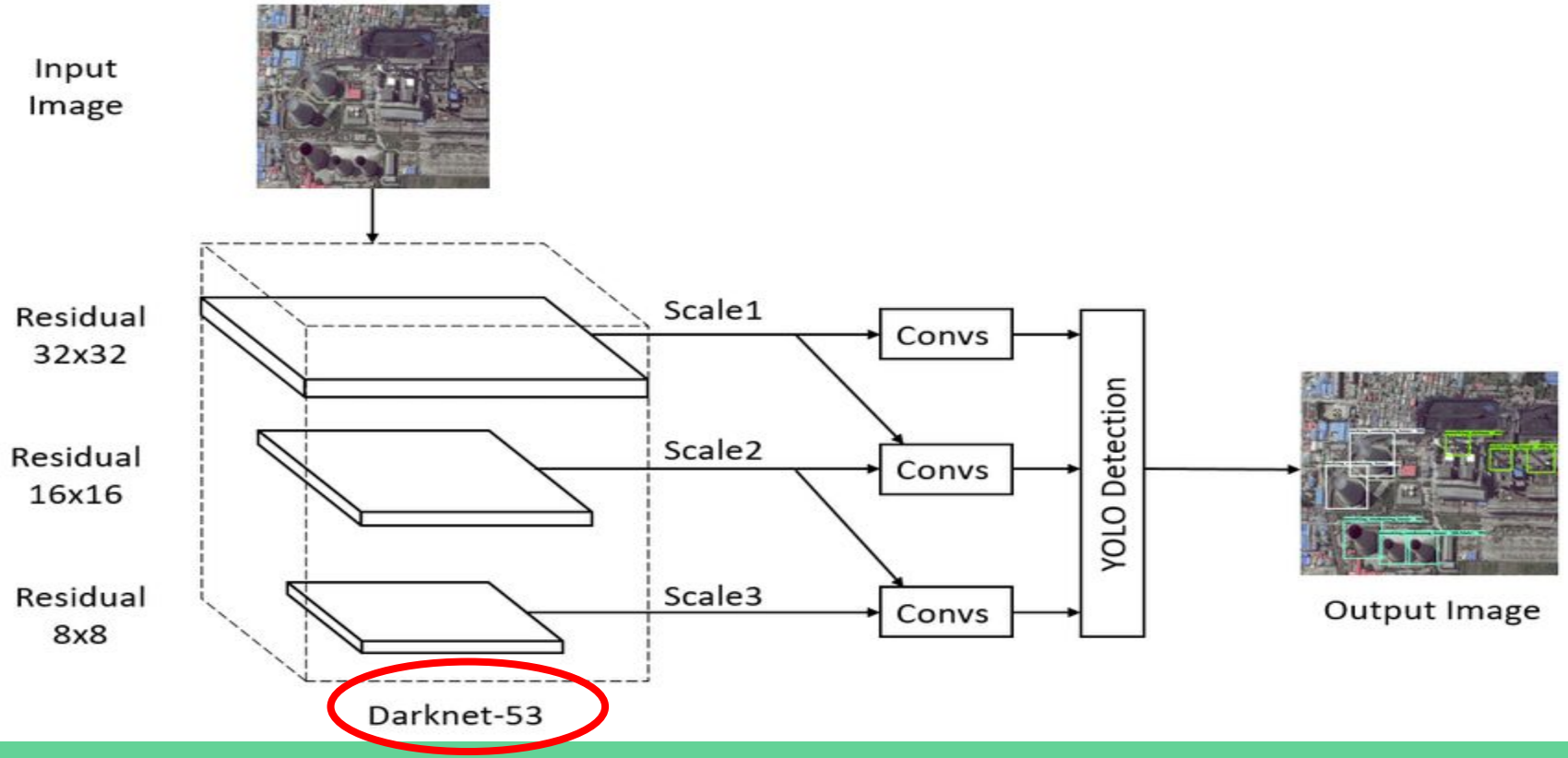
YOLOv3 (2018)


Input Image
Residual 32x32
Residual 16x16
Residual 8x8
Darknet-53
Scale1
Scale2
Scale3
Convs
Convs
Convs
YOLO Detection
Output Image

YOLOv3 (2018)


Input Image
Residual 32x32
Residual 16x16
Residual 8x8
Darknet-53
Scale1
Scale2
Scale3
Convs
Convs
Convs
YOLO Detection
Output Image

YOLOv3 (2018)


Input Image
Residual 32x32
Residual 16x16
Residual 8x8
Darknet-53
Scale1
Scale2
Scale3
Convs
Convs
Convs
YOLO Detection
Output Image

Trabajos Más Recientes

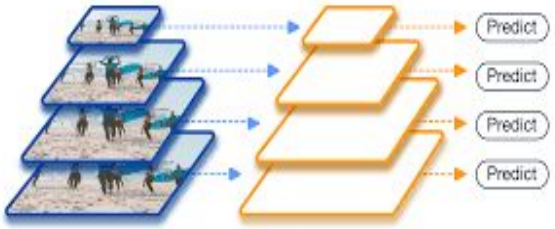


Imagen piramidal

Trabajos Más Recientes

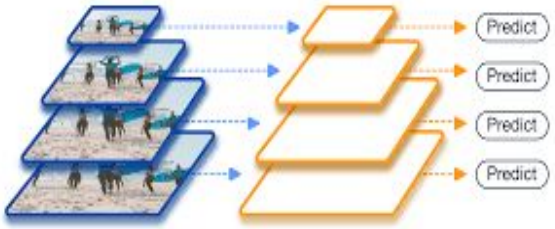
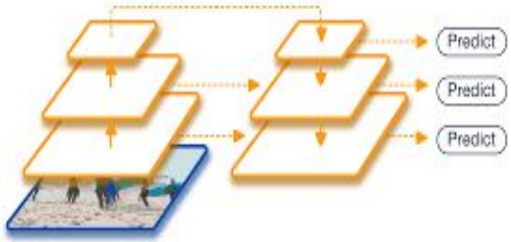


Imagen piramidal



Característica piramidal

Trabajos Más Recientes

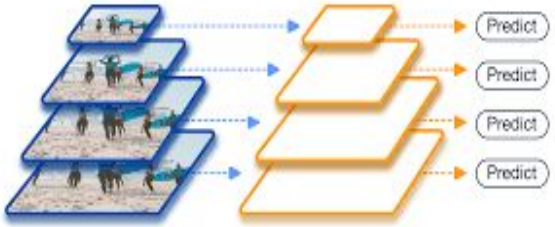
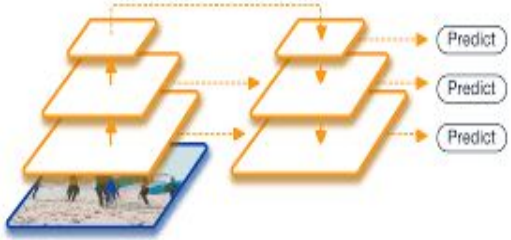
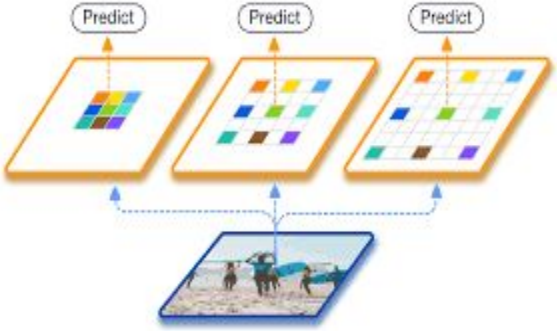


Imagen piramidal

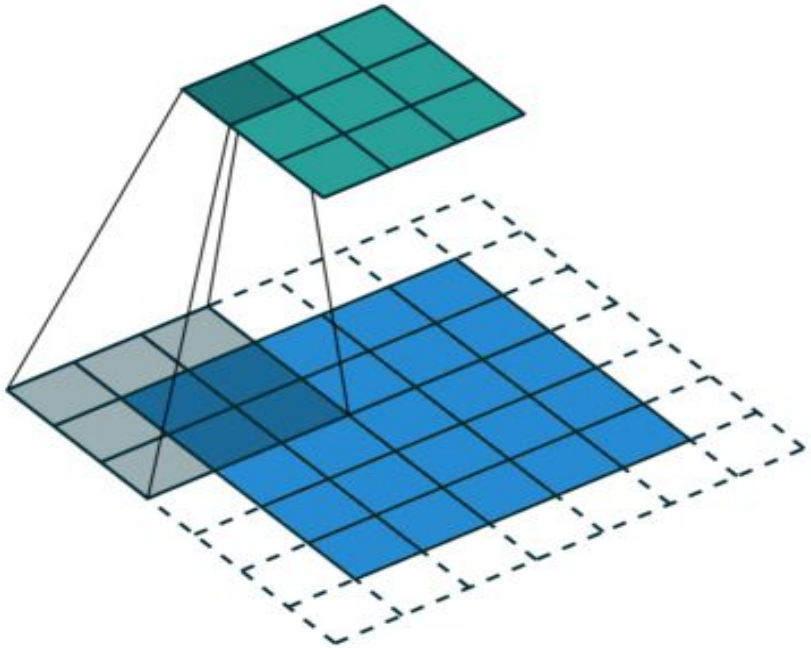


Característica piramidal

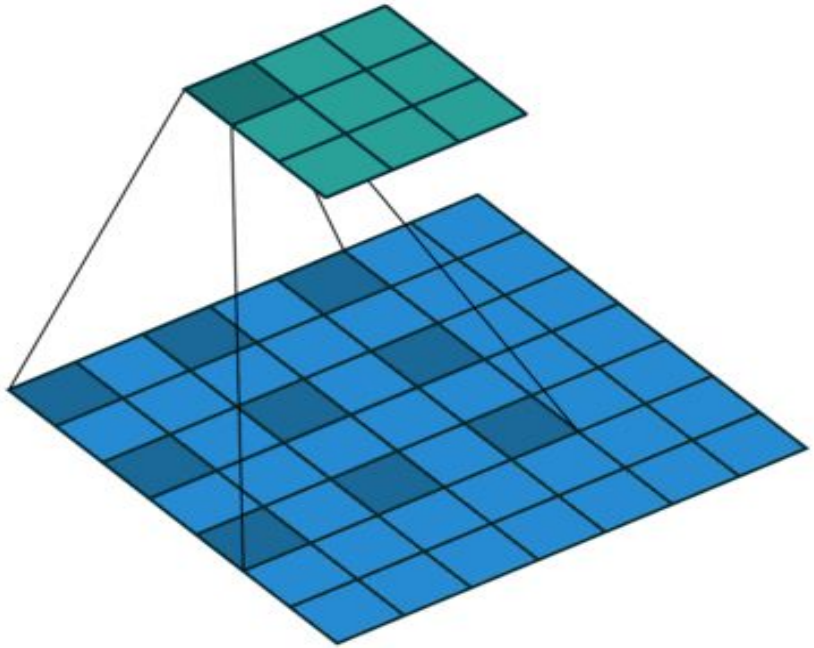


Arquitectura Tridente

TridentNet(2019*)



Convolución estándar



Convolución extendida

Bibliografía

1. Girshick, R. B., Donahue, J., Darrell, T., & Malik, J. (2013). Rich feature hierarchies for accurate object detection and semantic segmentation. CoRR, abs/1311.2524.
2. Uijlings, Jasper & Sande, K. & Gevers, T. & Smeulders, Arnold. (2013). Selective Search for Object Recognition. International Journal of Computer Vision. 104. 154-171.
3. Ren, S., He, K., Girshick, R. B., & Sun, J. (2015). Faster R-CNN: towards real-timeobject detection with region proposal networks. CoRR, abs/1506.01497.
4. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & Lecun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional net-works.
5. Redmon, J., Divvala, S. K., Girshick, R. B., & Farhadi, A. (2015). You only look once: Unified, real-time object detection. CoRR, abs/1506.02640.
6. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C., & Berg, A. C. 71(2015). SSD: single shot multibox detector. CoRR, abs/1512.02325.
7. Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. CoRR,abs/1804.02767.
8. Lin, T., Goyal, P., Girshick, R. B., He, K., & Dollár, P. (2017). Focal loss for dense object detection. CoRR, abs/1708.02002
9. Li, Yanghao & Chen, Yuntao & Wang, Naiyan & Zhang, Zhaoxiang. (2019). Scale-Aware Trident Networks for Object Detection.