上海交通大学

计算机视觉

教师：赵旭

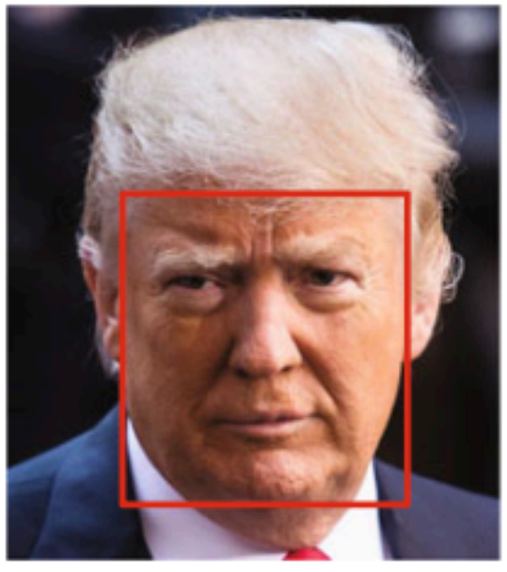
班级：AI4701

2024 春

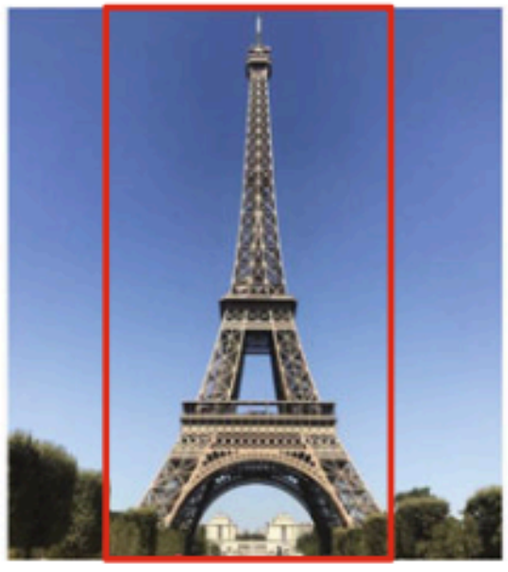
17. 物体检测

内容

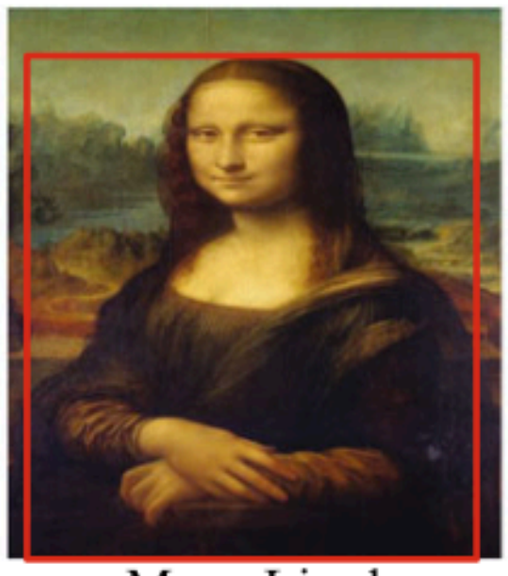
- 传统方法
- 数据集与评估准则
- 基于深度学习的方法
- 关键问题



Donald Trump's face



Eiffel Tower



Mona Lisa by
Leonardo da Vinci

My neigbour's dog

Specific Objects

Car

Car

Cat

Cat

Cat

Cat

Generic Object Categories

Li Liu et al. ***Deep Learning for Generic Object Detection: A Survey*****. IJCV, 2019**

Technique Evolution

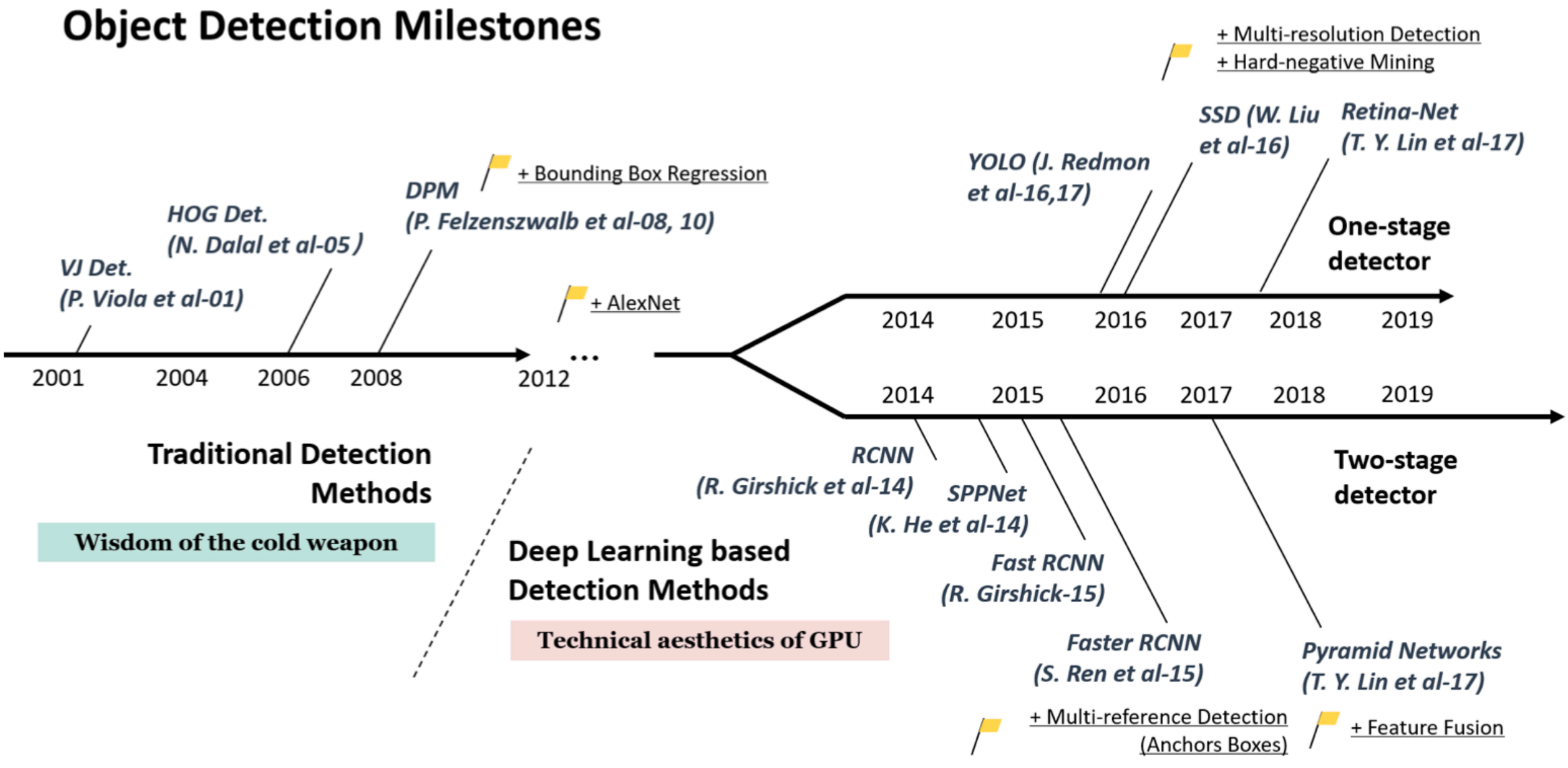
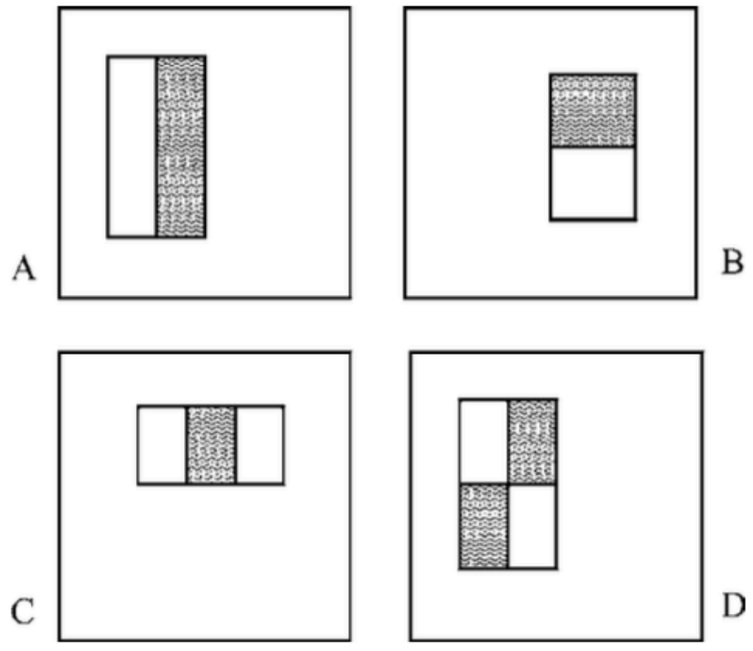


Fig. 2. A road map of object detection. Milestone detectors in this figure: VJ Det. [10, 11], HOG Det. [12], DPM [13–15], RCNN [16], SPPNet [17], Fast RCNN [18], Faster RCNN [19], YOLO [20], SSD [21], Pyramid Networks [22], Retina-Net [23].

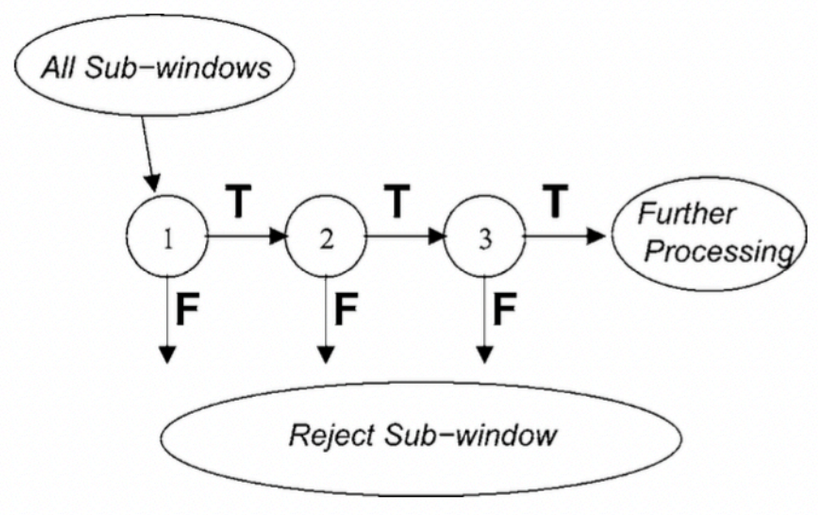
Zhengxia Zou et al. *Object Detection in 20 Years: A Survey*. **arXiv, 2019**

Traditional detectors - Viola Jones Detectors

- Sliding windows: to go through all possible locations and scales in an image to see if any window contains a human face

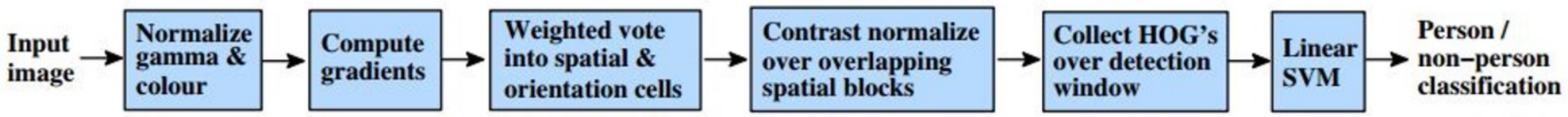


- Three contributions
  - Integral image
  - Feature selection
  - Detection cascades



Paul Viola & Michael J. Jones, Robust Real-Time Face Detection, IJCV, 2004

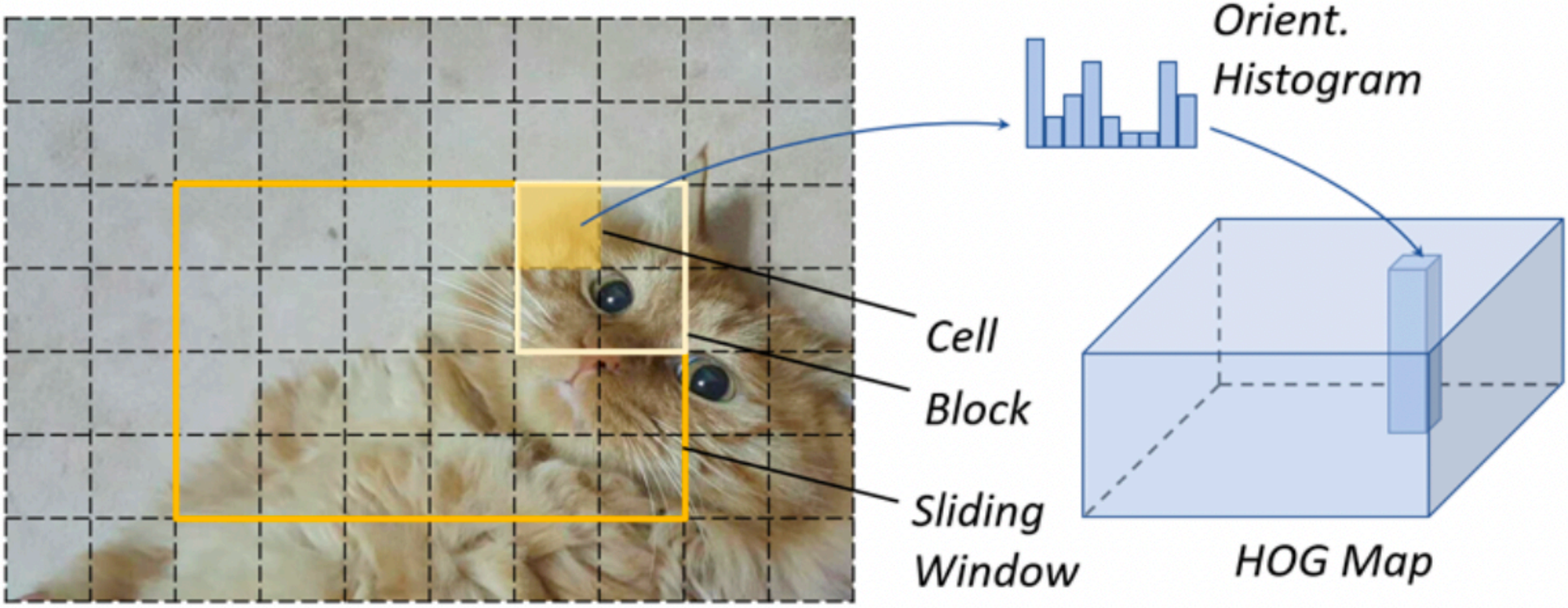
Traditional detectors - HOG



- Sliding Window (with HOG feature extraction) + Classifier (SVM)
- Designed to be computed on a dense grid of uniformly spaced cells and use overlapping local contrast normalization for improving accuracy.
- Rescale the input image for multiple times while keeping the size of a detection window unchanged.

Dalal, Navneet and Triggs, Bill, *Histograms of oriented gradients for human detection*. CVPR, 2005.

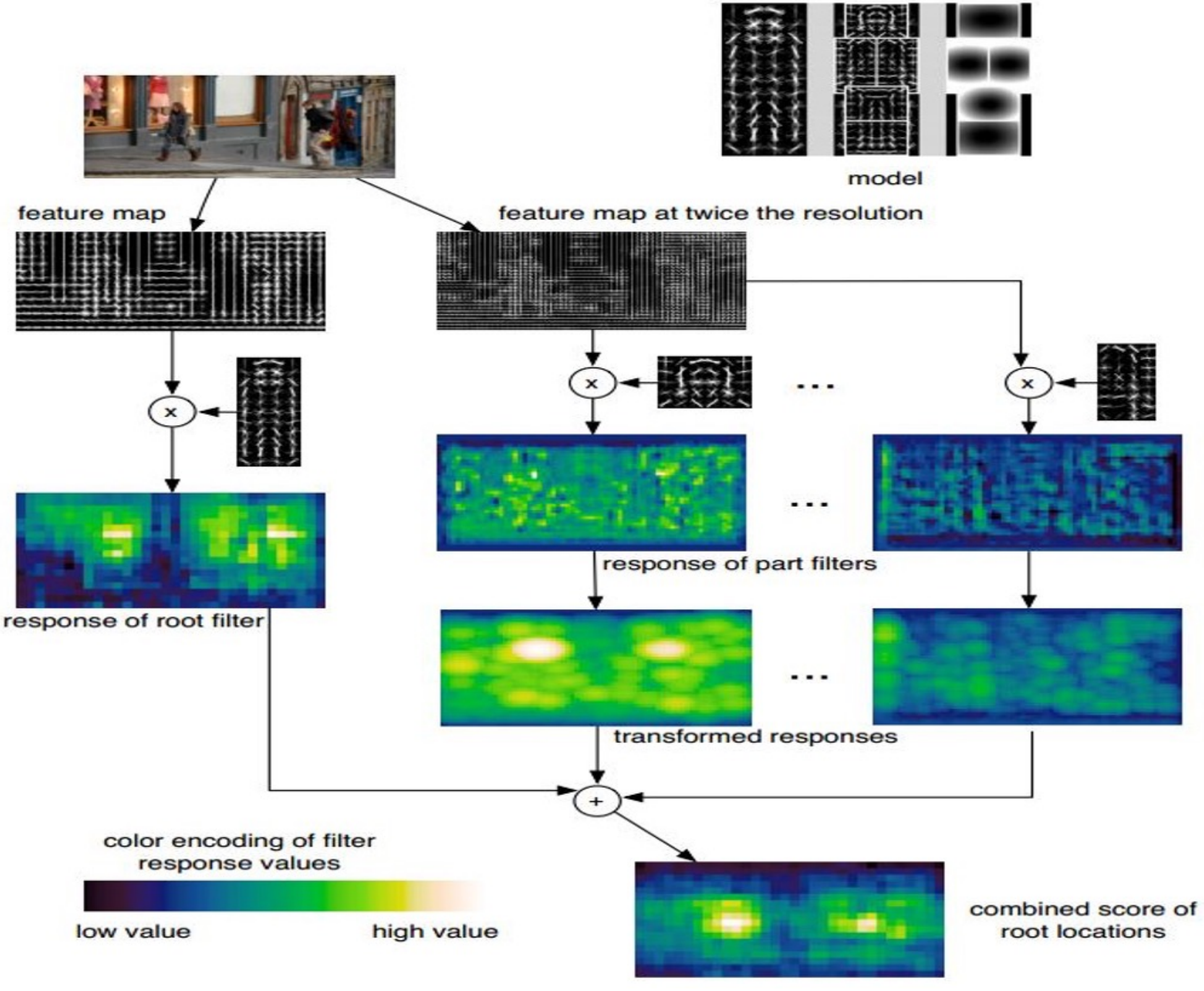
From HOG to HOG Map



Zhengxia Zou et al. *Object Detection in 20 Years: A Survey*. arXiv, 2019

Traditional detectors - DPM

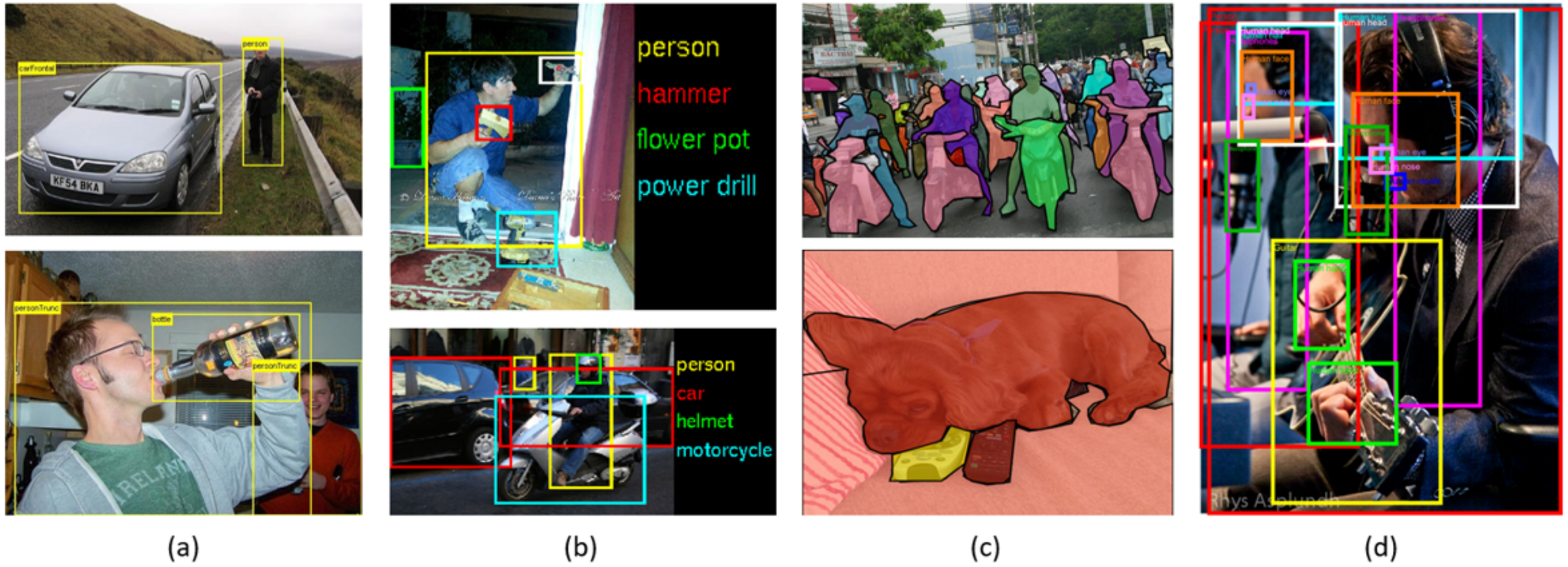
- DPM is the most successful traditional detector in recent years and is firstly proposed in 2008.
- It won the first place in Pascal VOC object detection challenge in 07,08,09.
- Learning of a proper way of decomposing an object, and the inference can be considered as an ensemble of detections on different object parts.
- A typical DPM detector consists of a root-filter and a number of part-filters.
- Propose some insightful techs: mixture models, hard negative mining, bounding box regression.


model
feature map
feature map at twice the resolution
x
x
x
response of part filters
response of root filter
transformed responses
+
color encoding of filter
response values
low value
high value
combined score of
root locations

Datasets

Dataset	train		validation		trainval		test	
	images	objects	images	objects	images	objects	images	objects
VOC-2007	2,501	6,301	2,510	6,307	5,011	12,608	4,952	14,976
VOC-2012	5,717	13,609	5,823	13,841	11,540	27,450	10,991	-
ILSVRC-2014	456,567	478,807	20,121	55,502	476,688	534,309	40,152	-
ILSVRC-2017	456,567	478,807	20,121	55,502	476,688	534,309	65,500	-
MS-COCO-2015	82,783	604,907	40,504	291,875	123,287	896,782	81,434	-
MS-COCO-2018	118,287	860,001	5,000	36,781	123,287	896,782	40,670	-
OID-2018	1,743,042	14,610,229	41,620	204,621	1,784,662	14,814,850	125,436	625,282

TABLE 1
Some well-known object detection datasets and their statistics.

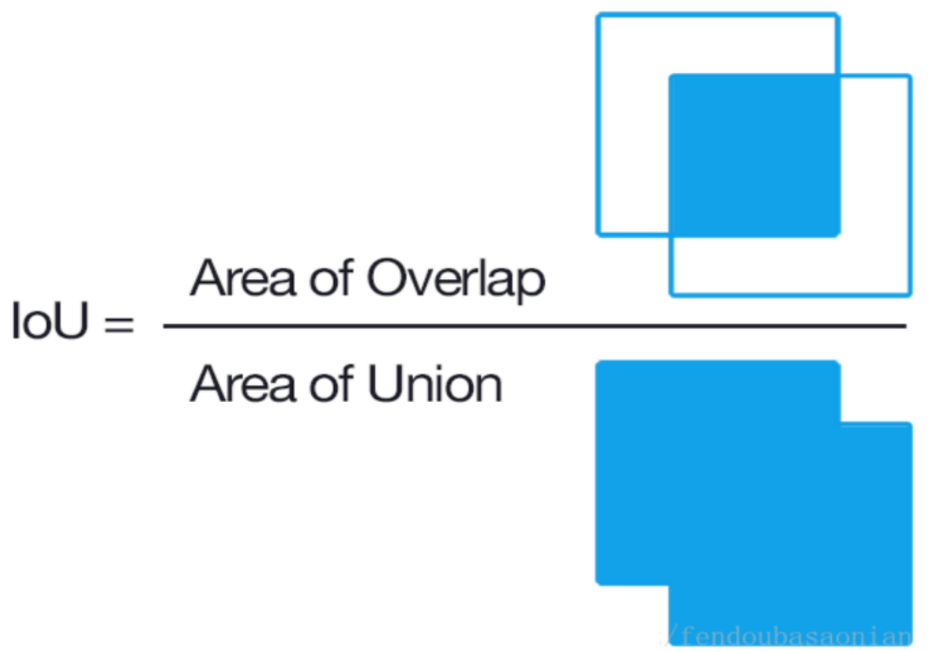


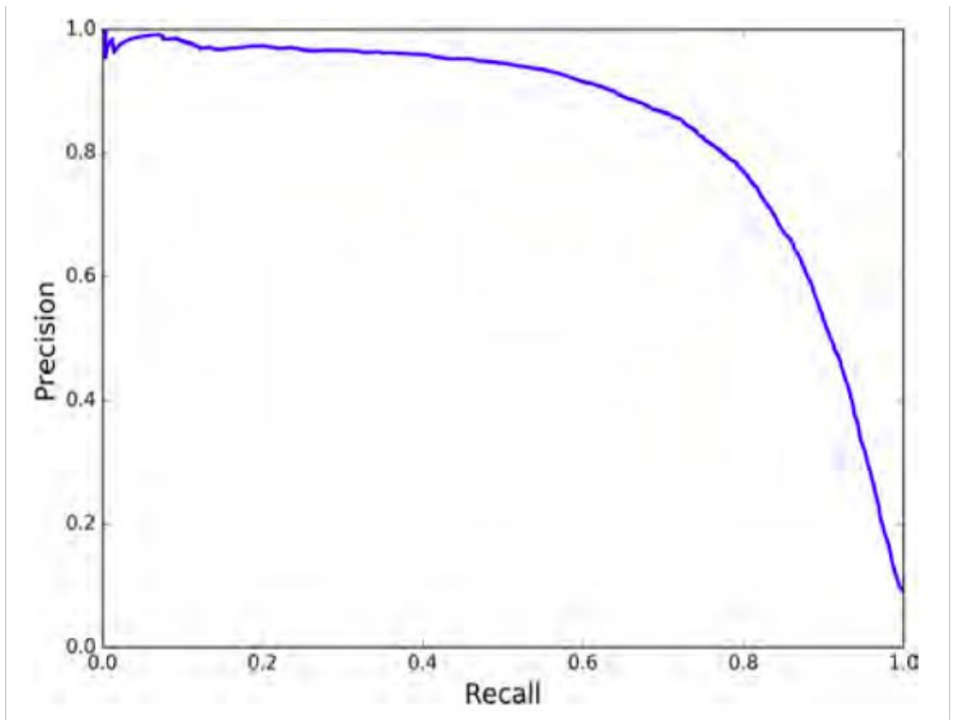
Some example images and annotations in (a) PASCAL-VOC07, (b) ILSVRC, (c) MS-COCO, and (d) Open Images.

Zhengxia Zou et al. *Object Detection in 20 Years: A Survey*. arXiv, 2019

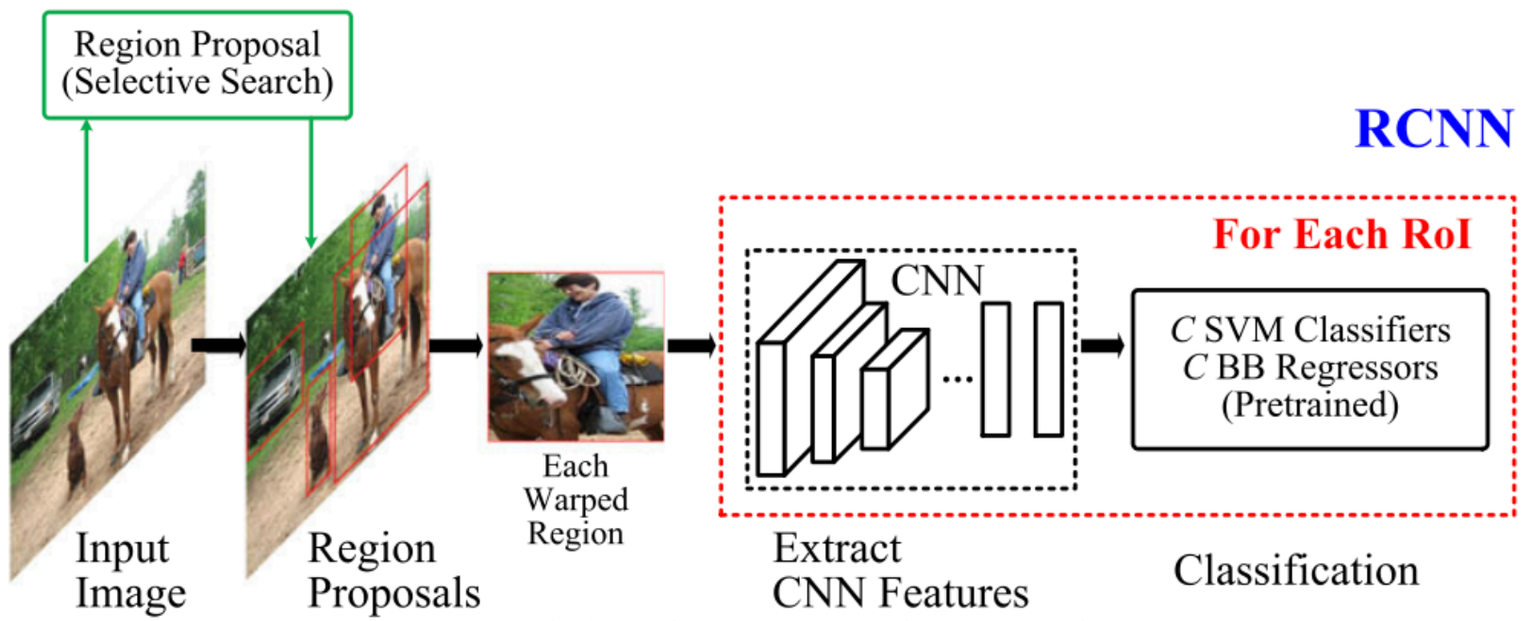
Evaluation metrics

- How fast?
  - FPS - Frames Per Second
- How good?
  - Localization: **IoU** - measure **correctness (positiveness)** of a detection bounding box.
  - Classification: **confidence** represented with probability
  - Recall = TP / (TP+FN)
  - Precision = TP / (TP+FP)
  - AP - average detection precision under different recalls (average over different IoUs)
  - mAP - average over all the object categories





CNN based two-stage detectors - RCNN



Li Liu et al. *Deep Learning for Generic Object Detection: A Survey*. IJCV, 2019

- Use the traditional selective search to obtain the proposals / RoI (Region of Interests)
- Crop the image patches within each proposal, and employ CNN to extract features.
- Use the extracted CNN features for SVM classification learning and inference.

R. Girshick et al. *Rich feature hierarchies for accurate object detection and semantic segmentation*, CVPR, 2014

Training

Testing

Stage 1: Region Proposals Computation

Stage 2: CNN Model Training and Finetune

Stage 3: Class Specific SVM Classifiers Training

Stage 4: Bounding Box Regressors Training

Testing of Input Images

Region Proposals Extraction

Preparing Data for Finetuning

Class 1

Class 2

Class *N*

Non Objects

Warped Region Proposals (Positives: RPs of IoU≥ 0.5)

Preparing Data for SVM Training

Not Planes

Not Horses

Ground Truths vs Negatives (RPs of IoU<0.3 and Backgrounds)

ImageNet

PreTraining

Finetune

Initialization

CNN Feature Extraction

Test Image

Region Proposals

Warped RPs

CNN Features

CNN Features

Bounding Box Regression 1

Bounding Box Regression *N*

SVM Classifier 1

SVM Classifier *N*

SVM Classification

Bounding Box Regression

CNN based two-stage detectors - RCNN

- Pros:
  - Introducing CNN into object detection which actives excellent performances.
- Cons:
  - Training is a multistage pipeline, slow and hard to optimize because each individual stage must be trained separately.
  - The CNN computation is redundant and time-consuming.
  - Testing is slow, since CNN features are extracted per object proposal in each test image, without shared computation.

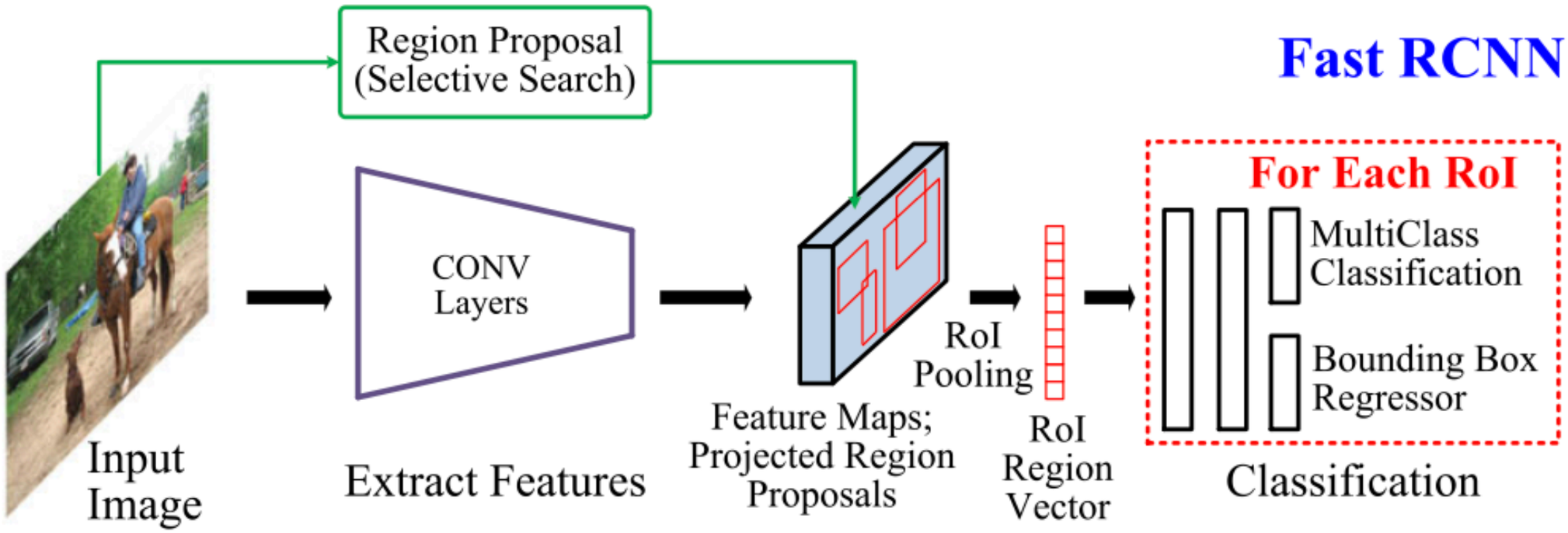
R. Girshick et al. *Rich feature hierarchies for accurate object detection and semantic segmentation*, CVPR, 2014

CNN based two-stage detectors - SPPNet

- Pros:
  - Enables a CNN to generate a fixed-length representation regardless of the size of image/region of interest.
  - The feature maps can be computed from the entire image only once
    - More than 20 times faster than R-CNN without sacrificing any detection accuracy (VOC07 mAP=59.2%).
- Cons:
  - Training still is multistage
  - Only fine-tunes its fully connected layers while simply ignores all previous layers

He, K et al. *Spatial pyramid pooling in deep convolutional networks for visual recognition. ECCV,* 2014

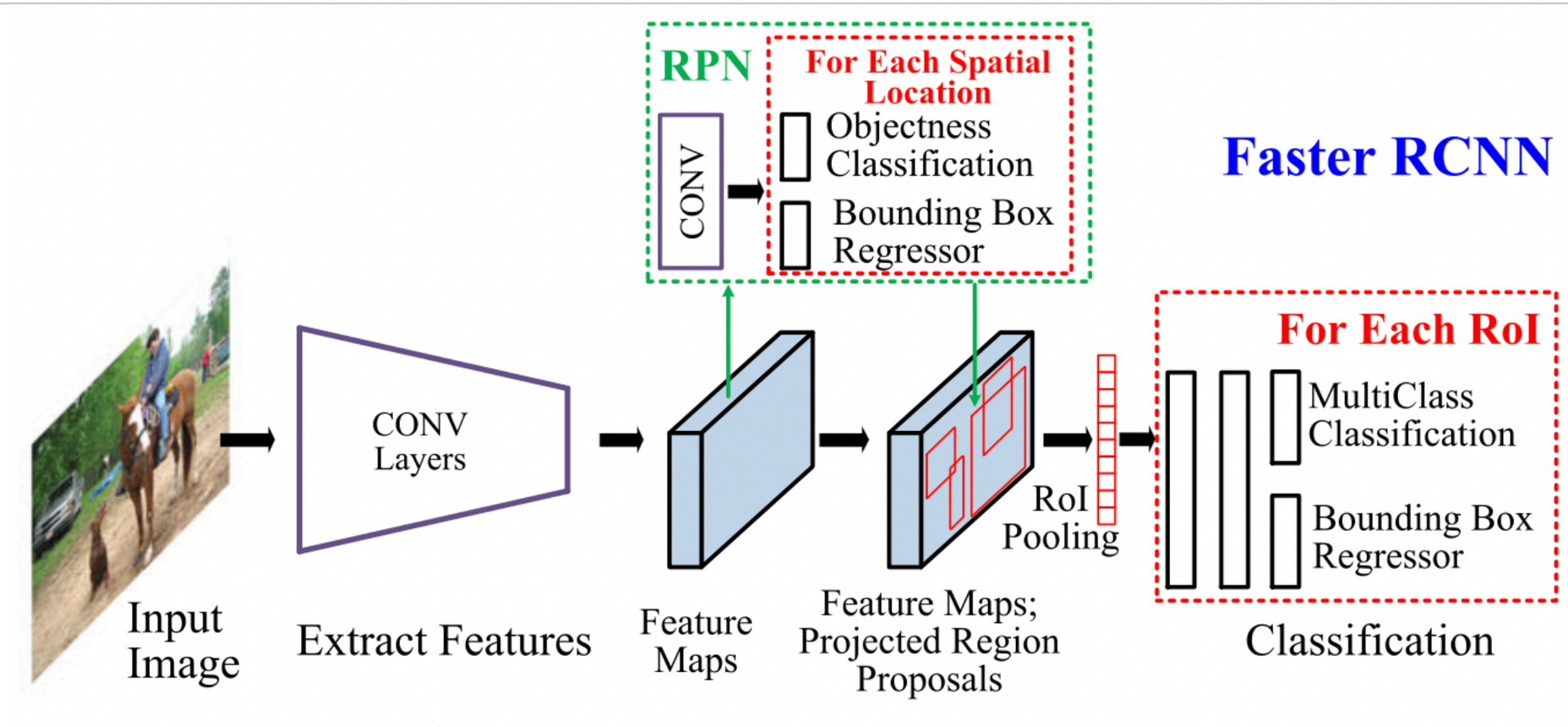
CNN based two-stage detectors - FastRCNN



- Pros:
  - Enables simultaneously training of a detector and a bounding box regressor under the same network configurations.
  - Fast RCNN increased the mAP from 58.5% (RCNN) to 70.0% while with a detection speed over 200 times faster than R-CNN.
- Cons:
  - Selective search still is used to generate proposals, so still slow.

R. Girshick et al. *Fast RCNN*, ICCV, 2015.

CNN based two-stage detectors - FasterRCNN



- Pros:
  - The selective search is replaced by a CNN in producing region proposals. An efficient and accurate Region Proposal Network (RPN) for generating region proposals.
  - **Propose the anchor mechanism**, set anchor boxes with various scales and aspect ratios at each location on the feature map and use Network to learn the classification and regression
  - ~5 FPS (including all stages) on a GPU, achieving good accuracy on PASCAL VOC 2007 using 300 proposals per image
- Cons
  - Still slow with some fully connected layers

Ren et al. Faster r-cnn: Towards real-time object detection with region proposal networks," NeurIPS 2015.

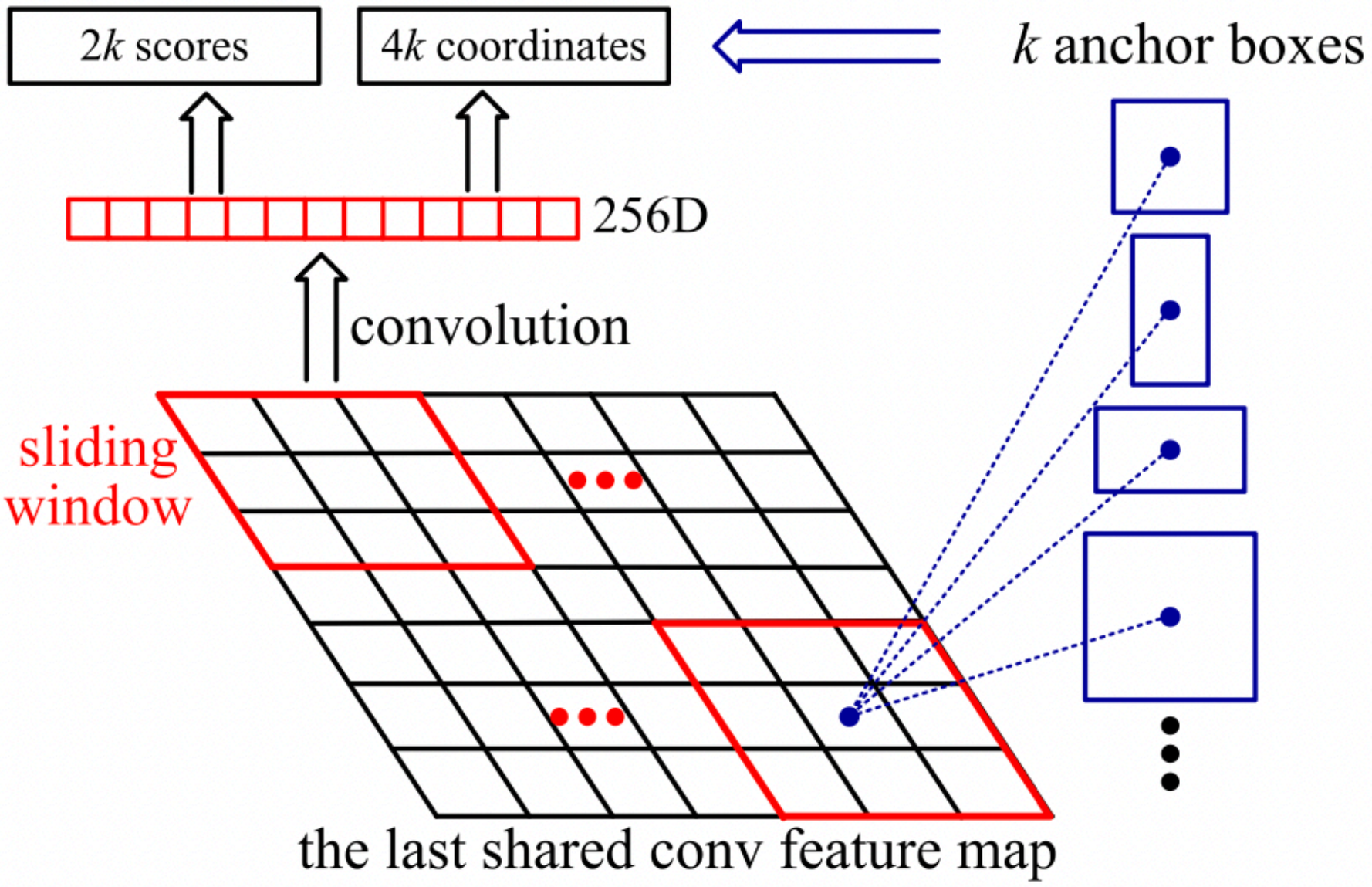
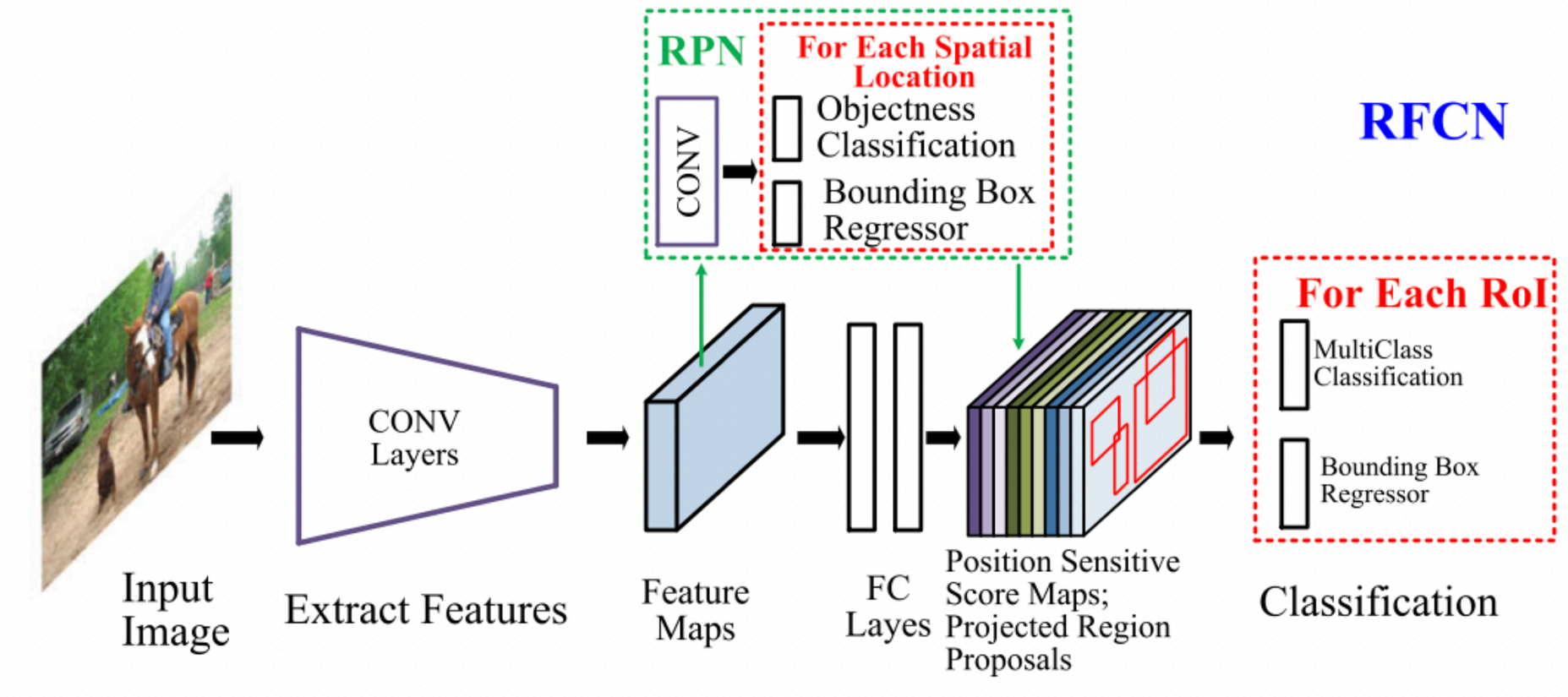


Illustration of the region proposal network (RPN)

Li Liu et al. ***Deep Learning for Generic Object Detection: A Survey*****. IJCV, 2019**

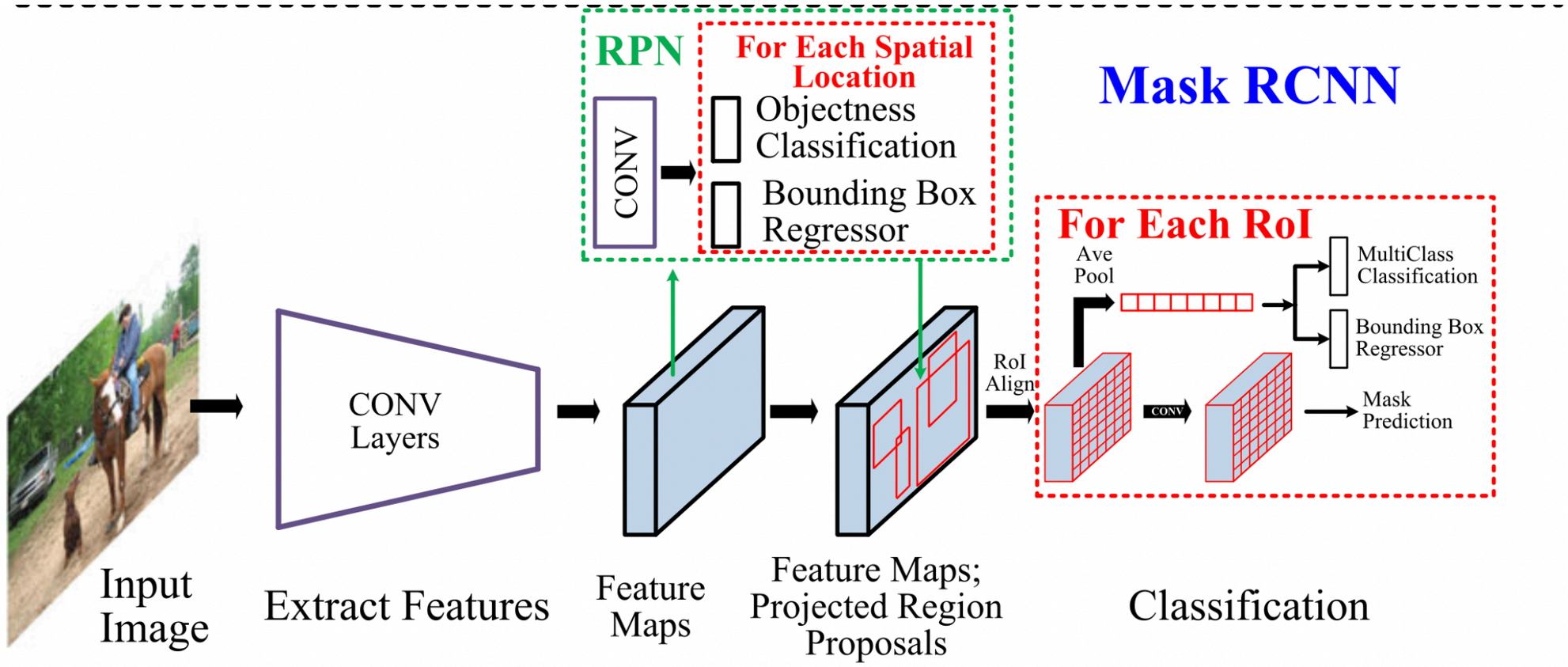
CNN based two-stage detectors - RFCN



- Pros:
  - No hidden FC layers
  - All CONV layers to construct a shared RoI sub-network

J. Dai et al. *R-fcn: Object detection via region-based fully convolutional networks*, NeurIPS 2016.

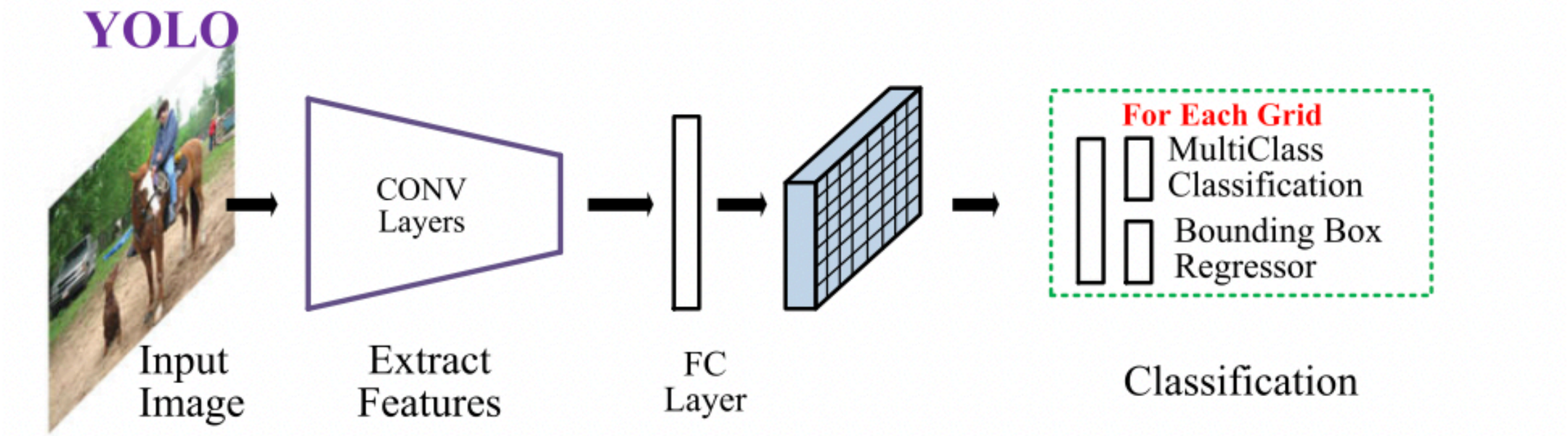
CNN based two-stage detectors - MaskRCNN



- Predict the spatial segmentation mask along with classification and bounding box regression
- Segmentation loss、classification loss and bounding box regression loss are optimized together.
- RoI Align is proposed to use interpolation operation to obtain the precise RoI pooling feature value.
- 5FPS

He, K. et al., Mask r-cnn. ICCV, 2017.

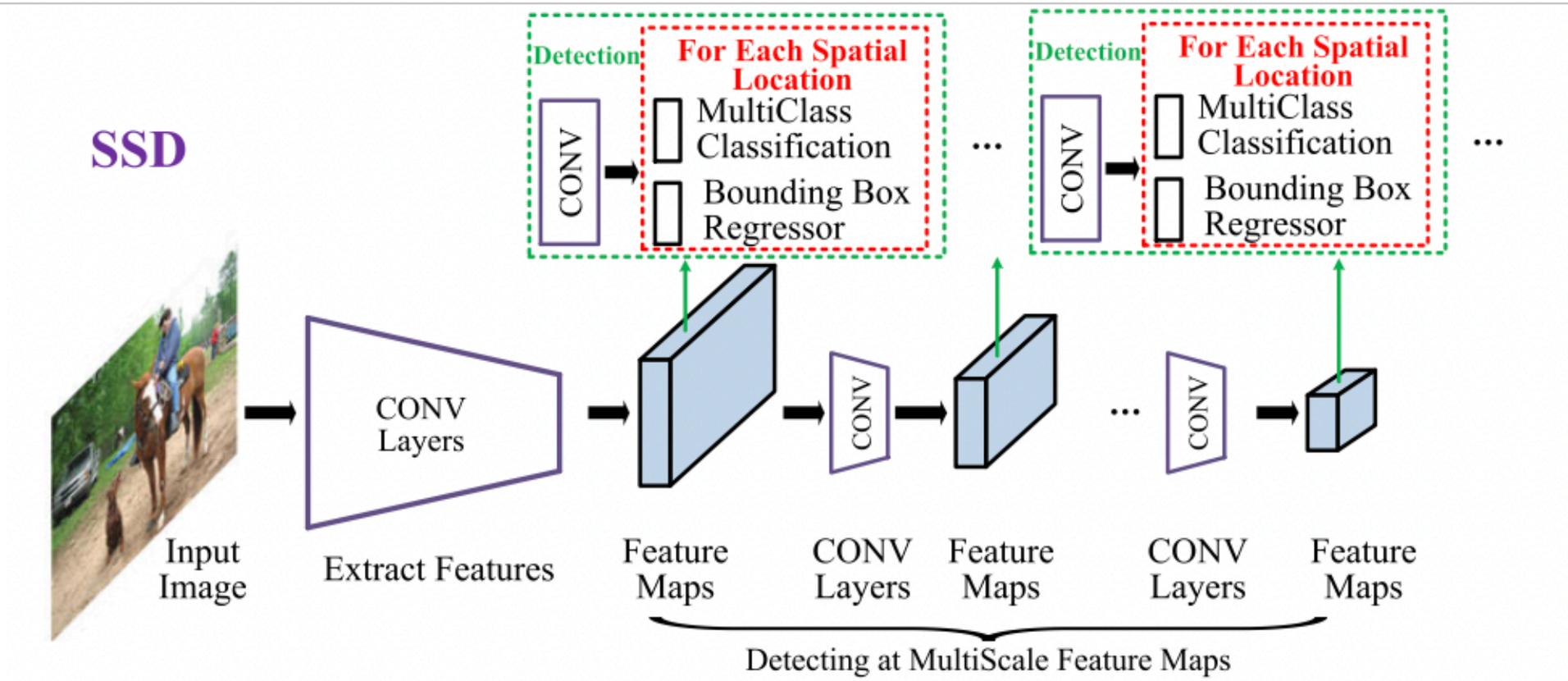
CNN based one-stage detectors - YOLO



- The first one-stage detector in deep learning era
- Directly predicts detections using a small set of candidate regions
- Divides an image into an S × S grid, each predicting C class probabilities, B bounding box locations, and confidence scores.
- 45FPS-155FPS
- YOLO-v2, v3,….v5, v8
- YOLO9000 (allows weakly supervised detection)
- Cons: corse localization

J. Redmon et al., *You only look once: Unified, real-time object detection*, CVPR, 2016.

CNN based one-stage detectors - SSD



- Performs detection over multiple scales by operating on multiple CONV feature maps, each of which predicts category scores and box offsets for bounding boxes of appropriate sizes.
- For a 300 × 300 input, SSD achieves 74.3% mAP on the VOC2007 test at 59 FPS versus Faster RCNN 7 FPS / mAP 73.2% or YOLO 45 FPS / mAP 63.4%.

Liu et al. *SSD Single Shot Multibox Detector*, ECCV, 2016.

Popular DCNN architectures

No.	DCNN architecture	#Paras ($\times 10^6$)	#Layers (CONV+FC)	Test error (Top 5)	First used in	Highlights
1	AlexNet (Krizhevsky et al. 2012b)	57	$5+2$	15.3%	Girshick et al. (2014)	The first DCNN found effective for ImageNet classification; the historical turning point from hand-crafted features to CNN; Winning the ILSVRC2012 Image classification competition
2	ZFNet (fast) (Zeiler and Fergus 2014)	58	$5+2$	14.8%	He et al. (2014)	Similar to AlexNet, different in stride for convolution, filter size, and number of filters for some layers
3	OverFeat (Sermanet et al. 2014)	140	$6+2$	13.6%	Sermanet et al. (2014)	Similar to AlexNet, different in stride for convolution, filter size, and number of filters for some layers
4	VGGNet (Simonyan and Zisserman 2015)	134	$13+2$	6.8%	Girshick (2015)	Increasing network depth significantly by stacking 3×3 convolution filters and increasing the network depth step by step
5	GoogLeNet (Szegedy et al. 2015)	6	22	6.7%	Szegedy et al. (2015)	Use Inception module, which uses multiple branches of convolutional layers with different filter sizes and then concatenates feature maps produced by these branches. The first inclusion of bottleneck structure and global average pooling
6	Inception v2 (Ioffe and Szegedy 2015)	12	31	4.8%	Howard et al. (2017)	Faster training with the introduce of batch normalization
7	Inception v3 (Szegedy et al. 2016)	22	47	3.6%		Inclusion of separable convolution and spatial resolution reduction
8	YOLONet (Redmon et al. 2016)	64	$24+1$	–	Redmon et al. (2016)	A network inspired by GoogLeNet used in YOLO detector
9	ResNet50 (He et al. 2016)	23.4	49	3.6% (ResNets)	He et al. (2016)	With identity mapping, substantially deeper networks can be learned

Li Liu et al. *Deep Learning for Generic Object Detection: A Survey*. **IJCV, 2019**

Important Issues - Scale

Important Issues - Scale

- Scale problems: “size” and “aspect ratio”
- Feature pyramids and sliding windows (2004-2014)
  - VJ\HOG\DPM
  - mixture model for variations in aspect ratio
- Object proposals: generate boxes at varies size and aspect ratio (2010-2015)
- Deep regression: directly predict the coordinates of a bounding box based on the deep features (2013-2016)
- Multi-reference/-resolution detection (after 2015)
  - Anchor box
  - Faster RCNN, SSD

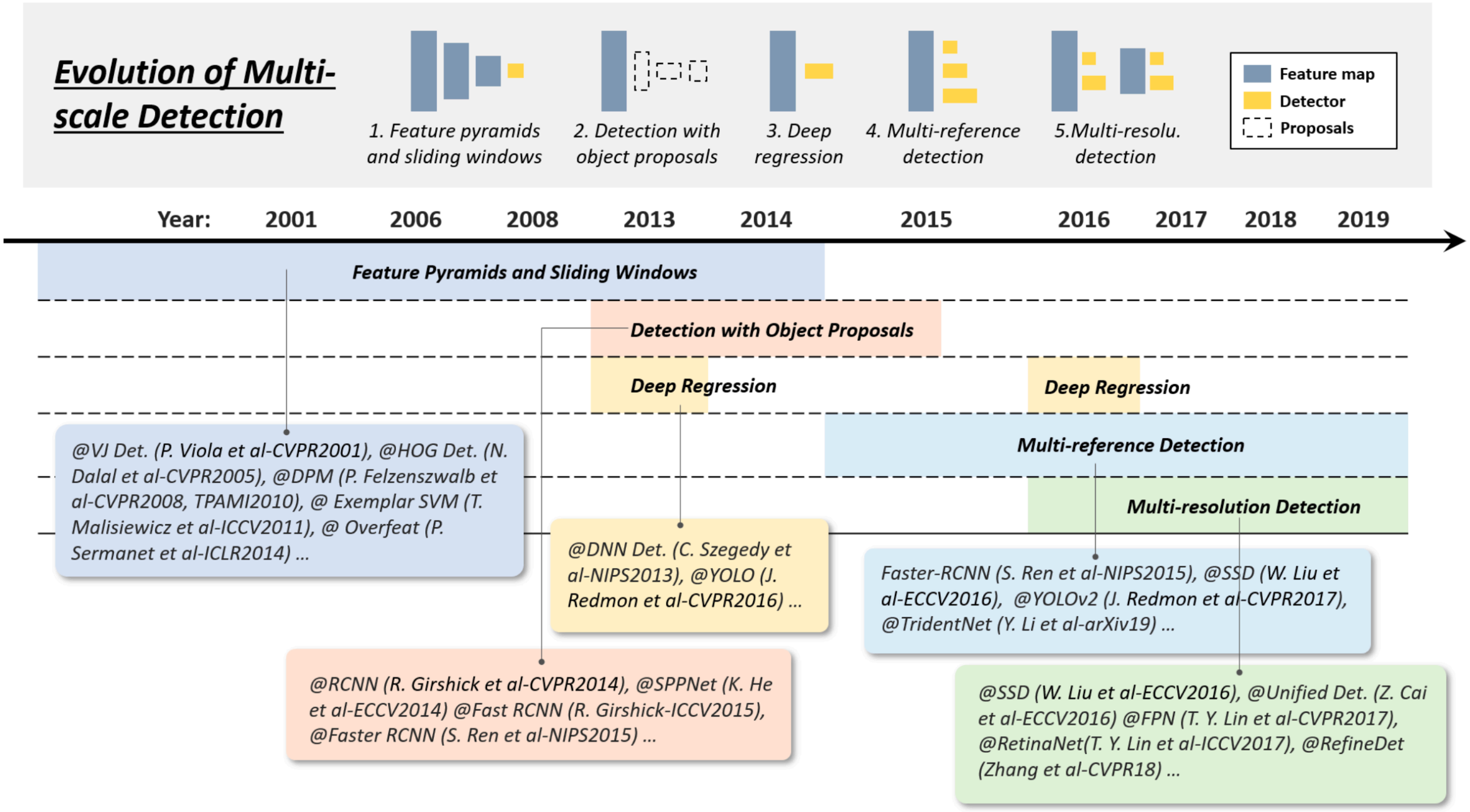
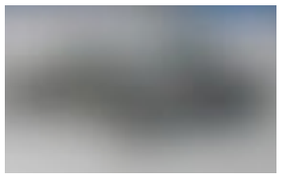
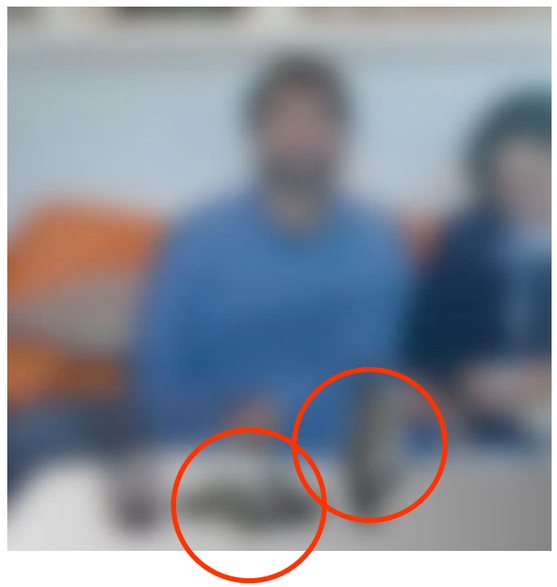
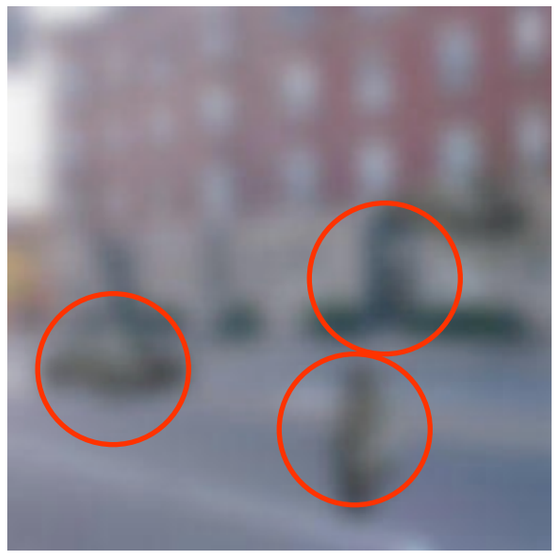


Fig. 6. Evolution of multi-scale detection techniques in object detection from 2001 to 2019: 1) feature pyramids and sliding windows, 2) detection with object proposals, 3) deep regression, 4) multi-reference detection, and 5) multi-resolution detection. Detectors in this figure: VJ Det. [10], HOG Det. [12], DPM [13, 15], Exemplar SVM [36], Overfeat [103], RCNN [16], SPPNet [17], Fast RCNN [18], Faster RCNN [19], DNN Det. [104], YOLO [20], YOLO-v2 [48], SSD [21], Unified Det. [105], FPN [22], RetinaNet [23], RefineDet [55], TridentNet [56].

Important Issues - Context



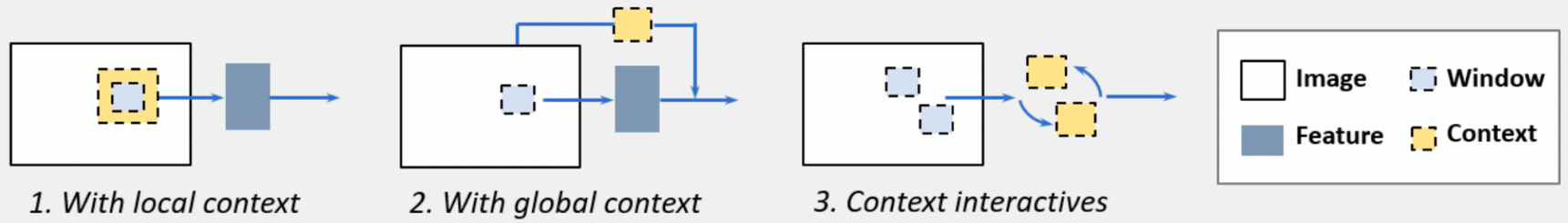




Important Issues - Context

- Local context: Local context refers to the visual information in the area that surrounds the object to detect
  - enlarge the networks' receptive field or the size of object proposals
  - incorporate a small amount of background information
  - inclusion of local contextual regions
- Global context: exploits scene configuration as an additional source of information for object detection
  - integrate a statistical summary of the scene
  - large receptive field or global pooling operation
  - think of the global context as a kind of sequential information and to learn it with the recurrent neural networks
- Context interactive: object-object, object-scene

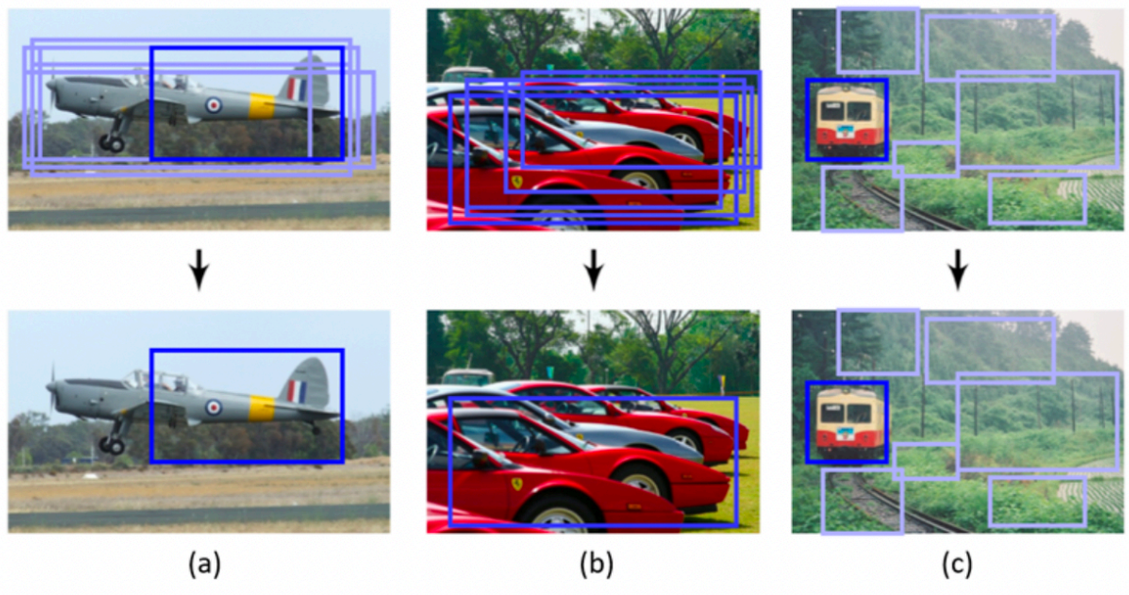
Evolution of Context Priming in Object Detection



Zhengxia Zou et al. *Object Detection in 20 Years: A Survey*. **arXiv, 2019**

Important Issues - NMS

- NMS - Non-Maximum Suppression
  - post-processing step to remove the replicated boxes
- Greedy selection: the bounding box with the maximum detection score is selected while its neighboring boxes are removed according to a predefined overlap threshold



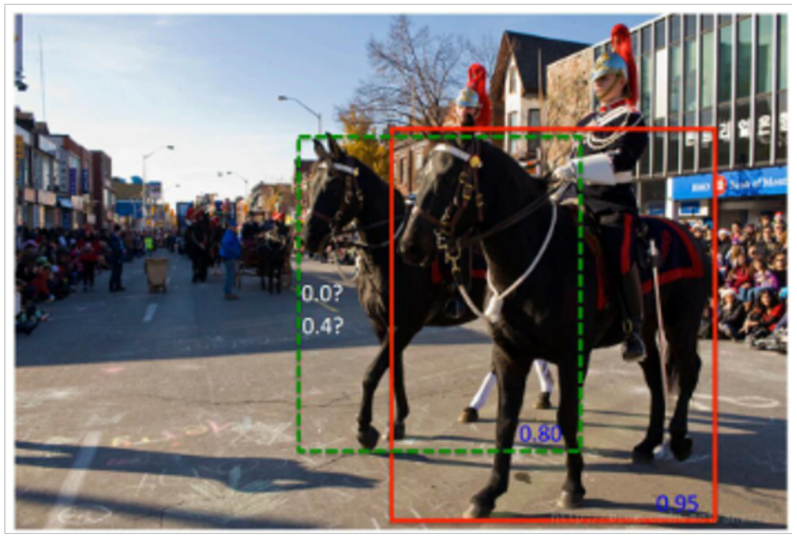
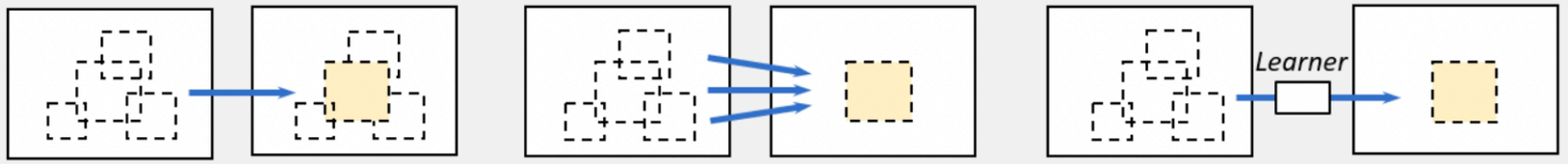


Fig. 11. Examples of possible failures when using a standard greedy selection based non-max suppression: (a) the top-scoring box may not be the best fit, (b) it may suppress nearby objects, and (c) it does not suppress false positives. Images from *R. Rothe et al. ACCV2014* [156].

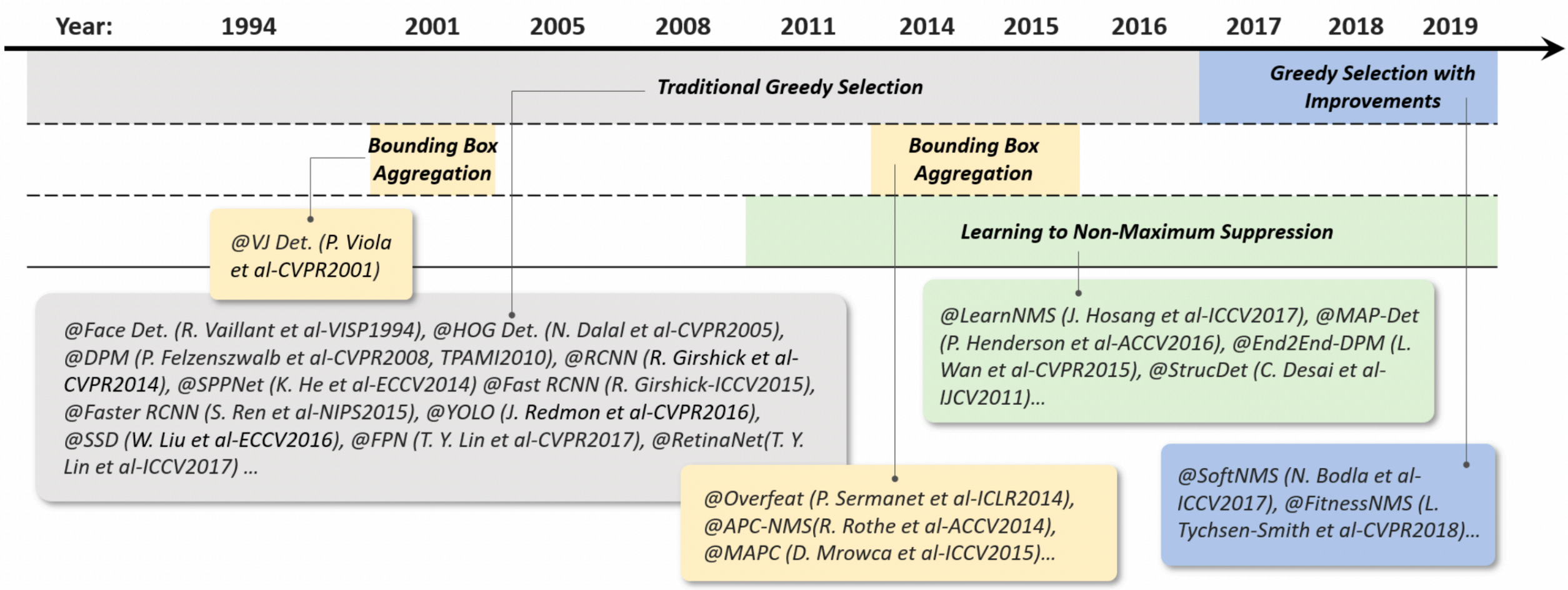
Important Issues - NMS

- NMS - Non-Maximum Suppression
  - post-processing step to remove the replicated boxes
- BB aggregation:
  - combining or clustering multiple overlapped bounding boxes into one final detection
- Learning to NMS:
  - think of NMS as a filter to re-score all raw detections and to train the NMS as part of a network in an end-to-end fashion

Evolution of Non-Max Suppression



1. Greedy selection　　2. Bounding box aggregation　　3. Learning to NMS

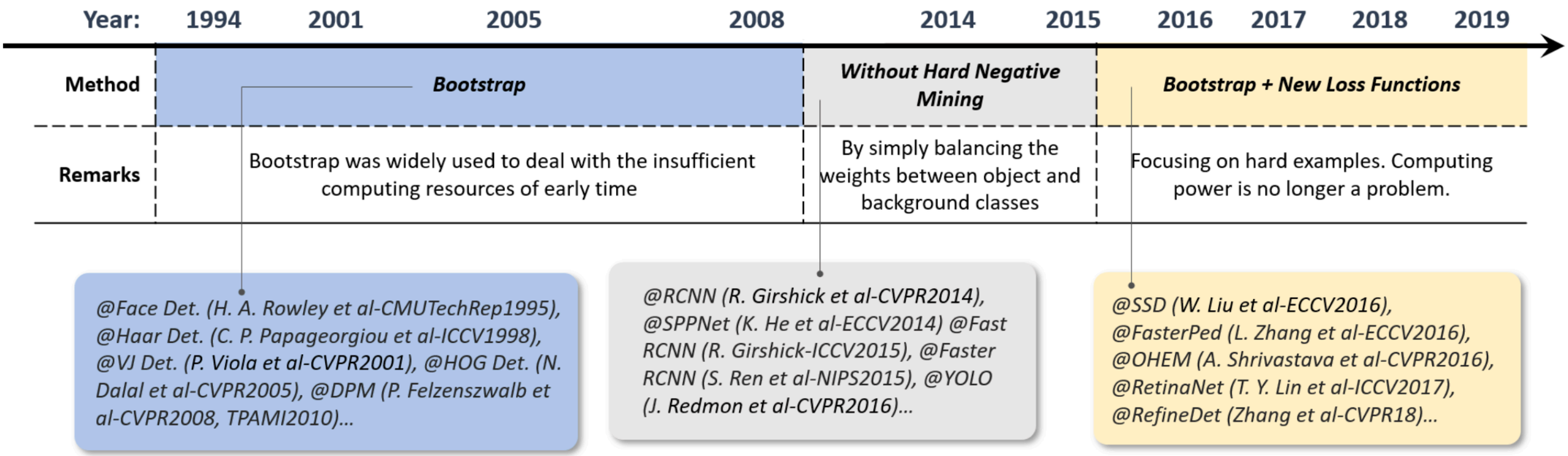


Zhengxia Zou et al. ***Object Detection in 20 Years: A Survey*****. arXiv, 2019**

Important Issues - HNM

- Aims to deal with the problem of imbalanced data during training
  - Easy negatives: background samples, too many
- Bootstrap: the training starts with a small part of background samples and then iteratively add new misclassified backgrounds during the training process
- In DNN:
  - Weight-balancing
  - Redesign the loss function, so that more focus on hard, misclassified examples
  - Anchor refinement module

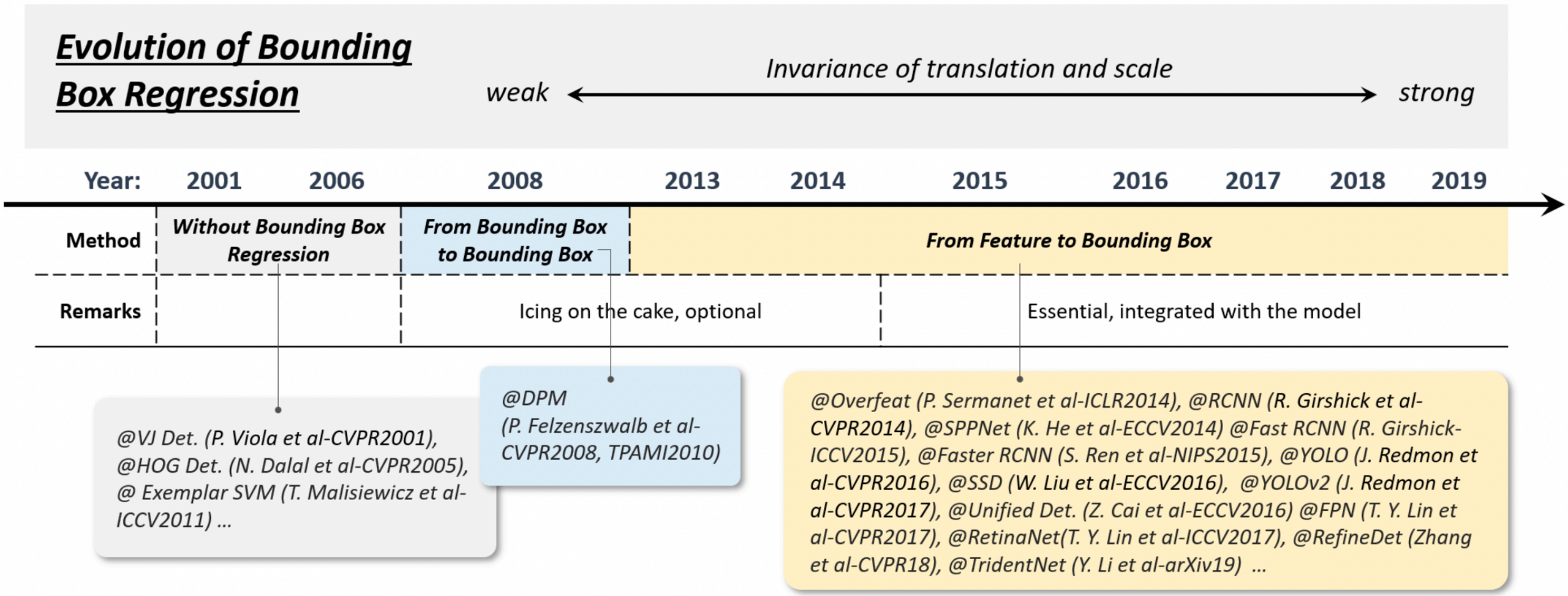
Evolution of Hard Negative Mining



Zhengxia Zou et al. ***Object Detection in 20 Years: A Survey*****. arXiv, 2019**

Important Issues - BB Regression

- Bounding Box regression: refine the location of a predicted bounding box based on the initial proposal or the anchor box
- Without BB regression: VJ
- From BB to BB: DPM
- From feature to BB



Zhengxia Zou et al. ***Object Detection in 20 Years: A Survey*****. arXiv, 2019**

Important Issues - Backbones

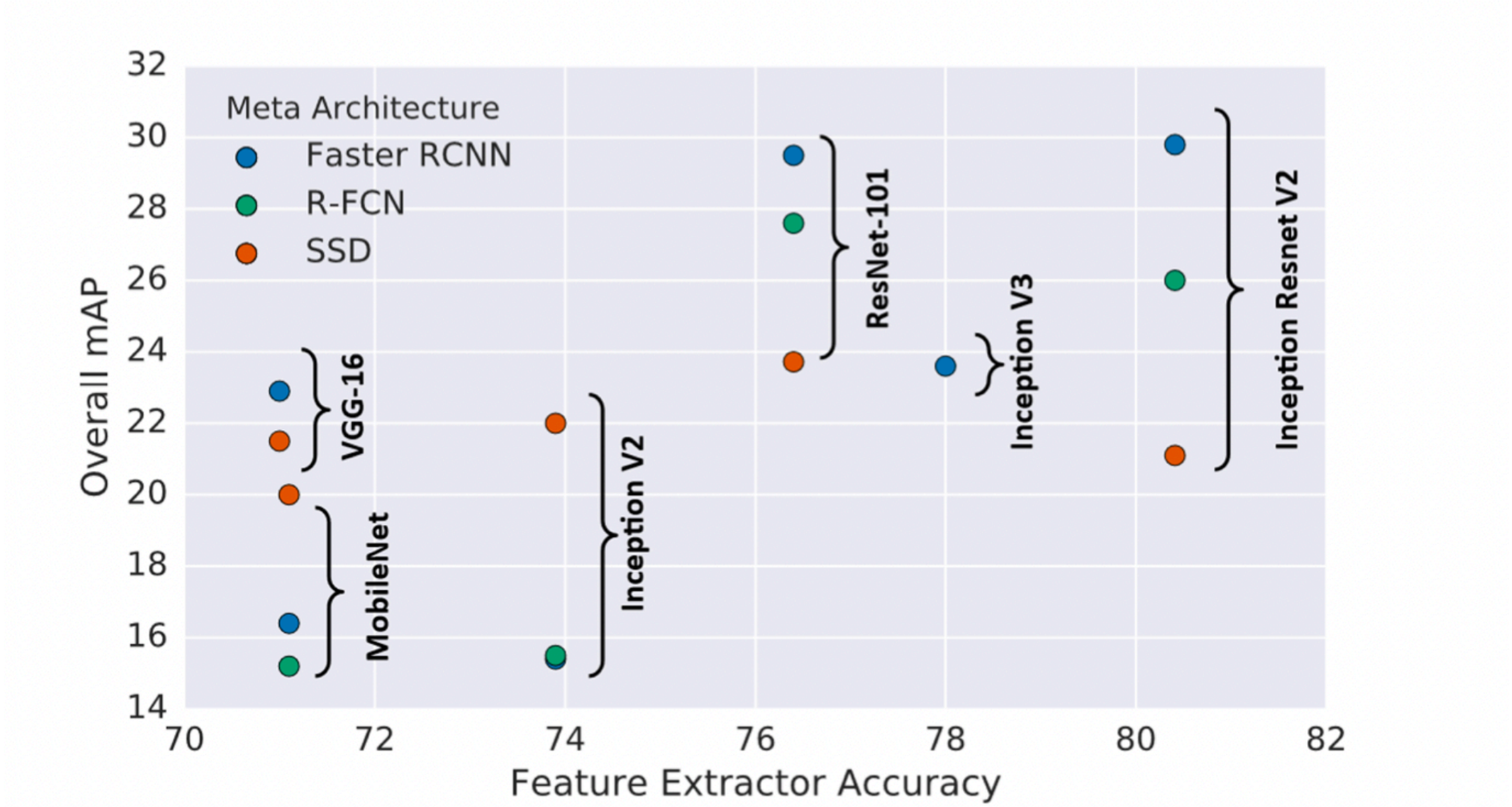


Fig. 17. A comparison of detection accuracy of three detectors: Faster RCNN [19], R-FCN [46] and SSD [21] on MS-COCO dataset with different detection engines. Image from *J. Huang et al. CVPR2017* [27].

Zhengxia Zou et al. ***Object Detection in 20 Years: A Survey*****. arXiv, 2019**

Important Issues - Feature Fusion

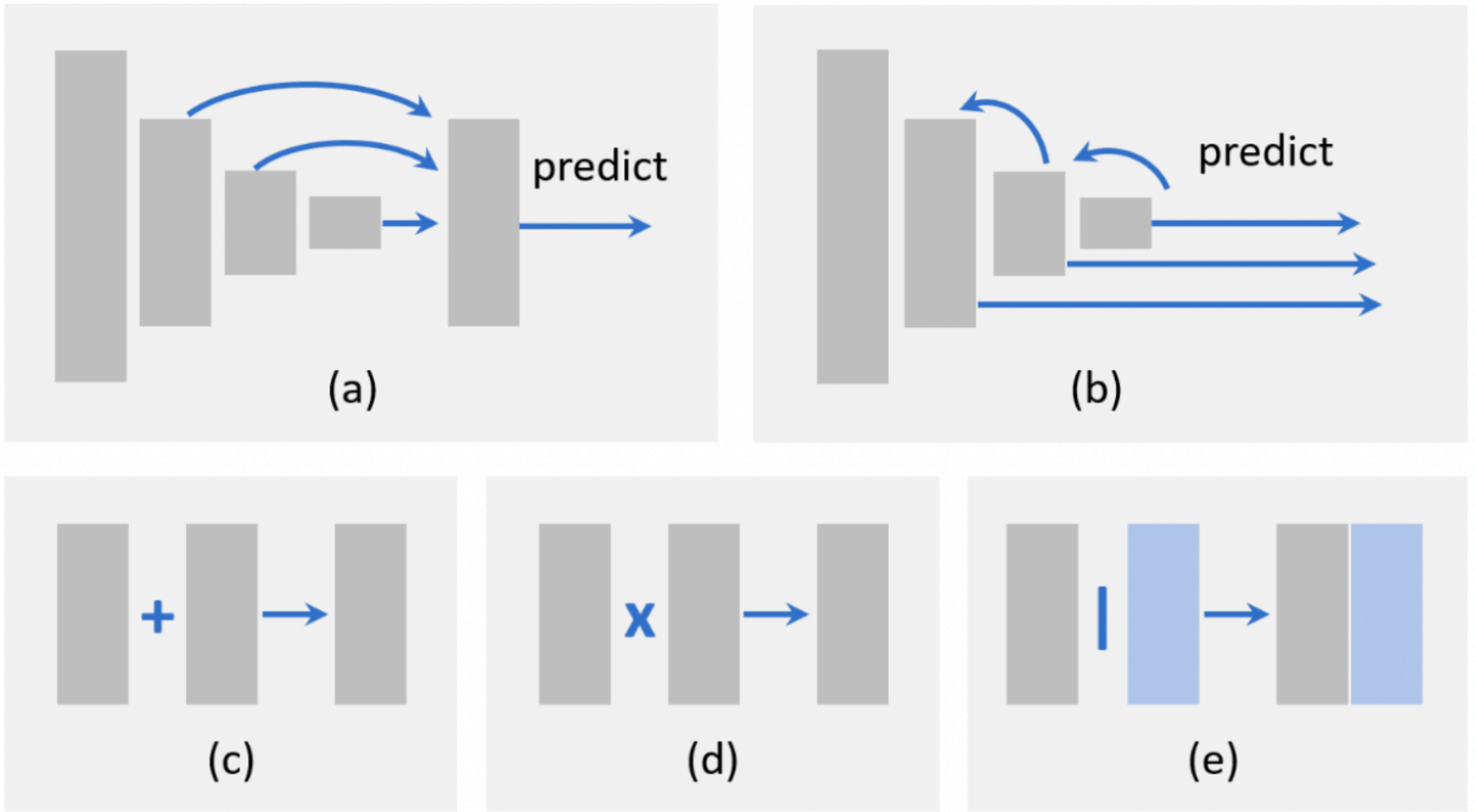
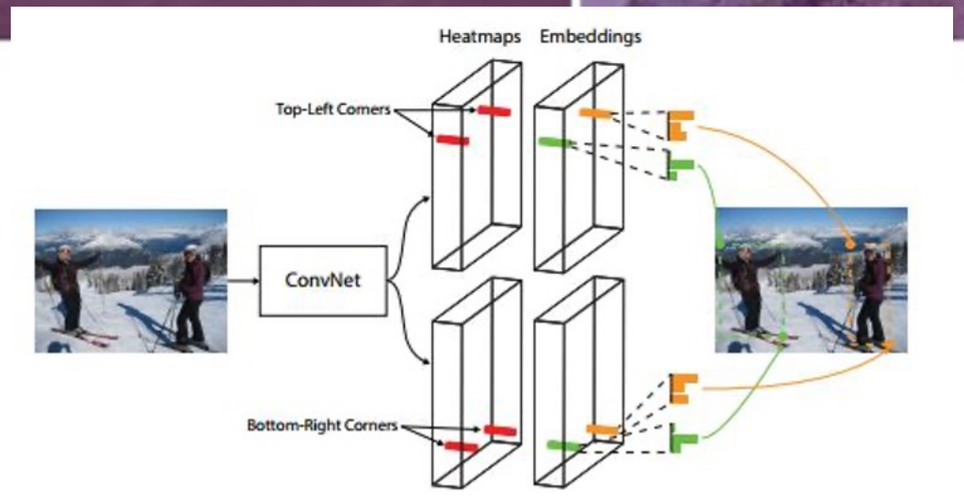


Fig. 18. An illustration of different feature fusion methods: (a) bottom-up fusion, (b) top-down fusion, (c) element-wise sum, (d) element-wise product, and (e) concatenation.

Zhengxia Zou et al. ***Object Detection in 20 Years: A Survey*****. arXiv, 2019**

Important Issues - Detection Paradigm

- Detection as key points localization



H. Law and J. Deng, *Cornernet: Detecting objects as paired keypoints*, ECCV, 2018.

Important Issues - Rotation

- How to achieve rotation invariant object detection
  - Data augmentation
  - Rotation invariant loss functions: introduced a constraint on the original detection loss function so that to make the features of rotated objects unchanged
  - Rotation calibration: make geometric transformations. Spatial Transformer Networks (STN)
  - Rotation RoI Pooling: from Cartesian coordinates to polar coordinates

Important Issues - Adversarial training

- Small object detection: GAN has been used to enhance the detection on small objects by narrowing the representations between small and large ones.
- Occlusion handling: generate occlusion masks by using adversarial training
- Adversarial attack

Applications

- Pedestrian detection
- Face detection
- Text detection
- Traffic Sign and Traffic Light Detection
- Remote Sensing Target Detection

Zhengxia Zou et al. ***Object Detection in 20 Years: A Survey*****. arXiv, 2019**

Fig. 23. Challenges in traffic sign detection and traffic light detection: (a) Illumination changes. Image from *pxhere* (free of copyrights). (b) Motion blur. Image from GTSRB Dataset [81]. (c) Detection under bad weather. Image from *Flickr* and *Max Pixel* (free of copyrights).

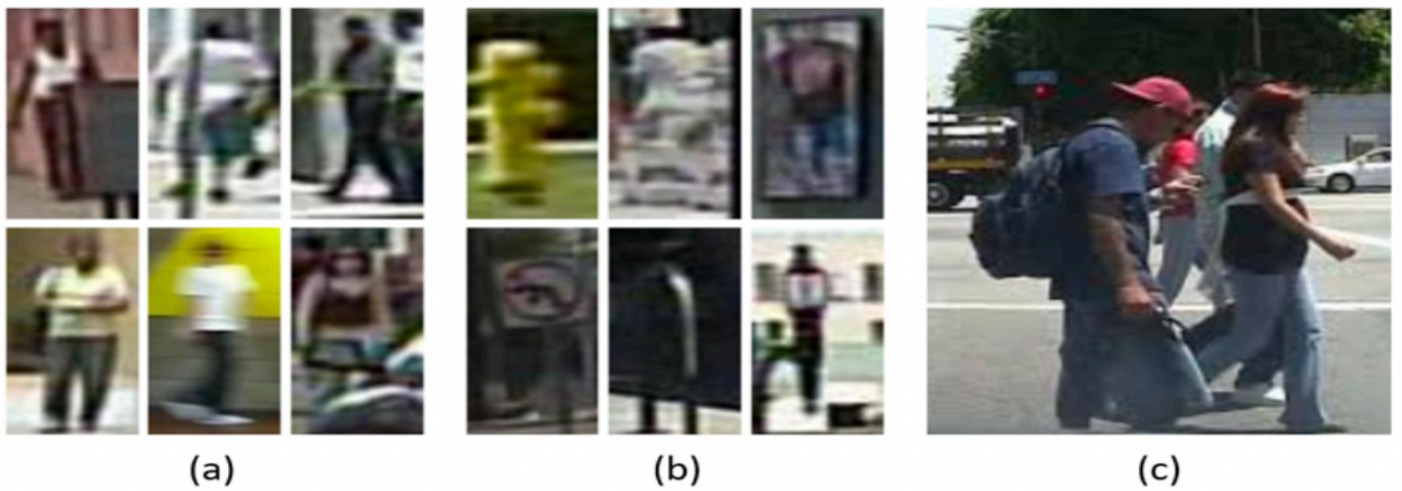
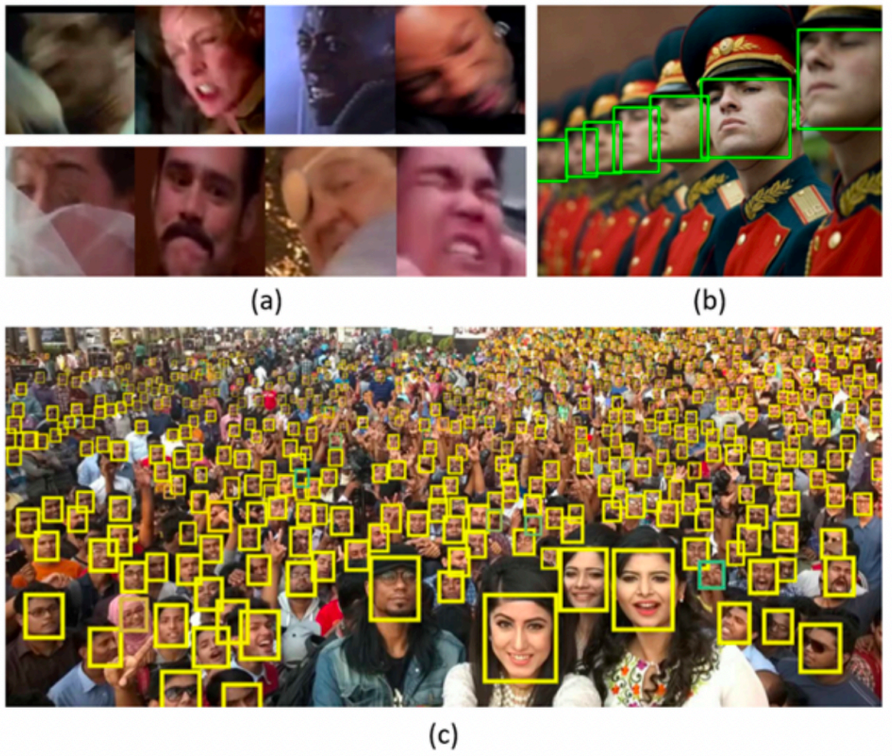


Fig. 20. Some hard examples of pedestrian detection from Caltech dataset [59, 60]: (a) small pedestrians, (b) hard negatives, and (c) dense and occluded pedestrians.

Fig. 22. Challenges in text detection and recognition: (a) Variation of fonts, colors and languages. Image from *maxpixel* (free of copyrights). (b) Text rotation and perspective distortion. Image from *Y. Liu* et al. *CVPR2017* [336]. (c) Densely arranged text localization. Image from *Y. Wu et al. ICCV2017* [337].



g. 21. Challenges in face detection: (a) Intra-class variation, image m WildestFaces Dataset [70]. (b) Face occlusion, image from UFDD taset [69]. (c) Multi-scale face detection. Image from *P. Hu et al. /PR2017* [322].

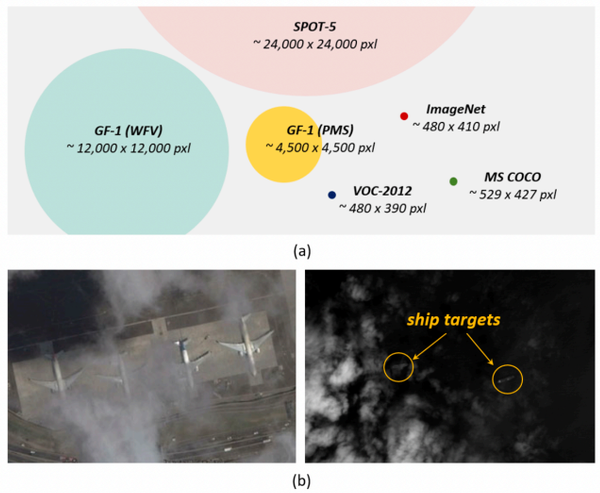


Fig. 24. Challenges in remote sensing target detection: (a) Detection in "big data": data volume comparison between a single-view remote sensing imagery and an average image size of VOC, ImageNet, and MS-COCO. (b) Targets occluded by cloud. Images from *S. Qiu et al. JSTARS2017* [380] and *Z. Zou et al. TGRS2016* [381].

Object detection: future direction

- Small object detection
- Weakly supervised detection
- Fusion of multiple modalities
- Lightweight object detection
- Real-time object detection/tracking
- Few / Zero Shot Object Detection
- …