ECG782: MULTIDIMENSIONAL DIGITAL SIGNAL PROCESSING

DEEP RECOGNITION

http://www.ee.unlv.edu/~b1morris/ecg782

OUTLINE

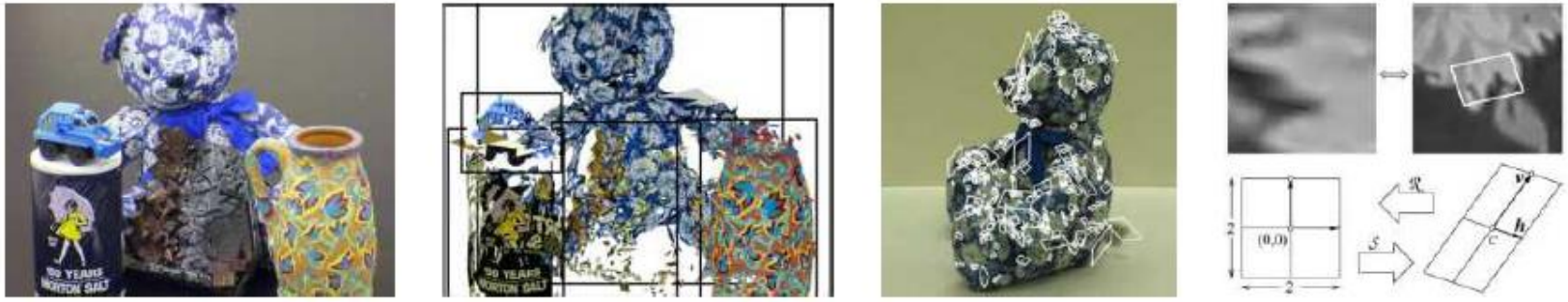
- Recognition Overview
- Instance Recognition
- Image Classification
- Object Detection
- Semantic Segmentation

RECOGNITION OVERVIEW

- Undergone largest changes and fastest developments in the last decade
  - Availability of larger labeled datasets
  - Breakthroughs in deep learning
- Historically , recognition was a “high-level task” built on top of lower-level components (e.g. feature detection and matching)
- With deep learning, there is little distinction between high- and low-level tasks → end-to-end learning

INSTANCE RECOGNITION I

- Re-recognize a known 2D/3D rigid object (exemplar)
  - Potentially with novel viewpoint, cluttered background, and partial occlusion



(a) (b) (c) (d)

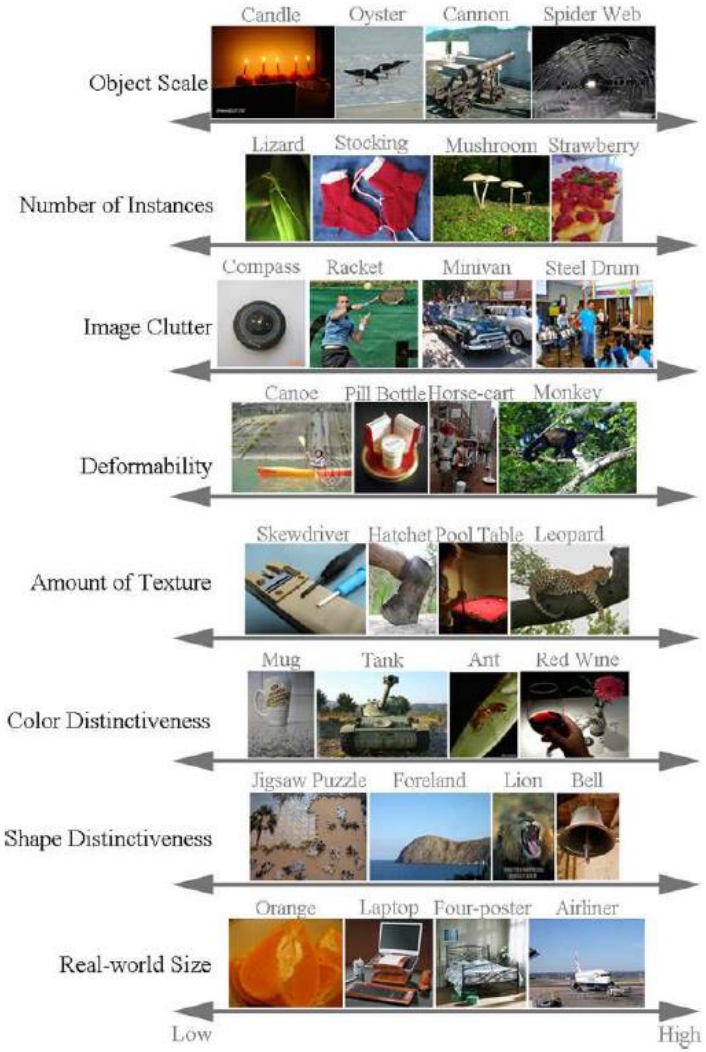
Figure 6.3 *3D object recognition with affine regions (Rothganger, Lazebnik* et al. *2006)* © *2006 Springer: (a) sample input image; (b) five of the recognized (reprojected) objects along with their bounding boxes; (c) a few of the local affine regions; (d) local affine region (patch) reprojected into a canonical (square) frame, along with its geometric affine transformations.*

INSTANCE RECOGNITION II

- General approach:
  - Find distinctive features while dealing with local appearance variation
  - Check for co-occurrence and relative positions (e.g. affine transformation)
- More challenging version: instance retrieval (content-based image retrieval) where the number of images to search is very large

IMAGE CLASSIFICATION

- Also known as category/class recognition
  - Must recognize members of highly variable categories
- Much more challenging than instance recognition
  - Same challenges but without known object
- Extensively studied area of CV
  - Where CNNs have dominated
- Note this is whole image classification



CLASSICAL APPROACHES: BOW

- Bag-of-words (features) – simple approach based co-occurrence of collected features
  - Detect features/keypoints
  - Describe keypoints = words
  - Compute histogram (distribution) of words
  - Compare histogram to database for matching
- Note: no geometric verification since not applicable to general objects

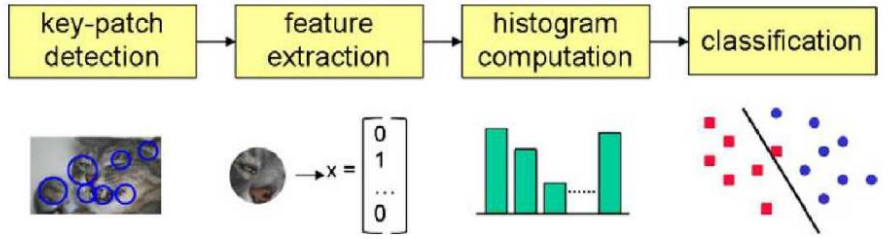


Figure 6.6 *A typical processing pipeline for a bag-of-words category recognition system (Csurka, Dance* et al. *2006) © 2007 Springer. Features are first extracted at keypoints and then quantized to get a distribution (histogram) over the learned* visual words *(feature cluster centers). The feature distribution histogram is used to learn a decision surface using a classification algorithm, such as a support vector machine.*

CLASSICAL APPROACHES: PARTS

- Approach to find constituent parts and measuring geometric relationships
  - Spring-like connections between subparts that have structure but allow variation
- Basic idea is to have an energy minimization function for subpart arrangements
- Common (graph) structures/topologies include threes and stars for efficiency
- Popular model: Deformable Part Model (DPM) of Felzenszwalb
  - Star model on HOG parts

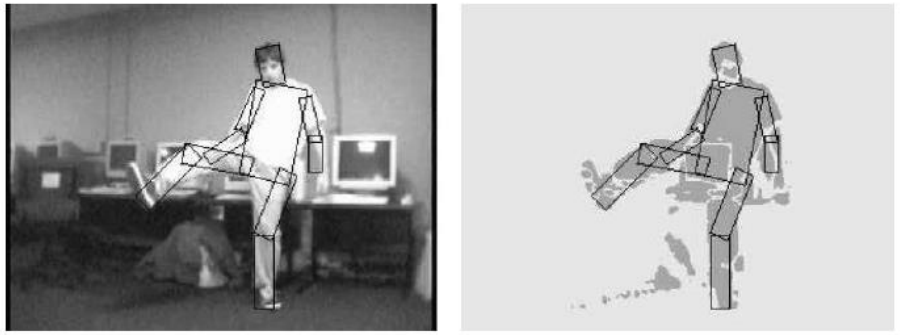


Figure 6.7 *Using pictorial structures to locate and track a person (Felzenszwalb and Huttenlocher 2005) © 2005 Springer. The structure consists of articulated rectangular body parts (torso, head, and limbs) connected in a tree topology that encodes relative part positions and orientations. To fit a pictorial structure model, a binary silhouette image is first computed using background subtraction.*

CLASSICAL APPROACHES: CONTEXT/SCENE

- Previous approaches were object-centric which limits recognition
  - Scene context is very important for disambiguation (e.g. lemon vs. tennis ball)
- Context models combine objects into scenes
  - Number of constituent objects is not known a priori
- The idea of context has been important for deep techniques

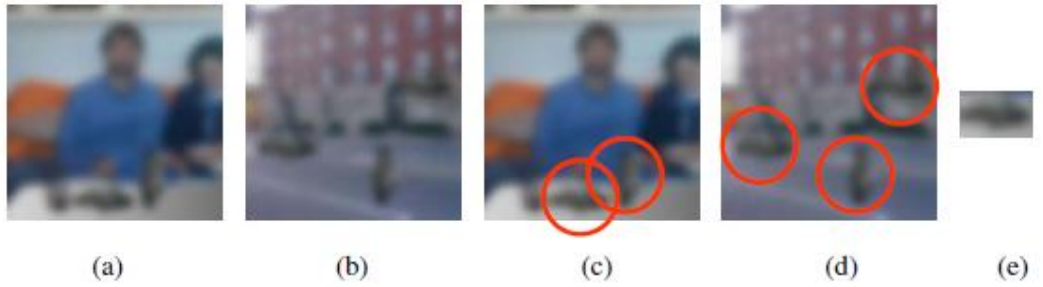


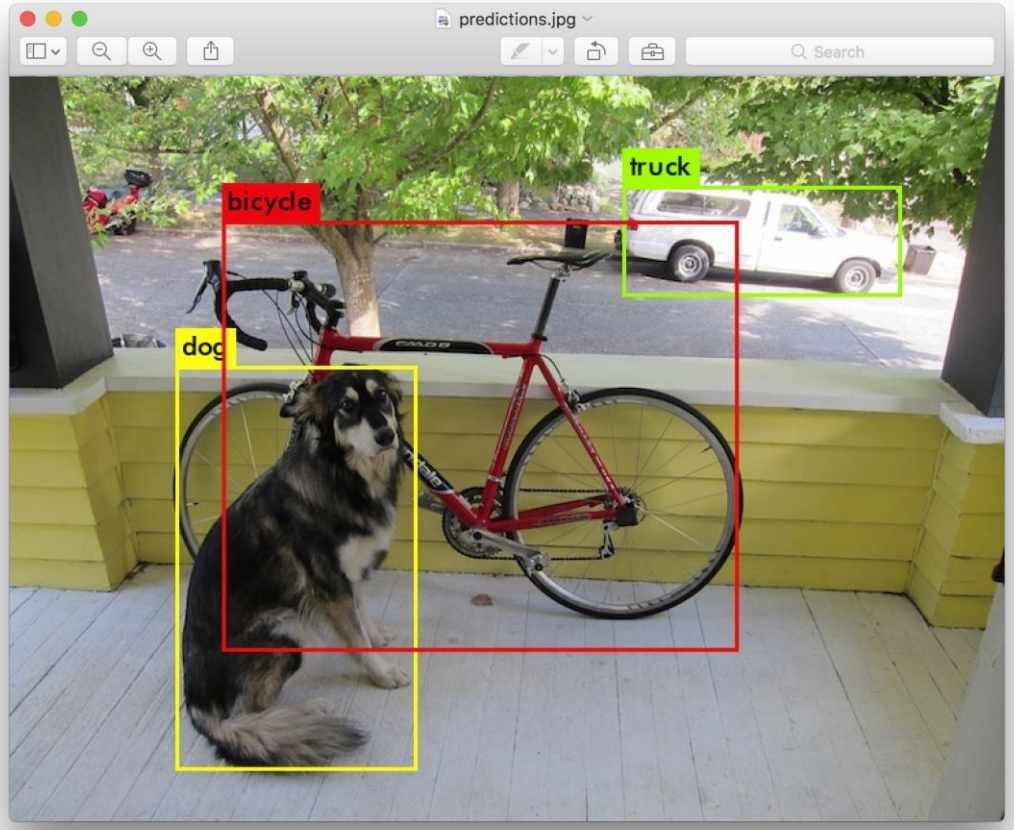
Figure 6.8 *The importance of context (images courtesy of Antonio Torralba). Can you name all of the objects in images (a–b), especially those that are circled in (c–d). Look carefully at the circled objects. Did you notice that they all have the same shape (after being rotated), as shown in column (e)?*

OBJECT DETECTION

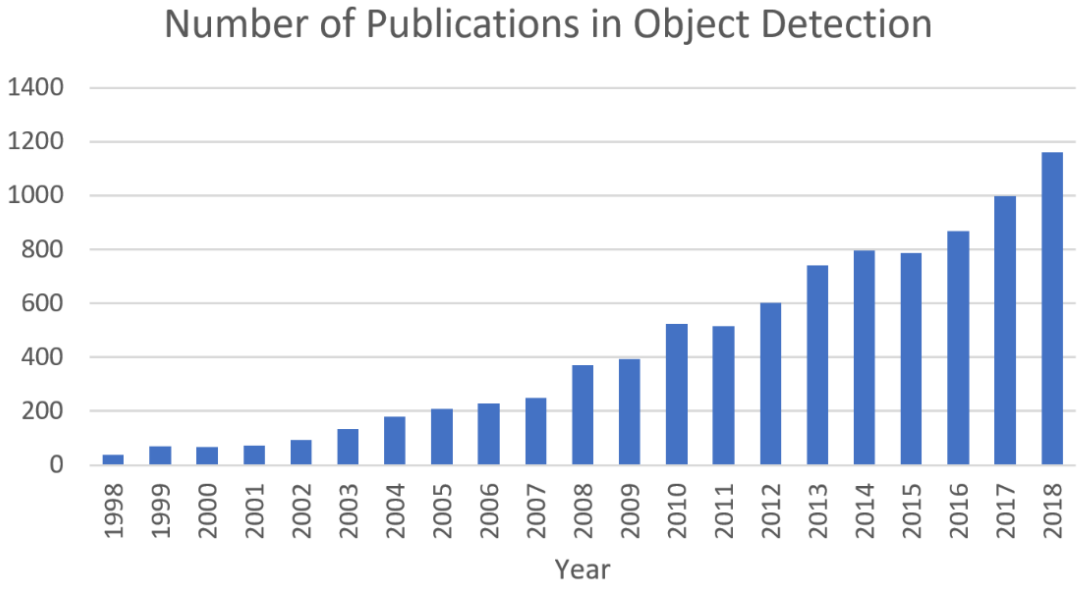
OBJECT DETECTION WITH DEEP LEARNING: A REVIEW
ZHAO, ZHENG, XU, AND WU, T-NNLS 2019

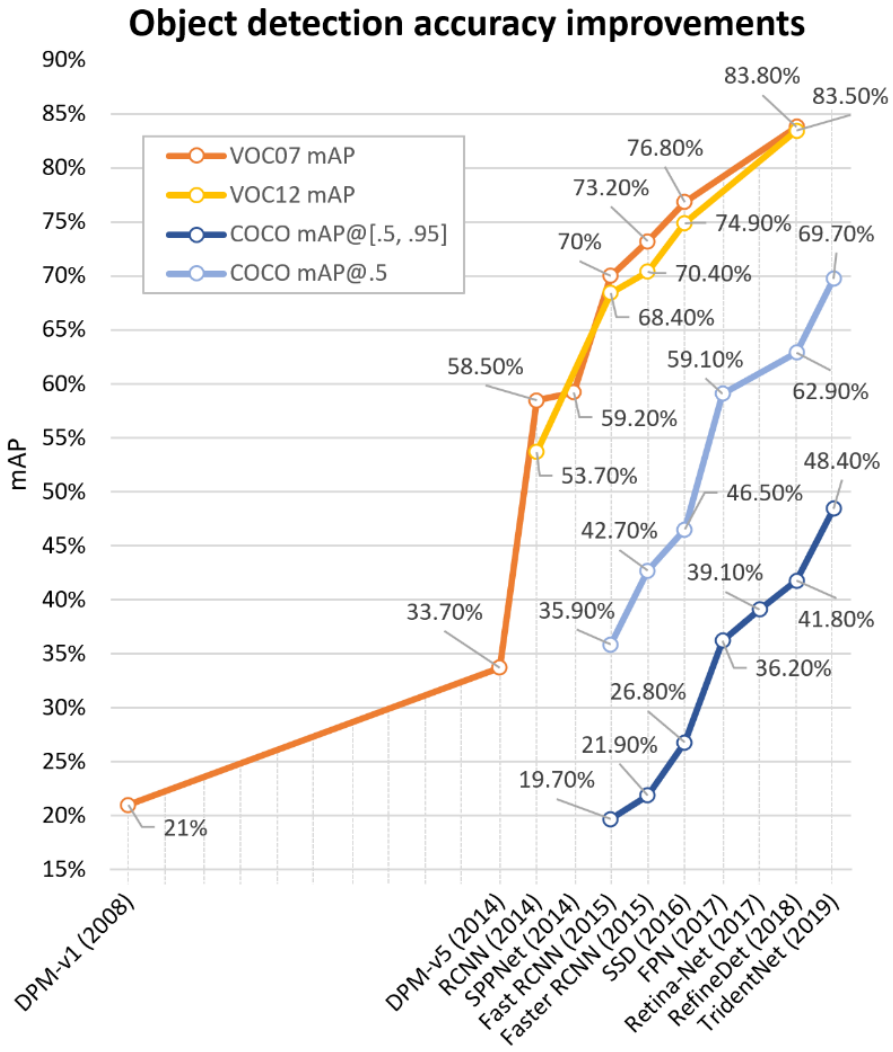
OBJECT DETECTION OVERVIEW

- Fundamental computer vision problem
- Categorize not just the whole image but delineate (with bounding boxes) where various objects are located (object localization)
  - Localization is viewed as a bounding box regression task
- Provides a semantic understanding of images (video)
- Related tasks: image classification, human behavior analysis, face recognition, autonomous driving



DEEP CNN DOMINANCE IN DETECTION





Zou et al., "Object Detection in 20 Years: A survey, 2019

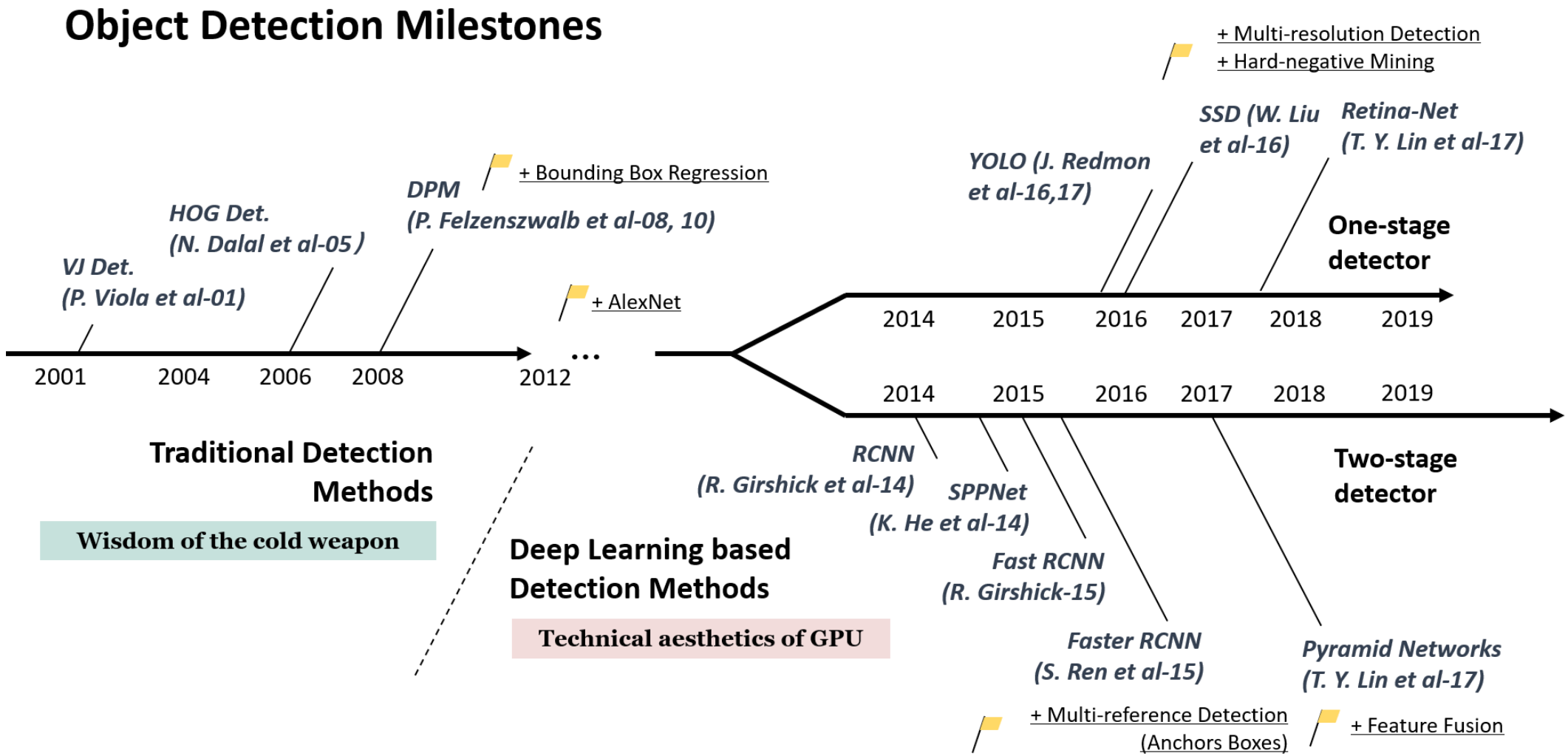
DEEP LEARNING AND CNNS

- Deep learning dominance:
  - Large scale annotated training datasets
  - Fast development of high performance parallel computing (GPUs)
  - Advances in network structures
    - Initialization: pre-training
    - Overfitting: Dropout and data augmentation
    - Efficiency: batch normalization
    - Architectures: AlexNet, Inception, ResNet
- CNN advantages:
  - Hierarchical feature representation
  - Deeper architecture for increased expressive capability
  - Can jointly optimize several related tasks (multi-task learning)
  - Classical CV can be recast as high-D data transform problems

GENERIC OBJECT DETECTION

- Locate and classify all objects (of interest) in an image
  - Label each object with a rectangular bounding box
  - Have a measure of confidence in detection
- Two major approaches:
  - Two-stage: i) generate region proposals and ii) classify each proposal into different object categories
  - One-stage: detection as a regression or classification to get both categories and locations directly at once

OBJECT DETECTION MILESTONES



Zou et al., "Object Detection in 20 Years: A survey, 2019

TRADITIONAL DETECTOR REVIEW

- Viola Jones cascade detector
  - Viola and Jones, 1999
- Histogram of Oriented Gradients (HOG) detector
  - Dalal and Triggs, 2005
- Deformable Part-based Model (DPM)
  - Felzenszwalb, 2008

VIOLA JONES

- Real-time face detection with sliding window for position and scale
- Integral image: speeded up Haar-like feature computation (speeded up filtering)
- Feature selection: Adaboost to automatically select a small but useful set of features (application driven filters)
- Detection cascades: multi-stage detector to avoid heavy computation on background windows but on faces

HOG

- Designed for pedestrian detection
- Improvement over SIFT and shape contexts
  - Balances feature invariance (translation, scale, illumination) and nonlinearity (different object categories)
- Descriptor computed on dense grid of uniformly spaced cells
- Used overlapping local contrast normalization over blocks
- Resizes input image while keeping detection window fixed for scale

DPM

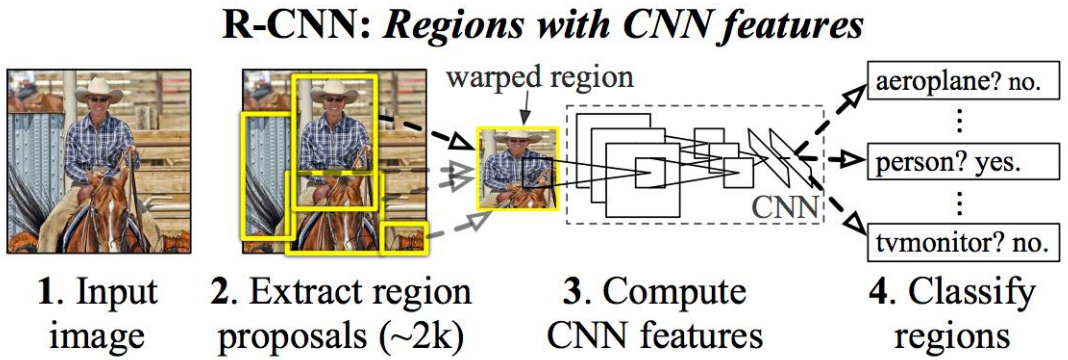
- Extension of HOG and was winner of VOC 07-09
- Divide and conquer detection – object built from smaller parts to detect (bike has wheels, body, etc.)
  - Use of a star-model for connections – a root filter and part-filters
- Important contributions:
  - Extended with mixture models for more real-world variation (e.g. bike from front or side)
  - Hard negative mining – create negative examples on the margin
  - Bounding box regression

TWO-STAGE DETECTOR MILESTONES

- Region proposal based frameworks
  - "Coarse-to-fine" process somehow similar to human brain – scan full scene and then focus on region of interest
- Approaches
  - Overfeat – sliding window
  - Region CNN (R-CNN)
  - Spatial Pyramid Pooling Networks (SPPNet)
  - Fast R-CNN
  - Faster R-CNN
  - Feature pyramid network (FPN)

R-CNN (GIRSHICK 2013)

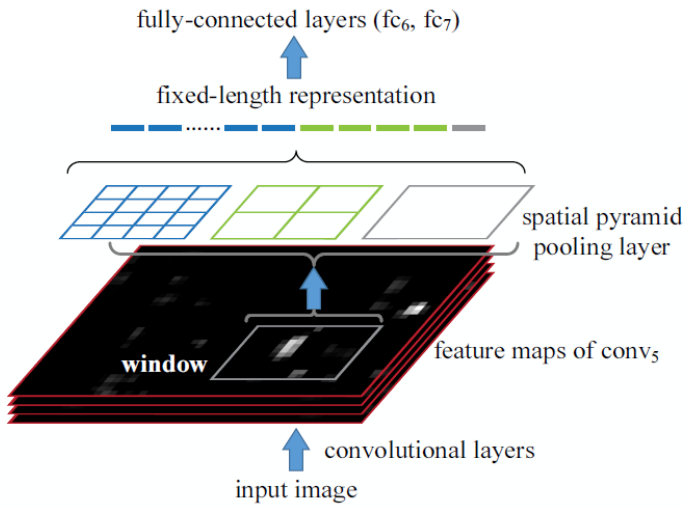
- Use selective search (Uijlings 2011) to generate a small set of potential object regions
  - Bottom-up grouping and saliency for proposals of various size
- Rescale proposals to fixed size and evaluate ImageNet pretrained CNN for feature extraction
- Multi-class linear SVM for classification



- Advantages: significant performance boost on VOC07
- Shortcomings: Redundant feature computations on overlapping regions make this slow

SPPNET (HE 2014)

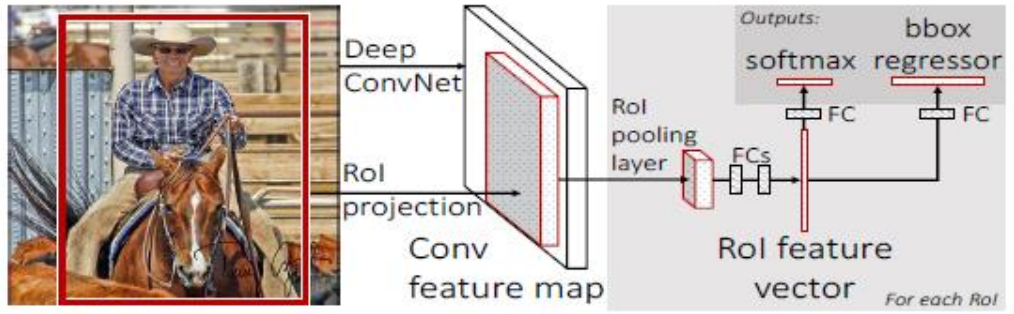
- Spatial pyramid pooling (SPP) layer enables a CNN to generate a fixed-length representation regardless of image size/ROI without rescaling
- Feature maps computed once for entire image and fixed-length representation can be made of arbitrary region
  - Use conv5 layer for SPP layer



- Advantage: 20x faster than R-CNN without accuracy loss
- Shortcomings: Training is still multi-stage and only FC layers are trained

FAST R-CNN (GIRSHICK 2015)

- Simultaneously train detector and bounding box regressor
  - No need for linear SVM layers
- Like SPPNet, image is only processed with convolutions once
  - RoI pooling layer to generate fixed-length feature vector
- FC layers branch to outputs:
  - Softmax class probabilities
  - Refined bounding box positions
- Optimized jointly with multitask loss (classification + localization)



- Advantages: Increased VOC mAP from by 11.5% from R-CNN
- Shortcomings: speed still limited by region proposals

FASTER R-CNN (REN 2015)

- Generate object proposals with a CNN model
  - First end-to-end and near real-time deep learning detector
- Introduced region proposal network (RPN)
  - Nearly cost-free region proposals as opposed to selective search
  - Produces object boundaries and scores for all positions simultaneously
  - Sliding window across conv layer
- Use of reference boxes (anchors) that match popular object dimensions
  - Later regressed for final bbox

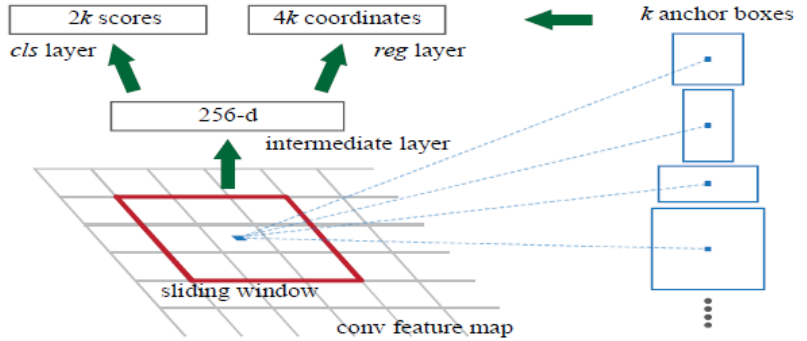
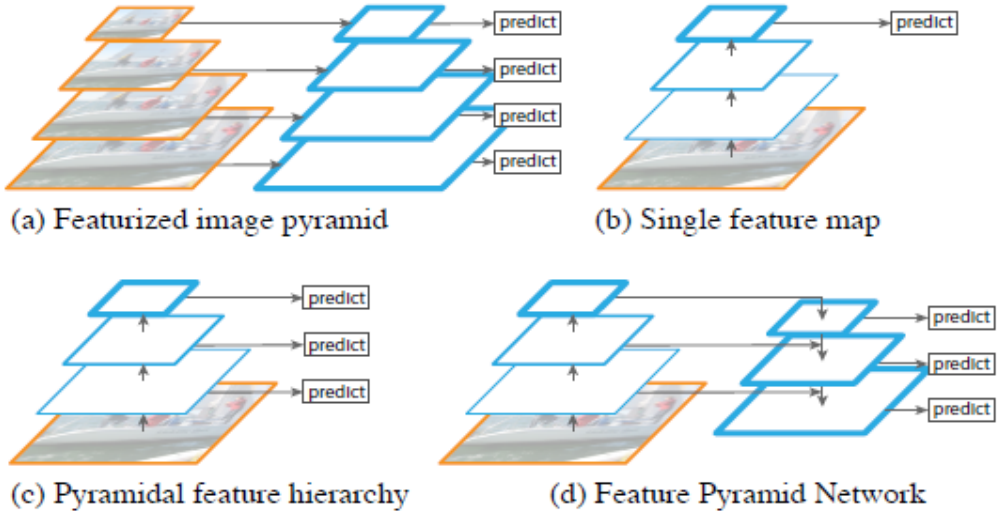


Fig. 6. The RPN in Faster R-CNN [18]. K predefined anchor boxes are convoluted with each sliding window to produce fixed-length vectors which are taken by cls and reg layer to obtain corresponding outputs.

- Advantages: trained end-to-end (all layers) and high 5 fps on GPU with SOTA VOC results
- Shortcomings: long training time, poor performance on extreme scales/shapes, object regions rather than instances

FPN (LIN 2017)

- Handle wide scale variation through use of image pyramid
  - Deeper CNN layers useful for category recognition but poor for localization
- Top-down architecture with lateral connections to share high level features with higher resolution of lower layers
  - Avoid expensive explicit image pyramid computation
- General approach for efficient multi-scale representation
  - Extensively used in semantic segmentation



(a) Featurized image pyramid (b) Single feature map

(c) Pyramidal feature hierarchy (d) Feature Pyramid Network

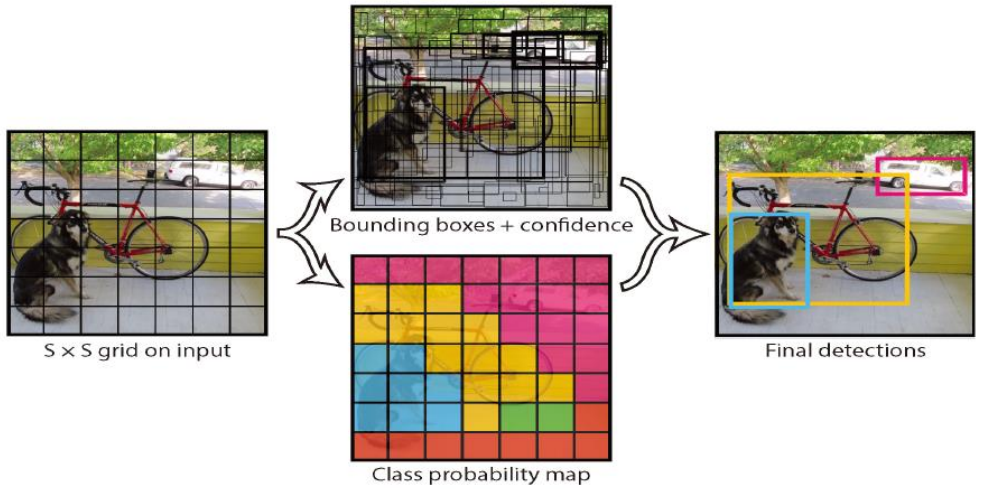
Fig. 7. The main concern of FPN [66]. (a) It is slow to use an image pyramid to build a feature pyramid. (b) Only single scale features is adopted for faster detection. (c) An alternative to the featurized image pyramid is to reuse the pyramidal feature hierarchy computed by a ConvNet. (d) FPN integrates both (b) and (c). Blue outlines indicate feature maps and thicker outlines denote semantically stronger features.

ONE-STAGE DETECTOR MILESTONES

- End-to-end regression/classification methods
  - Single step to produce detections
- Approaches
  - MultiBox
  - AttentionNet
  - Grid-based object detector (G-CNN)
  - You Only Look Once (YOLO)
  - Single Shot Multi-box Detector (SSD)

YOLO (REDMOND 2015)

- First one-stage detector
  - Extremely fast by abandoning proposal detection + verification approach
- Divides an image into regions and predicts bounding boxes and probabilities for all regions simultaneously
  - Each grid region predicts objects centered within that grid cell
  - B bounding boxes are predicted with associated confidence score



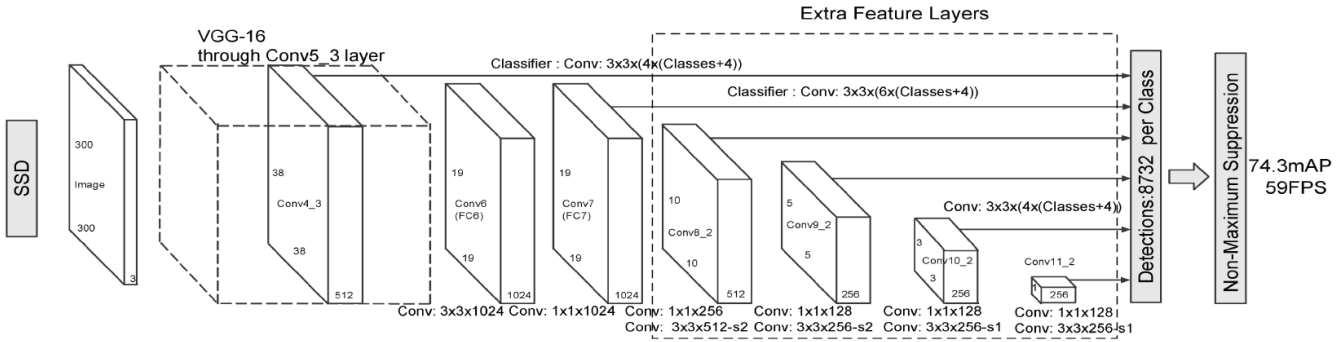
- Advantages:
  - Extremely fast (45-155 fps VOC)
- Shortcomings:
  - Poorer localization than two-stage detectors
  - Difficulty with small scale objects

YOLO II

- Customized CNN architecture from scratch
  - Inception-like modules
- Divide image into $S \times S$ grid
- Each grid cell predicts an object centered with the cell
  - Local search with relative coordinates (scale for image size)
  - B bounding boxes predicted for each cell with confidence
  - Conditional class probabilities predicted for each of the C
- Training loss
  - Bounding box localization
    - Box center relative to grid
    - Normalized height/width relative to image size
  - Confidence score
  - Classification error
    - Only when object is in cell
- Upgrades (v2, v3, etc.)
  - Batch normalization
  - Anchor boxes
  - Dimension cluster
  - Multi-scale training

SSD (LIU 2015)

- Multi-reference and multi-resolution detection technique
  - Detects at different scales at different layers of network
  - Better handles small objects
- Inspired by anchors of MultiBox RPN, and multi-scale representation
- Add feature layers at the end of standard backbone (VGG16)
  - Predict offsets to default bounding boxes of different scales and aspect ratios and confidences
  - Final detection after NMS on multi-scale refined boxes



- Advantages:
  - Fast (59 fps) while more accurate than YOLO
- Shortcomings:
  - Still issues with small objects (better backbone e.g. ResNet101)

SSD II

- MultiBox (Szegedy 2014)
  - Inception-like structure to reduce dimensionality but not spatial resolution (height x width)
  - Confidence loss to measure objectiveness of bounding box (categorical cross-entry)
  - Location loss to measure how far a predicted bounding box (L2 but SSD uses smooth L1)
- Used anchors to get good prediction starting point for regression
  - 11 priors/feature map = 1420 anchors/image for images at multiple scales and sizes
  - SSD extended idea to each cell in feature map to avoid explicit anchor pre-train (6/cell)
- Hard negative mining - 3:1 ratio of neg:pos train examples
  - Need to keep low IoU predictions
- Data augmentation – random flipping and patches of original image at different IoU ratios
- Non-maximum suppression (NMS) – discard low confidence and IoU
- 80% of time is spent on base VGG16
  - Can improve speed/performance with better backbone

TECHNIQUES FOR BASE IMPROVEMENT

- Multi-task learning – learn better representation from multiple correlated tasks
  - Train conv layers for e.g. region proposal, classification, and segmentation
- Multi-scale representation – combine activations from multiple layers with skip-layer connections
  - Provide semantic information of different spatial resolutions
- Contextual modeling – exploit features from surround
  - Provide features from different support regions/resolutions which help with occlusion and local similarities (e.g. tennis ball versus lemon when a racket is nearby)

REFERENCES

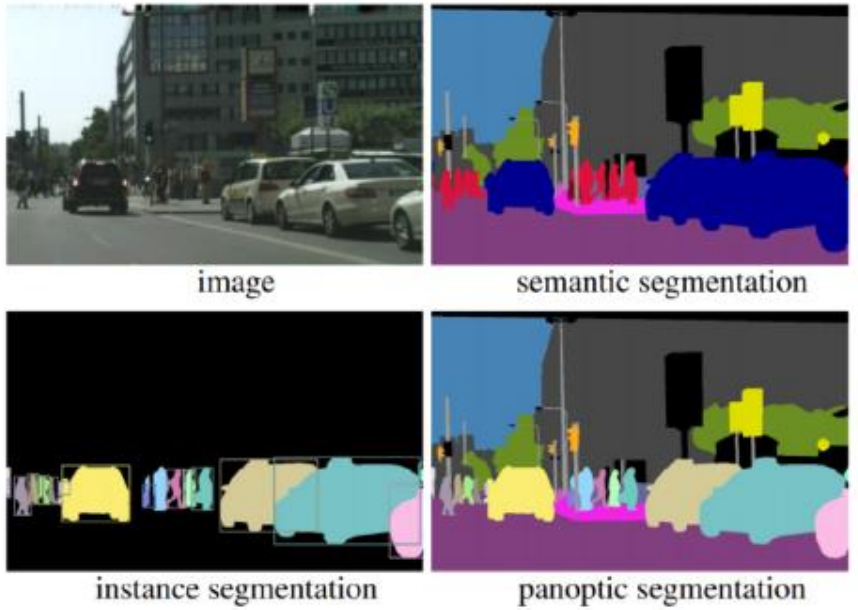
- For more complete overview, see recent surveys
- Object Detection with Deep Learning: A Review
- Object Detection in 20 Years: A Survey

IMAGE SEGMENTATION

EVOLUTION OF IMAGE SEGMENTATION USING DEEP CONVOLUTIONAL NEURAL NETWORKS: A SURVEY, SULTANA, SUFIAN, AND DUTTA, KBS 2020

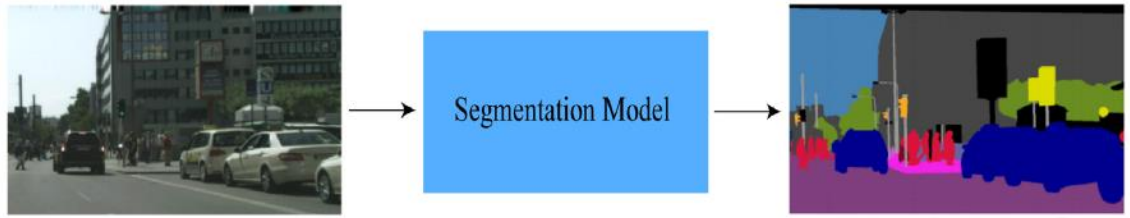
SEGMENTATION TASKS

- Segmentation – CV task of segregating an image into multiple regions according to different properties of pixels (e.g. color, intensity, texture)
  - Typically a low-level task that relies on spatial information (neighborhood)
- Semantic segmentation – associate a class label for every pixel in an image
- Instance segmentation – mask (segment) each instance of an object in an image independently
- Panoptic segmentation – combination of semantic segmentation and instance segmentation
  - Label both class and separate instances (detection)



SEMANTIC SEGMENTATION

- Pixel level class labels
- Have relied heavily on CNNs since 2012
- Popular approaches:
  - Fully convolutional network
  - Dilated/atrous convolution
  - Top-down/bottom-up approach
  - Global context
  - Receptive field enlargement and multi-scale context



FCN [LONG 2017]

- Fully convolutional network (FCN) was proposed for semantic segmentation
- Use standard CNN backbone but remove dense FC layers
  - Use of 1x1 convolution instead
  - Produces a class presence heatmap in low-resolution
- Bilinear interpolation used to upsample coarse output to pixel resolution
- Skip connections (deep jet) to combine final prediction layer with higher res/feature-rich lower layers

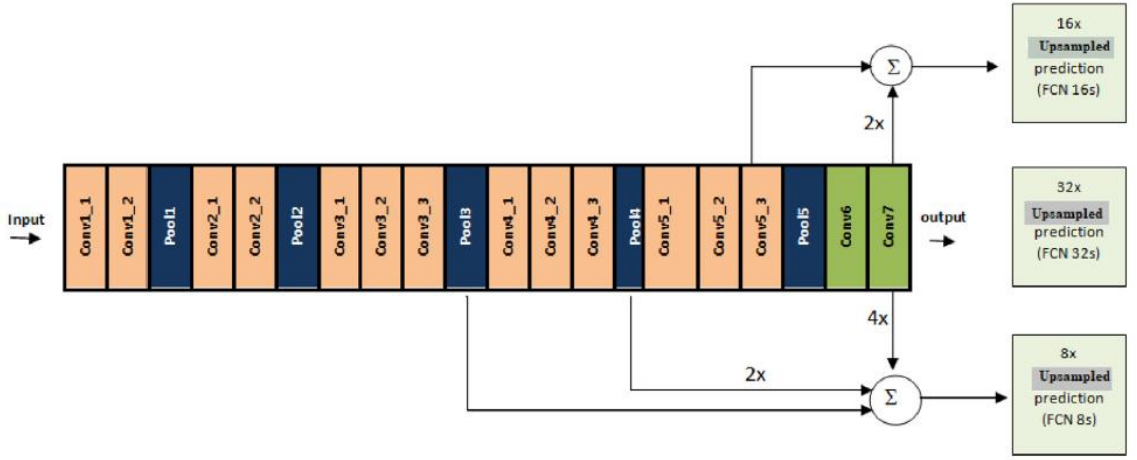
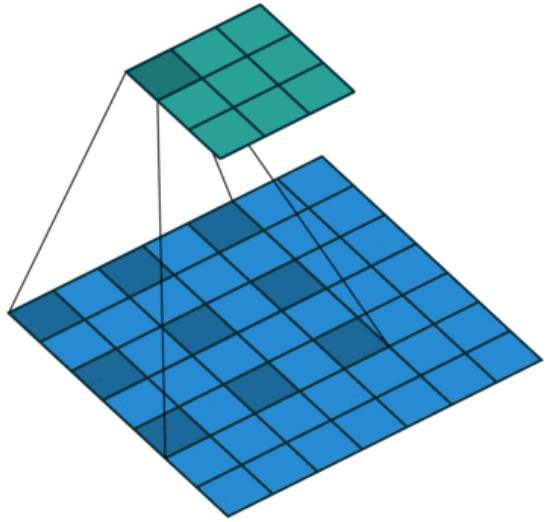


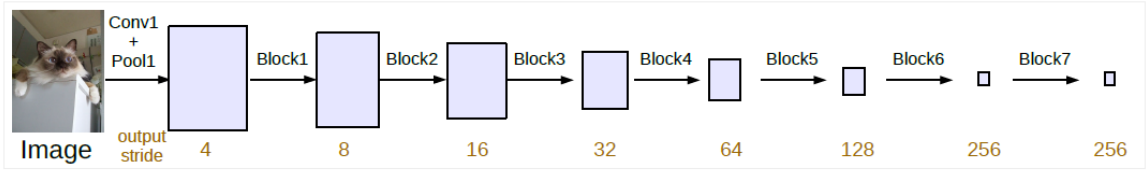
Fig. 4. Architecture of FCN32s, FCN16s, FCN8s.

DILATED/ATROUS CONVOLUTION

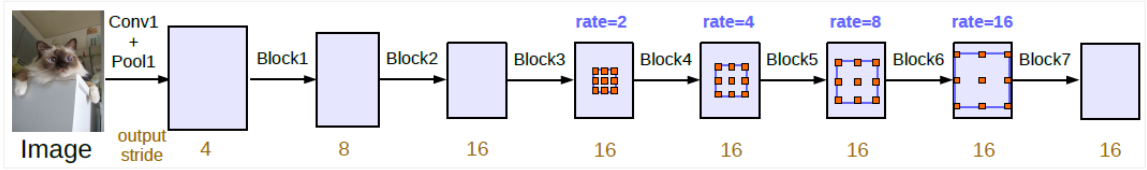
- Context is important for segmentation but Traditional convolution is expensive for larger field-of-view (kernel size)
- Atrous convolution introduces a dilation rate
  - Trade-off context vs localization
- Traditional CNN loses resolution while atrous can keep it
  - Larger feature map is better for segmentation (less interpolation)
  - However, isolates pixel from context
- Key architectures: DilatedNet and DeepLab (CRF for fine details)

source





(a) Going deeper without atrous convolution.

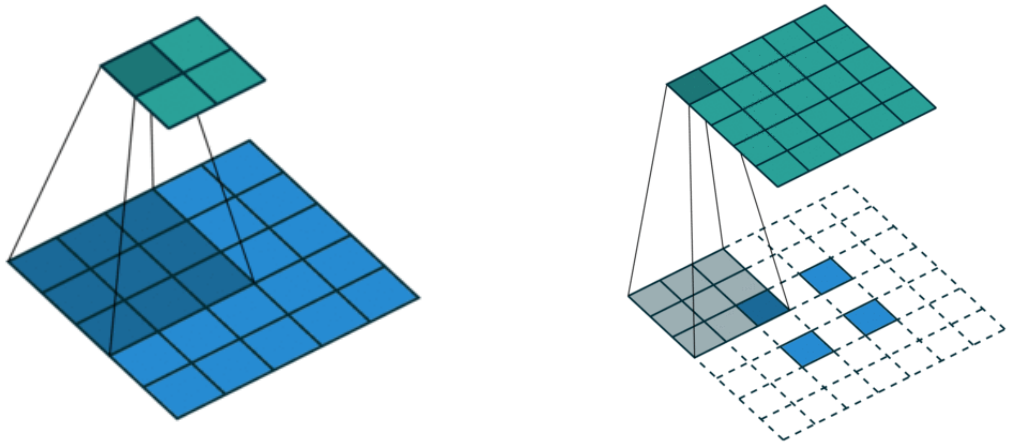


(b) Going deeper with atrous convolution. Atrous convolution with $rate > 1$ is applied after block3 when $output\_stride = 16$.

TOP-DOWN/BOTTOM-UP APPROACH

- Encoder-decoder architecture
  - Convolution encodes image features
  - Deconvolutional network to decode features into pixels/labels
- Deconvolution (transposed convolution) reconstructs spatial resolution
  - Upscaling convolution operation
- Both encoder and decoder extract features
- Generally lose fine-grained information in encoding process
  - Skip connections utilized to pass higher-resolution features
- Key architectures: Deconvnet, U-Net, SegNet, FC-DenseNet, HRNet

- conv de-conv [source]



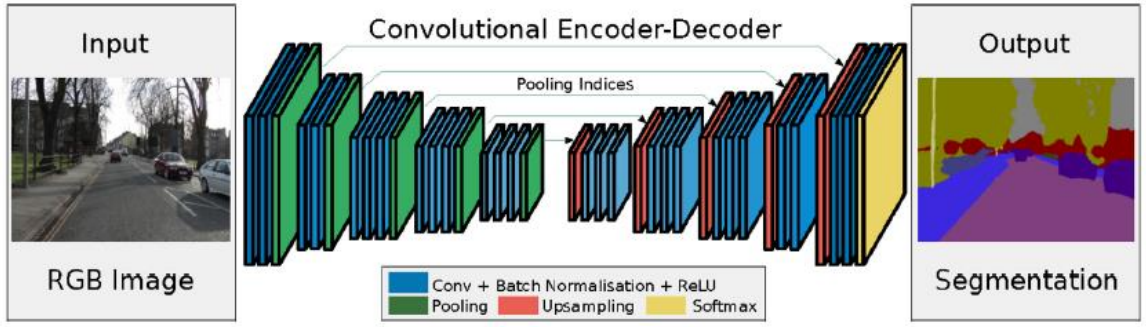
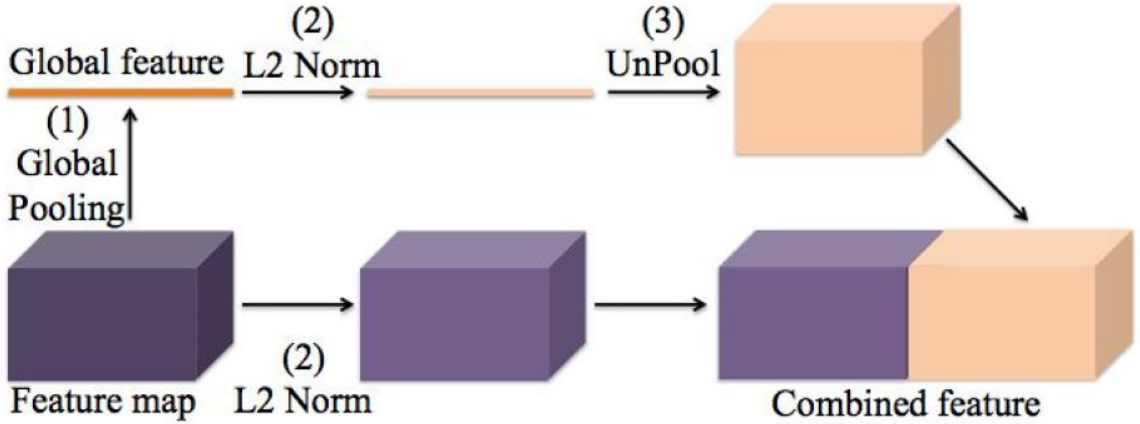


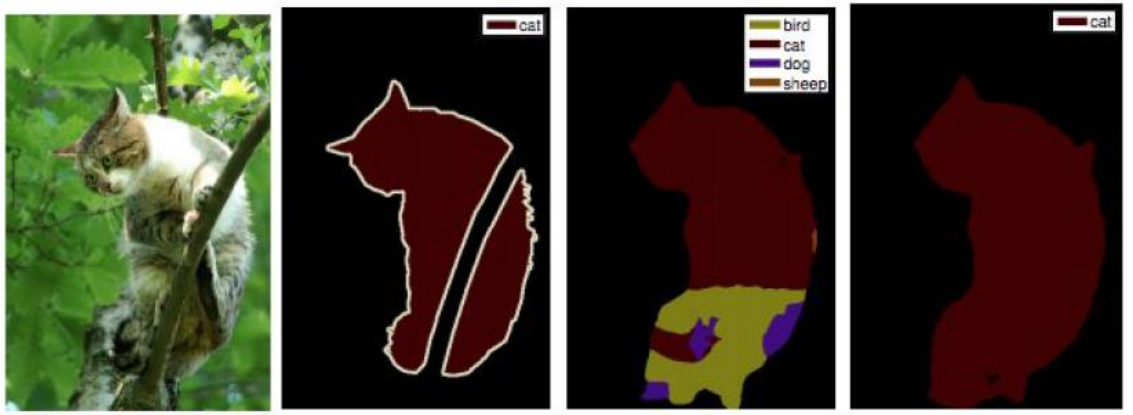
Fig. 9. Encoder–decoder architecture of SegNet.
Source: From [93].

GLOBAL CONTEXT

- Most segmentation relies on just local information but global context is important
  - Add global features or global context information
- Global features
  - Global average pool (final layers)
  - Large convolution kernels
- Context
  - Use of class mapping
- Helps resolve inaccuracies but lacks scaling information of multiscale objects
- Key architectures: ParseNet, GCN, EncNet



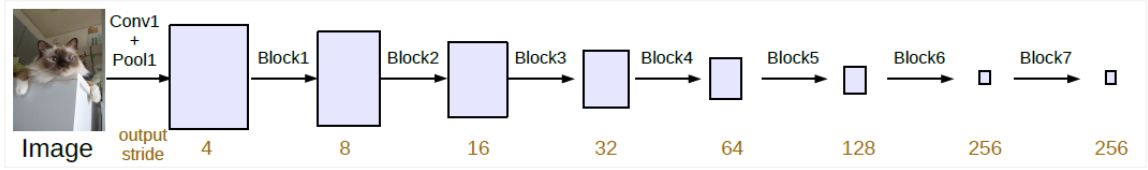
(e) ParseNet contexture module overview.



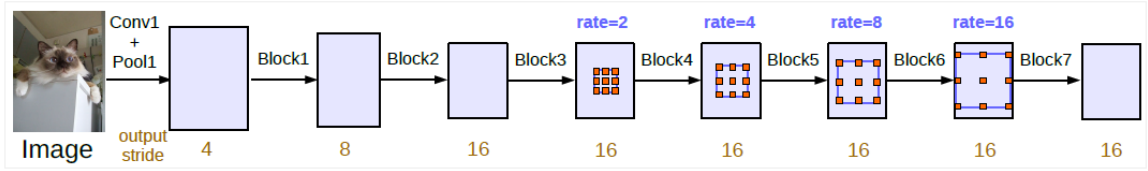
(a) Image (b) Truth (c) FCN (d) ParseNet

RECEPTIVE FIELD ENLARGEMENT AND MULTI-SCALE CONTEXT

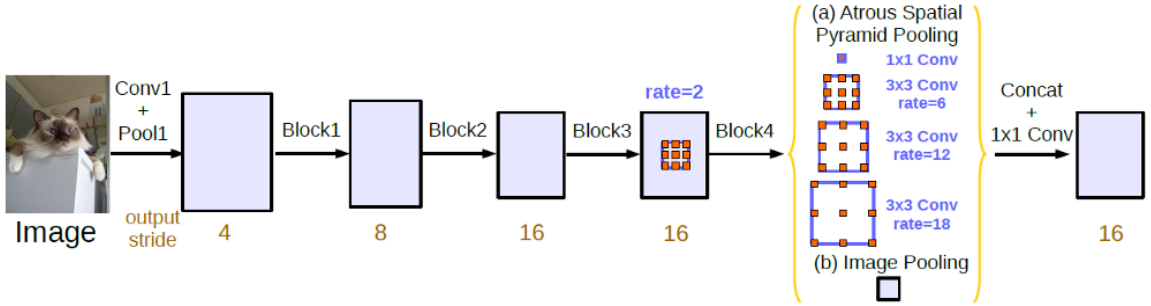
- Use of feature pyramid techniques for multi-resolution representation
  - Atrous Special Pooling Pyramid (ASPP)
  - Pyramid pooling module
- Provides better localization
- Helps incorporate scale information of objects for fine-grained segmentation
- Key architectures: DeepLabv2, DeepLabv3, PSPNet, Gated-SCNN



(a) Going deeper without atrous convolution.



(b) Going deeper with atrous convolution. Atrous convolution with $rate > 1$ is applied after block3 when $output\_stride = 16$.



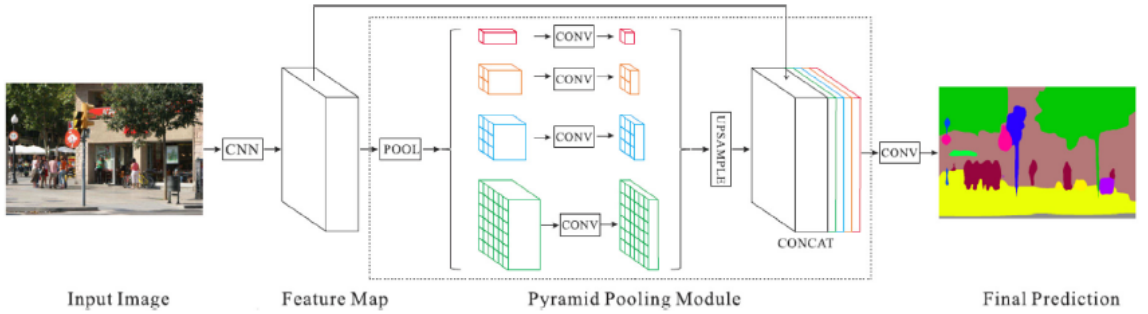


Fig. 15. PSPNet Model Design.

INSTANCE SEGMENTATION

- Each instance of a particular object is masked independently
- Task is intertwined with object detection
  - Detection gives bounding box while instance segmentation further refines with mask
- General approach is to give proposals of objects/masks and refine
- Mask R-CNN as example
  - Faster R-CNN extension
  - RPN for object proposals – classification and bounding box regression
  - Separate segmentation network for each ROI

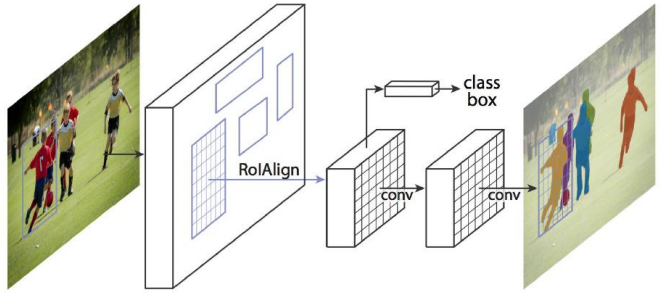


Fig. 17. Mask R-CNN architecture for instance segmentation. From [64].

Fig. 18. Mask R-CNN results on sample images from the COCO test set. From [64].

PANOPTIC SEGMENTATION

- Combination of instance segmentation and semantic segmentation
  - Newer segmentation task
- General approach:
  - Heads for semantic segmentation
  - Head for instance segmentation
  - Panoptic head to combine
- Key architectures: OANet, UPSNet, Multitask Network

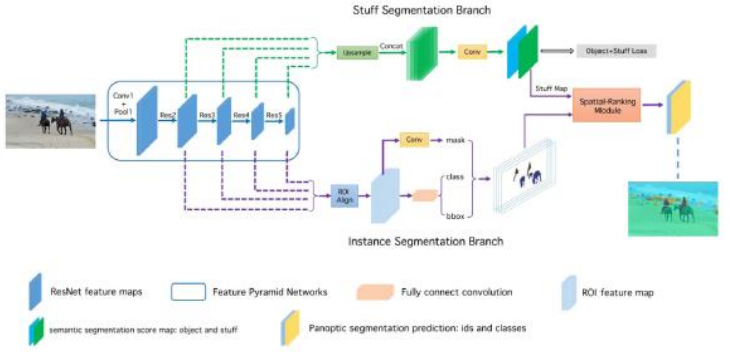


Fig. 27. Architecture of Occlusion Aware Network (OANet).
Source: From [183].

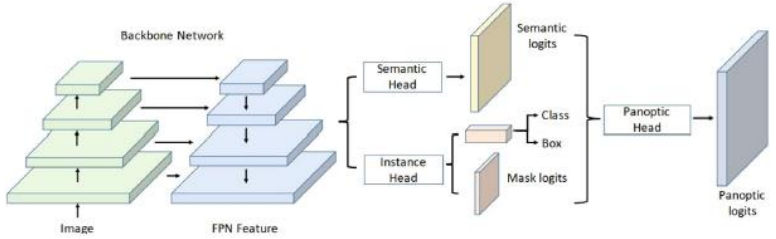


Fig. 28. Architecture of unified panoptic segmentation network (UPSNet).
Source: From [185].

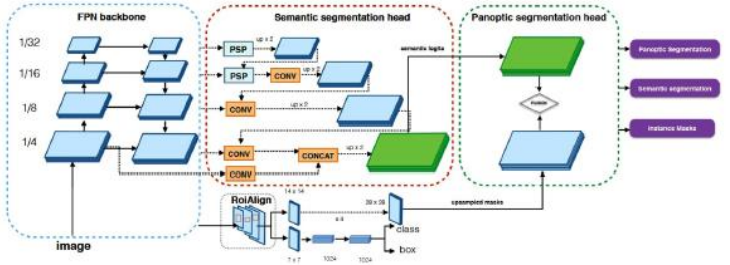


Fig. 29. Architecture of Multitask Network for Panoptic Segmentation.
Source: From [186].

REFERENCES

- For more complete overview, see recent surveys

- Evolution of Image Segmentation using Deep Convolutional Neural Network: A Survey

- Image Segmentation Using Deep Learning: A Survey