Lec5 Object Detection

·AIPKU·

人工智能引论实践课 计算机视觉小班

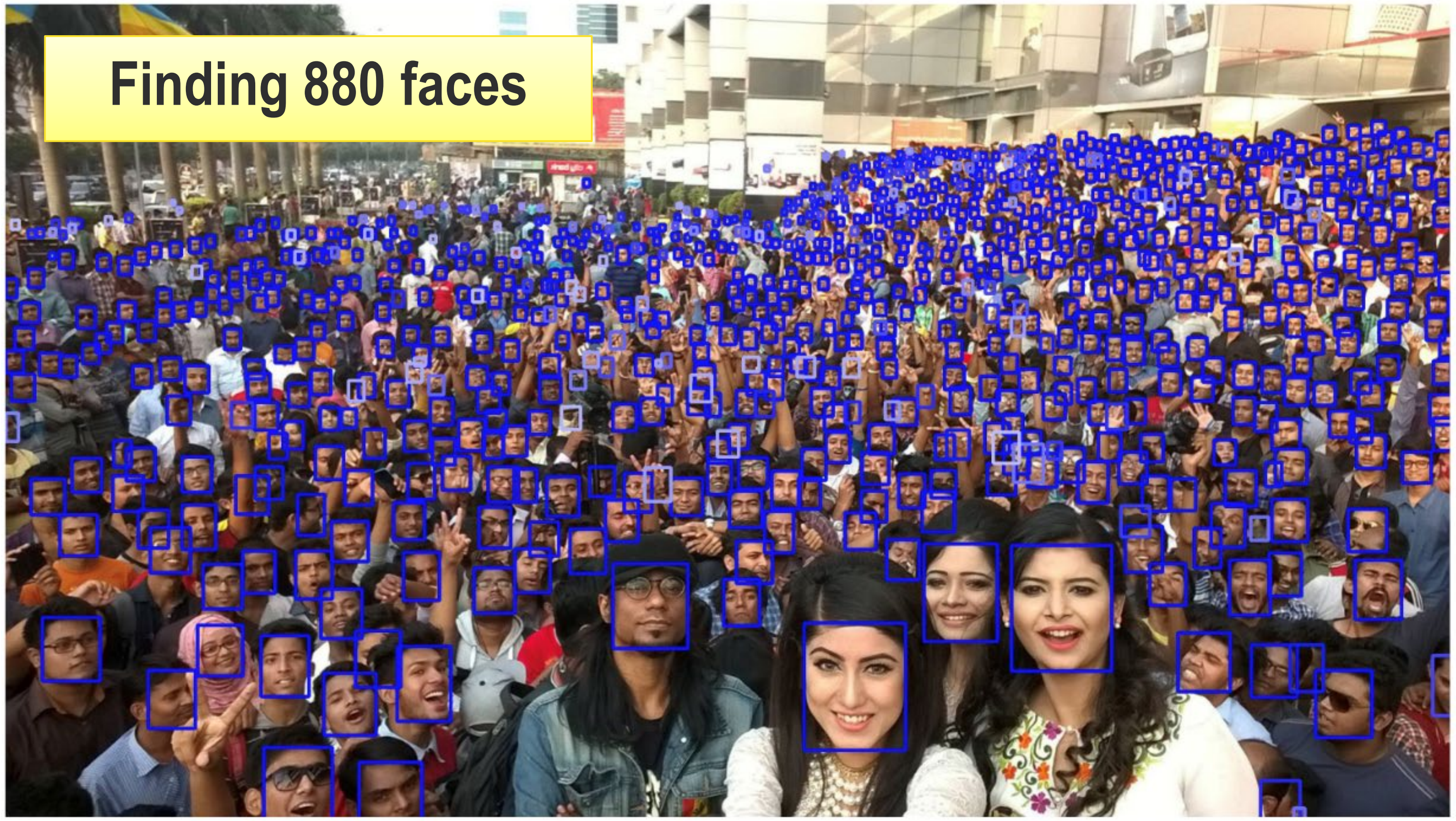
主讲人：刘家瑛

部分讲义来源与致谢

1. Fei-Fei Li, Justin Johnson, Serena Yeung. Stanford University CS231n: Deep Learning for Computer Vision
2. Li Liu, Wanli Ouyang, Xiaogang Wang, Paul W. Fieguth, Jie Chen, Xinwang Liu, Matti Pietikäinen. Deep Learning for Generic Object Detection: A Survey. IJCV 2020
3. Ross B. Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv 2013
4. Ross B. Girshick. Fast R-CNN. ICCV 2015
5. Shaoqing Ren, Kaiming He, Ross B. Girshick, Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. NIPS 2015
6. Deep learning object detection. https://github.com/hoya012/deep_learning_object_detection

Generic Object Detection

- Given an arbitrary image, determine whether or not there are any instances of semantic objects from predefined categories and, if present, to return the spatial location and extent
  - Also called *object class detection* or *object category detection*
  - One of the most fundamental and challenging problems in computer vision



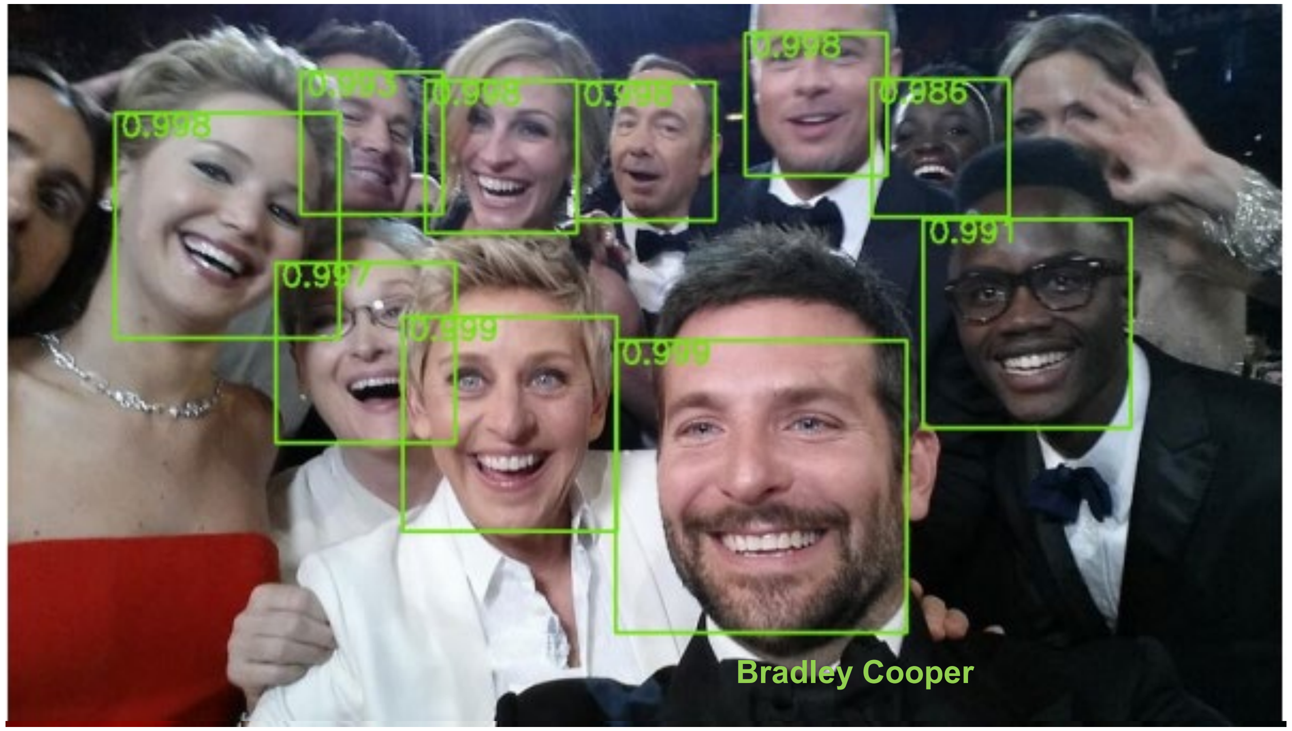
AIPKU

License Plate Detection & Recognition

- In **Unconstrained Scenarios**
- **Task**: find and recognize license plates in images

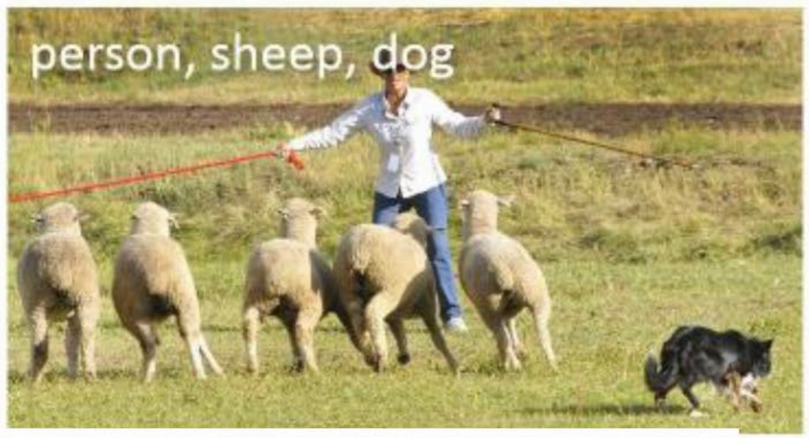
Types of Object Detection

- **Detection of specific categories**
  - Detecting different instances of predefined object categories, *such as human, cars*
- **Detection of specific instance**
  - Detecting instances of a particular object, *such as Donald Trump's face, the Bradley Cooper's face*

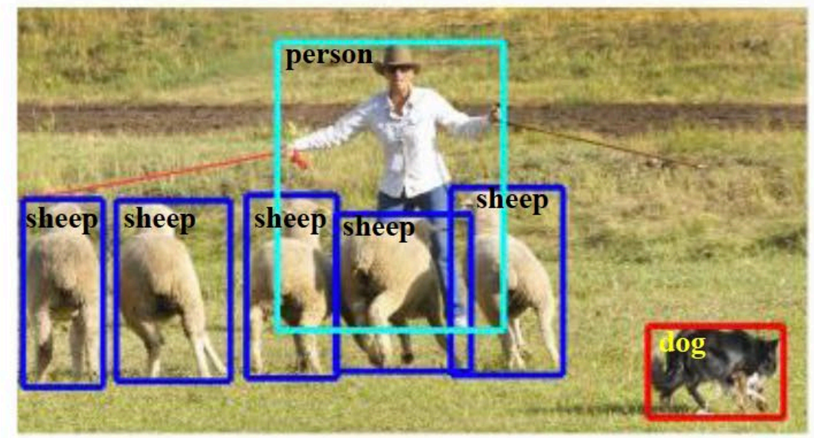


Recognition Problems related to Detection

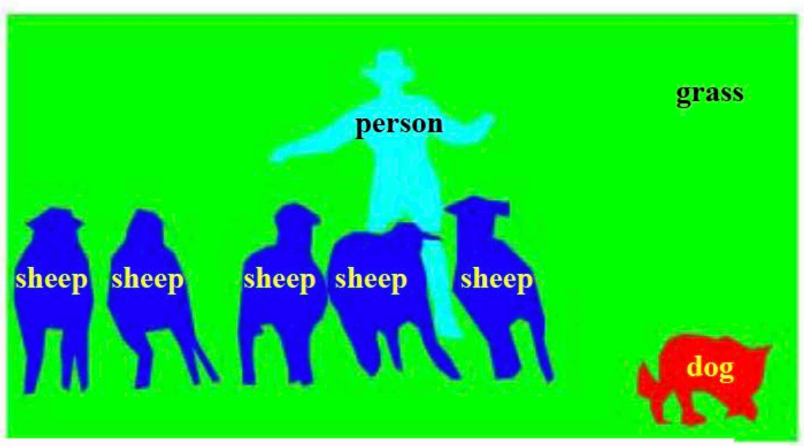
Assigning one or more object class labels to a given image, determining presence without the need of location



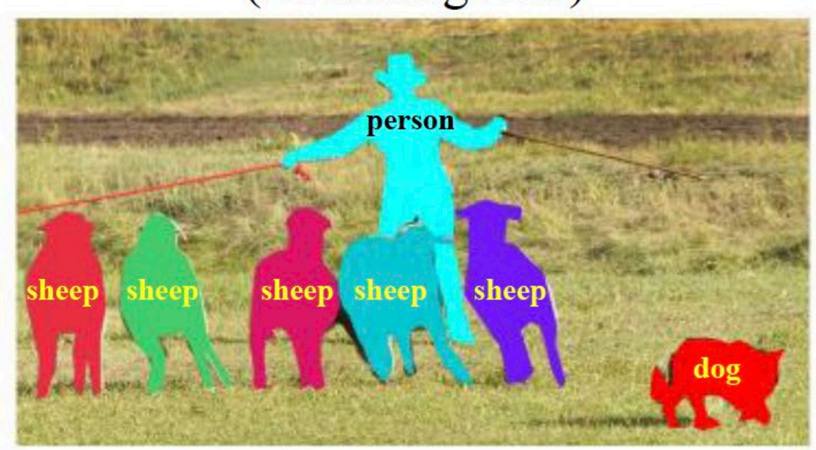
(a) Object Classification



(b) Generic Object Detection (Bounding Box)



(c) Semantic Segmentation



(d) Object Instance Segmetation

Assigning each pixel in an image to a semantic class label

Distinguishing different instances of the same object class, while semantic segmentation does not distinguish different instances

Object Classification

This image is CC0 public domain

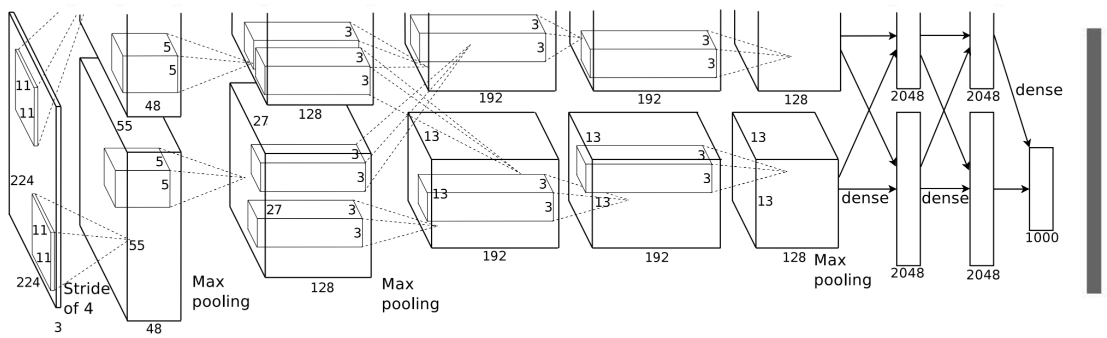


Figure copyright Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, 2012. Reproduced with permission.

Vector:
4096

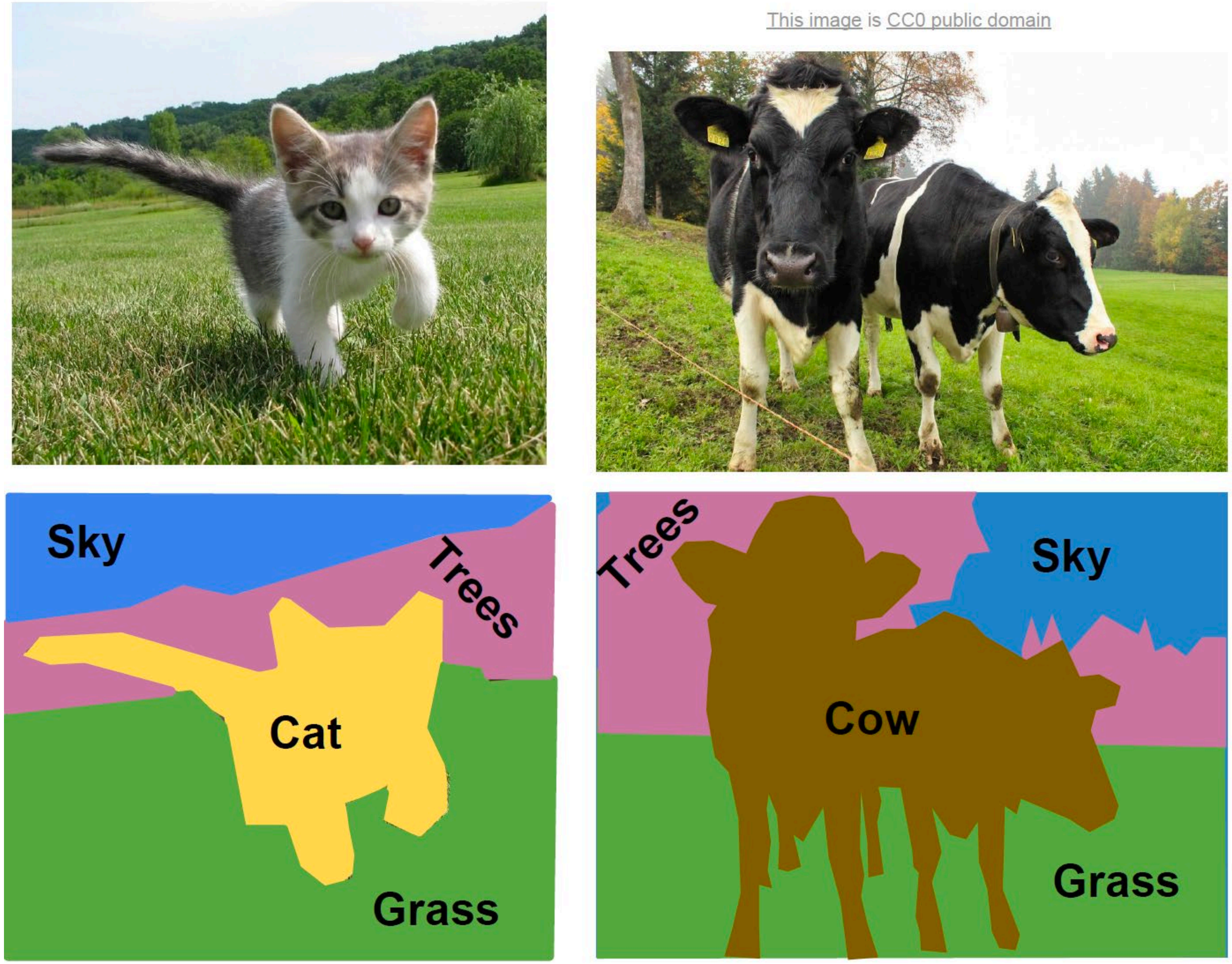
Fully-Connected:
4096 to 1000

Class Scores
Cat: 0.9
Dog: 0.05
Car: 0.01
...

Source: http://cs231n.stanford.edu/2017/

Semantic Segmentation

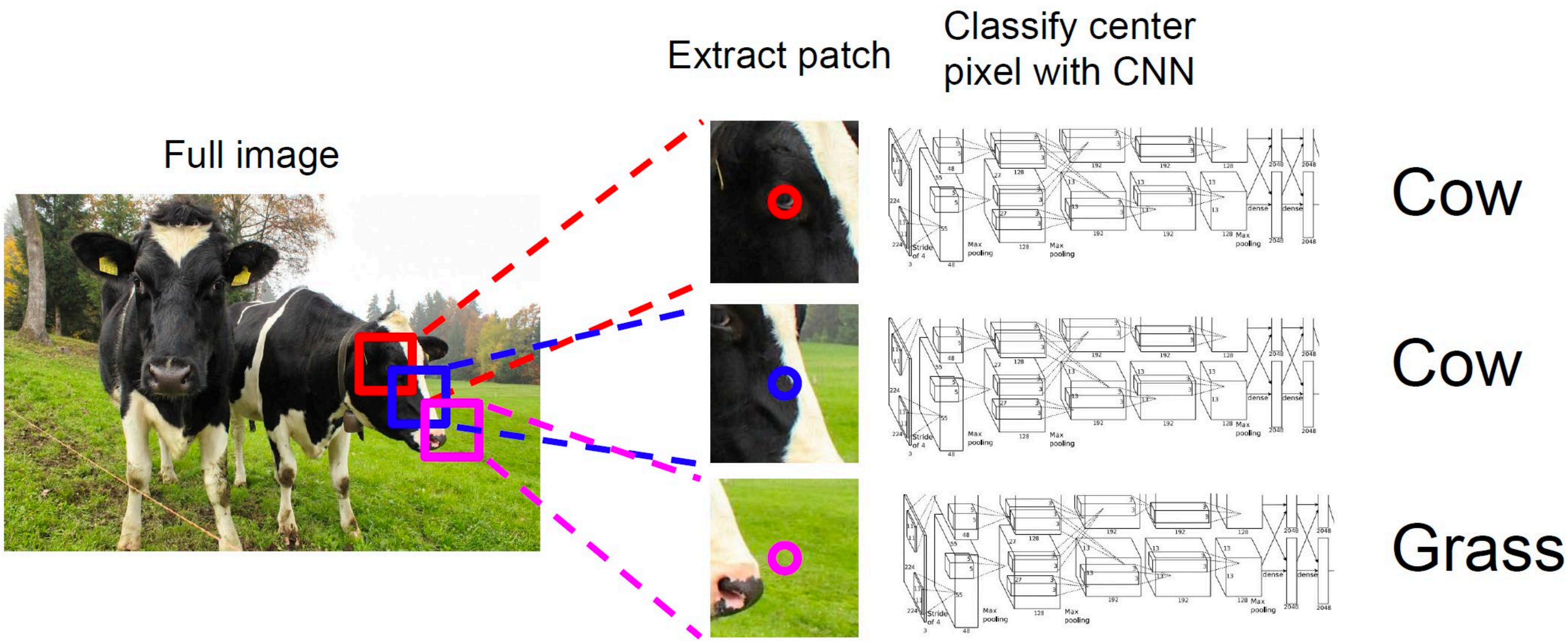
- Label each pixel in the image with a category Label
- Don’t differentiate instances, only care about pixels



Source: http://cs231n.stanford.edu/2017/

Semantic Segmentation Idea

- **Semantic Segmentation Idea: Sliding Window**

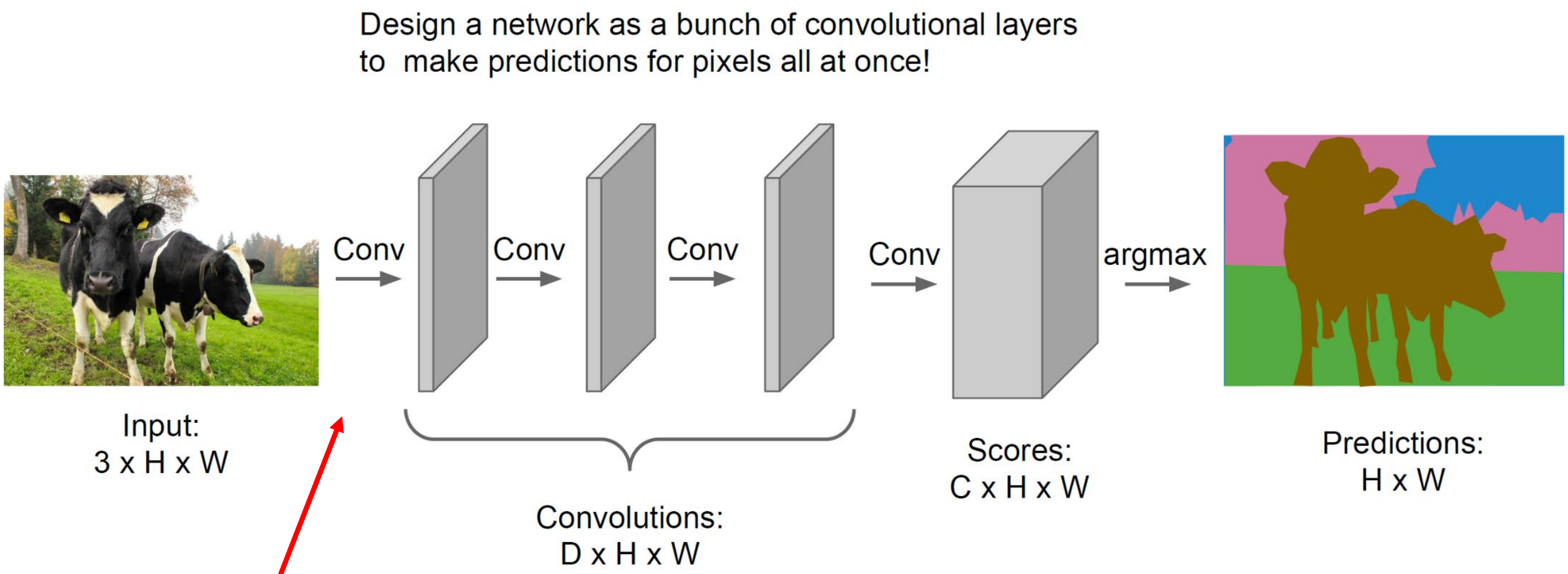


Problem: Very inefficient! Not reusing shared features between overlapping patches!

Source: http://cs231n.stanford.edu/2017/

Semantic Segmentation Idea

- **Semantic Segmentation Idea: Fully Convolutional**

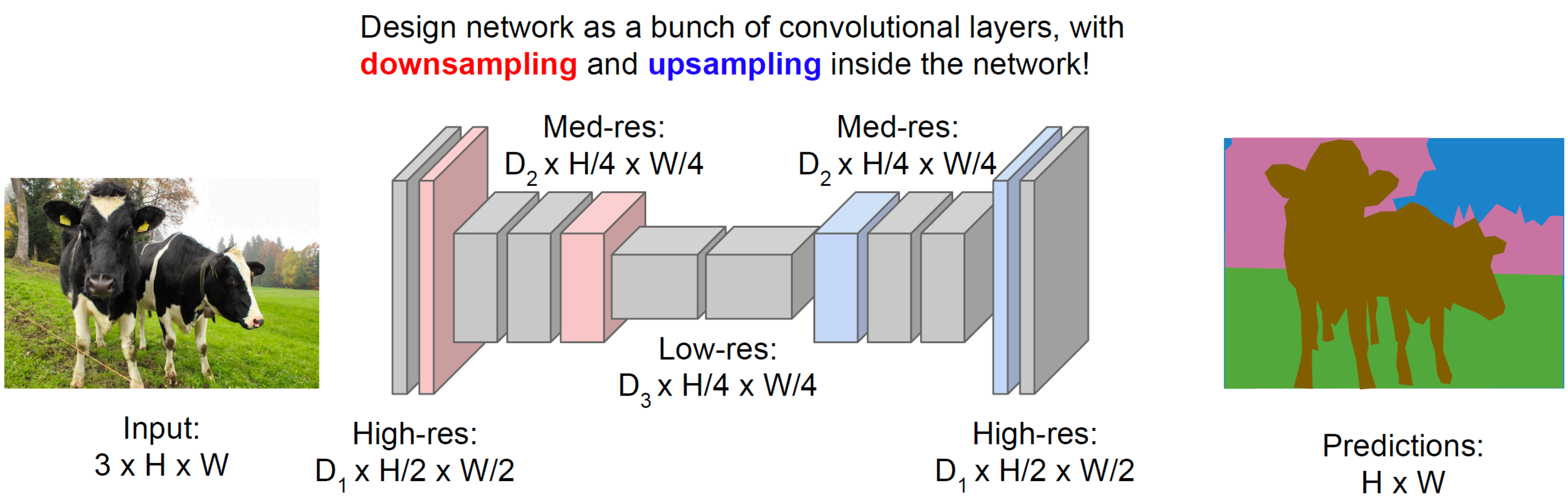


Problem: convolutions at original image resolution will be very expensive ...

Source: http://cs231n.stanford.edu/2017/

Semantic Segmentation Idea

- **Semantic Segmentation Idea: Fully Convolutional**



Source: http://cs231n.stanford.edu/2017/

Diversity of Definitions

- There is no universal agreement in the literature on the definitions of various vision subtasks

- Often encountered terms such as:
  - Detection
  - Localization
  - Recognition
  - Classification
  - Categorization
  - Verification
  - Identification
  - Annotation
  - Labeling
  - Understanding

  are often differently defined

Object Detection

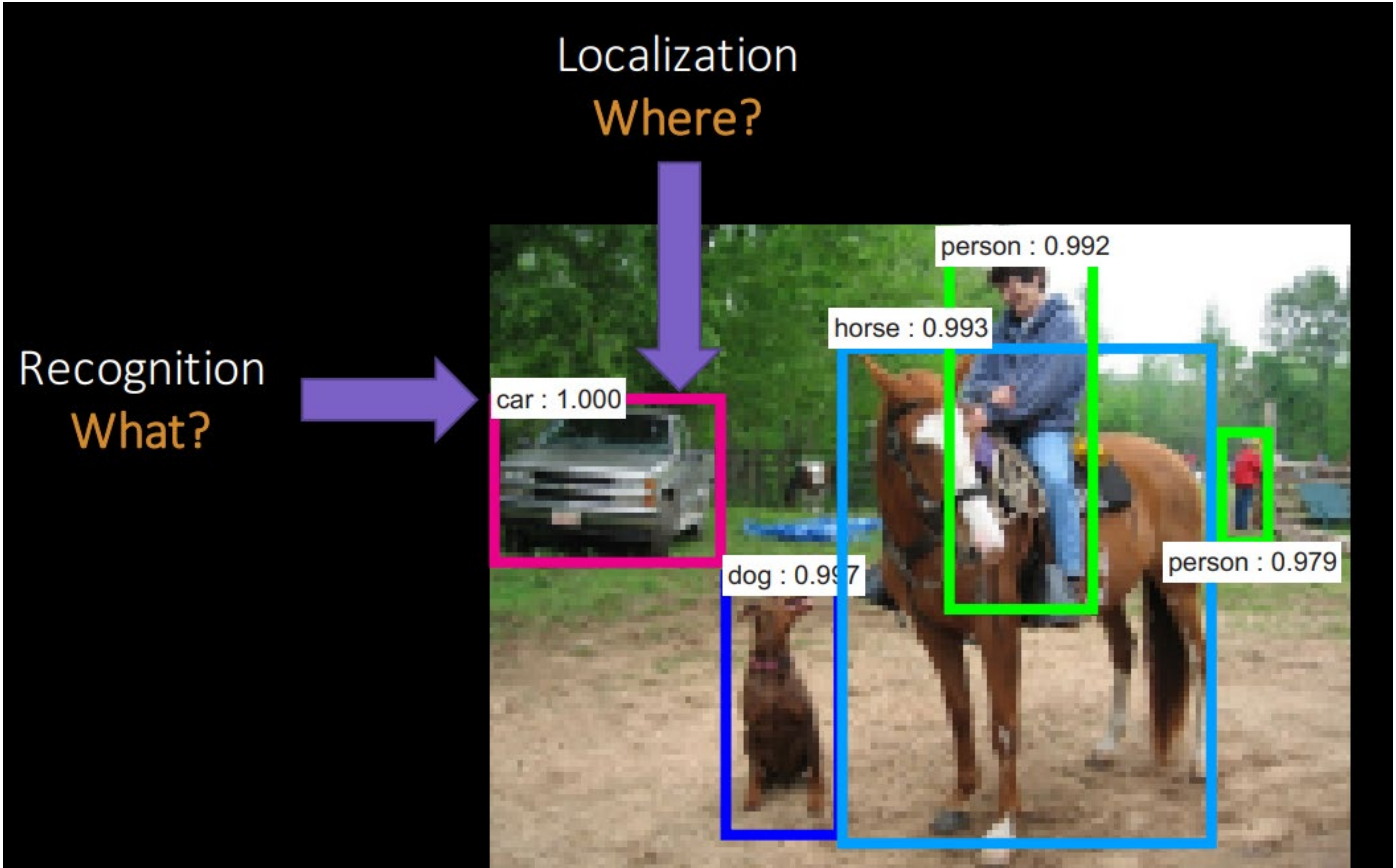
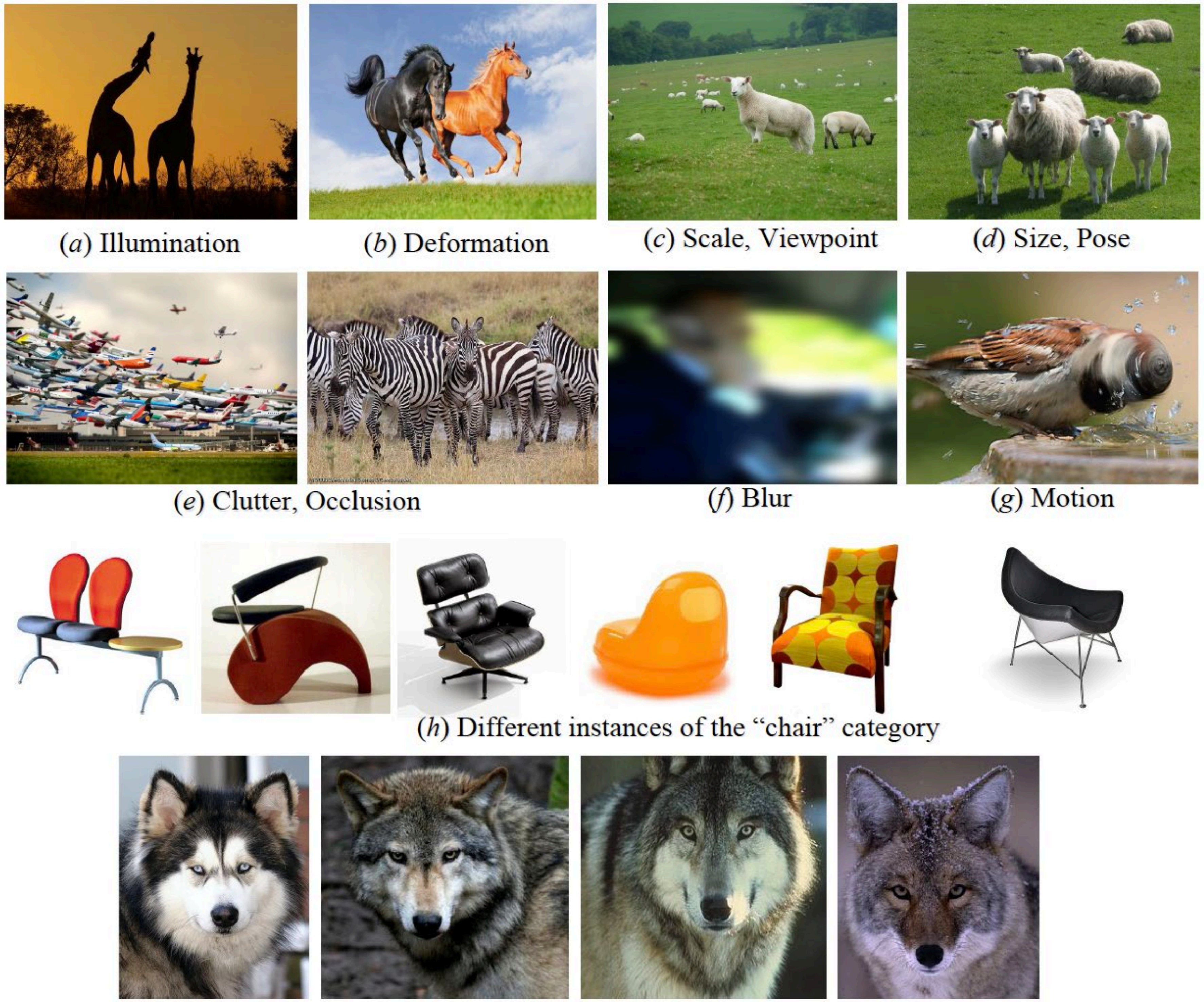


Figure adapted from Kaiming He

Challenges in Generic Object Detection



(*a*) Illumination (*b*) Deformation (*c*) Scale, Viewpoint (*d*) Size, Pose

(*e*) Clutter, Occlusion (*f*) Blur (*g*) Motion

(*h*) Different instances of the "chair" category

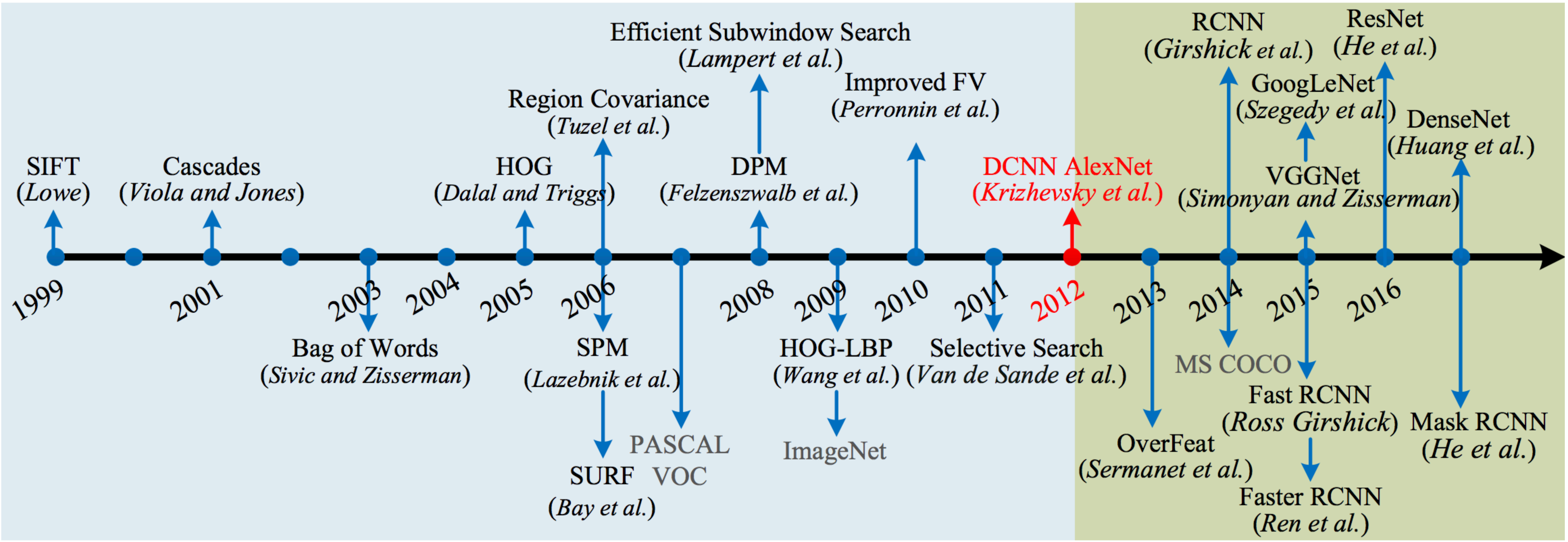
(*i*) Small Interclass Variations: four different categories

Challenges in Generic Object Detection

Ideal Detector

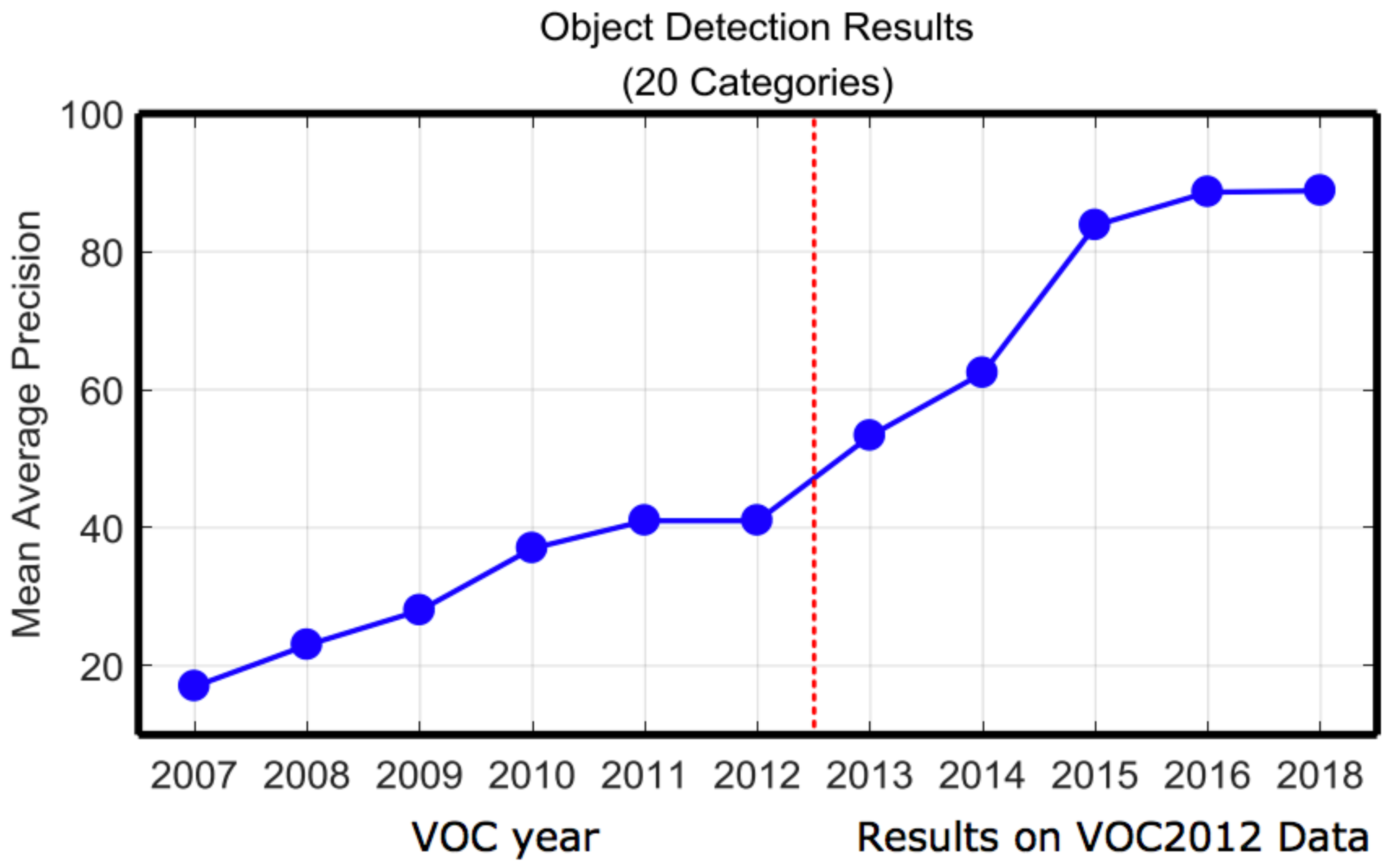
- **High Accuracy**
  - Localization Acc.
  - Recognition Acc.
  - **High Robustness To Intraclass Variations**
    - **Many different instances in each class**
      e.g. different color, texture, material and shape *etc.*
    - **Object instance diversities**
      pose, nonrigid deformations
    - **Imaging conditions and unconstrained environments**
      e.g. illumination, view point, scale, occlusion, shading, clutter, blur, motion, weather condition
    - **Image noise**
      e.g. imaging noise, filter distortions, compression noise
  - **High Distinctiveness**
    - **Interclass ambiguities**
    - **Thousands of real-world object classes**
      structured and unstructured
- **High Efficiency**
  - Time Efficiency
  - Memory Efficiency
  - Storage Efficiency
  - **Thousands of object categories in real world**
  - **Requiring localizing and recognizing objects**
  - **Large number of possible locations of objects**
  - **Large-scale image/video data**

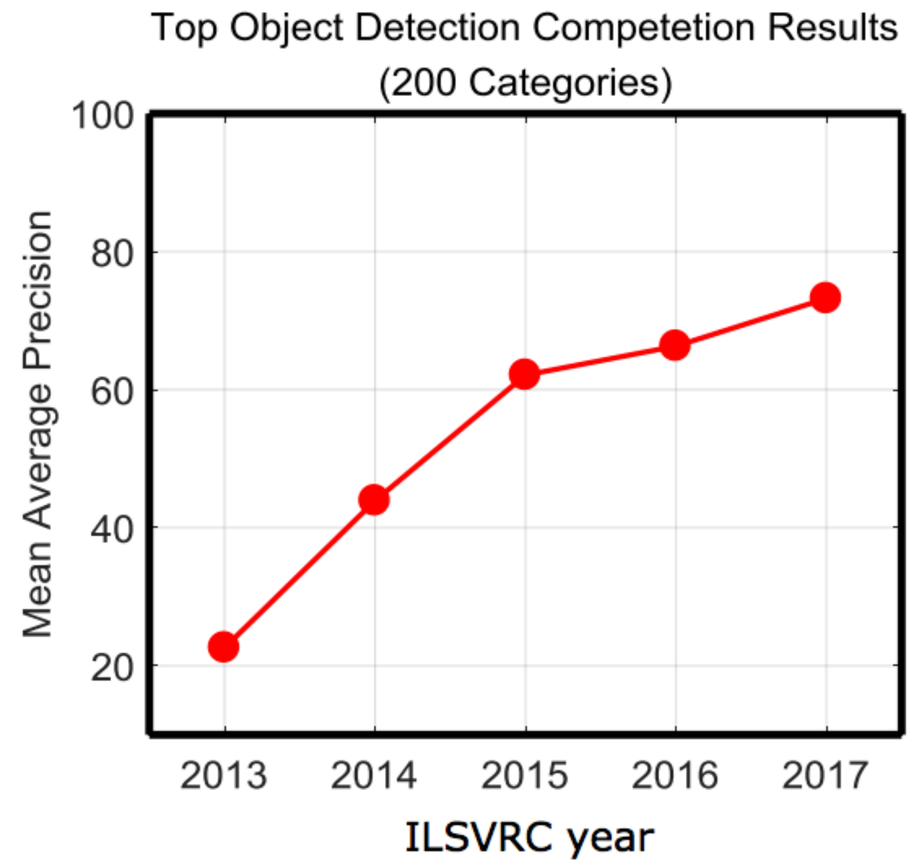
Progress in the Past Two Decades



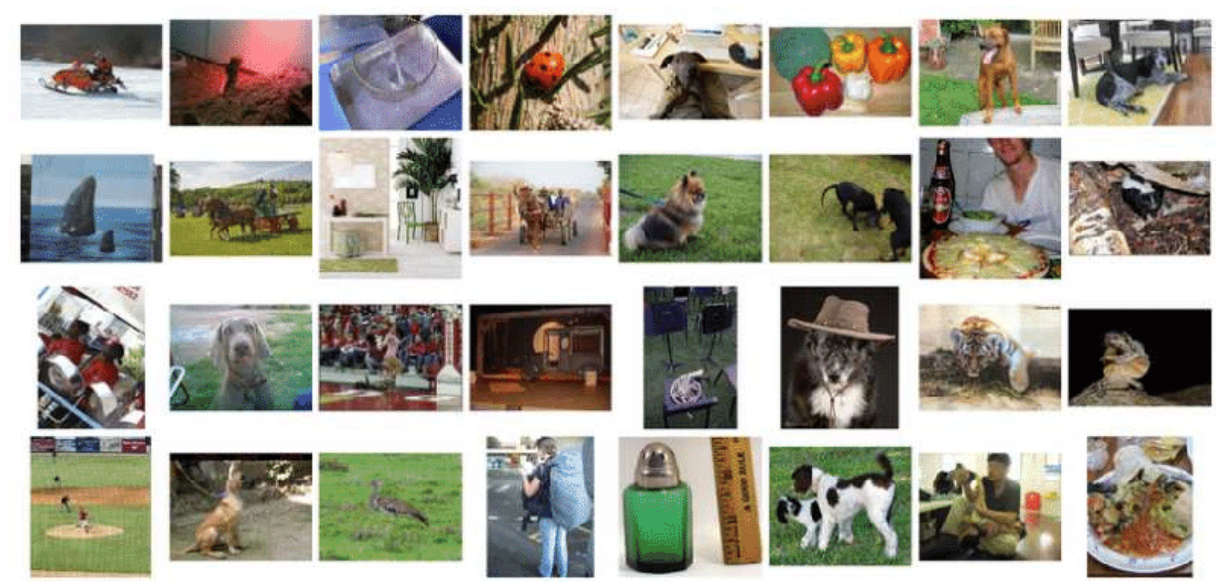
Recent Evolution of Object Detection Performance

Turning Point in 2012: Deep Learning Achieved Record Breaking Image Classification Result





Results on VOC2012 Data



Summarization of Related Surveys (1/4)

- **Since 2000**

No.	Survey Title	Ref.	Year	Published	Content
1	Monocular Pedestrian Detection: Survey and Experiments	[51]	2009	PAMI	Evaluating three detectors with additional experiments integrating the detectors into full systems
2	Survey of Pedestrian Detection for Advanced Driver Assistance Systems	[60]	2010	PAMI	A survey of pedestrian detection for advanced driver assistance systems
3	Pedestrian Detection: An Evaluation of the State of The Art	[48]	2012	PAMI	Focus on a more thorough and detailed evaluation of detectors in individual monocular images
4	Detecting Faces in Images: A Survey	[226]	2002	PAMI	First survey of face detection from a single image
5	A Survey on Face Detection in the Wild: Past, Present and Future	[232]	2015	CVIU	A survey of face detection in the wild since 2000
6	On Road Vehicle Detection: A Review	[196]	2006	PAMI	A review of vision based onroad vehicle detection systems where the camera is mounted on the vehicle

51. Enzweiler M., Gavrila D. M. (2009) Monocular pedestrian detection: Survey and experiments. IEEE TPAMI 31(12):2179–2195

60. Geronimo D., Lopez A. M., Sappa A. D., Graf T. (2010) Survey of pedestrian detection for advanced driver assistance systems. IEEE TPAMI 32(7):1239–1258

48. Dollar P., Wojek C., Schiele B., Perona P. (2012) Pedestrian detection: An evaluation of the state of the art. IEEE TPAMI 34(4):743–761

226. Yang M., Kriegman D., Ahuja N. (2002) Detecting faces in images: A survey. IEEE TPAMI 24(1):34–58

232. Zafeiriou S., Zhang C., Zhang Z. (2015) A survey on face detection in the wild: Past, present and future. Computer Vision and Image Understanding 138:1–24

196. Sun Z., Bebis G., Miller R. (2006) On road vehicle detection: A review. IEEE TPAMI 28(5):694–711

Summarization of Related Surveys (2/4)

7	Text Detection and Recognition in Imagery: A Survey	[227]	2015	PAMI	A survey of text detection and recognition in color imagery
8	Toward Category Level Object Recognition	[169]	2007	Book	Collects a series of representative papers on object categorization, detection, and segmentation
9	The Evolution of Object Categorization and the Challenge of Image Abstraction	[46]	2009	Book	A trace of the evolution of object categorization in the last four decades
10	Context based Object Categorization: A Critical Survey	[59]	2010	CVIU	A review of different ways of using contextual information for object categorization
11	50 Years of Object Recognition: Directions Forward	[5]	2013	CVIU	A review of the evolution of object recognition systems in the last five decades
12	Visual Object Recognition	[69]	2011	Tutorial	Covers fundamental and time tested approaches for both instance and category object recognition techniques

227. Ye Q., Doermann D. (2015) Text detection and recognition in imagery: A survey. IEEE TPAMI 37(7):1480–1500

169. Ponce J., Hebert M., Schmid C., Zisserman A. (2007) Toward Category Level Object Recognition. Springer

46. Dickinson S., Leonardis A., Schiele B., Tarr M. (2009) The Evolution of Object Categorization and the Challenge of Image Abstraction in *Object Categorization: Computer and Human Vision Perspectives*. Cambridge University Press

59. Galleguillos C., Belongie S. (2010) Context based object categorization: A critical survey. Computer Vision and Image Understanding 114:712–722

5. Andreopoulos A., Tsotsos J. (2013) 50 years of object recognition: Directions forward. Computer Vision and Image Understanding 117(8):827–891

69. Grauman K., Leibe B. (2011) Visual object recognition. Synthesis lectures on artificial intelligence and machine learning 5(2):1–181

Summarization of Related Surveys (3/4)

13	Object Class Detection: A Survey	[240]	2013	ACM CS	First survey of generic object detection methods before 2011
14	Feature Representation for Statistical Learning based Object Detection: A Review	[125]	2015	PR	A survey on feature representation methods in statistical learning based object detection, including handcrafted and a few deep learning based features
15	Salient Object Detection: A Survey	[17]	2014	arXiv	A survey for Salient object detection
16	Representation Learning: A Review and New Perspectives	[12]	2013	PAMI	A review of unsupervised feature learning and deep learning, covering advances in probabilistic models, autoencoders, manifold learning, and deep networks
17	Deep Learning	[116]	2015	Nature	An introduction to deep learning and its typical applications
18	A Survey on Deep Learning in Medical Image Analysis	[133]	2017	MIA	A survey of deep learning for image classification, object detection, segmentation, registration, and others in medical image analysis

240. Zhang X., Yang Y., Han Z., Wang H., Gao C. (2013) Object class detection: A survey. ACM Computing Surveys 46(1):10:1–10:53 1, 2, 3, 4,

125. Li Y., Wang S., Tian Q., Ding X. (2015) Feature representation for statistical learning based object detection: A review. Pattern Recognition 48(11):3542–3559 3

17. Borji A., Cheng M., Jiang H., Li J. (2014) Salient object detection: A survey. arXiv: 14115878v1 1:1–26 3

12. Bengio Y., Courville A., Vincent P. (2013) Representation learning: A review and new perspectives. IEEE TPAMI 35(8):1798–1828 2, 3, 10

116. LeCun Y., Bengio Y., Hinton G. (2015) Deep learning. Nature 521:436–444 1, 2, 3, 10

133. Litjens G., Kooi T., Bejnordi B., Setio A., Ciompi F., Ghafoorian M., J. van der Laak B. v., Sánchez C. (2017) A survey on deep learning in medical image analysis. Medical Image Analysis 42:60–88 2, 3

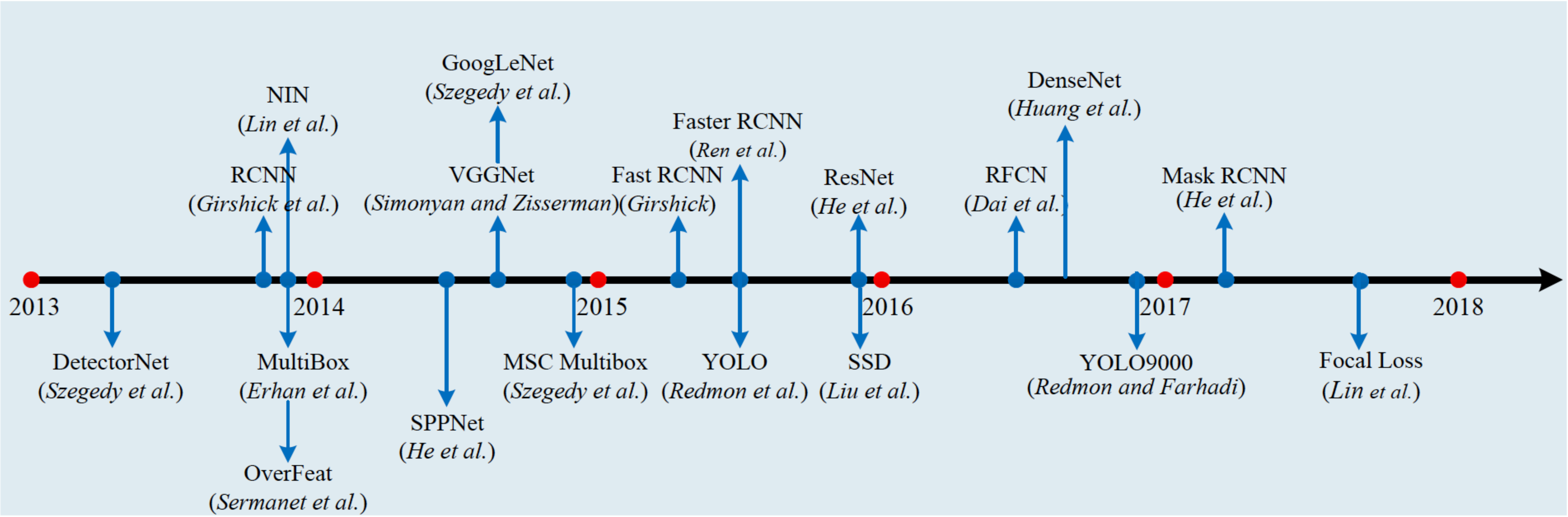
Summarization of Related Surveys (4/4)

19	Recent Advances in Convolutional Neural Networks	[71]	2017	PR	A broad survey of the recent advances in CNN and its applications in computer vision, speech and natural language processing
20	Tutorial: Tools for Efficient Object Detection	–	2015	ICCV15	A short course for object detection only covering recent milestones
21	Tutorial: Deep Learning for Objects and Scenes	–	2017	CVPR17	A high level summary of recent work on deep learning for visual recognition of objects and scenes
22	Tutorial: Instance Level Recognition	–	2017	ICCV17	A short course of recent advances on instance level recognition, including object detection, instance segmentation and human pose prediction
23	Tutorial: Visual Recognition and Beyond	–	2018	CVPR18	This tutorial covers methods and principles behind image classification, object detection, instance segmentation, and semantic segmentation.
24	**Deep Learning for Generic Object Detection**	–	**2018**	**Ours**	**A comprehensive survey of deep learning for generic object detection**

71. Gu J., Wang Z., Kuen J., Ma L., Shahroudy A., Shuai B., Liu T., Wang X., Wang G., Cai J., Chen T. (2017) Recent advances in convolutional neural networks. Pattern Recognition pp. 1–24 2, 3, 10

Milestones in Generic Object Detection

- Nearly all detectors proposed over the last several years are based on one of these milestone detectors, attempting to improve on one or more aspects



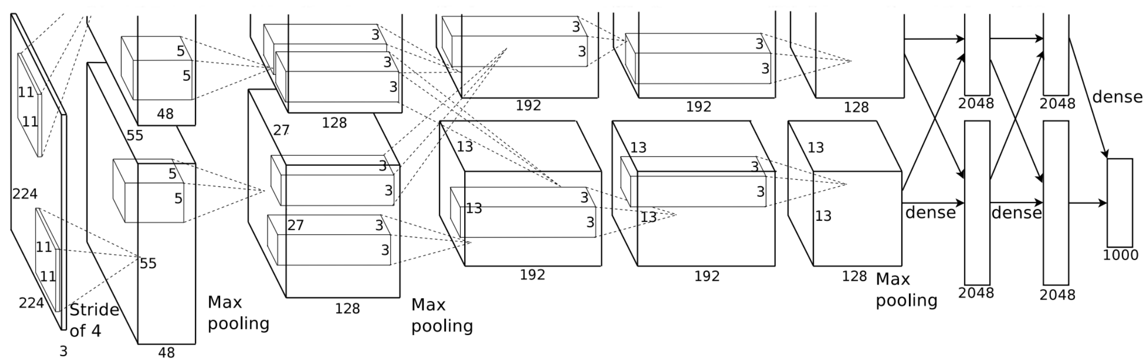
Generic Object Detection Two Main Categories

- Region proposal based (**two stage) framework**
  - Category-independent region proposals are generated from an image
  - Category-specific classifiers are used to determine the category labels of the proposals

- Region proposal free (one **stage) framework**
  - which is a single proposed method which does not separate detection proposal, making the overall pipeline single-stage

Object Detection as Classification: Sliding Window

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background

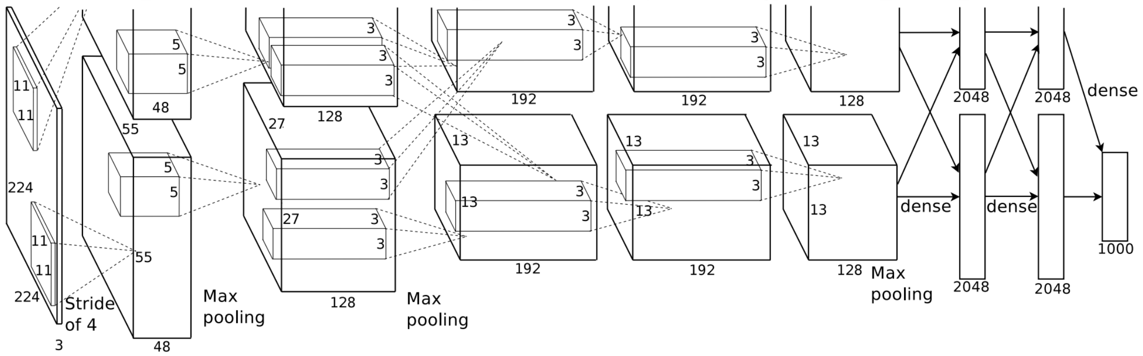


Dog? NO
Cat? NO
Background? YES

Source: http://cs231n.stanford.edu/2017/

Object Detection as Classification: Sliding Window

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background

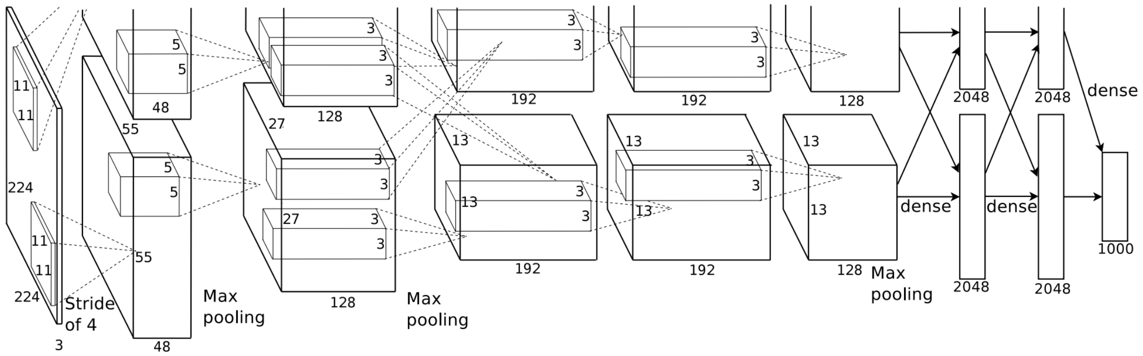


Dog? YES
Cat? NO
Background? NO

Source: http://cs231n.stanford.edu/2017/

Object Detection as Classification: Sliding Window

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background

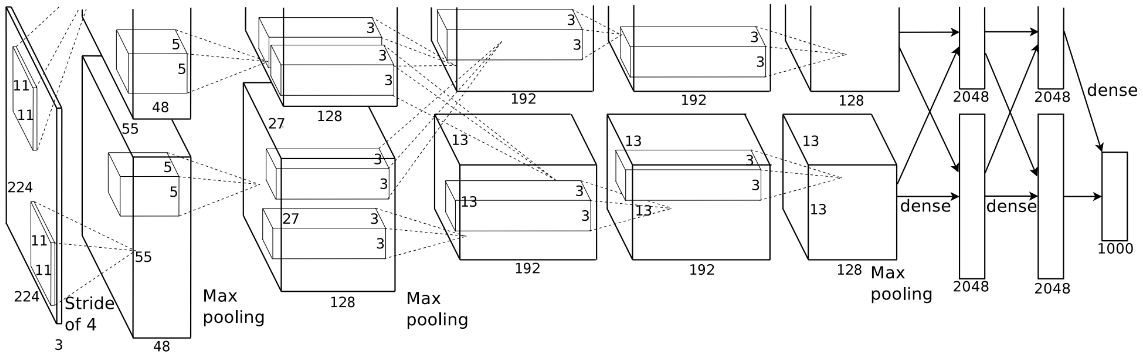


Dog? YES
Cat? NO
Background? NO

Source: http://cs231n.stanford.edu/2017/

Object Detection as Classification: Sliding Window

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background

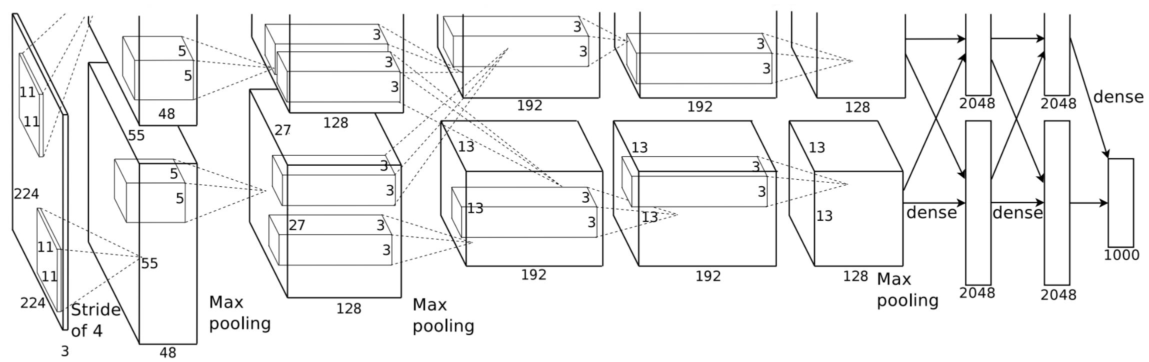


Dog? NO
Cat? YES
Background? NO

Source: http://cs231n.stanford.edu/2017/

Object Detection as Classification: Sliding Window

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



Dog? NO
Cat? YES
Background? NO

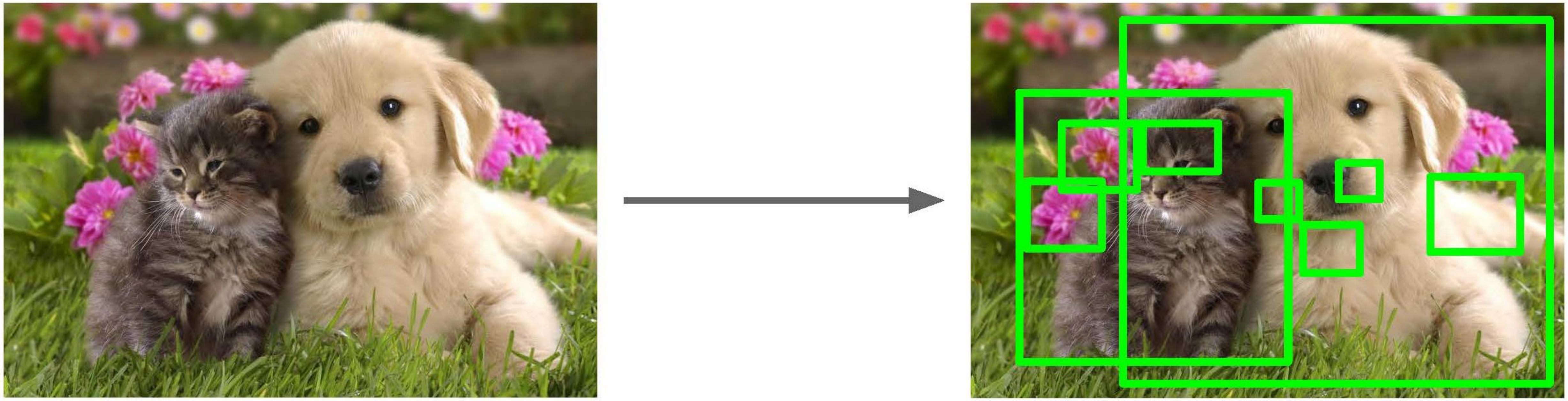
Problem: Need to apply CNN to huge number of locations and scales, very computationally expensive!

Source: http://cs231n.stanford.edu/2017/

Region Proposals

- Find “blobby” image regions that are likely to contain objects
- Relatively fast to run;

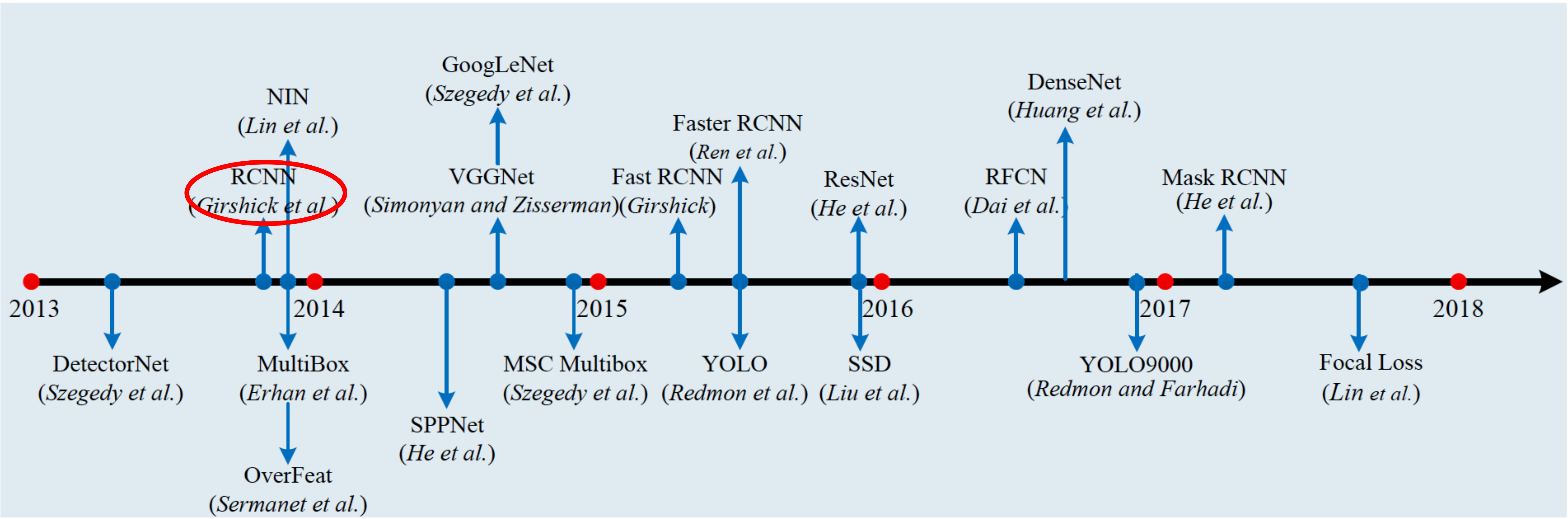
e.g. Selective Search gives 1000 region proposals in a few seconds on CPU



Source: http://cs231n.stanford.edu/2017/

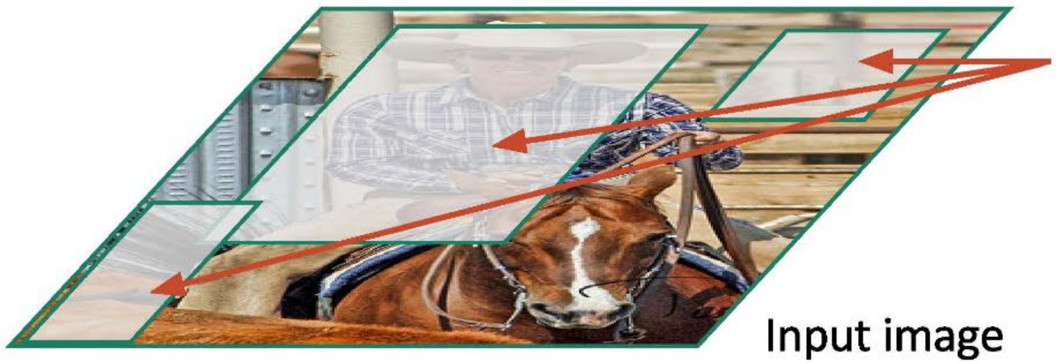
Milestones in Generic Object Detection

- Nearly all detectors proposed over the last several years are based on one of these milestone detectors, attempting to improve on one or more aspects



R-CNN

J. Uijlings *et al*. "Selective search for object recognition," IJCV, 2013.

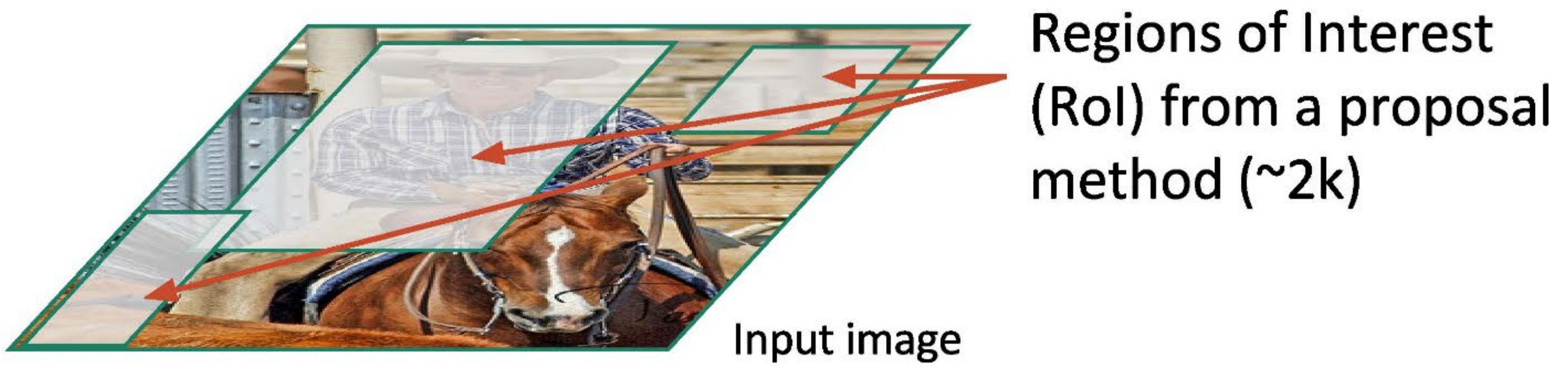


Regions of Interest (RoI) from a proposal method (~2k)

Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

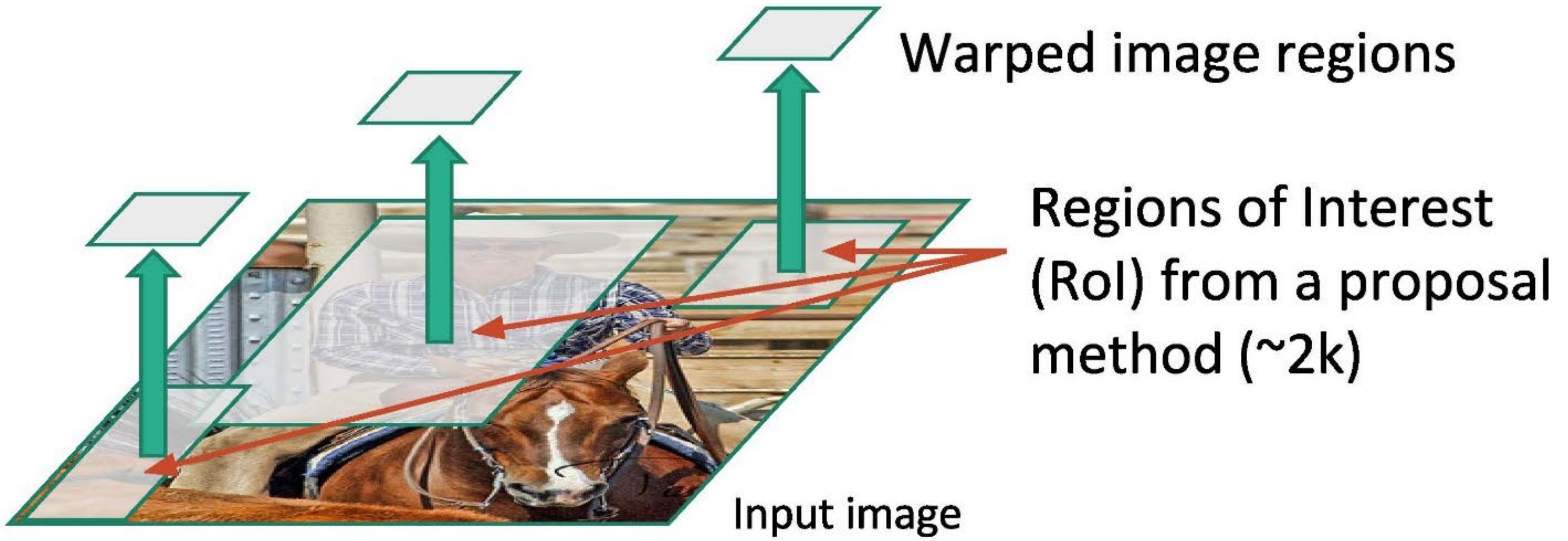
Source: http://cs231n.stanford.edu/2017/

R-CNN



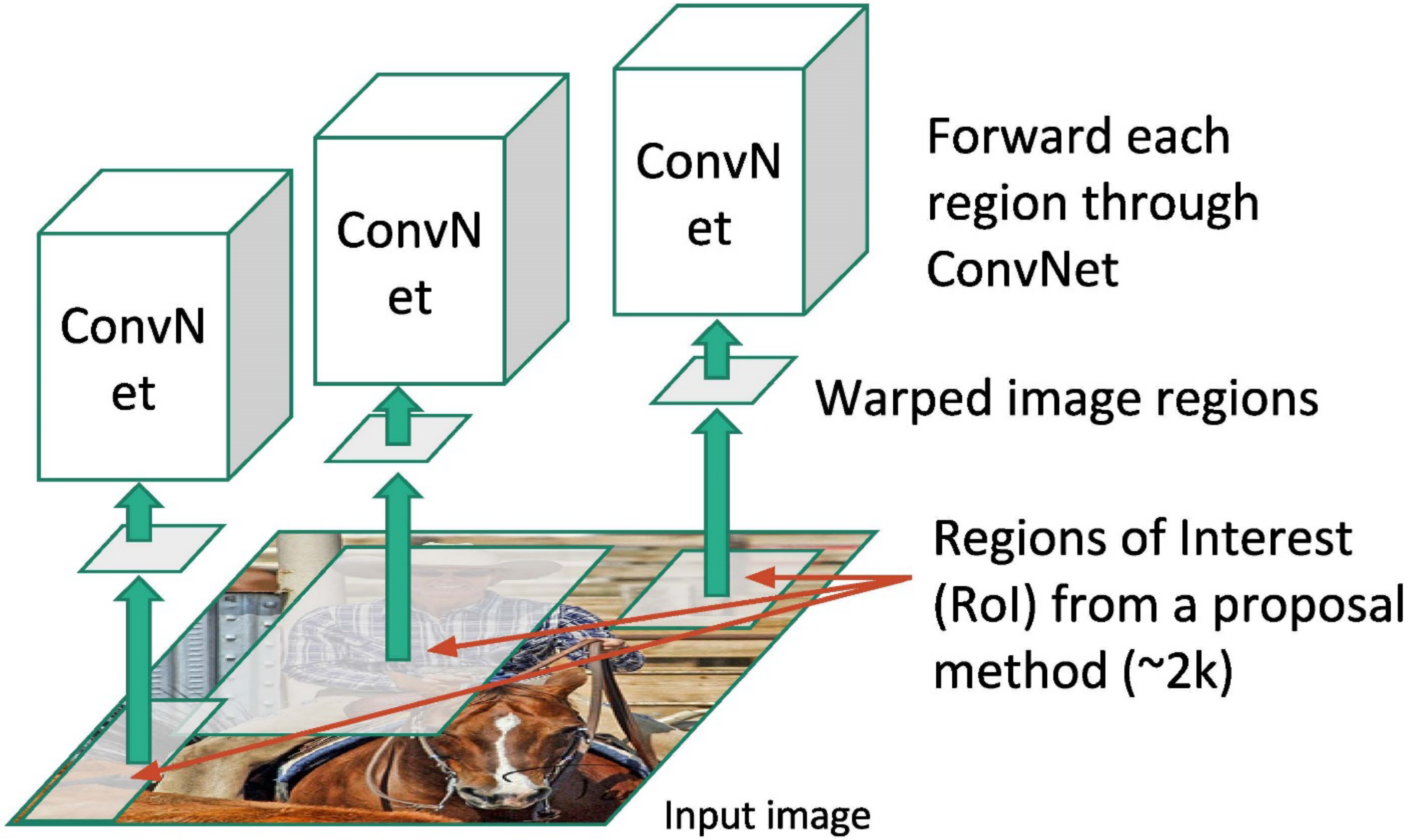
Source: http://cs231n.stanford.edu/2017/

R-CNN



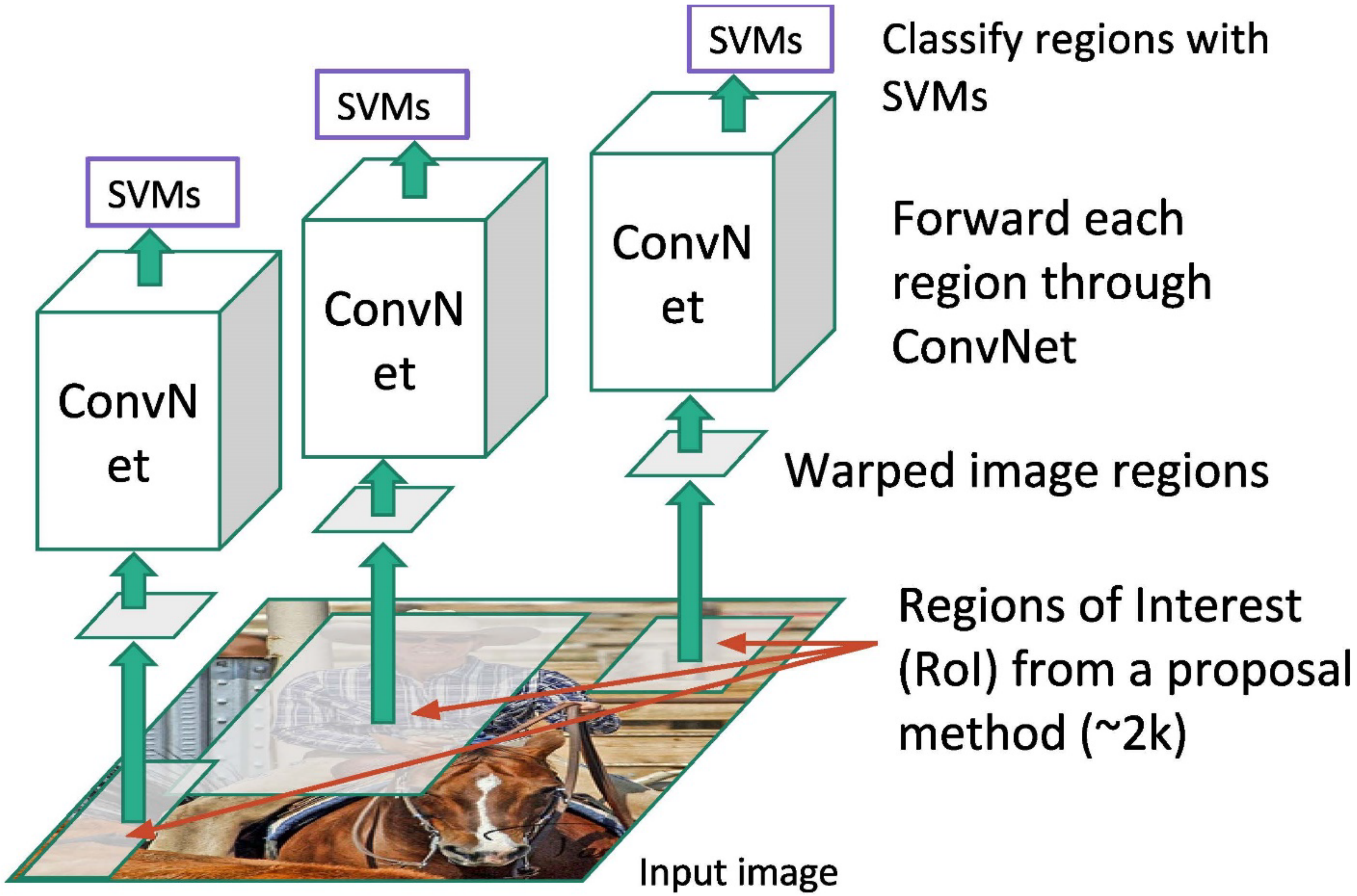
Source: http://cs231n.stanford.edu/2017/

R-CNN



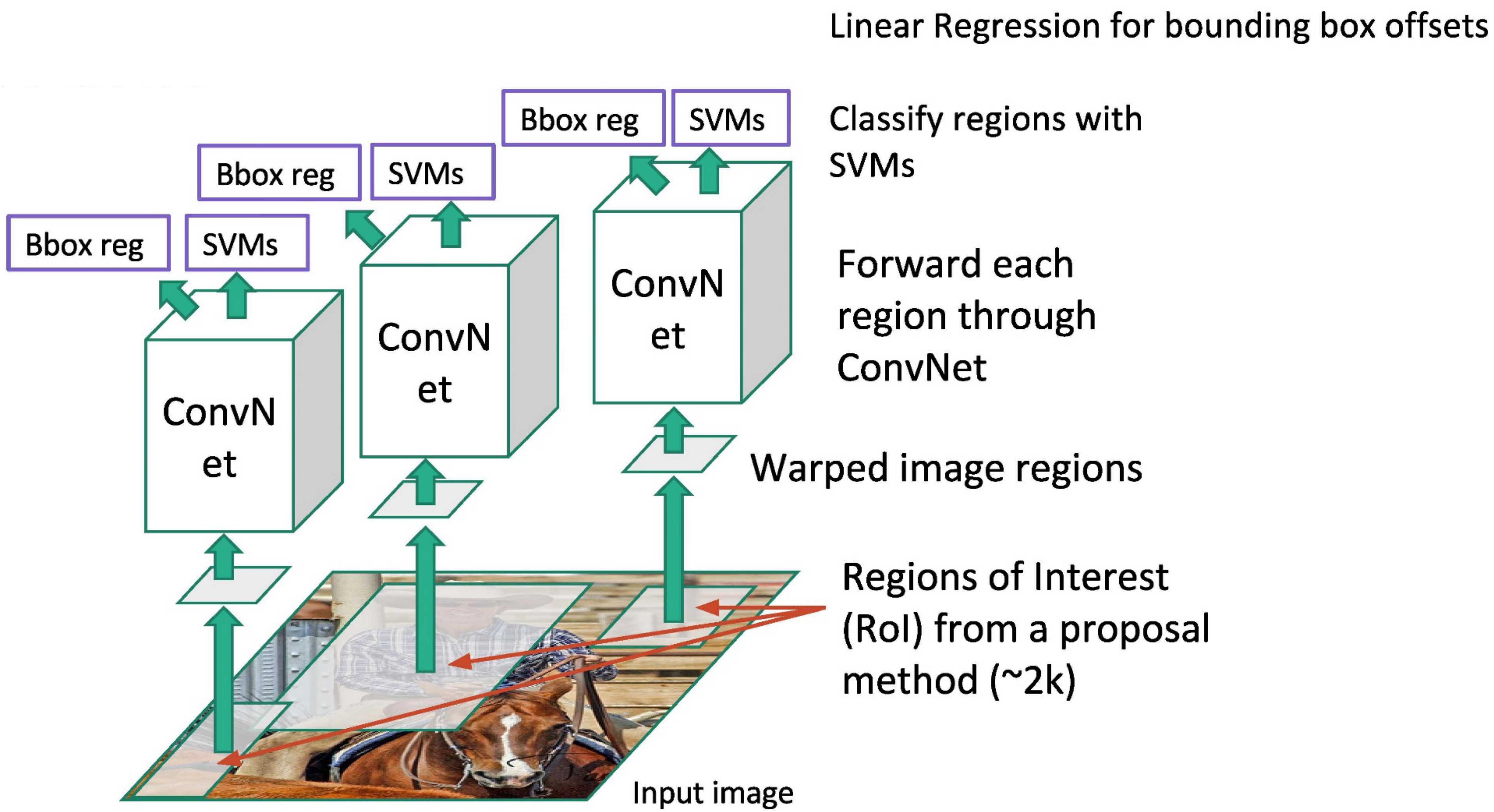
Source: http://cs231n.stanford.edu/2017/

R-CNN



Source: http://cs231n.stanford.edu/2017/

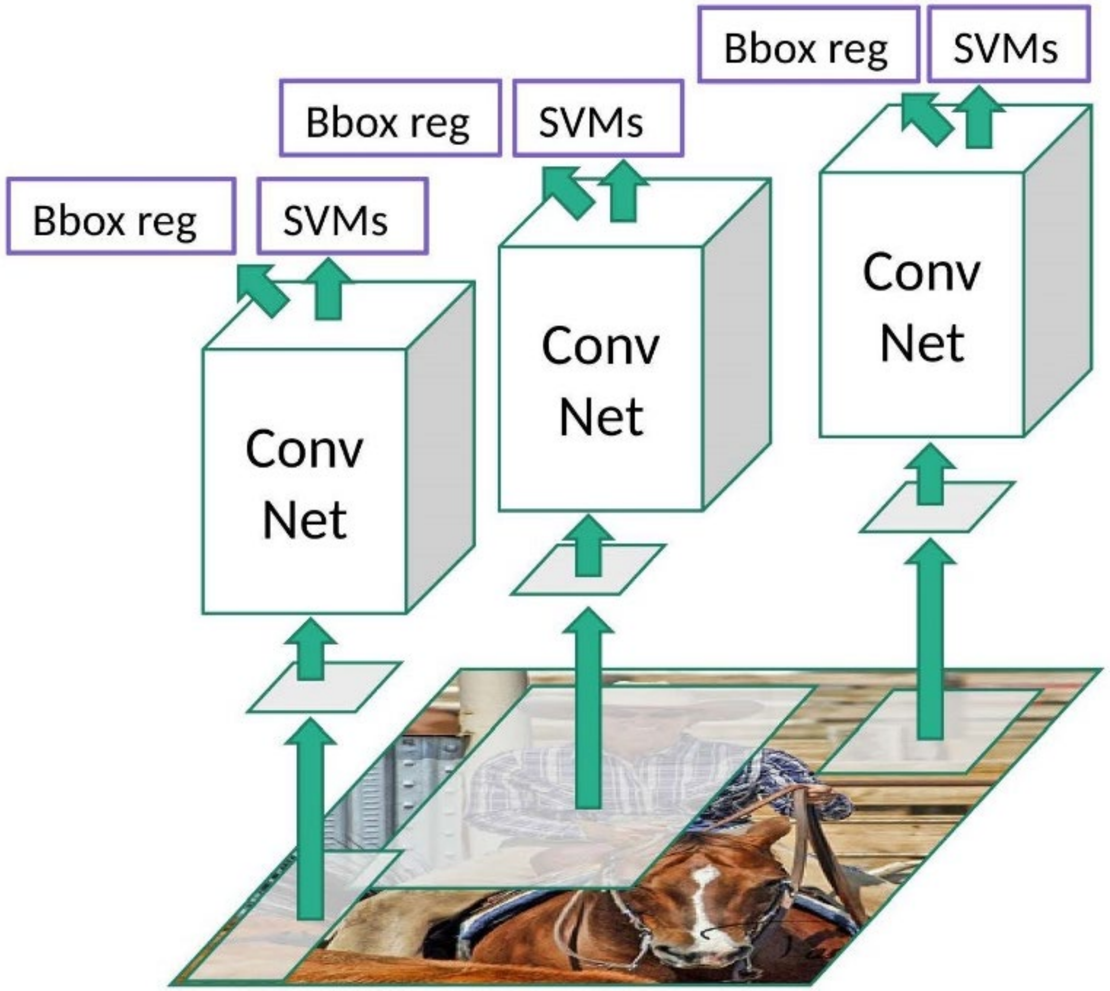
R-CNN



Source: http://cs231n.stanford.edu/2017/

R-CNN: Problems

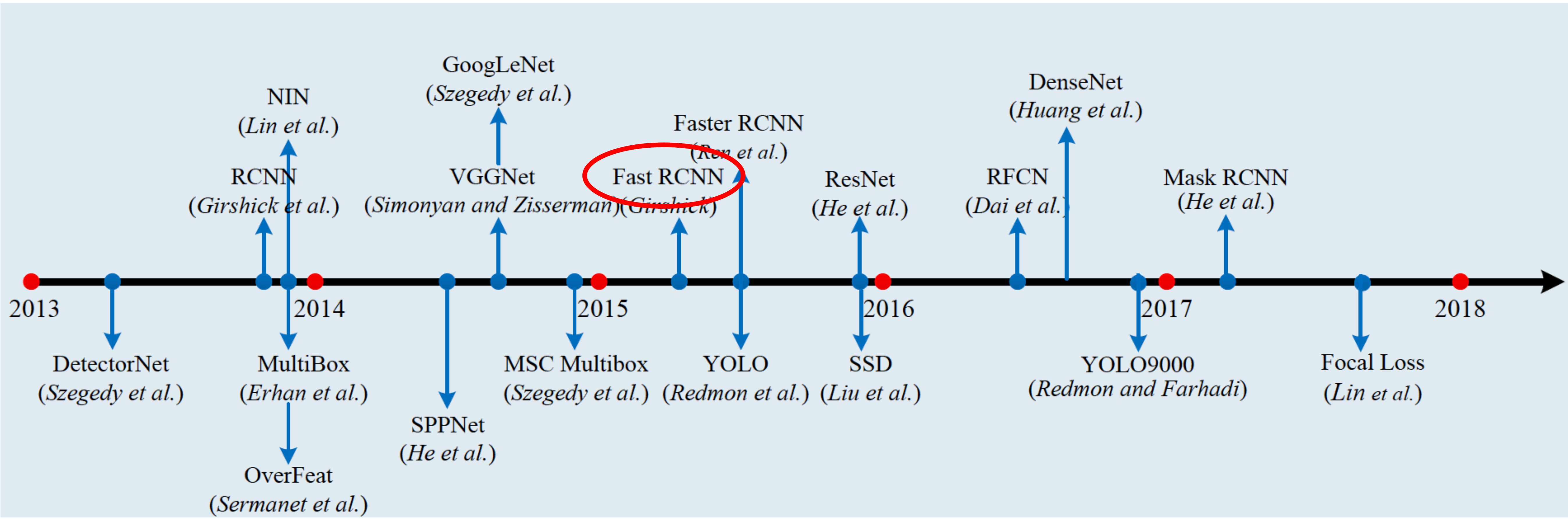
- Training is slow (84h), taking a lot of disk space
- Inference (detection) is slow
- 47s / image with VGG16 [Simonyan, ICLR15]



Source: http://cs231n.stanford.edu/2017/

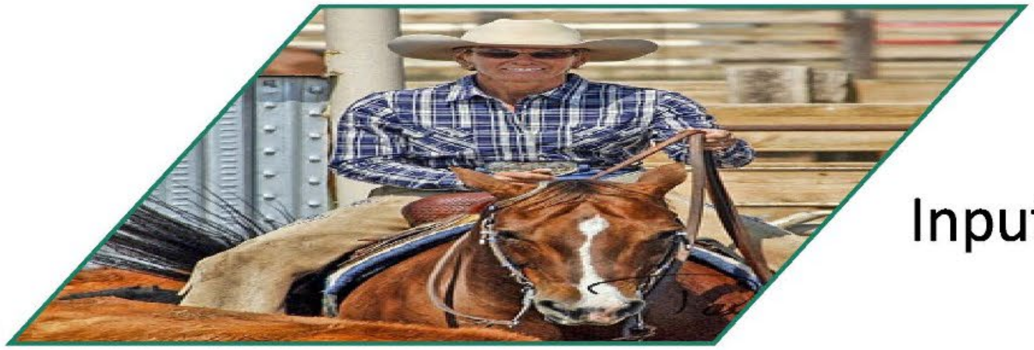
Milestones in Generic Object Detection

- Nearly all detectors proposed over the last several years are based on one of these milestone detectors, attempting to improve on one or more aspects



Fast R-CNN

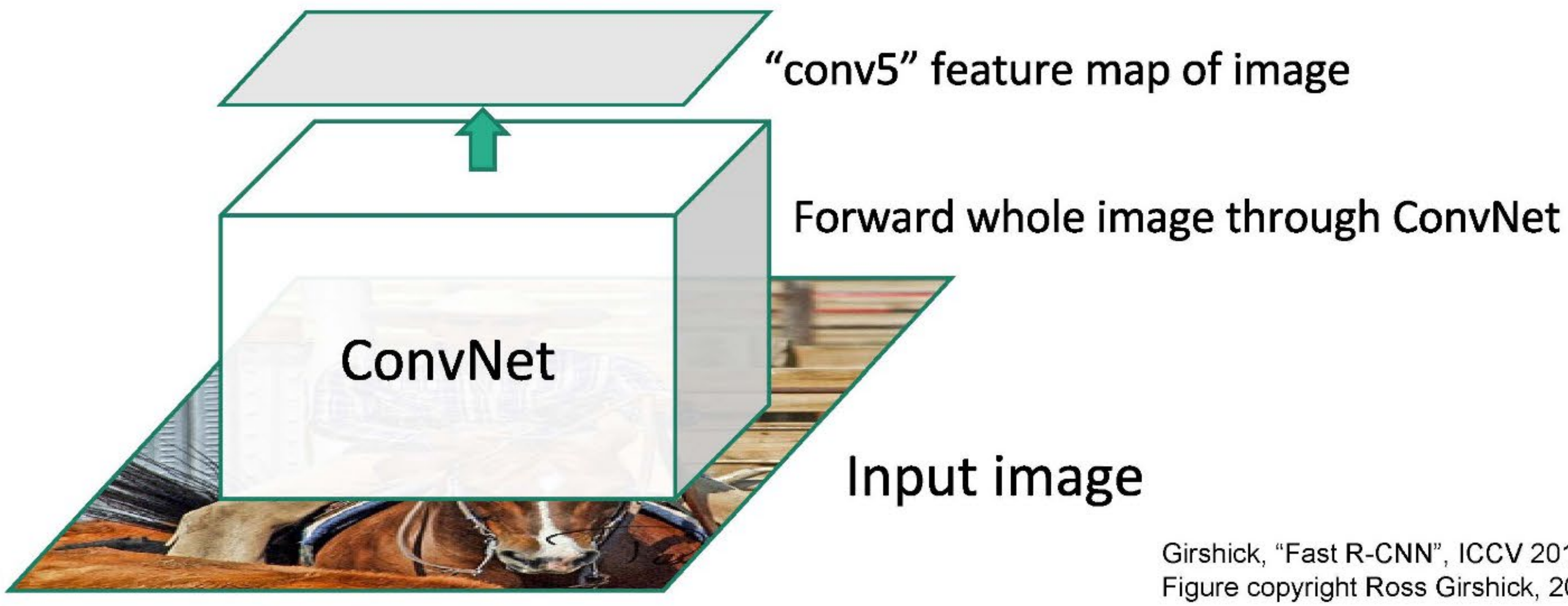
R. Girshick, “Fast R-CNN,” ICCV, 2015.



Input image

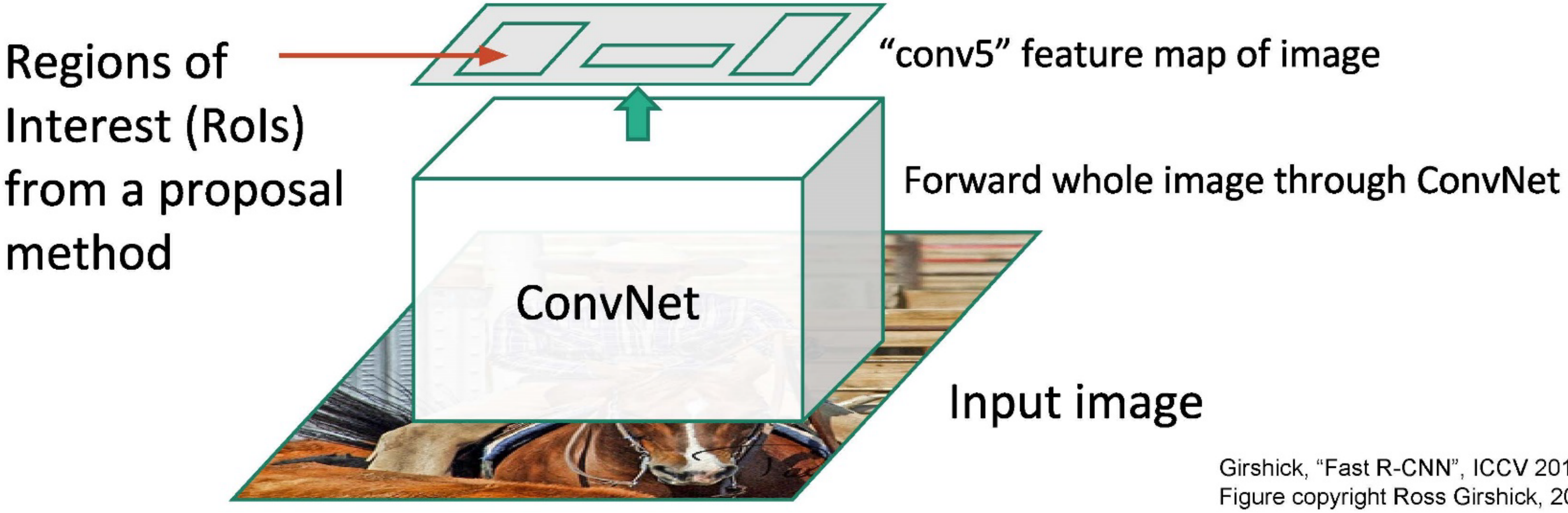
Girshick, “Fast R-CNN”, ICCV 2015.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Fast R-CNN



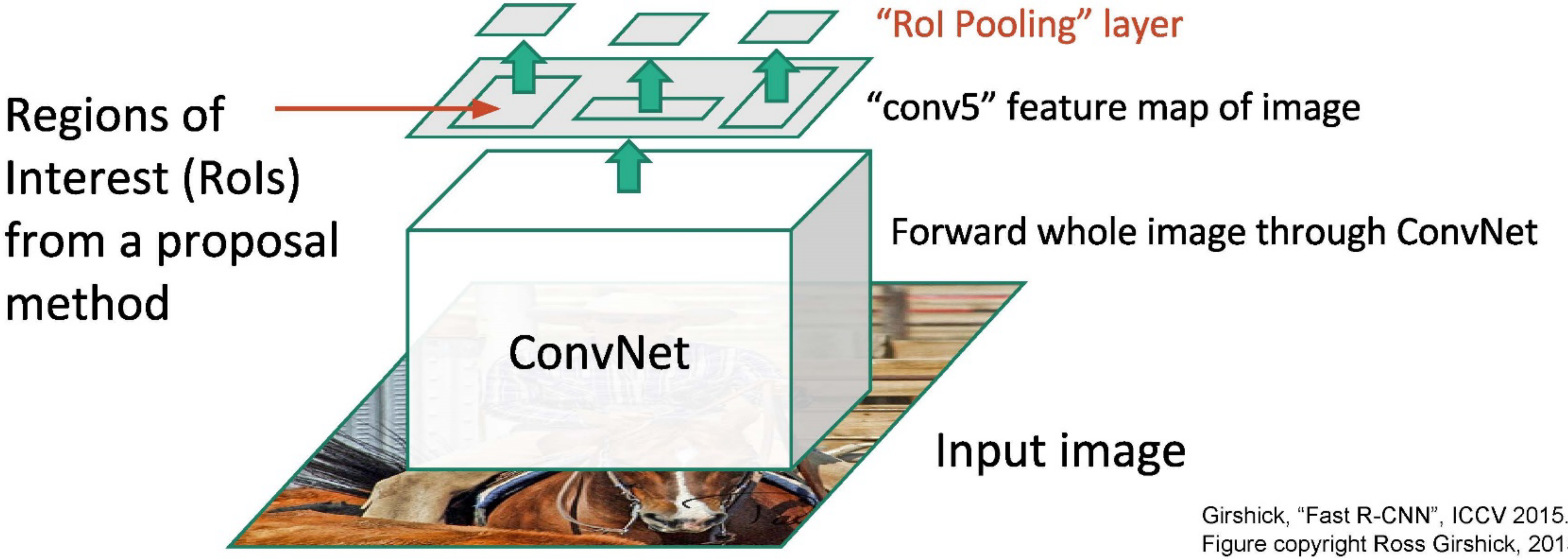
Girshick, “Fast R-CNN”, ICCV 2015.

Fast R-CNN



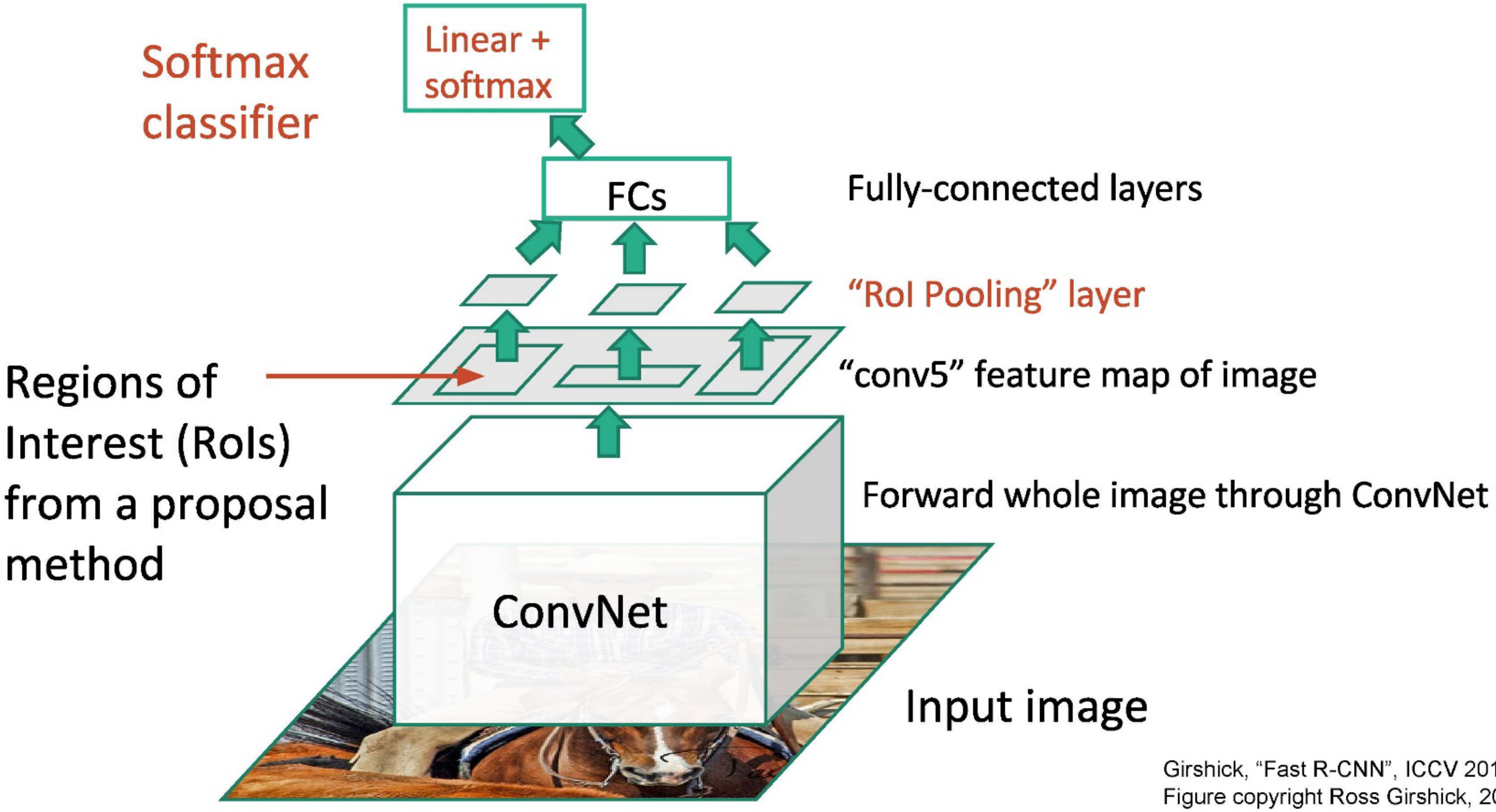
Girshick, "Fast R-CNN", ICCV 2015.

Fast R-CNN



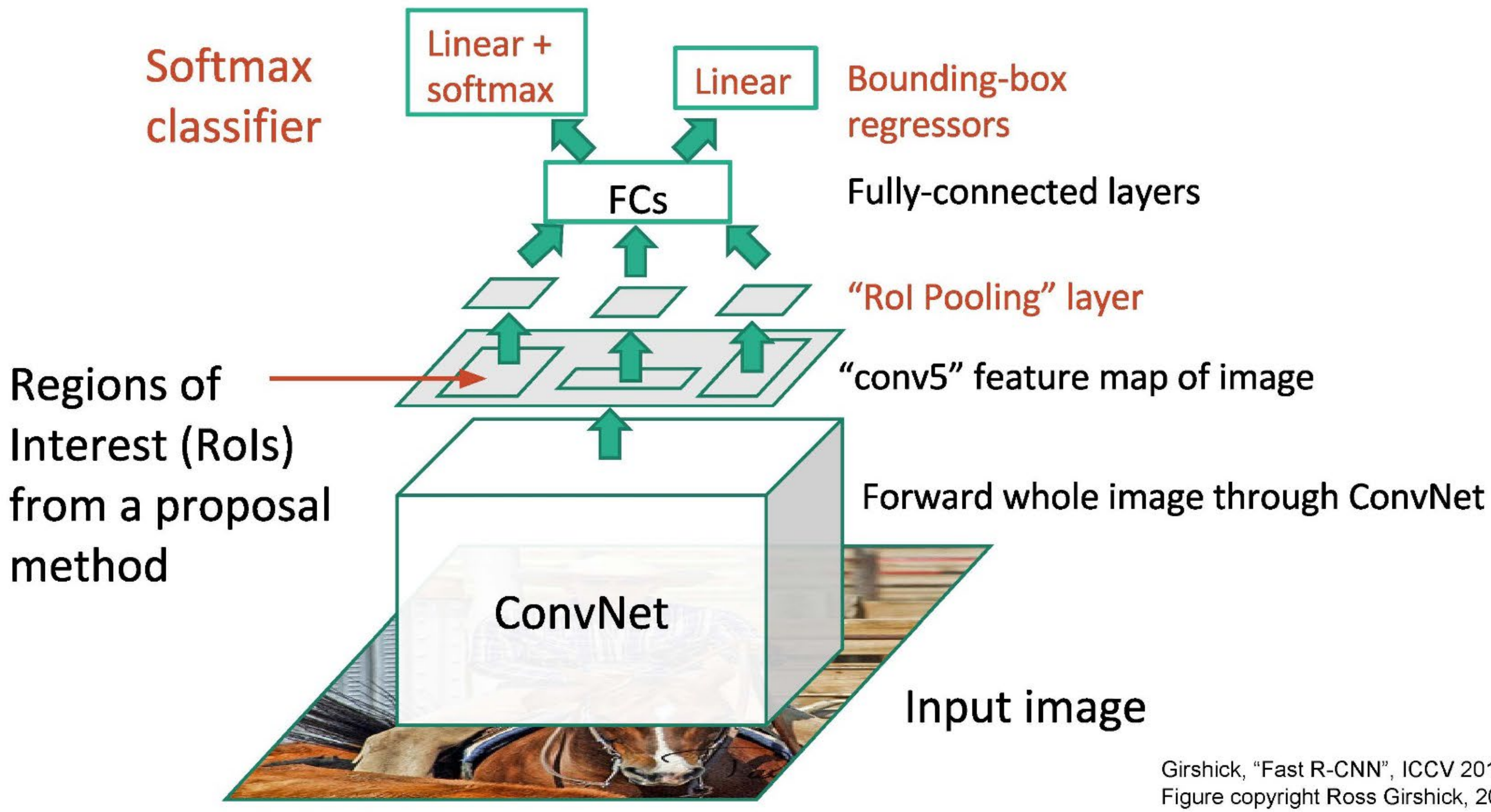
Girshick, "Fast R-CNN", ICCV 2015.

Fast R-CNN



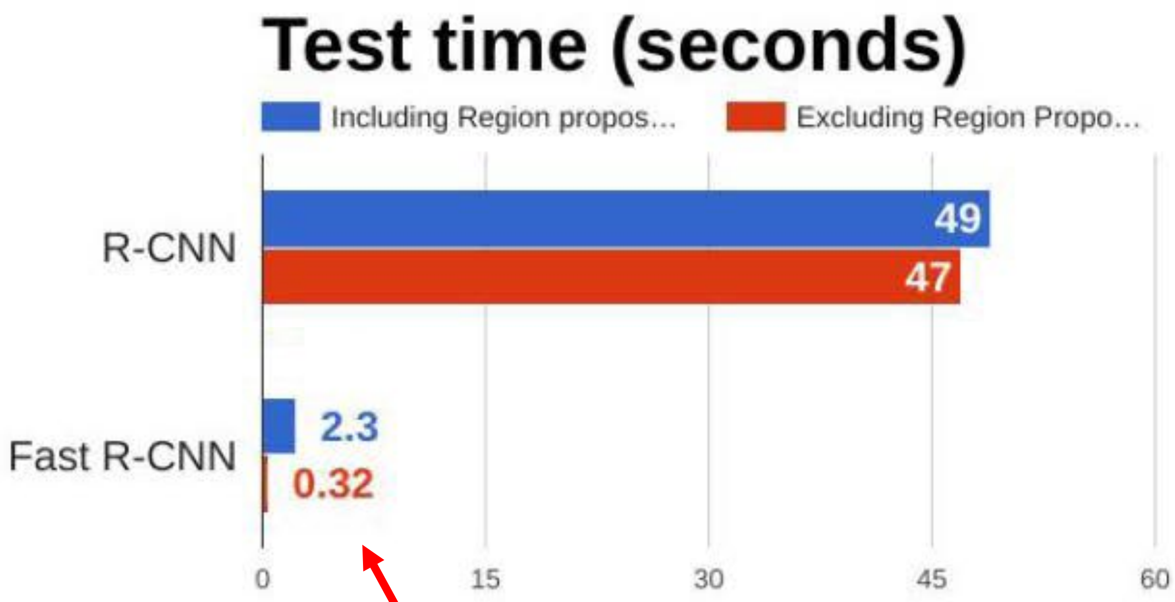
Girshick, "Fast R-CNN", ICCV 2015.

Fast R-CNN



Girshick, "Fast R-CNN", ICCV 2015.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

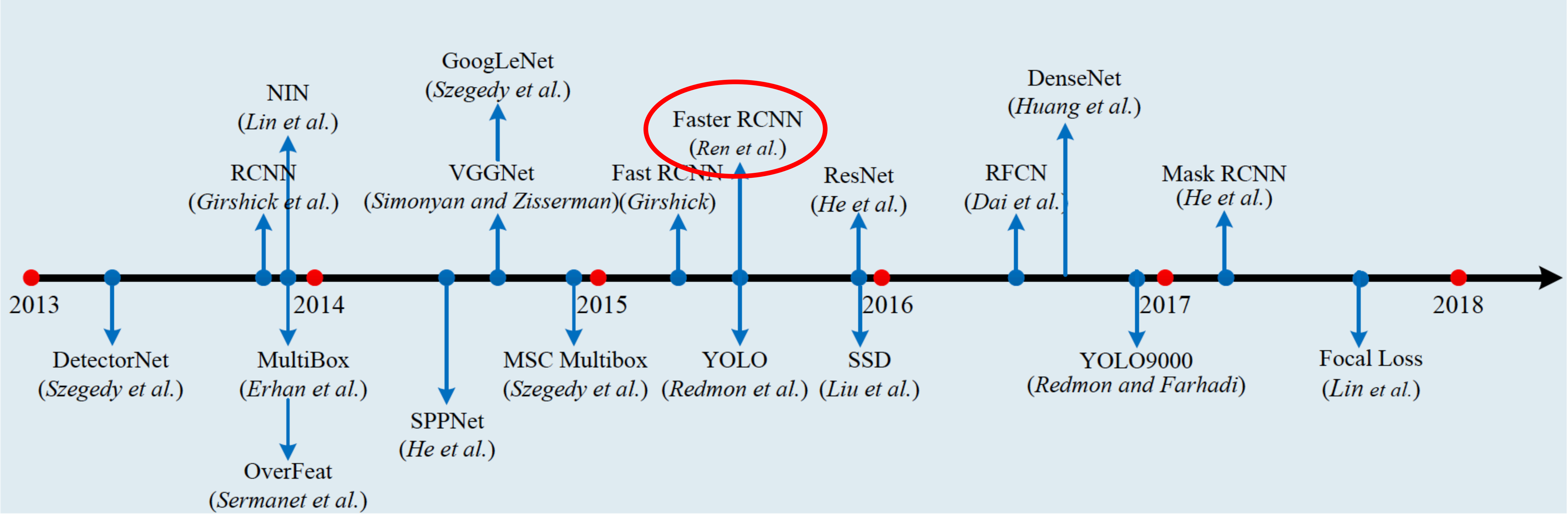
R-CNN Vs. Fast R-CNN



Problem: Runtime dominated by region proposals!

Milestones in Generic Object Detection

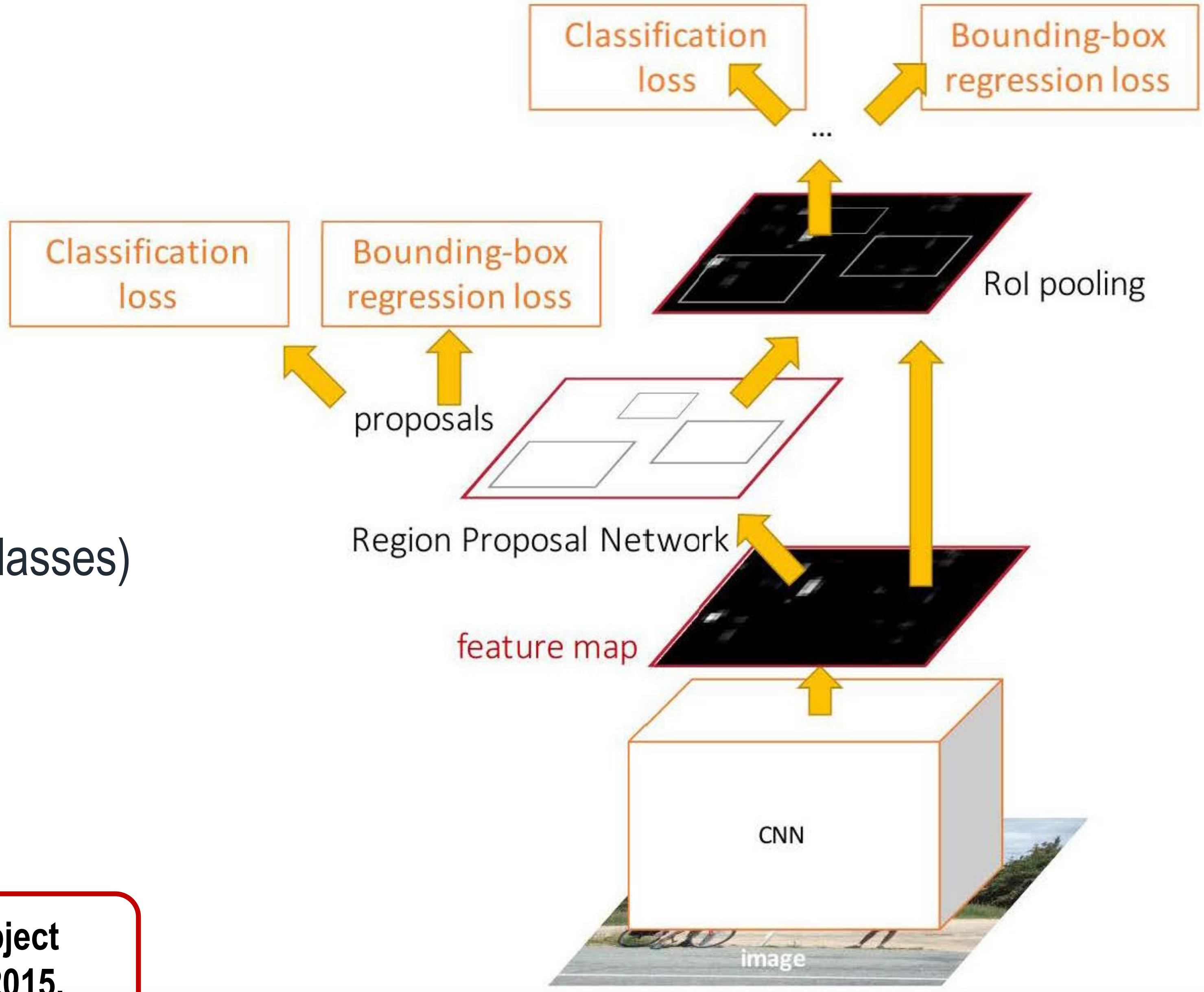
- Nearly all detectors proposed over the last several years are based on one of these milestone detectors, attempting to improve on one or more aspects.



Faster R-CNN: Make CNN do proposals!

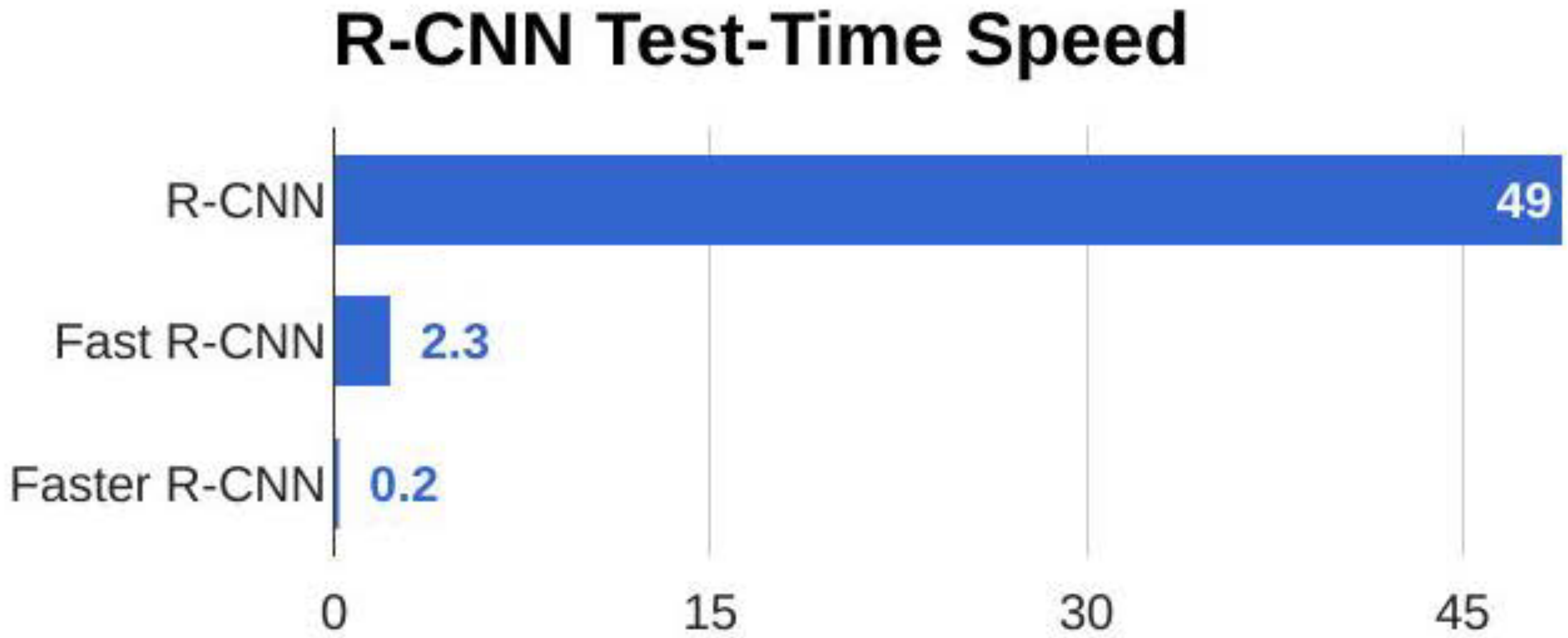
- Insert Region Proposal Network (RPN) to predict proposals from features

- Jointly train with 4 losses:
  1. RPN classify object / not object
  2. RPN regress box coordinates
  3. Final classification score (object classes)
  4. Final box coordinates



S. Ren *et al.* "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," NIPS, 2015.

R-CNN Vs. Fast R-CNN Vs Faster R-CNN



Source: http://cs231n.stanford.edu/2017/

More CNN based Architectures (1/2)

No.	DCNN Architecture	#Paras ($\times 10^6$)	#Layers (CONV+FC)	Test Error (Top 5)	First Used In	Highlights
1	AlexNet [110]	57	$5+2$	15.3%	[65]	The first DCNN; The historical turning point of feature representation from traditional to CNN; In the classification task of ILSVRC2012 competition, achieved a winning Top 5 test error rate of 15.3%, compared to 26.2% given by the second best entry.
2	OverFeat [183]	140	$6+2$	13.6%	[183]	Similar to AlexNet, differences including a smaller stride for CONV1 and 2, different filter size for some layers, more filters for some layers.
3	ZFNet (fast) [234]	58	$5+2$	14.8%	[77]	Highly similar to AlexNet, with a smaller filter size in CONV1 and a smaller stride for CONV1 and 2.
4	VGGNet16 [191]	134	$13+2$	6.8%	[64]	Increasing network depth significantly with small 3×3 convolution filters; Significantly better performance.
5	GoogLeNet [200]	6	22	6.7%	[200]	With the use of Inception module which concatenates feature maps produced by filters of different sizes, the network goes wider and parameters are much less than those of AlexNet *etc*.
6	Inception v2 [99]	12	31	4.8%	[88]	Faster training with the introduce of Batch Normalization.
7	Inception v3 [201]	22	47	3.6%		Going deeper with Inception building blocks in efficient ways.
8	YOLONet [174]	64	$24+1$	–	[174]	A network inspired by GoogLeNet used in YOLO detector.
9	ResNet50 [79]	23.4	49	3.6% (ResNets)	[79]	With the use of residual connections, substantially deeper but with fewer parameters than previous DCNNs (except for GoogLeNet).
10	ResNet101 [79]	42	100		[79]	

More CNN based Architectures (2/2)

11	InceptionResNet v1 [202]	21	87	3.1% (Ensemble)	[96]	A residual version of Inception with similar computational cost of Inception v3, but with faster training process.
12	InceptionResNet v2 [202]	30	95			A costlier residual version of Inception, with significantly improved recognition performance.
13	Inception v4 [202]	41	75			A Inception variant without residual connections with roughly the same recognition performance as InceptionResNet v2, but significantly slower.
14	ResNeXt50 [223]	23	49	3.0%	[223]	Repeating a building block that aggregates a set of transformations with the same topology.
15	DenseNet201 [94]	18	200	–	[246]	Design dense block, which connects each layer to every other layer in a feed forward fashion; Alleviate the vanishing gradient problem, strengthen feature propagation, encourage feature reuse, and substantially reduce the number of parameters.
16	DarkNet [173]	20	19	–	[173]	Similar to VGGNet, but with significantly less parameters due to the use of fewer filters at each layer.
17	MobileNet [88]	3.2	$27 + 1$	–	[88]	Light weight deep CNNs using depthwise separable convolutions for mobile applications.
18	SE ResNet50 [91]	26	50	2.3% (SENets)	[91]	Proposing a novel block called *Squeeze and Excitation* to model feature channel relationship; Can be flexibly used in all existing CNNs to improve recognition performance at minimal additional computational cost.

Deep Learning Object Detection

- https://github.com/hoya012/deep_learning_object_detection

R-CNN (2013.11) → **OverFeat** (ICLR' 14) → MultiBox (CVPR' 14) → SPP-Net (ECCV' 14) → MR-CNN (ICCV' 15) → DeepBox (ICCV' 15) → AttentionNet (ICCV' 15) →

Fast R-CNN (ICCV' 15) → DeepProposal (ICCV' 15) → **RPN** (NIPS' 15) → **Faster R-CNN** (NIPS' 15) → **YOLO v1** (CVPR' 16) → G-CNN (CVPR' 16) → AZNet (CVPR' 16) →

Inside-OutsideNet(ION) (CVPR' 16) → HyperNet (CVPR' 16) → OHEM (CVPR' 16) → CRAFT (CVPR' 16) → MultiPathNet(MPN) (BMVC' 16) → **SSD** (ECCV' 16) →

GBDNet (ECCV' 16) → CPF (ECCV' 16) → MS-CNN (ECCV' 16) → R-FCN (NIPS' 16) → PVANET (NIPSW' 16) → DeepID-Net (PAMI' 16) → NoC (TPAMI' 16) → DSSD (Arxiv' 17) → TDM (CVPR' 17) →

Feature Pyramid Net(**FPN**) (CVPR' 17) → **YOLO v2** (CVPR' 17) → RON (CVPR' 17) → DCN (ICCV' 17) → DeNet (ICCV' 17) → CoupleNet (ICCV' 17) → **RetinaNet** (ICCV' 17) →

Mask R-CNN (ICCV' 17) → DSOD (ICCV' 17) → SMN (ICCV' 17) → **YOLO v3** (Arxiv' 18) → SIN (CVPR' 18) → STDN (CVPR' 18) → RefineDet (CVPR' 18) → RFBNet (ECCV' 18) → ···

Collections of Paper and Code: 2014

2014

- **[R-CNN]** Rich feature hierarchies for accurate object detection and semantic segmentation | Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik | **[CVPR' 14]** | `[pdf]` `[official code - caffe]`
- **[OverFeat]** OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks | Pierre Sermanet, et al. | **[ICLR' 14]** | `[pdf]` `[official code - torch]`
- **[MultiBox]** Scalable Object Detection using Deep Neural Networks | Dumitru Erhan, et al. | **[CVPR' 14]** | `[pdf]`
- **[SPP-Net]** Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition | Kaiming He, et al. | **[ECCV' 14]** | `[pdf]` `[official code - caffe]` `[unofficial code - keras]` `[unofficial code - tensorflow]`

Collections of Paper and Code: 2015

2015

- **[MR-CNN]** Object detection via a multi-region & semantic segmentation-aware CNN model | Spyros Gidaris, Nikos Komodakis | **[ICCV' 15]** | `[pdf]` `[official code - caffe]`
- **[DeepBox]** DeepBox: Learning Objectness with Convolutional Networks | Weicheng Kuo, Bharath Hariharan, Jitendra Malik | **[ICCV' 15]** | `[pdf]` `[official code - caffe]`
- **[AttentionNet]** AttentionNet: Aggregating Weak Directions for Accurate Object Detection | Donggeun Yoo, et al. | **[ICCV' 15]** | `[pdf]`
- **[Fast R-CNN]** Fast R-CNN | Ross Girshick | **[ICCV' 15]** | `[pdf]` `[official code - caffe]`
- **[DeepProposal]** DeepProposal: Hunting Objects by Cascading Deep Convolutional Layers | Amir Ghodrati, et al. | **[ICCV' 15]** | `[pdf]` `[official code - matconvnet]`
- **[Faster R-CNN, RPN]** Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks | Shaoqing Ren, et al. | **[NIPS' 15]** | `[pdf]` `[official code - caffe]` `[unofficial code - tensorflow]` `[unofficial code - pytorch]`

2016

- **[YOLO v1]** You Only Look Once: Unified, Real-Time Object Detection | Joseph Redmon, et al. | **[CVPR' 16]** | `[pdf]` `[official code - c++]`
- **[G-CNN]** G-CNN: an Iterative Grid Based Object Detector | Mahyar Najibi, et al. | **[CVPR' 16]** | `[pdf]`
- **[AZNet]** Adaptive Object Detection Using Adjacency and Zoom Prediction | Yongxi Lu, Tara Javidi. | **[CVPR' 16]** | `[pdf]`
- **[ION]** Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks | Sean Bell, et al. | **[CVPR' 16]** | `[pdf]`
- **[HyperNet]** HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection | Tao Kong, et al. | **[CVPR' 16]** | `[pdf]`
- **[OHEM]** Training Region-based Object Detectors with Online Hard Example Mining | Abhinav Shrivastava, et al. | **[CVPR' 16]** | `[pdf]` `[official code - caffe]`
- **[CRAPF]** CRAFT Objects from Images | Bin Yang, et al. | **[CVPR' 16]** | `[pdf]` `[official code - caffe]`
- **[MPN]** A MultiPath Network for Object Detection | Sergey Zagoruyko, et al. | **[BMVC' 16]** | `[pdf]` `[official code - torch]`

Collections of Paper and Code: 2016

- **[SSD]** SSD: Single Shot MultiBox Detector | Wei Liu, et al. | **[ECCV' 16]** | `[pdf]` `[official code - caffe]` `[unofficial code - tensorflow]` `[unofficial code - pytorch]`
- **[GBDNet]** Crafting GBD-Net for Object Detection | Xingyu Zeng, et al. | **[ECCV' 16]** | `[pdf]` `[official code - caffe]`
- **[CPF]** Contextual Priming and Feedback for Faster R-CNN | Abhinav Shrivastava and Abhinav Gupta | **[ECCV' 16]** | `[pdf]`
- **[MS-CNN]** A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection | Zhaowei Cai, et al. | **[ECCV' 16]** | `[pdf]` `[official code - caffe]`
- **[R-FCN]** R-FCN: Object Detection via Region-based Fully Convolutional Networks | Jifeng Dai, et al. | **[NIPS' 16]** | `[pdf]` `[official code - caffe]` `[unofficial code - caffe]`
- **[PVANET]** PVANET: Deep but Lightweight Neural Networks for Real-time Object Detection | Kye-Hyeon Kim, et al. | **[NIPSW' 16]** | `[pdf]` `[official code - caffe]`
- **[DeepID-Net]** DeepID-Net: Deformable Deep Convolutional Neural Networks for Object Detection | Wanli Ouyang, et al. | **[PAMI' 16]** | `[pdf]`
- **[NoC]** Object Detection Networks on Convolutional Feature Maps | Shaoqing Ren, et al. | **[TPAMI' 16]** | `[pdf]`

Collections of Paper and Code: 2017

2017

- **[DSSD]** DSSD : Deconvolutional Single Shot Detector | Cheng-Yang Fu1, et al. | **[Arxiv' 17]** | `[pdf]` `[official code - caffe]`
- **[TDM]** Beyond Skip Connections: Top-Down Modulation for Object Detection | Abhinav Shrivastava, et al. | **[CVPR' 17]** | `[pdf]`
- **[FPN]** Feature Pyramid Networks for Object Detection | Tsung-Yi Lin, et al. | **[CVPR' 17]** | `[pdf]` `[unofficial code - caffe]`
- **[YOLO v2]** YOLO9000: Better, Faster, Stronger | Joseph Redmon, Ali Farhadi | **[CVPR' 17]** | `[pdf]` `[official code - c++]` `[unofficial code - caffe]` `[unofficial code - tensorflow]` `[unofficial code - tensorflow]` `[unofficial code - pytorch]`
- **[RON]** RON: Reverse Connection with Objectness Prior Networks for Object Detection | Tao Kong, et al. | **[CVPR' 17]** | `[pdf]` `[official code - caffe]` `[unofficial code - tensorflow]`
- **[DCN]** Deformable Convolutional Networks | Jifeng Dai, et al. | **[ICCV' 17]** | `[pdf]` `[official code - mxnet]` `[unofficial code - tensorflow]` `[unofficial code - pytorch]`

Collections of Paper and Code: 2017

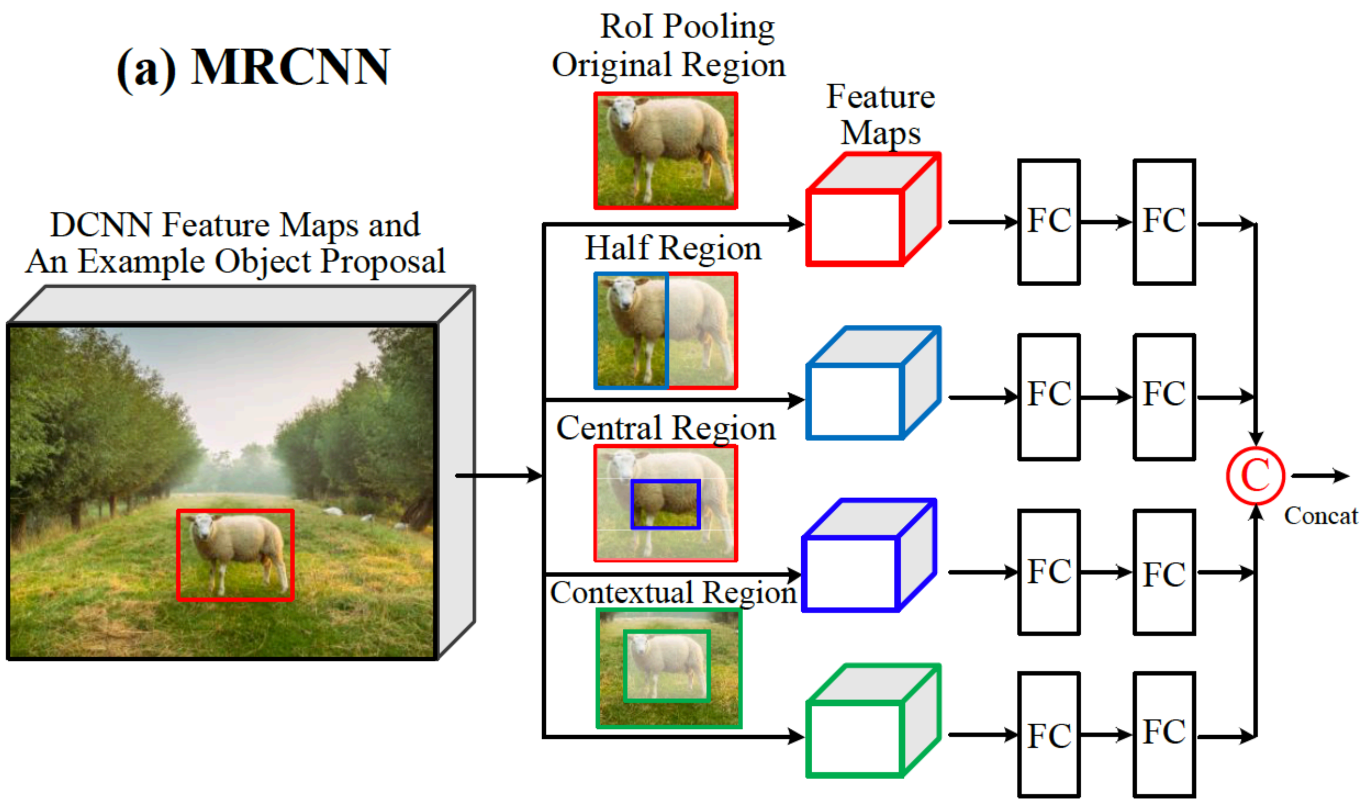
- **[DeNet]** DeNet: Scalable Real-time Object Detection with Directed Sparse Sampling | Lachlan Tychsen-Smith, Lars Petersson | **[ICCV' 17]** | `[pdf]` `[official code - theano]`
- **[CoupleNet]** CoupleNet: Coupling Global Structure with Local Parts for Object Detection | Yousong Zhu, et al. | **[ICCV' 17]** | `[pdf]` `[official code - caffe]`
- **[RetinaNet]** Focal Loss for Dense Object Detection | Tsung-Yi Lin, et al. | **[ICCV' 17]** | `[pdf]` `[official code - keras]` `[unofficial code - pytorch]` `[unofficial code - mxnet]` `[unofficial code - tensorflow]`
- **[Mask R-CNN]** Mask R-CNN | Kaiming He, et al. | **[ICCV' 17]** | `[pdf]` `[official code - caffe2]` `[unofficial code - tensorflow]` `[unofficial code - tensorflow]` `[unofficial code - pytorch]`
- **[DSOD]** DSOD: Learning Deeply Supervised Object Detectors from Scratch | Zhiqiang Shen, et al. | **[ICCV' 17]** | `[pdf]` `[official code - caffe]` `[unofficial code - pytorch]`
- **[SMN]** Spatial Memory for Context Reasoning in Object Detection | Xinlei Chen, Abhinav Gupta | **[ICCV' 17]** | `[pdf]`

2018

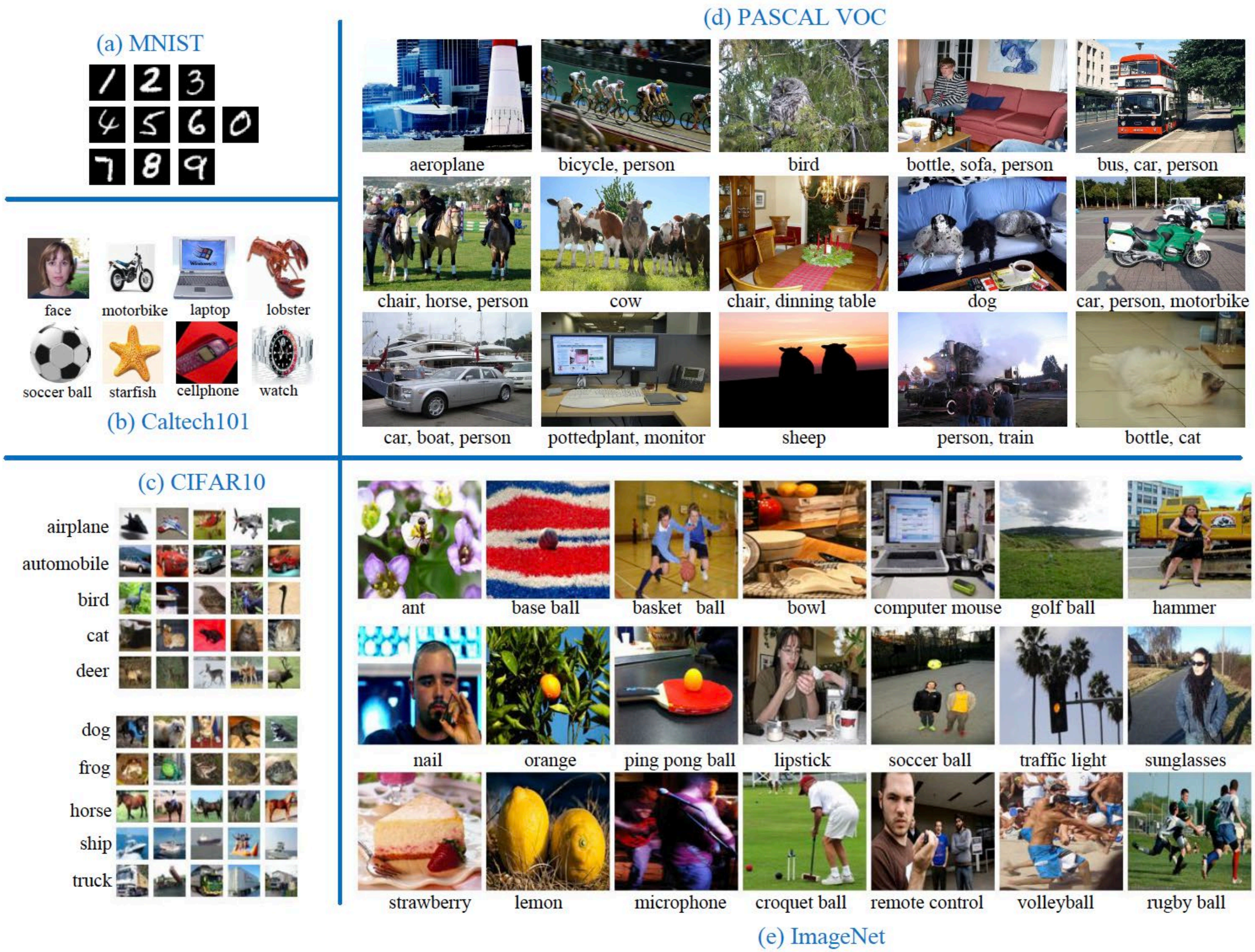
- **[YOLO v3]** YOLOv3: An Incremental Improvement | Joseph Redmon, Ali Farhadi | **[Arxiv' 18]** | `[pdf]` `[official code - c++]` `[unofficial code - pytorch]` `[unofficial code - pytorch]` `[unofficial code - keras]` `[unofficial code - tensorflow]`
- **[SIN]** Structure Inference Net: Object Detection Using Scene-Level Context and Instance-Level Relationships | Yong Liu, et al. | **[CVPR' 18]** | `[pdf]` `[official code - tensorflow]`
- **[STDN]** Scale-Transferrable Object Detection | Peng Zhou, et al. | **[CVPR' 18]** | `[pdf]`
- **[RefineDet]** Single-Shot Refinement Neural Network for Object Detection | Shifeng Zhang, et al. | **[CVPR' 18]** | `[pdf]` `[official code - caffe]` `[unofficial code - chainer]`
- **[RFBNet]** Receptive Field Block Net for Accurate and Fast Object Detection | Songtao Liu, et al. | **[ECCV' 18]** | `[pdf]` `[official code - pytorch]`

Improvement: Context Modeling

- Context can broadly be grouped into one of three categories:
- **Semantic context:** The likelihood of an object to be found in some scenes but not in others.
- **Spatial context:** The likelihood of finding an object in some position and not others with respect to other objects in the scene.
- **Scale context:** Objects have a limited set of sizes relative to other objects in the scene.



Benchmarking Datasets



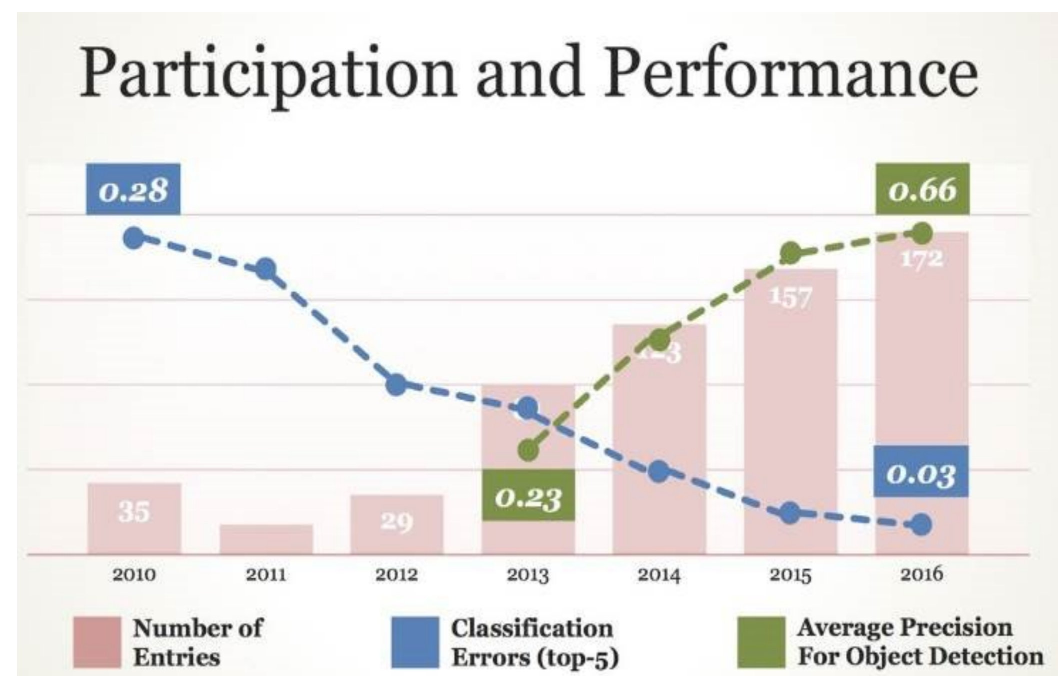
Benchmarking Datasets

Tab. Object Recognition Databases List

Dataset Name	Total Images	Categories	Images/ Category	Objects/ Image	Image Size	Started Year
MNIST	60,000	10	6,000	1	28x28	1998
Caltech101	9,145	101	40~800	1	300x200	2004
Caltech256	30,607	256	80+	1	300x200	2007
Scenes15	4,485	15	200~400	-	256x256	2006
PASCAL VOC	11,540	20	303~4087	2.4	470x380	2005
SUN	131,072	908	-	16.8	500x300	2010
ImageNet	14 M+	21,841	-	1.5	500x400	2009
MS COCO	328,000+	91	-	7.3	640x480	2014
Place	10 M+	434	-	-	256x256	2014
Open Images	9 M+	6000+	-	-	Varied	2017

IMAGENET Large Scale Visual Recognition Challenge (ILSVRC)

- ILSVRC evaluates algorithms for object detection and image classification at large scale.
- One high level motivation is to allow researchers to compare progress in detection across a wider variety of objects -- taking advantage of the quite expensive labeling effort.
- Another motivation is to measure the progress of computer vision for large scale image indexing for retrieval and annotation.



MS COCO

COCO
Common Objects in Context

info@cocodataset.org

Home People Dataset Tasks Evaluate

What is COCO?

COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

- Object segmentation
- Recognition in context
- Superpixel stuff segmentation
- 330K images (>200K labeled)
- 1.5 million object instances
- 80 object categories
- 91 stuff categories
- 5 captions per image
- 250,000 people with keypoints

Dataset examples

Tasks: Detection | Keypoints | Stuff | Panoptic | Captions

COCO 2018 Object Detection Task

1. Overview

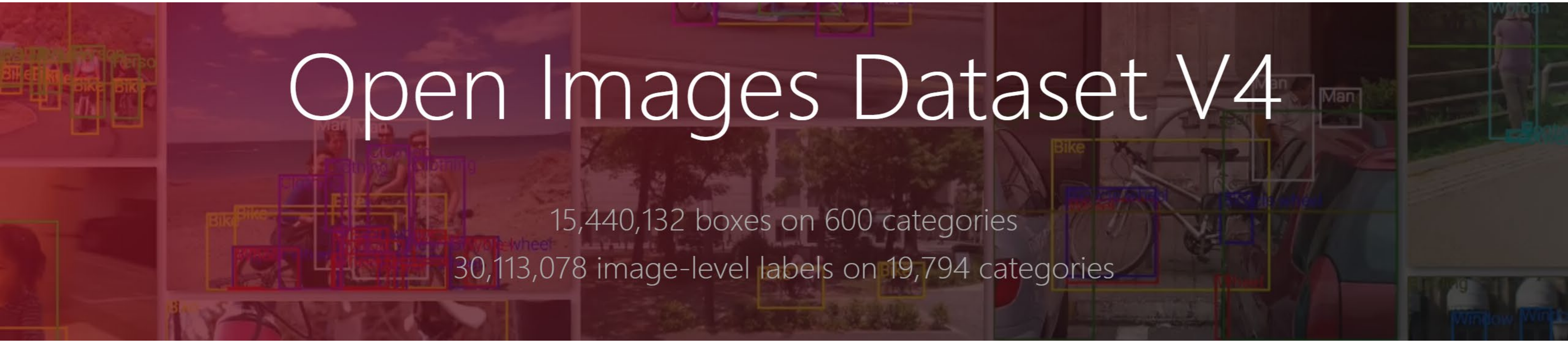
The COCO Object Detection Task is designed to push the state of the art in object detection forward. COCO features two object detection tasks: using either bounding box output or object segmentation output (the latter is also known as instance segmentation). For full details of this task please see the detection evaluation page. Note: **only the detection task with object segmentation output will be featured at the COCO 2018 challenge** (more details follow below).

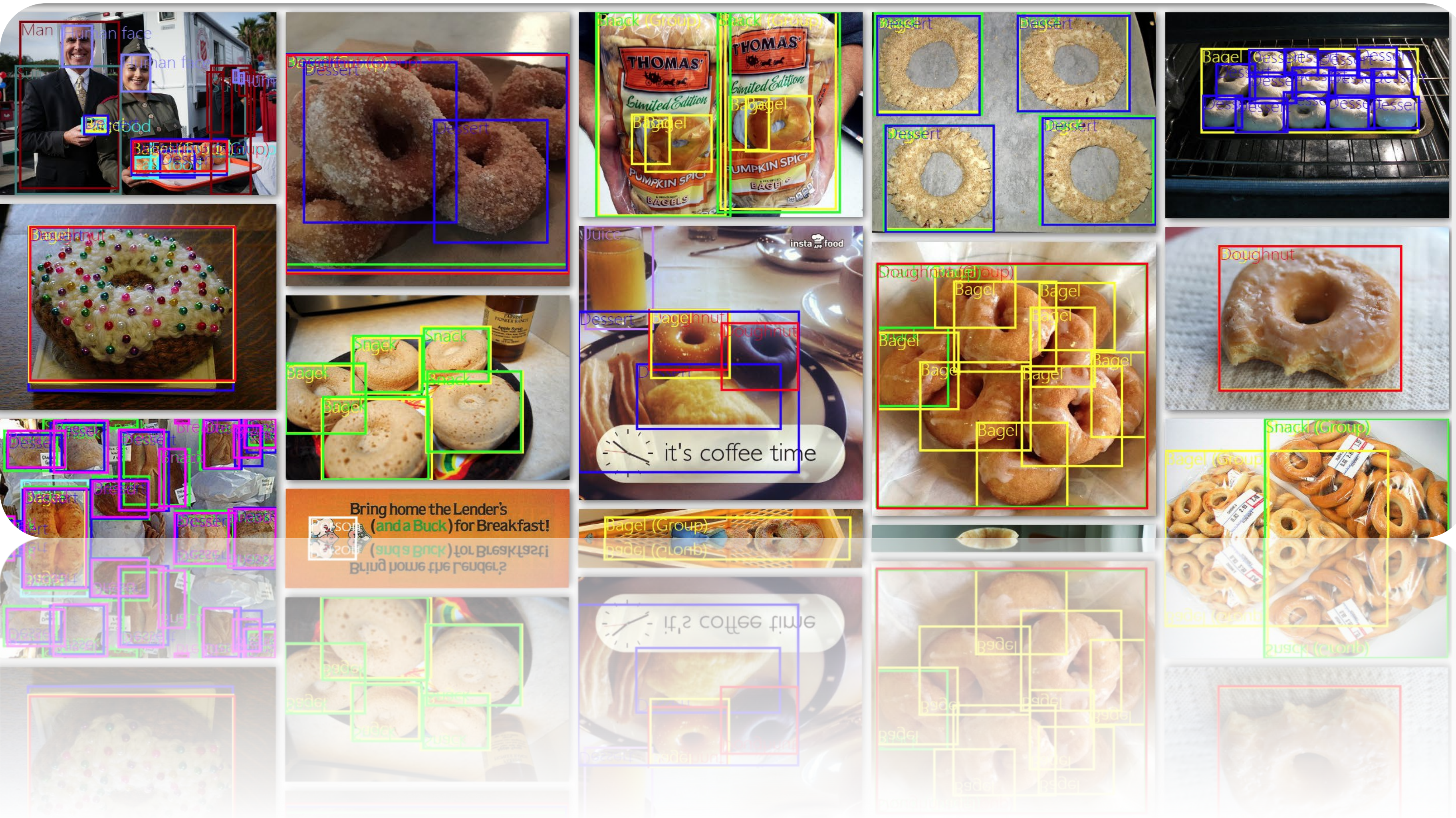
COCO 2018 Keypoint Detection Task

1. Overview

The COCO Keypoint Detection Task requires localization of person keypoints in challenging, uncontrolled conditions. The keypoint task involves simultaneously detecting people *and* localizing their keypoints (person locations are *not* given at test time). For full details of this task please see the keypoint evaluation page.

Open Images





Overview of Open Images

Open Images is a dataset of ~9 million images that have been annotated with image-level labels and object bounding boxes.

Evaluation (1/3)

Metric	Meaning	Definition and Description	
TP	True Positive	A true positive detection	
FP	False Positive	A false positive detection	
β	Confidence Threshold	A confidence threshold for computing $P(\beta)$ and $R(\beta)$.	
ε	IOU Threshold	VOC	Typically around 0.5
		ILSVRC	$\min(0.5, \frac{wh}{(w+10)(h+10)})$; $w \times h$ is the size of a GT box.
		MS COCO	Ten IOU thresholds $\varepsilon \in \{0.5 : 0.05 : 0.95\}$
$P(\beta)$	Precision	The fraction of correct detections out of the total detections returned by the detector with confidence of at least β.	
$R(\beta)$	Recall	The fraction of all N_c objects detected by the detector having a confidence of at least β.	
AP	Average Precision	Computed over the different levels of recall achieved by varying the confidence β.	

Evaluation (2/3)

mAP	mean Average Precision	VOC	AP at a single IOU and averaged over all classes.
		ILSVRC	AP at a modified IOU and averaged over all classes.
		MS COCO	• AP_{coco}: mAP averaged over ten IOUs: $\{0.5 : 0.05 : 0.95\}$; • $AP_{coco}^{\text{IOU}=0.5}$: mAP at IOU=0.50 (PASCAL VOC metric); • $AP_{coco}^{\text{IOU}=0.75}$: mAP at IOU=0.75 (strict metric); • AP_{coco}^{small}: mAP for small objects of area smaller than 32^2; • $AP_{coco}^{\text{medium}}$: mAP for objects of area between 32^2 and 96^2; • AP_{coco}^{large}: mAP for large objects of area bigger than 96^2;
AR	Average Recall	The maximum recall given a fixed number of detections per image, averaged over all categories and IOU thresholds.	
AR	Average Recall	MS COCO	• $AR_{coco}^{\text{max}=1}$: AR given 1 detection per image; • $AR_{coco}^{\text{max}=10}$: AR given 10 detection per image; • $AR_{coco}^{\text{max}=100}$: AR given 100 detection per image; • AR_{coco}^{small}: AR for small objects of area smaller than 32^2; • $AR_{coco}^{\text{medium}}$: AR for objects of area between 32^2 and 96^2; • AR_{coco}^{large}: AR for large objects of area bigger than 96^2;

Evaluation (3/3)

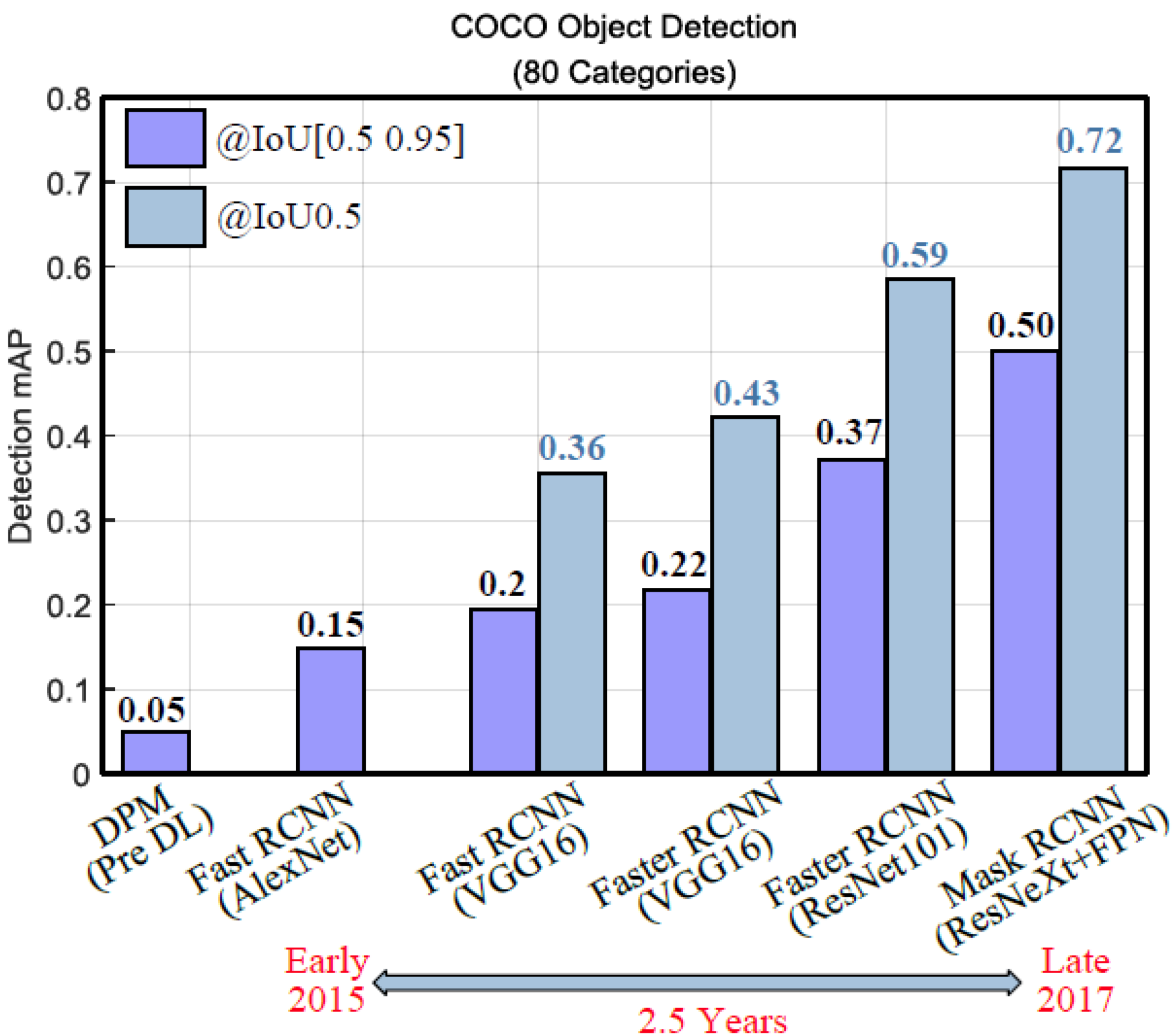


Fig. Evolution of object detection performance on COCO.
The backbone network, the design of detection framework and the availability of good and large scale datasets are the three most important factors in detection.

Remark

As a longstanding, fundamental and challenging problem in computer vision, object detection has been an active area of research for several decades.

L. Liu *et al.*, "Deep Learning for Generic Object Detection: A Survey," 2018.