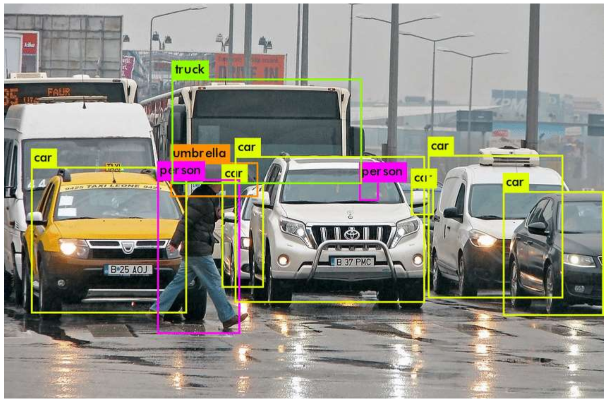
Chapter 4

Region based CNNs

1. CNNs for Object Detection

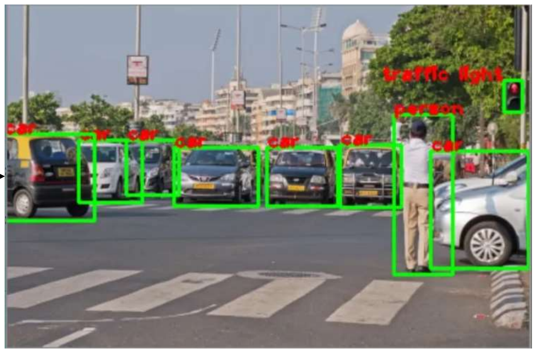


2. CNNs for Object Segmentation

1. CNNs for Object Detection

Convolutional Neural Networks which detect different objects, their sizes, and their locations in an image

Convolutional Neural Network

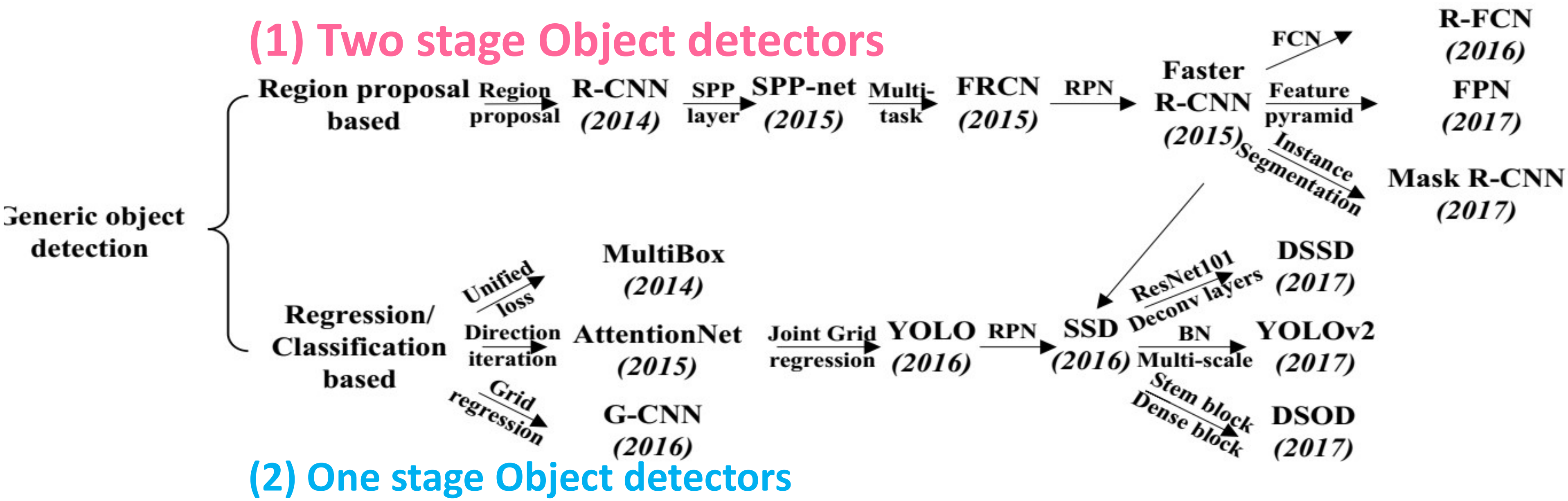


Detected Objects :
1- Labels
2- Bounding box (BB)
(center+ width + height)

A Hybrid of classification and Regression problem

Evolutionary history of RCNNs

*Zhong-Qiu Zhao, 2019



(1) Two stage Object detectors

- RCNN
- *SPP-Net* (Spatial Pyramid Pooling)
- Fast RCNN
- Faster RCNN (FRCN)
- FPN
- RFCN (Region Fully Connected Network)
- Mask RCNN

- A network which has a separate module to generate region proposals is termed as a two-stage detector.
- These models try to find an arbitrary number of objects proposals in an image during the first stage and then classify and localize them in the second.
- As these systems have two separate steps, they generally take longer to generate proposals, have complicated architecture and lacks global context.

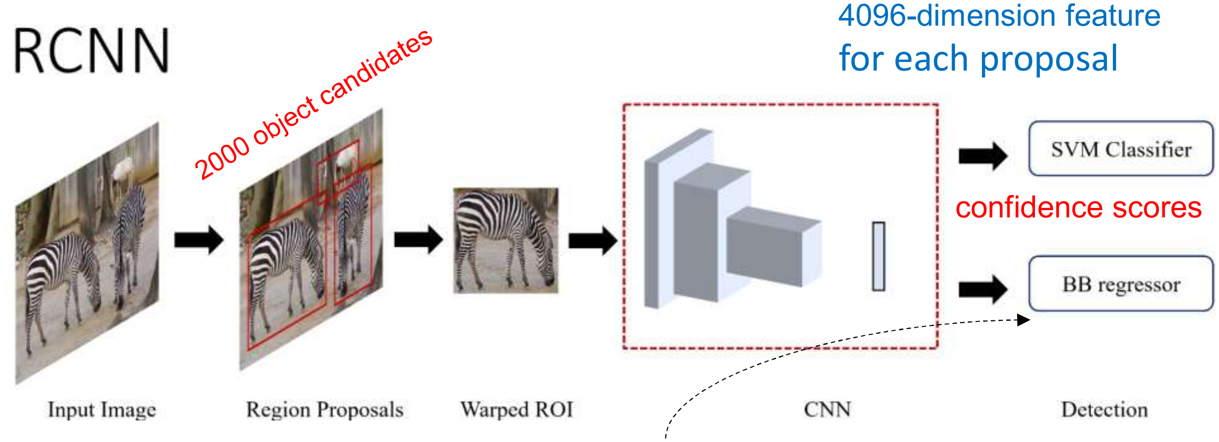
Region-based Convolutional Neural Network (RCNN)

*R. Girshick 2014

A mean-subtracted input image is first passed through the region proposal module, which produces 2000 object candidates.
This module find parts of the image which has a higher probability of finding an object using Selective Search.
These candidates are then warped and propagated through a CNN network, which extracts a 4096-dimension feature vector for each proposal.
Girshick et al. used AlexNet as the backbone architecture of the detector.
The feature vectors are then passed to the trained, class-specific Support Vector Machines (SVMs) to obtain confidence scores.
Non-maximum suppression (NMS) is later applied to the scored regions, based on its IoU and class.
Once the class has been identified, the algorithm predicts its bounding box using a trained bounding-box regressor which predicts four parameters i.e., center coordinates of box along with its width and height .

A multistage training process

4096-dimension feature
for each proposal



Bounding-Box regressor
four parameters i.e., center coordinates of box along with its width and height

Region Proposal Networks abbreviated as RPN.
To generate these so called "proposals" for the region where the object lies, a small network is slide over a convolutional feature map that is the output by the last convolutional layer.
This module find parts of the image which has a higher probability of finding an object using Selective Search.

SPP-Net (SPP-Net is considerably faster than the R-CNN model with comparable accuracy)

*K. He , 2015

SPP(Spatial Pyramid Pooling)-net only shifted the convolution layers of CNN before the region proposal module and added a pooling layer, thereby making the network independent of size/aspect ratio and reducing the computations.

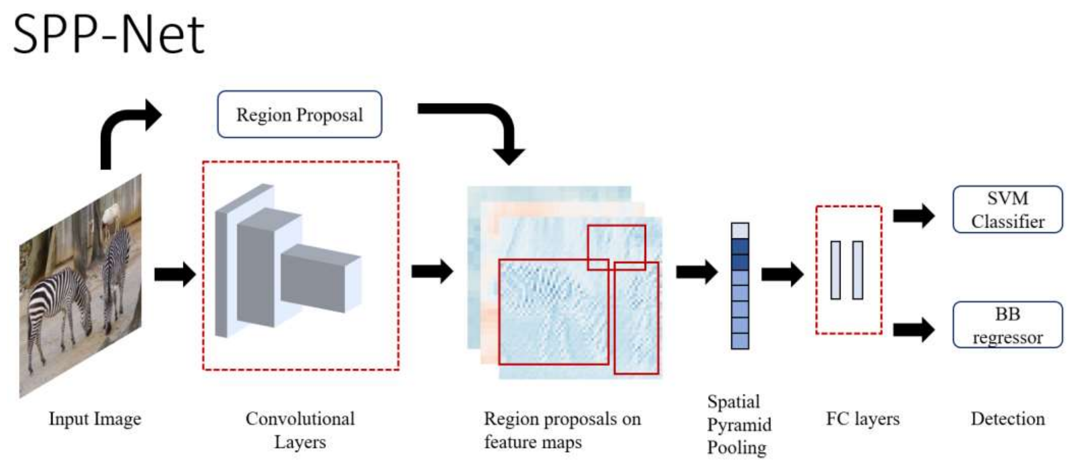
The selective search algorithm is used to generate candidate windows.

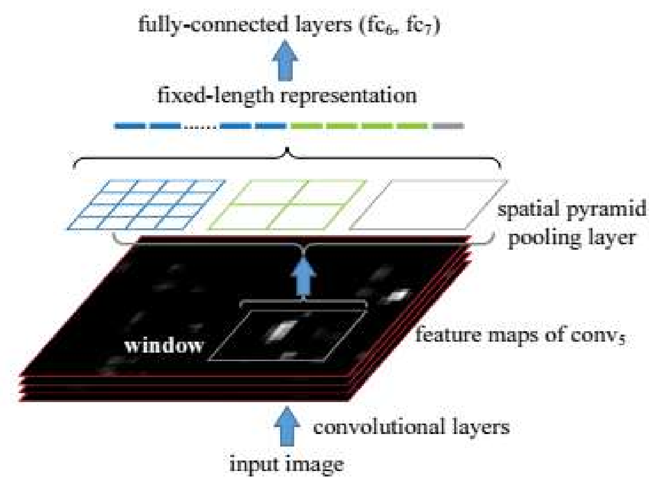
Feature maps are obtained by passing the input image through the convolution layers of a ZF-5 network.

The candidate windows are then mapped on to the feature maps, which are subsequently converted into fixed length representations by spatial bins of a pyramidal pooling layer.

This vector is passed to the fully connected layer and ultimately, to SVM classifiers to predict class and score.

Similar to R-CNN, SPP-net has as post processing layer to improve localization by bounding box regression.





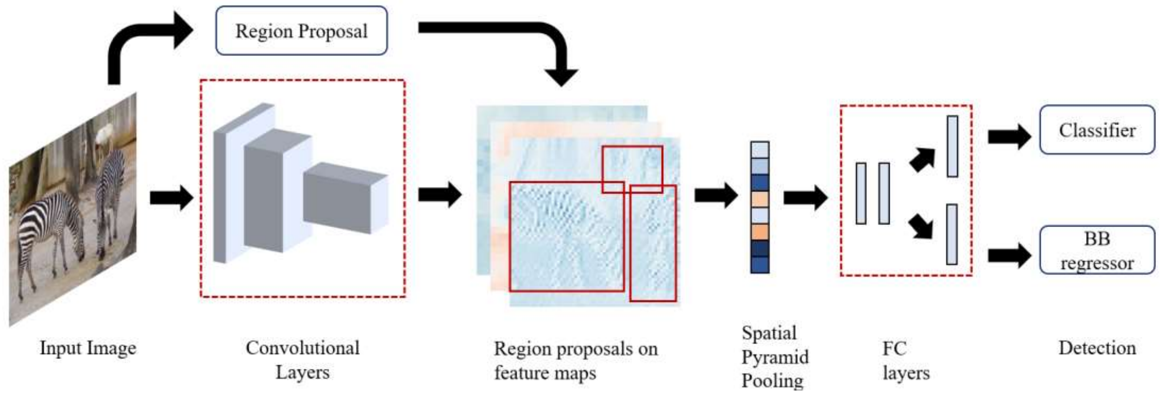
Fast RCNN

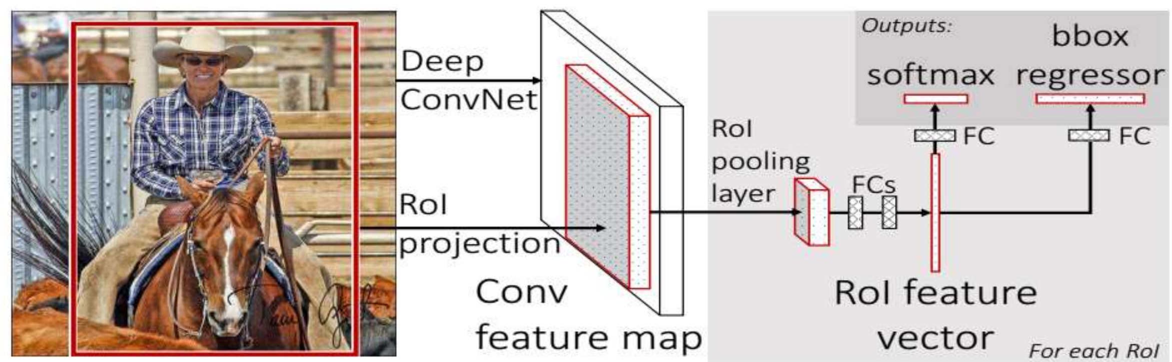
*R. Girshick 2015

- R-CNN/SPPNet need to train multiple systems separately.
- Fast R-CNN solved this by creating a single end-to-end trainable system.
- The network takes as input an image and its object proposals.
- The image is passed through a set of convolution layers and the object proposals are mapped to the obtained feature maps.
- Girshick replaced pyramidal structure of pooling layers from SPP-net with a single spatial bin, called RoI (**Region of interest)** pooling layer.
- The RoI pooling layer is a special case of the SPP layer, which has only one pyramid level.
- This layer is connected to 2 fully connected layer and then branches out into a *N+1*-class SoftMax layer and a bounding box regressor layer, which has a fully connected layer as well.
- The model also changed the loss function of bounding box regressor from L2 to smooth L1 to better performance, while introducing a multi-task loss to train the network.

Fast RCNN

a multi-task loss to train the network



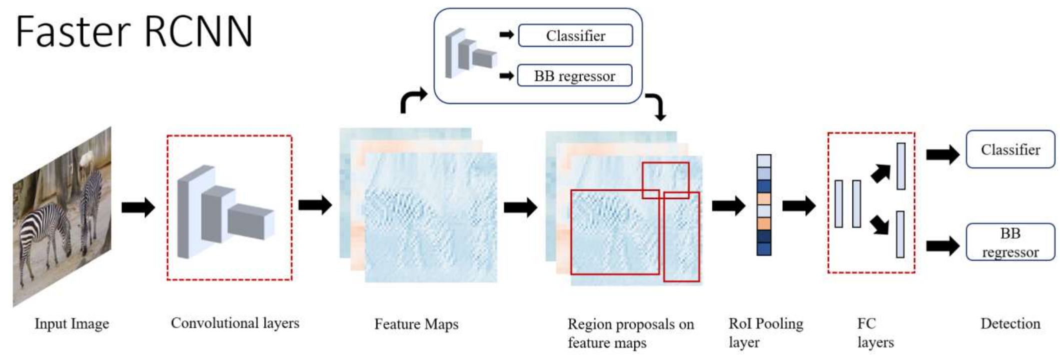


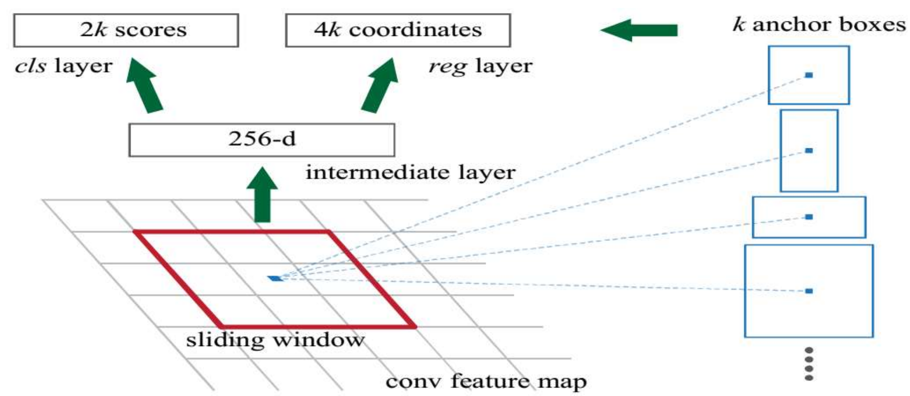
It simplified training procedure, removed pyramidal pooling and introduces a new loss function. The object detector, without the region proposal network, reported near real time speed with considerable accuracy

Faster RCNN

* S. Ren..2016

Faster RCNN takes an arbitrary input image and outputs a set of candidate windows.
Each such window has an associated *objectness score* which determines likelihood of an object.
Unlike its predecessors which used image pyramids to solve size variance of objects, RPN introduces Anchor boxes.
It used multiple bounding boxes of different aspect ratios and regressed over them to localize object.
The input image is first passed through the CNN to obtain a set of feature maps.
These are forwarded to the RPN, which produces bounding boxes and their classification.
Selected proposals are then mapped back to the feature maps obtained from previous CNN layer in RoI pooling layer, and ultimately fed to fully connected layer, which is sent to classifier and bounding box regressor.
Faster R-CNN is essentially Fast R-CNN with RPN as region proposal module.
Faster R-CNN improved the detection accuracy over the previous state-of-art by more than 3% . It fixed the bottleneck of slow region proposal and ran in near real time at 5 frames per second.

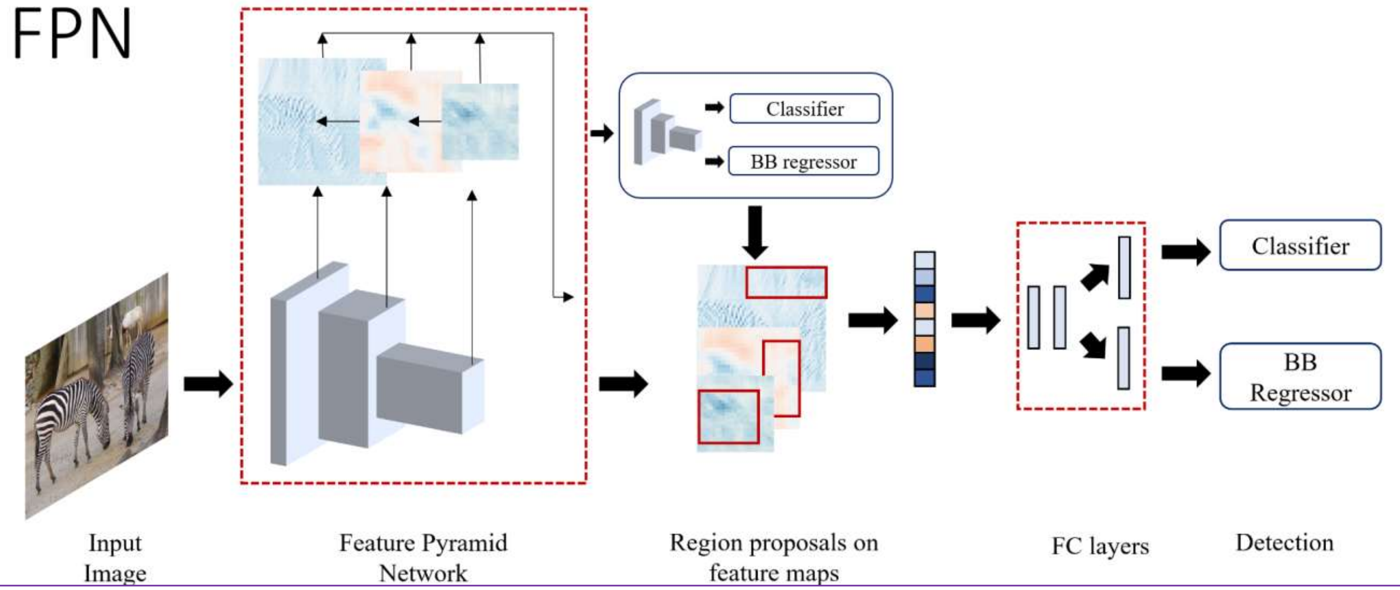




The RPN in Faster R-CNN K predefined anchor boxes are convoluted with each sliding window to produce fixed-length vectors which are taken by cls and reg layer to obtain corresponding outputs

FPN(Feature Pyramid Network)

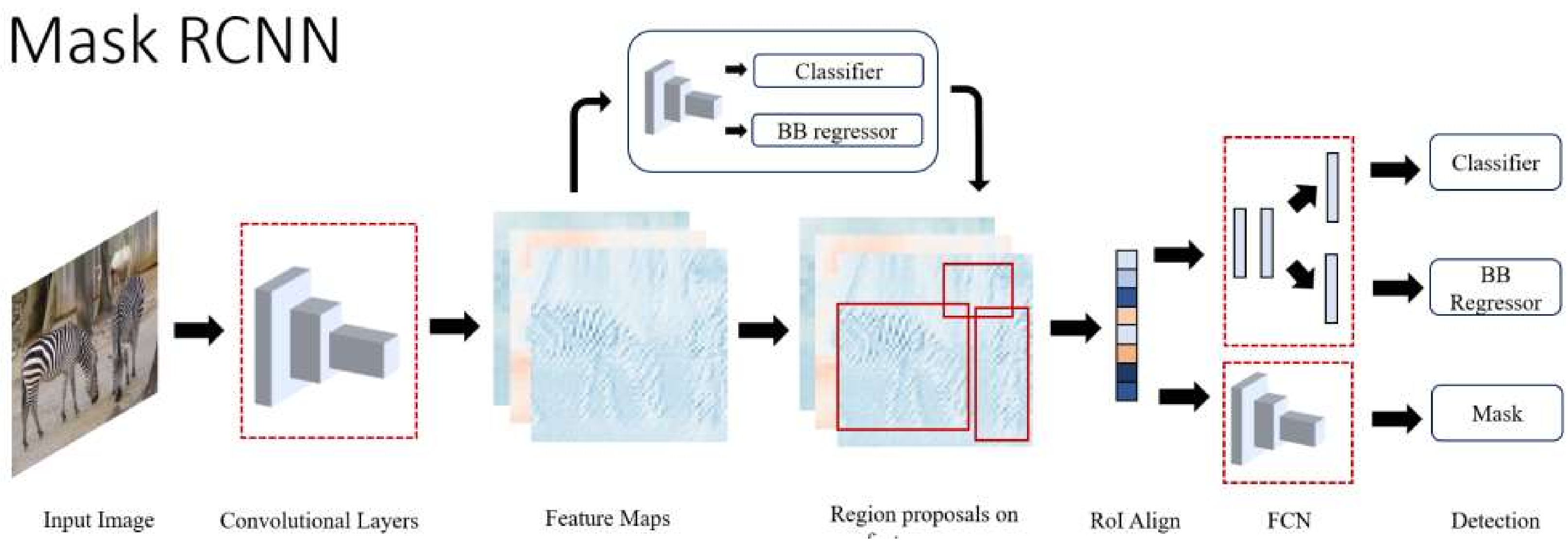
*T.-Y. Lin… 2017



FPN has a top-down architecture with lateral connections to build high-level semantic features at different scales.
The FPN has two pathways, a bottom-up pathway which is a ConvNet computing feature hierarchy at several scales and a top-down pathway which up samples coarse feature maps from higher level into high resolution features.
These pathways are connected by lateral connection by a *1x1* convolution operation to enhance the semantic information in the features.
FPN is used as a region proposal network (RPN) of a ResNet-101 based Faster R-CNN here.
FPN could provide high-level semantics at all scales, which reduced the error rate in detection.
It became a standard building block in future detections models and improved accuracy their accuracy across the table. It also lead to development of other improved networks

Mask RCNN

* K. He...2018



Mask R-CNN extends on the Faster R-CNN by adding another branch in parallel for pixel-level object instance segmentation.
The branch is a fully connected network applied on RoIs to classify each pixel into segments with little overall computation cost.
It uses similar basic Faster R-CNN architecture for object proposal, but adds a mask head parallel to classification and bounding box regressor head.
The authors chose the ResNeXt-101 as its backbone along with the feature Pyramid Network (FPN) for better accuracy and speed.
The loss function of Faster R-CNN is updated with the mask loss and as in FPN, it uses 5 anchor boxes with 3 aspect ratio. Overall training of Mask R-CNN is similar to faster R-CNN

NMS: Non-maximum Suppression

Non max suppression is **a technique used mainly in object detection that aims at selecting the best bounding box out of a set of overlapping boxes**.

The first step in NMS is to remove all the predicted bounding boxes that have a detection probability that is less than a given NMS threshold.

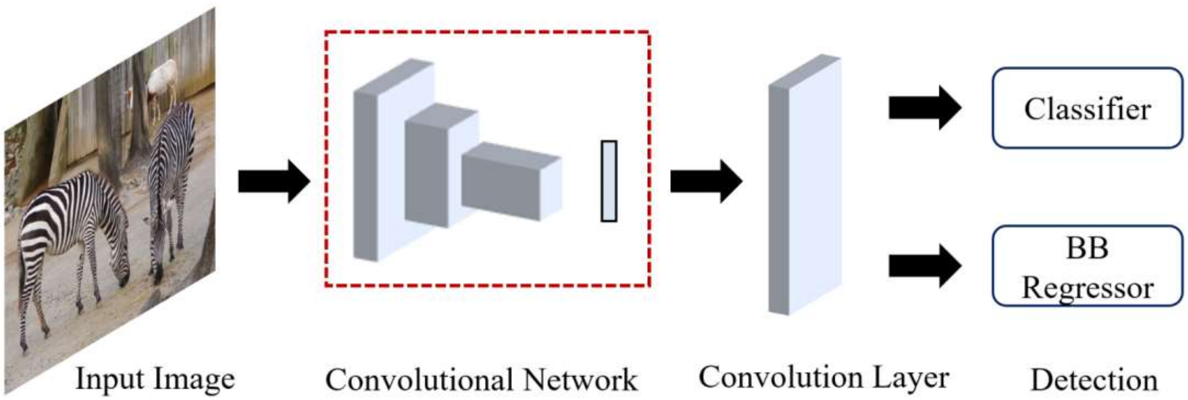
3. One stage object detectors

1. YOLO
2. SSD
3. YLOLO2, YOLO9000
4. Retina Net
5. YOLOv3
6. Center Net
7. EfficientDet
8. YOLOv4
9. Swin Transformer
10. YOLOx

- Single-stage detectors classify and localize semantic objects in a single shot using dense sampling.
- They use predefined boxes/keypoints of various scale and aspect ratio to localize objects.
- It edges two-stage detectors in real-time performance and simpler design.

YOLO (Y YOLO

* J. Redmon ...201



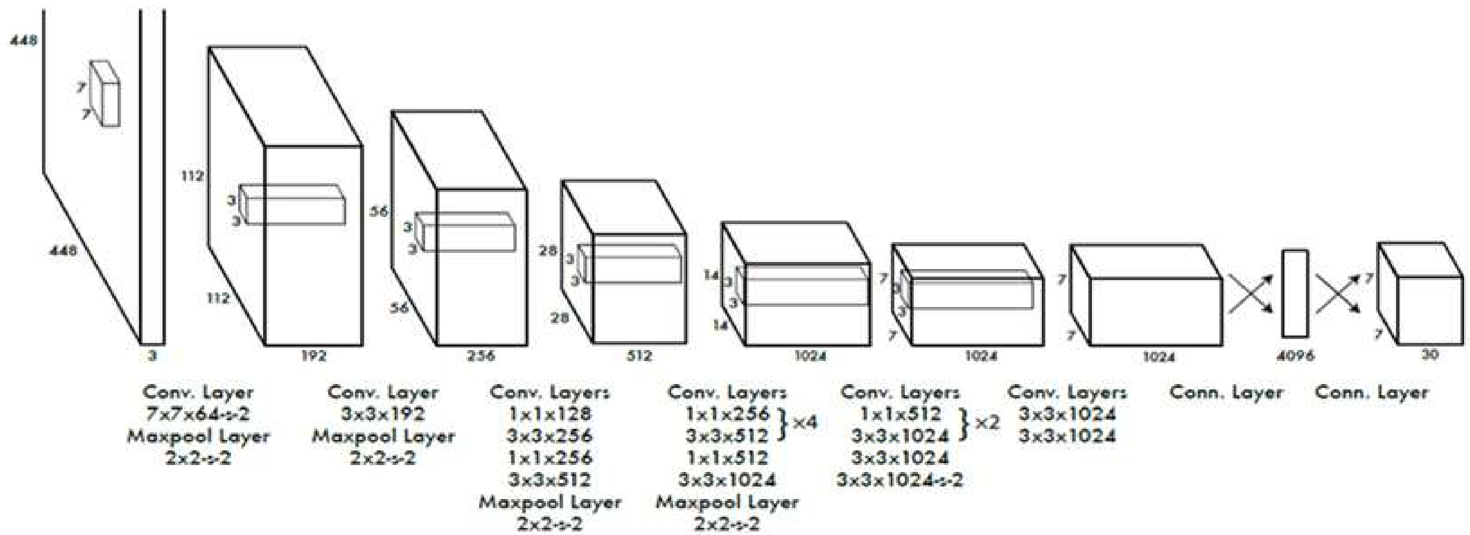
YOLO reframed it as a regression problem, directly predicting the image pixels as objects and its bounding box attributes. In YOLO, the input image is divided into a *S x S* grid and the cell where the object's center falls is responsible for detecting it.

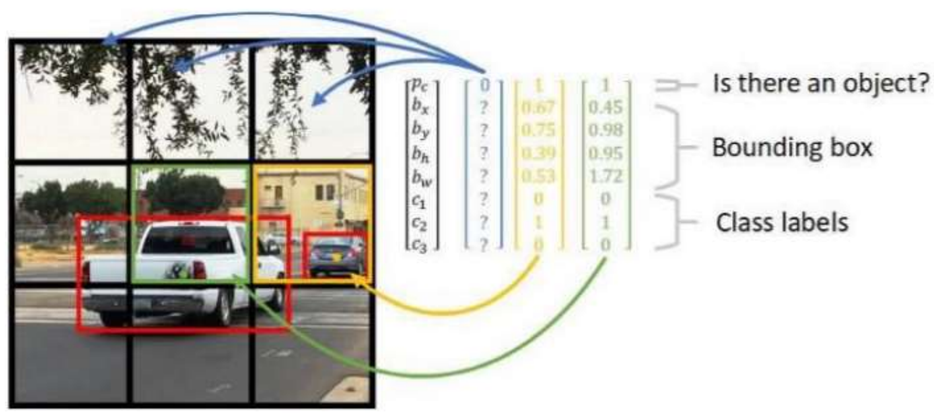
A grid cell predicts multiple bounding boxes, and each prediction array consists of 5 elements: center of bounding box – x and y, dimensions of the box – w and h, and the confidence score.

YOLO was inspired from the GoogLeNet model for image classification, which uses cascaded modules of smaller convolution networks.

It is pre-trained on ImageNet data till the model achieves high accuracy and then modified by adding randomly initialized convolution and fully connected layers.

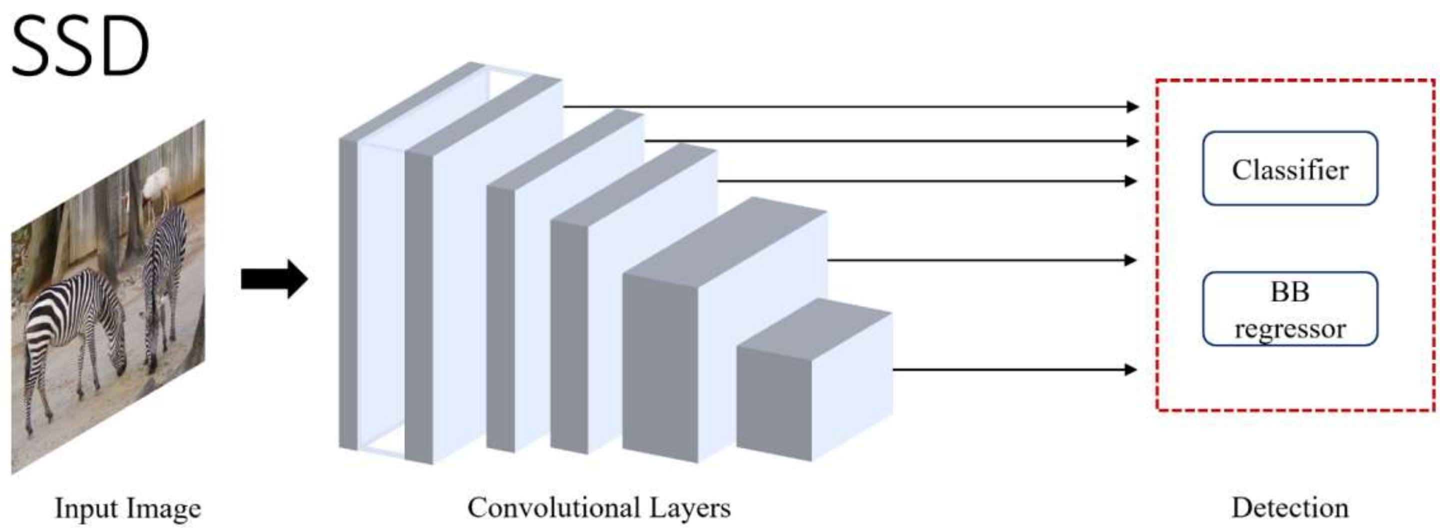
At training time, grid cells predict only one class as it converges better, but it is be increased during the inference time. Multitask loss, combined loss of all predicted components, is used to optimize the model.


448
7
7
448
3
112
3
3
112
192
56
3
3
56
256
28
3
3
28
512
14
3
3
14
1024
7
3
3
7
1024
7
7
1024
4096
7
7
30
Conv. Layer
7x7x64-s-2
Maxpool Layer
2x2-s-2
Conv. Layer
3x3x192
Maxpool Layer
2x2-s-2
Conv. Layers
1x1x128
3x3x256
1x1x256
3x3x512
Maxpool Layer
2x2-s-2
Conv. Layers
1x1x256
3x3x512
×4
1x1x512
3x3x1024
Maxpool Layer
2x2-s-2
Conv. Layers
1x1x512
3x3x1024
×2
3x3x1024
3x3x1024-s-2
Conv. Layers
3x3x1024
3x3x1024
Conn. Layer
Conn. Layer

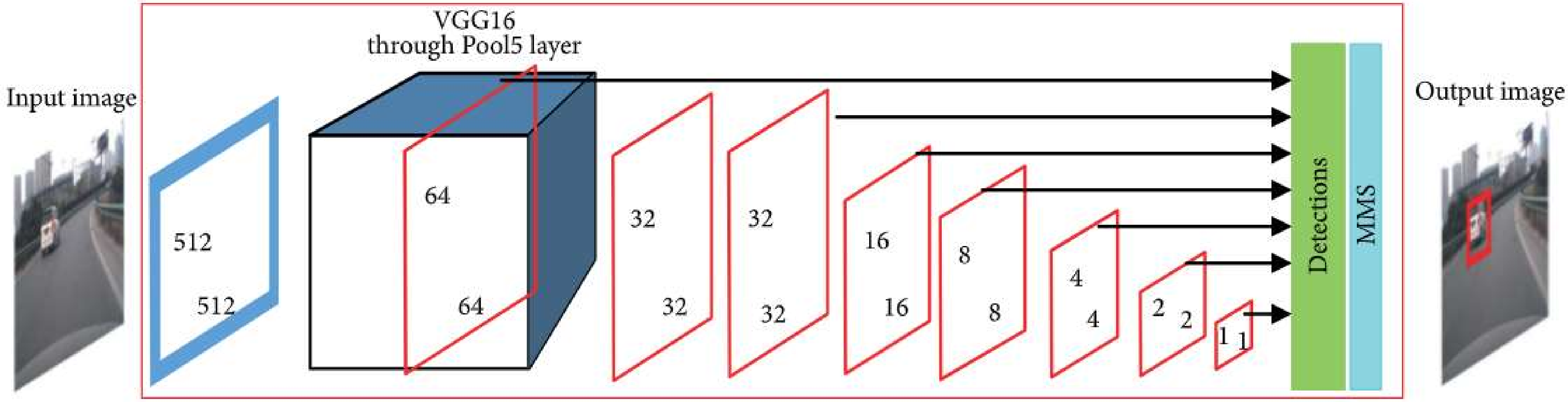

p_c b_x b_y b_h b_w c_1 c_2 c_3
0 ? ? ? ? ? ? ?
1 0.67 0.75 0.39 0.53 0 1 0
1 0.45 0.98 0.95 1.72 0 1 0
Is there an object?
Bounding box
Class labels

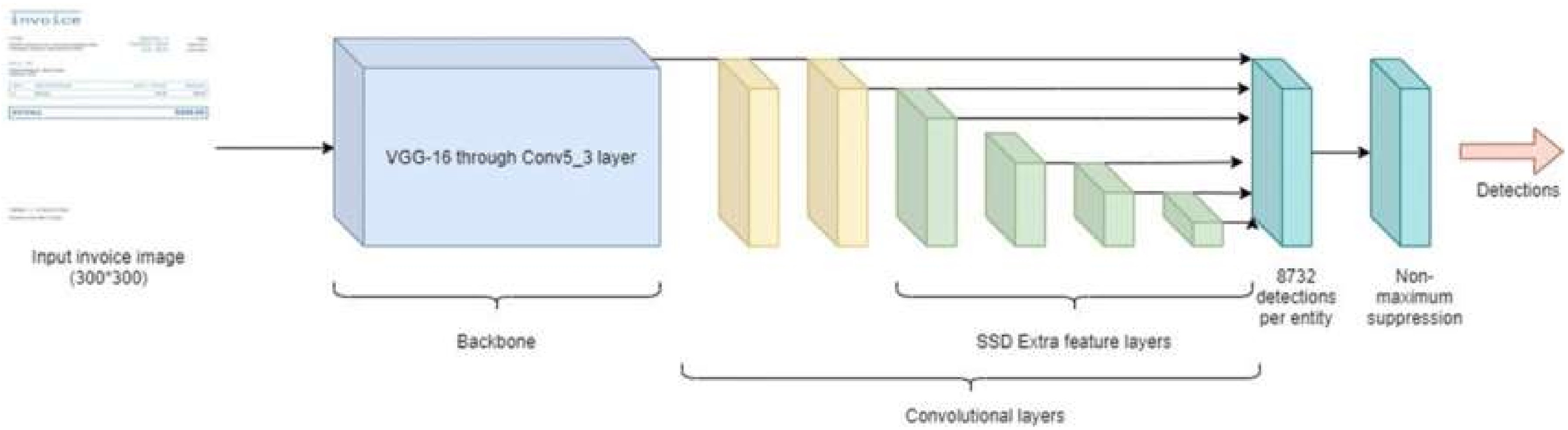
SSD(SingleShot Detector)

*W. Liu, 2016



SSD was the first single stage detector that matched accuracy of contemporary two stage detectors like Faster R-CNN [44], while maintaining real time speed.
SSD was built on VGG-16, with additional auxiliary structures to improve performance.
These auxiliary convolution layers, added to the end of the model, decrease progressively in size. SSD detects smaller objects earlier in the network when the image features are not too crude, while the deeper layers were responsible for offset of the default boxes and aspect ratios.
Even though SSD was significantly faster and more accurate than both state-of-art networks like YOLO and Faster R-CNN, it had difficulty in detecting small objects.
This issue was later solved by using better backbone architectures like ResNet and other small fixes


VGG16
through Pool5 layer
Input image
Output image
512
512
64
64
32
32
32
32
16
16
8
8
4
4
2
2
1
1
Detections
MMS

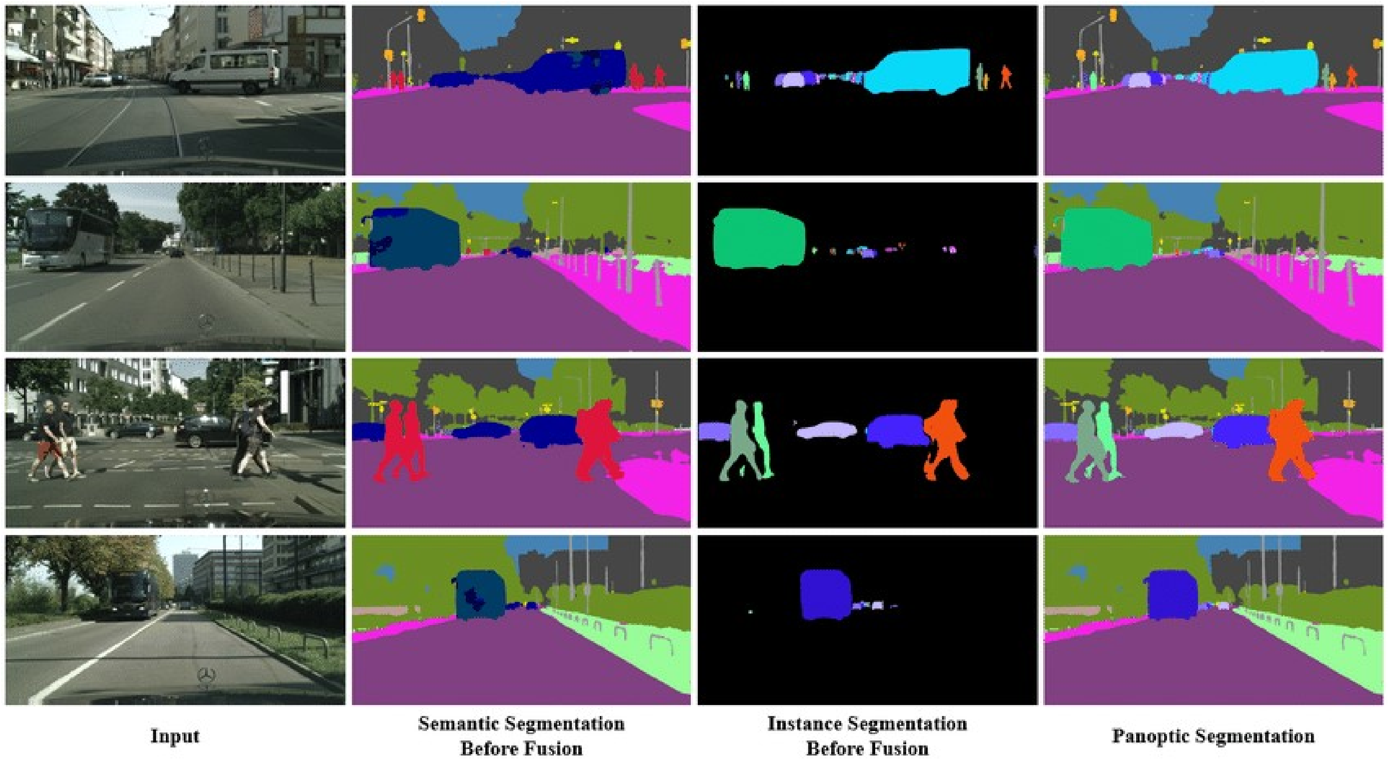

Input invoice image
(300*300)
VGG-16 through Conv5_3 layer
Backbone
SSD Extra feature layers
Convolutional layers
8732
detections
per entity
Non-
maximum
suppression
Detections

CNNs for Object Segmentation

CNNs which broke down an image into various subgroups called Image segments

There are three manners for segmentation: Semantic segmentation, Instance segmentation, and Panoptic segmentation

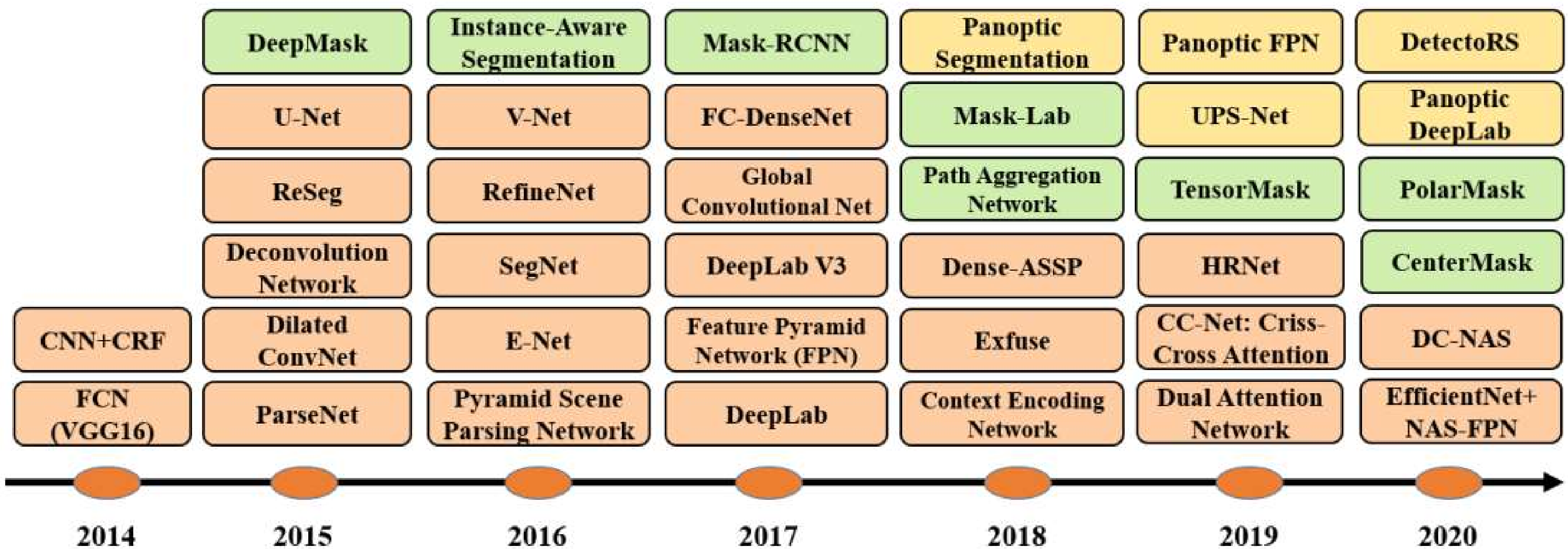
Panoptic segmentation is **proposed to unify the typically distinct tasks of semantic segmentation and instance segmentation**. The proposed task requires the generation of a rich and complete coherent scene segmentation, which is an important step towards a real-world visual system.



Semantic Seg. Instance Seg. Panoptic Seg.

A timeline of DL-based Segmentation algorithms

*S. Minaee 2020



The timeline of DL-based segmentation algorithms for 2D images, from 2014 to 2020.
Orange, green, and yellow blocks refer to semantic, instance, and panoptic segmentation algorithms respectively

Metrics For Segmentation Models

- **Pixel accuracy**

where P_{ij} is the number of pixels of class *i* predicted as belonging to class *j*.

$$\mathrm{PA} = \frac{\sum_{i=0}^{K} p_{ii}}{\sum_{i=0}^{K} \sum_{j=0}^{K} p_{ij}}$$

- **Mean Pixel Accuracy (MPA)**

MPA is the extended version of PA, in which the ratio of correct pixels is computed in a per-class manner and then averaged over the total number of classes

$$\mathrm{MPA} = \frac{1}{K+1} \sum_{i=0}^{K} \frac{p_{ii}}{\sum_{j=0}^{K} p_{ij}}$$

- **Intersection over Union (IoU)**

$$\mathrm{IoU} = J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

(IoU) or the **Jaccard Index** is one of the most commonly used metrics in semantic segmentation. It is defined as the area of intersection between the predicted segmentation map and the ground truth, divided by the area of union between the predicted segmentation map and the ground truth

- **Mean-IoU**

  the average IoU over all classes

- **Precision / Recall / F1 score**

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},$$

$$\text{F1-score} = \frac{2 \,\mathrm{Prec}\,\mathrm{Rec}}{\mathrm{Prec}+\mathrm{Rec}}$$

- **Dice coefficient**

$$\mathrm{Dice} = \frac{2|A \cap B|}{|A|+|B|} \qquad \mathrm{Dice} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} = \mathrm{F1}$$

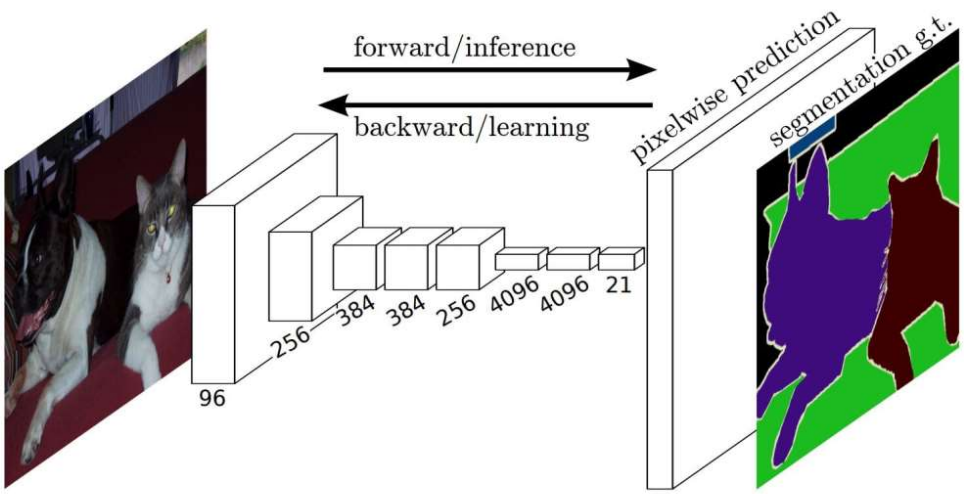
1. Fully Conventional Neural Networks

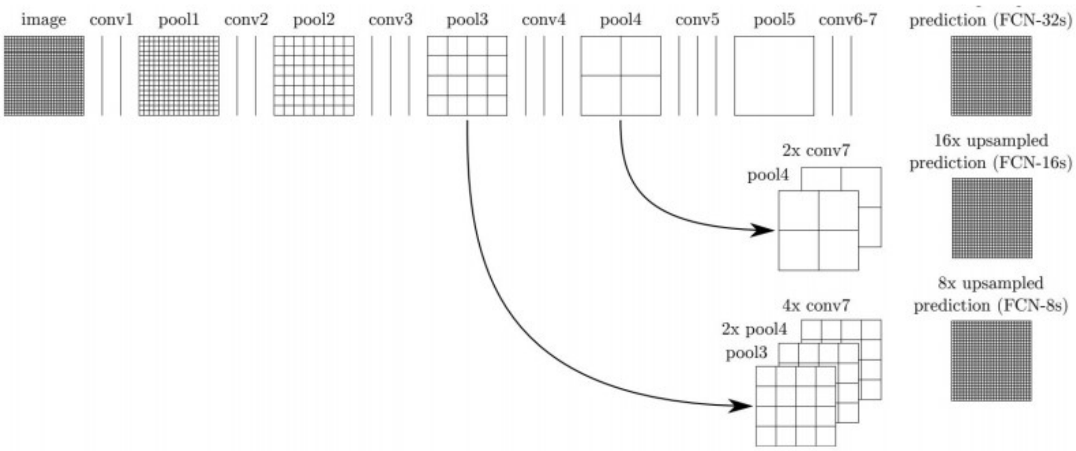
It includes only convolutional layers

The applied backbones is CNN architectures such as VGG16/GoogLeNet

FCN

J. Long... 2015

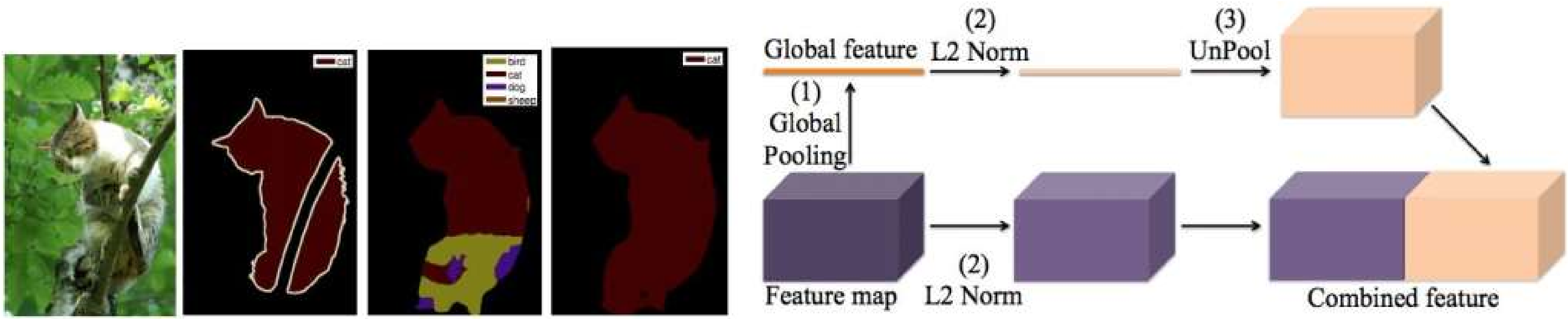




Through the use of skip connections in which feature maps from the final layers of the model are up-sampled and fused with feature maps of earlier layers, the model combines semantic information (from deep, coarse layers) and appearance information (from shallow, fine layers) in order to produce accurate and detailed segmentations.

ParseNet

W. Liu..2015



(a) Image (b) Truth (c) FCN (d) ParseNet (e) ParseNet contexture module overview.

ParseNet adds global context to FCNs by using the average feature for a layer to augment the features at each location.

The feature map for a layer is pooled over the whole image resulting in a context vector.

This context vector is normalized and un-pooled to produce new feature maps of the same size as the initial ones.

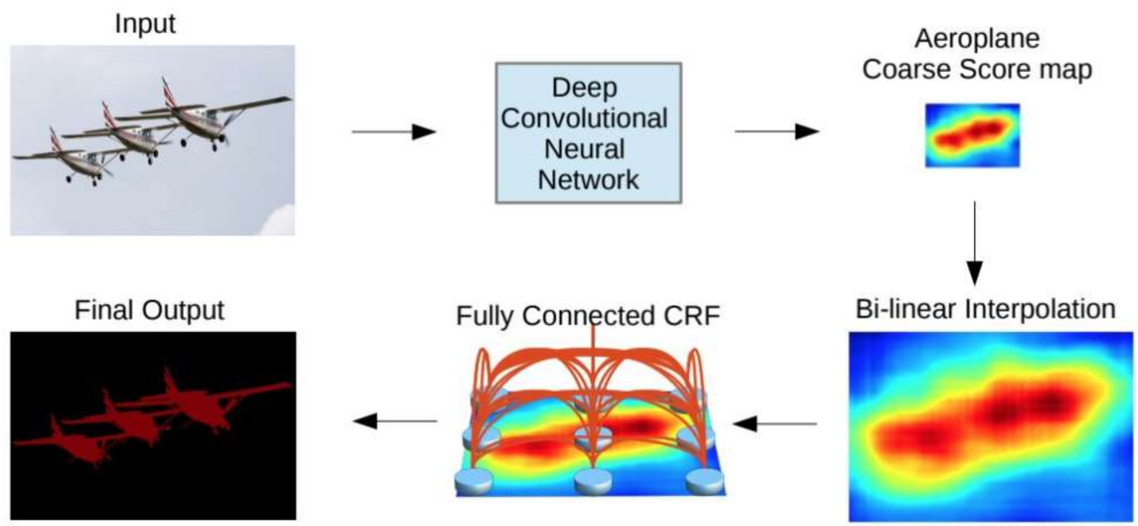
2. Convolutional Models With Graphical Models

As discussed, FCN ignores potentially useful scene-level semantic context. To integrate more context, several approaches incorporate probabilistic graphical models, such as Conditional Random Fields (CRFs) and Markov Random Field (MRFs), into DL architectures.

CNN+CRF (Condition Random Field)

Chen..2014

Chen proposed a semantic segmentation algorithm based on the combination of CNNs and fully connected CRFs. (CNN+CRF)
They showed that responses from the final layer of deep CNNs are not sufficiently localized for accurate object segmentation.
To overcome the poor localization property of deep CNNs, they combined the responses at the final CNN layer with a fully-connected CRF.
They showed that their model is able to localize segment boundaries at a higher accuracy rate than it was possible with previous methods.



The coarse score map of a CNN is upsampled via interpolated interpolation, and fed to a fully-connected CRF to refine the segmentation result.

Conditional random fields (CRFs) are **a class of statistical modeling methods used for structured prediction**. Whereas a classifier predicts a label for a single sample without onsidering "neighbouring" samples, a CRF can take context into account.

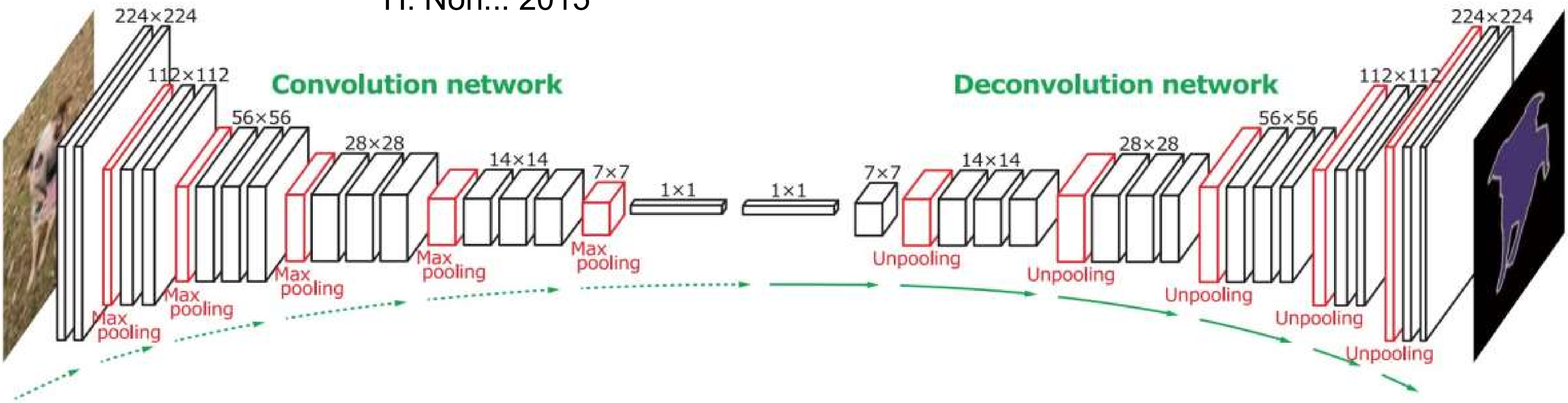
3. Encoder-Decoder Based Models

image segmentation based on the convolutional encoder-decoder architecture.

A. Encoder-Decoder Models for General Segmentation

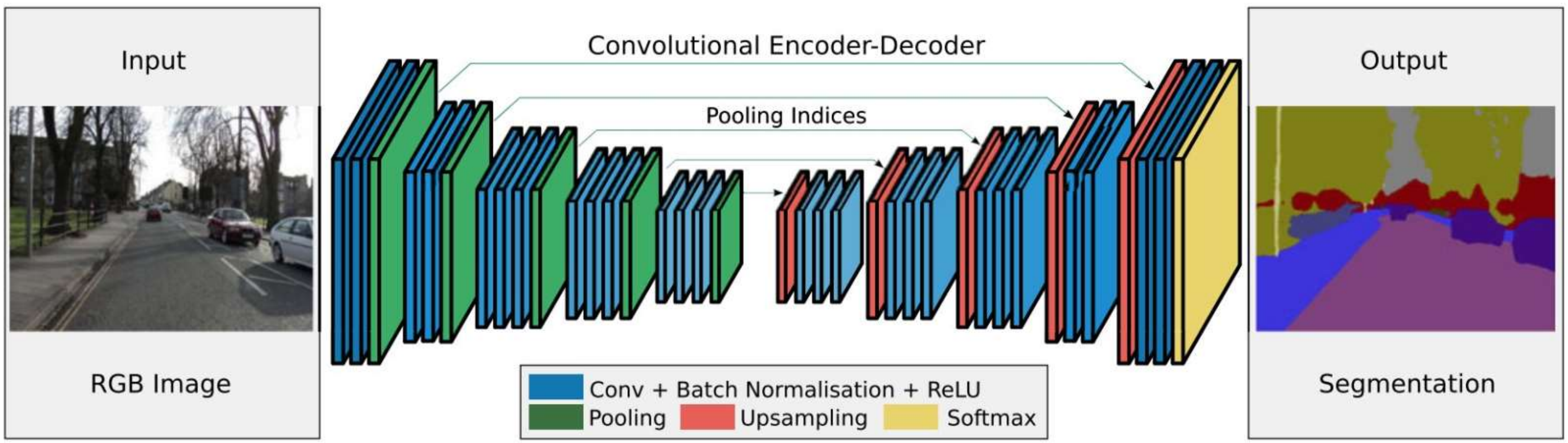
DSS (Deconvolutional semantic segmentation)

H. Noh... 2015



A convolution network based on the VGG 16-layer net, is a multi-layer deconvolution network to generate the accurate segmentation map.

SegNet
A. Kendall ...2015



SegNet has no fully-connected layers; hence, the model is fully convolutional. A decoder up-samples its input using the transferred pool indices from its encoder to produce a sparse feature map(s)

HRNET(high-resolution network)

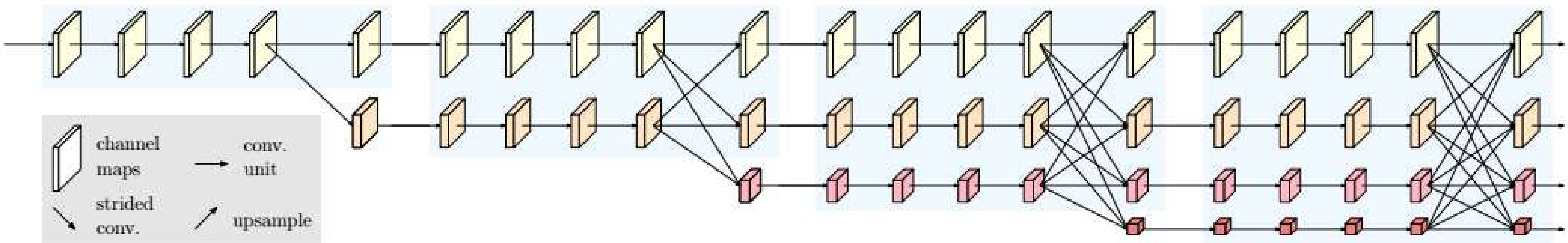
Y. Yuan...2019

Satge1
(high resolution Conv.)

Satge2
(Two resolution Blocks)

Satge3
(Three resolution Blocks)

Satge4
(Four resolution Blocks)



HRNet consists of parallel high-to-low resolution convolution streams with repeated information exchange across multi-resolution steams.

There are four stages:

The 1st stage consists of high-resolution convolutions.

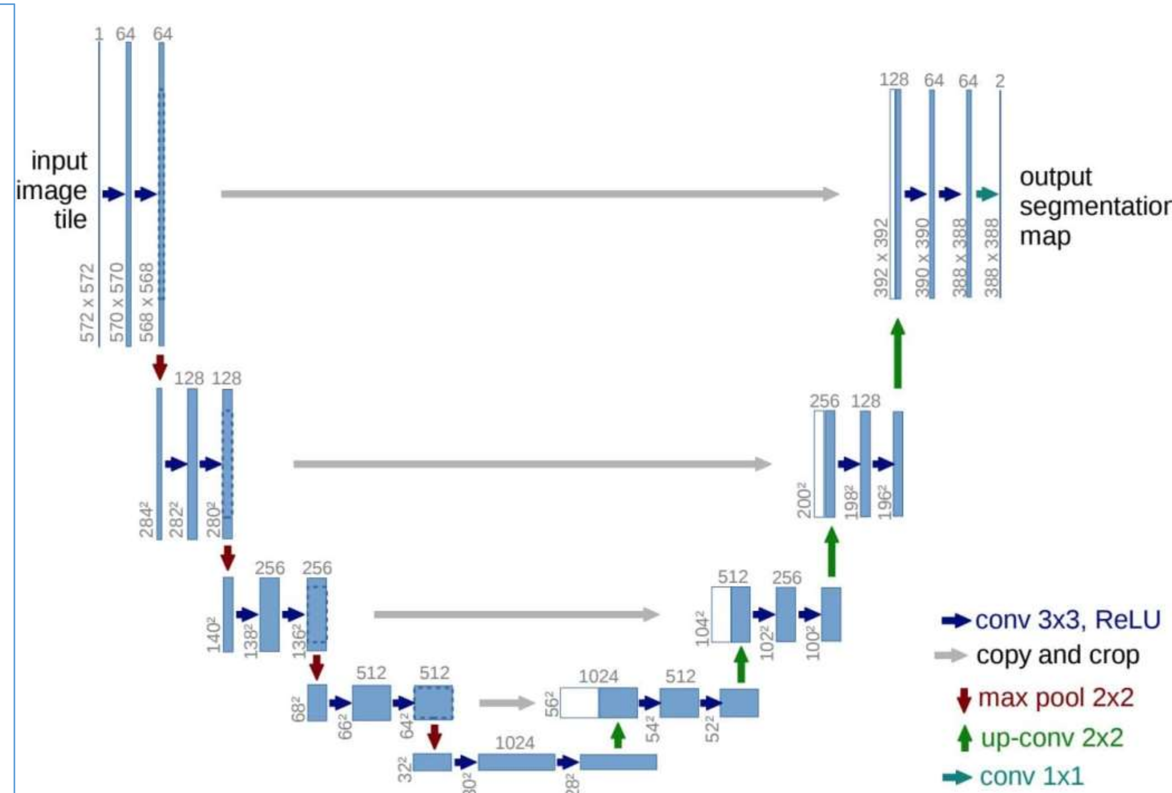
The 2nd (3rd, 4th) stage repeats two-resolution (three-resolution, four-resolution) blocks.

B. Encoder-Decoder Models for Medical and Biomedical Image Segmentation

UNet

Ronneberger....2015

U-Net is proposed for segmenting biological microscopy images.
Their network and training strategy relies on the use of data augmentation to learn from the very few annotated images effectively.
The U-Net architecture comprises two parts, a contracting path to capture context, and a symmetric expanding path that enables precise localization.
The down-sampling or contracting part has a FCN-like architecture that extracts features with 3 × 3 convolutions. The up-sampling or expanding part uses up-convolution (or deconvolution), reducing the number of feature maps while increasing their dimensions.
Feature maps from the down-sampling part of the network are copied to the up-sampling part to avoid losing pattern information.

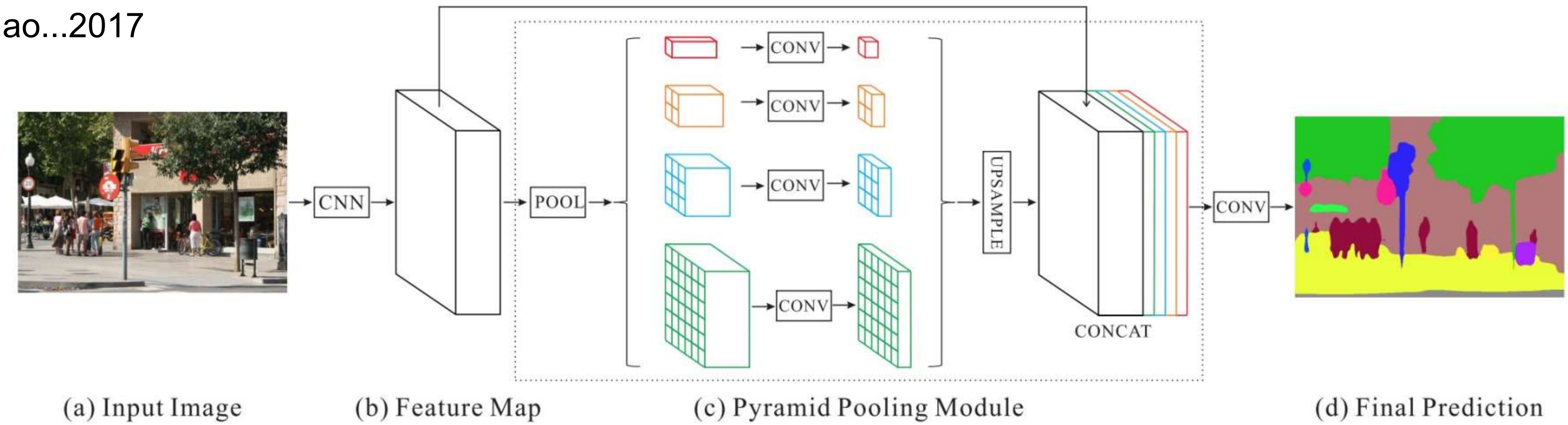


4. Multi-Scale and Pyramid Network Based Models

DNNs whose Backbone is from CNNs with Multi-Scale and Pyramid Network

PSPN Pyramid scene parsing network

Zhao...2017

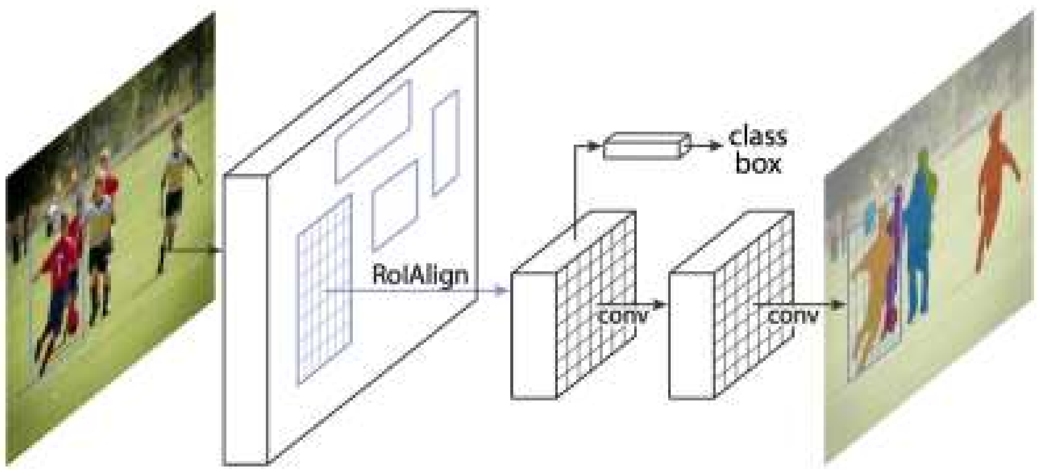


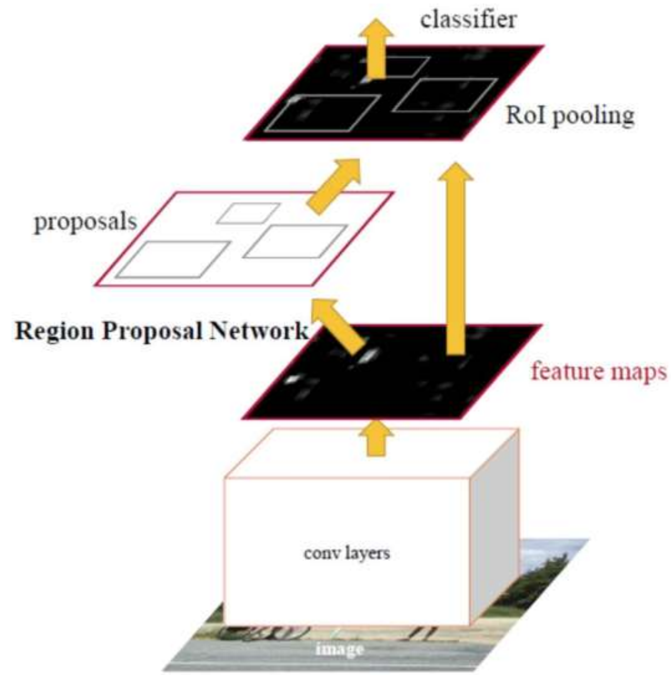
A CNN produces the feature map and a pyramid pooling module aggregates the different sub-region representations.
Up-sampling and concatenation are used to form the final feature representation from which, the final pixel-wise prediction is obtained through convolution.

5. R-CNN Based Models (for Instance Segmentation)

Mask R-CNN

K. He...2017



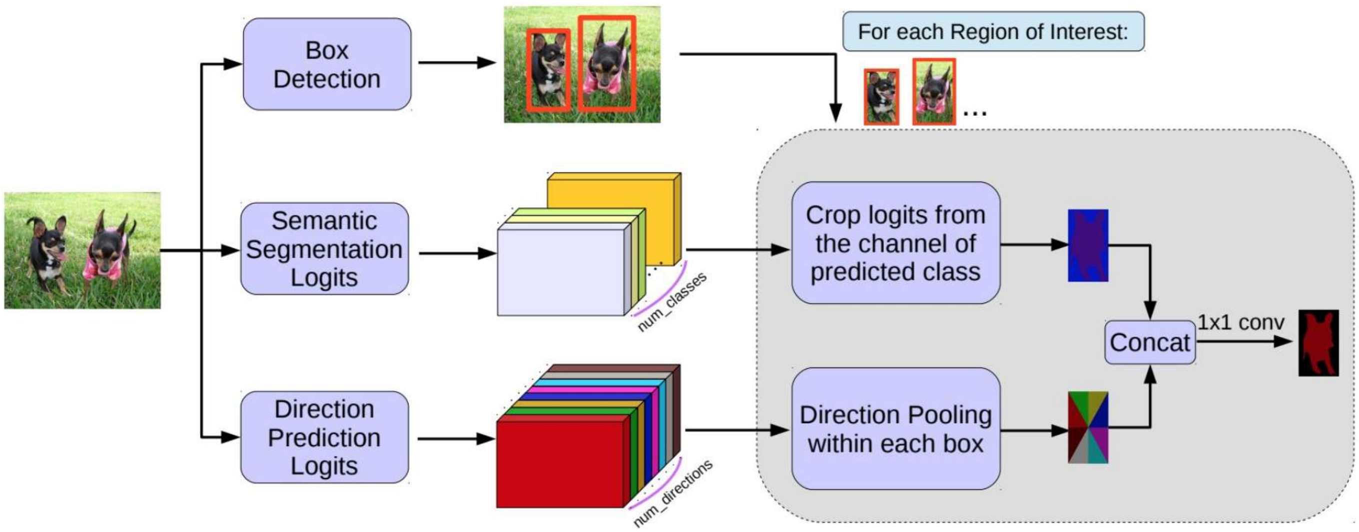


Mask R-CNN architecture for instance segmentation

MaskLab(Instance Segmentation)

L. Chen...2018

A model by refining object detection with semantic and direction features based on Faster R-CNN.
This model produces three outputs, box detection, semantic segmentation, and direction prediction.
Building on the FasterRCNN object detector, the predicted boxes provide accurate localization of object instances.
Within each region of interest, MaskLab performs foreground/background segmentation by combining semantic and direction prediction.



MaskLab generates three outputs—refined box predictions (from Faster R-CNN), semantic segmentation logits for pixel-wise classification, and direction prediction logits for predicting each pixel's direction toward its instance center.

End of Chapter4

Thank you