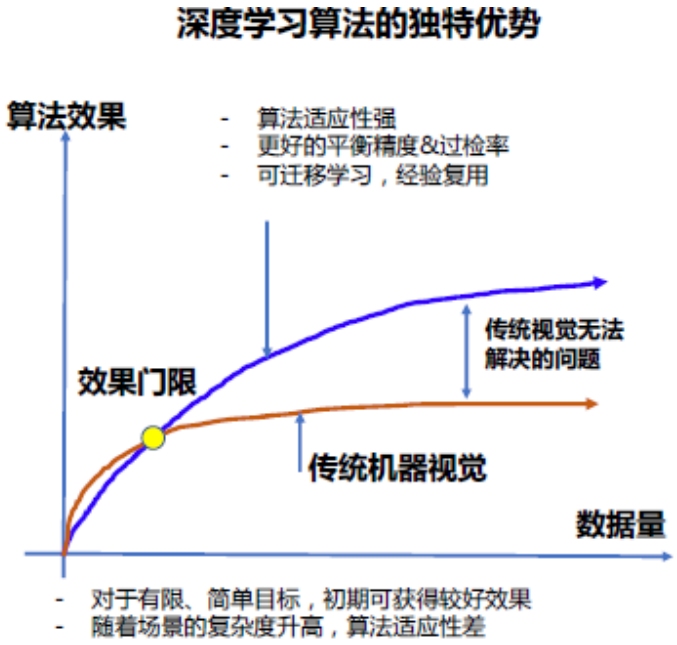
Object Detection

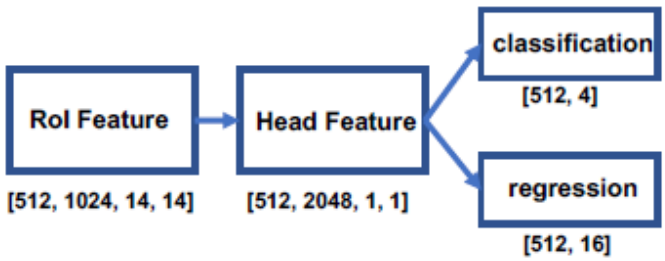
- Multi Objects classification + location
- environment： light + blur
- Existing challenge => training
  - crowded
  - occluded
  - highly overlapped
  - multi scale： extremely small + very large
  - few shot
  - rotate bounding box
- Existing challenge => deployment
  - how to decrease the size？
  - how to decrease the FLOP？
  - how to be real-time？
- traditional object detection
  - Process
    - Proposal Region(ROI)
    - extract feature （SIFT + HOG)
    - classification (SVM + Adaboost)
    - Postprocess(NMS)
  - limits
    - poor robust of algorithm
    - limit target area
    - huge cost to develop and maintain
-



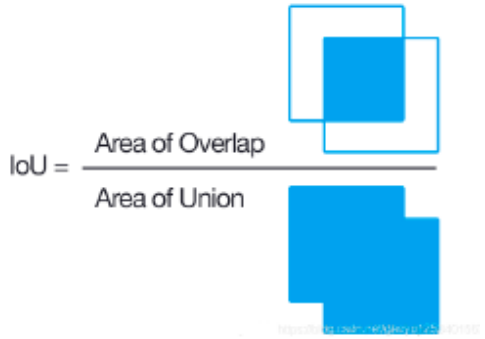
- Basic Definition
  - BBox: Bounding Box

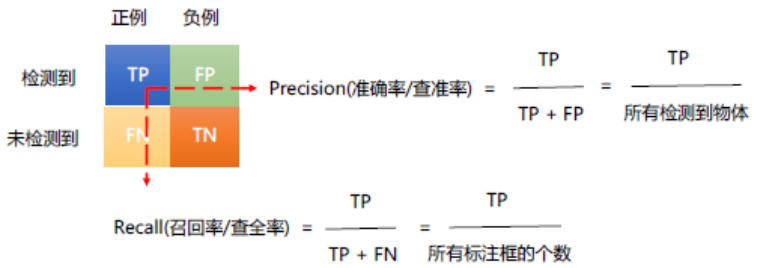
    - XYXY: left-top point + right-bottom point
    - XYWH: left + top +width + height
  - Anchor
    - manually set different size and area prior box
    - in SSD also called Prior box
    - each point on the feature map is used as a center point to generate multiple bounding boxes with different proportions
  - Region Proposal
  - RPN: Region Proposal Network in anchor-based two stage to extract candidate box
    - 1. Enter a piece of supervision information into the RPN network to determine whether the Anchor contains objects.
      - Anchor contains objects---positive samples
      - Anchor does not contain objects---negative samples
    - 2. Determine the positive or negative sample according to the value of Anchr and the real frame loU.
      - Positive sample: Anchor with the largest loU of a real frame,Anchor with any real frame loU>0.7 (1)

        - Negative sample: Anchor with IoU<0.3 with all real frames(0)
    - 3. Sampling rules: A total of 256 samples are sampled
        - Sampling randomly from positive samples, the number of samples does not exceed 128
        - Randomly sample from negative samples and fill in 256 samples
    - 4. Loss : classification (binary cross entropy) + regression loss(smooth L1)
- RoI: Region of Interest
    - ROI Pooling: candidate bbox share the feature map feature and keep the output shape the same (max pooling *4)
        - RoI Pooling will lead to mismatch between the detected information and the extracted features when it is rounded twice.
        - 1. The position of the candidate frame is rounded. When the position of RoI is not an integer, the position of RoI needs to be rounded.
        - 2. Round off when extracting features. Divide 4 sub-areas to do max pooling, and the length of the frame needs to be rounded approximately.
    - RoI Align: elimate the mismatch in ROI Pooling
        - Take N points uniformly in the area, find the four points closest to each point on the feature map; then obtain the output value of the point by means of bilinear interpolation. Finally, average the N points to get the output of the area.
- BBox Head training phase:
    - 1. Determine whether the proposals generated by the RPN network contain objects
    - 2. Determine the positive or negative sample based on the IoU of the RoI and the real frame.
        - Positive sample: IoU>0.5
        - Negative sample: IoU<0.5
    - 3. Sampling rules:
        - A total of 512 samples are sampled
        - Random sampling from positive samples, the number of samples does not exceed 128
        - Randomly sample from negative samples and fill in 512 samples
        -



- IoU: Intersaction over Union

- 

- mAP
  - 

    - TP: IoU>=阈值(如0.3)检测框数量
    - FP: IoU<阈值(如0.3)检测框数量
    - FN: 没有检测到的GT数量

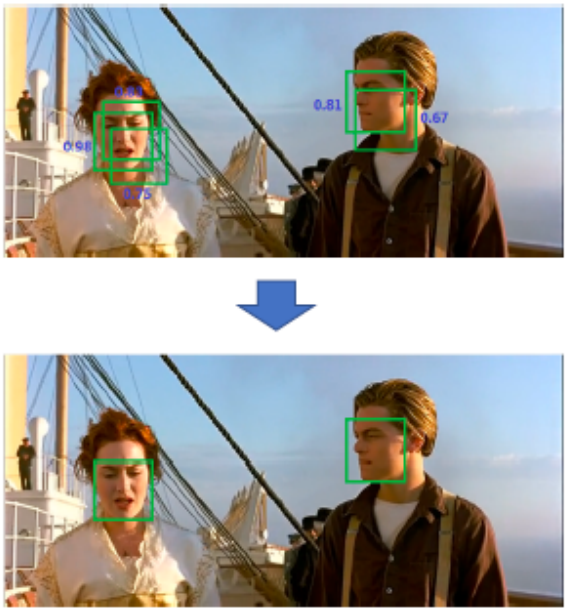
    - 举例:

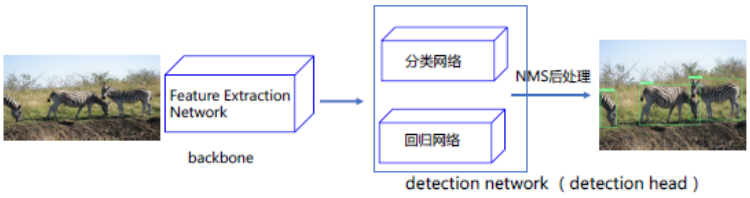
    输入样本: 某个类别10个框
    预测结果: 预测到8个框, 6个正确, 2个错误

    Precision: 6 / 8 = 0.75
    Recall: 6 / 10 = 0.6

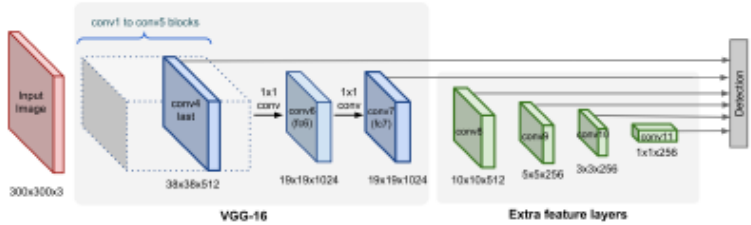
    - P-R曲线: 以Precision、Recall为纵、横坐标的曲线。
    - AP（Average Precision）: 某一类P-R 曲线下的面积
    - mAP（mean Average Precision）: 所有类别AP平均

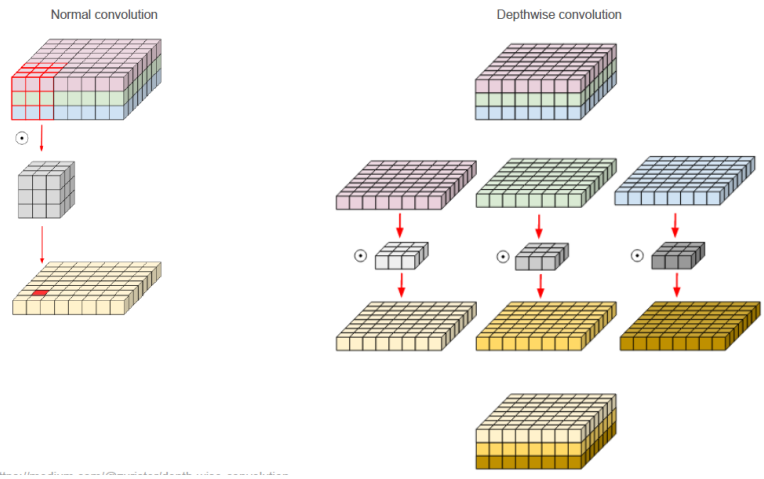
- NMS: Non-Maximum Suppression
  -



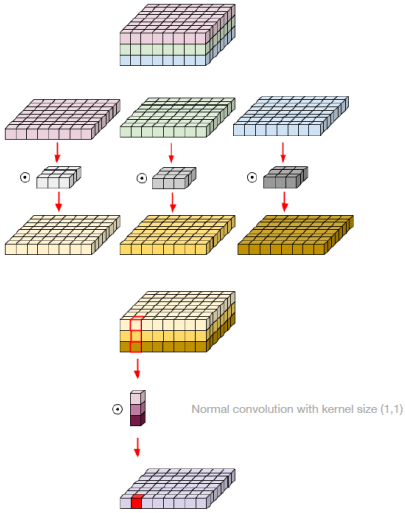
- One stage
  - 

  - directly to regress and classify the anchors
  - SSD



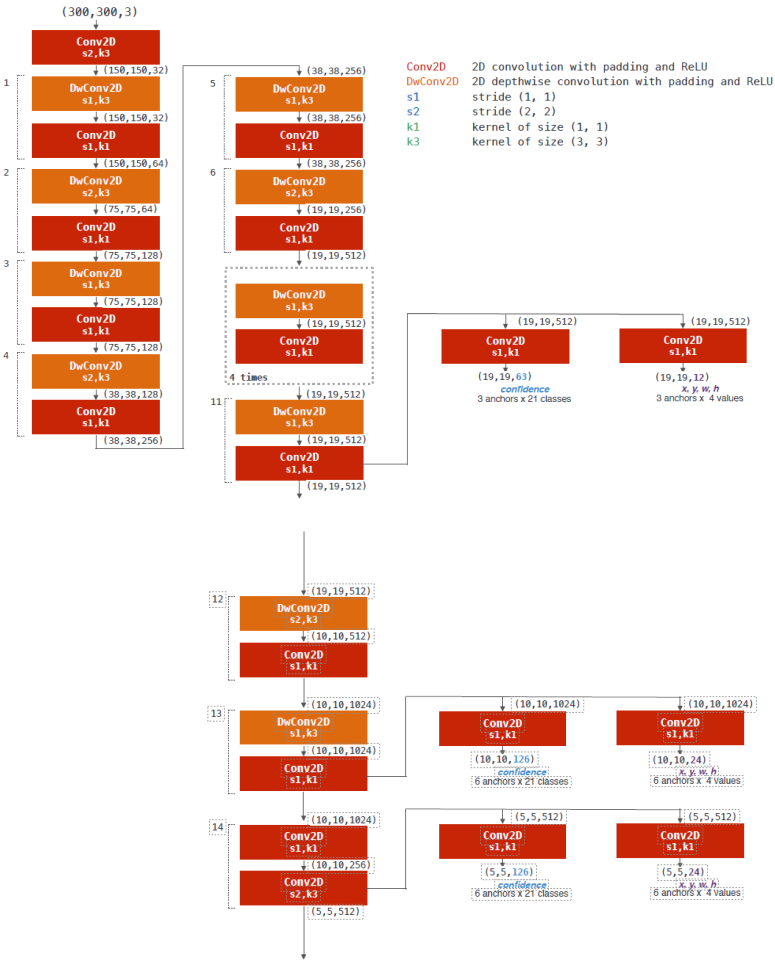


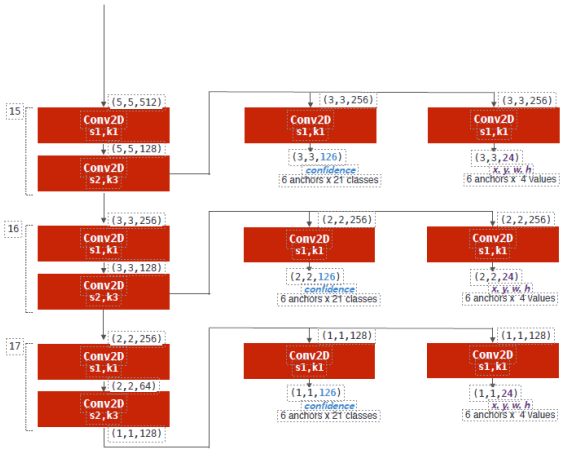
Depthwise separable convolution



Source: https://medium.com/@zurister/depth-wise-convolution

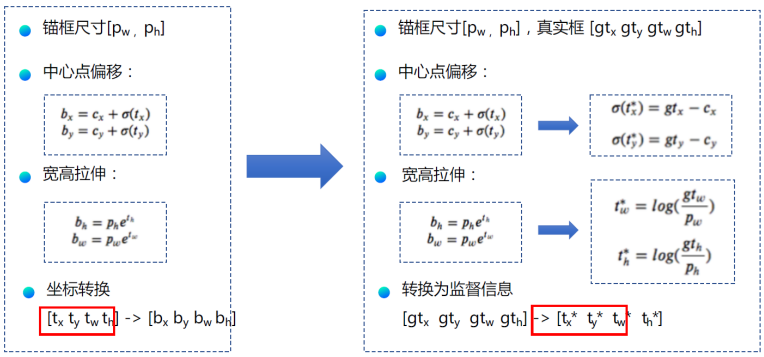
- Mobilenet SSD

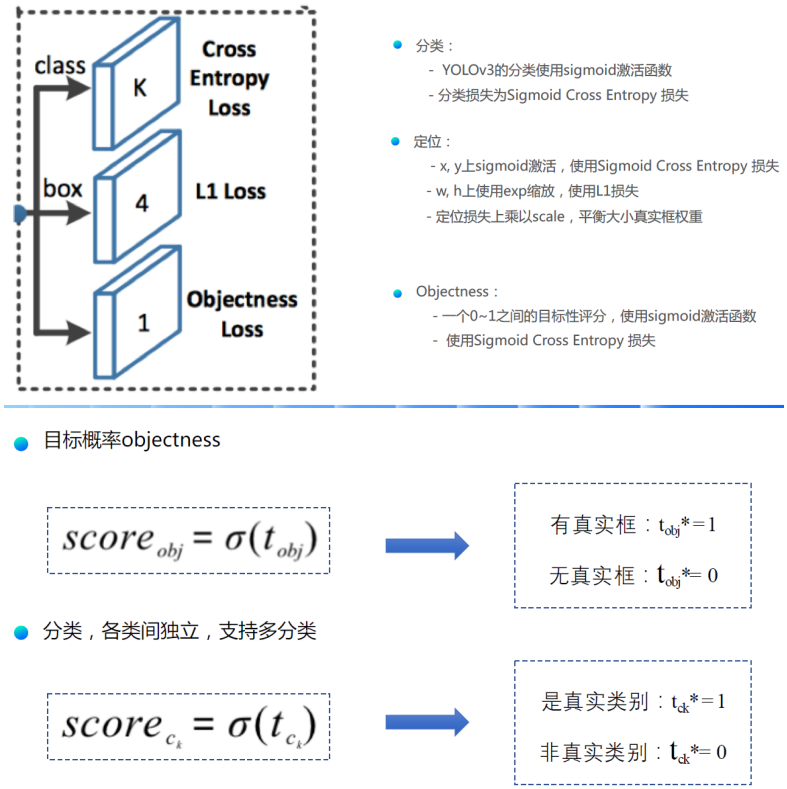




- YOLOV1
  - 1. Resize image
  - 2. Run convolutional network
    - Divide the image into S*S grids
    - On which grid the center of the object bbox falls on, the anchor box corresponding to the grid is responsible for detecting the object
    - The width and height (H, W) of the output feature map are S*S grids
    - Output feature map channel (C) dimensional organization x, y, w, h, Pobject, classification (one-hot) information
  - 3. NMS
- YOLOV2
  - Backbone 224*224 -> 448*448, DarkNet19
  - Conv+ Batch Norm
  - Kmeans
  - Multi-scale training
- YOLOV3
  - Backbone DarkNet53
  - Multi-scale prediction.
  - Cross-scale feature fusion
  - COCO data set clusters 9 different scale anchors, 3 for each scale
    - C=(4 + 1+ num of classes) *number of anchor on feature map
      - The feature map Cv represents the information of the prediction box
      - Coordinates: tx ty tw th
      - Target confidence: Pobj
      - Category: Pc
    -

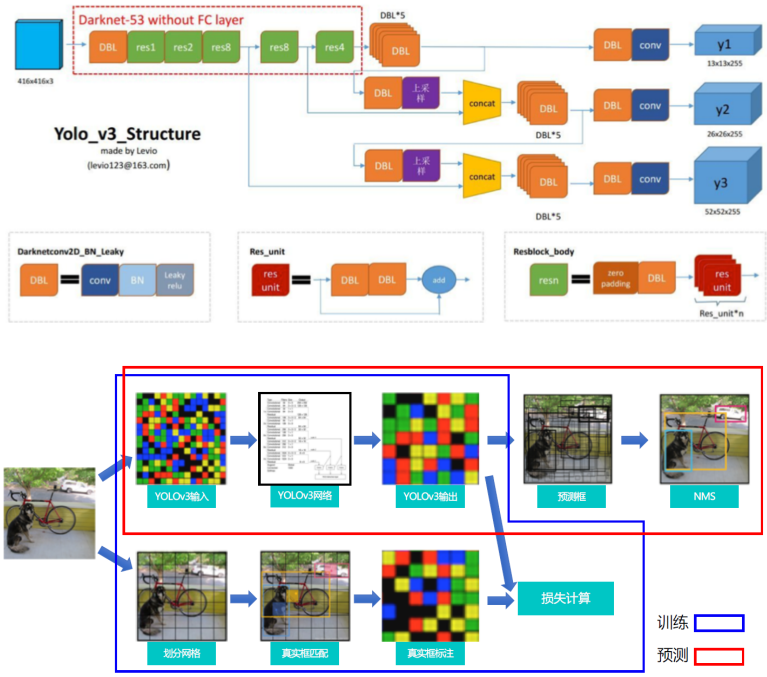
特征图	13*13			26*26			52*52		
感受野	大			中			小		
先验框	(116x90)	(156x198)	(373x326)	(30x61)	(62x45)	(59x119)	(10x13)	(16x30)	(33x23)

- The anchor frame only gives the width and height of the frame. How to represent the real frame of the object in the image?How many real frames are there? Where are they in the image?
  - Grid division, "frame sea" tactics
  - Grid coarse positioning + center point offset fine positioning
- The width and height of the anchor frame are fixed, and the width and height of the real frame of the object are various, how to express it?
  - Wide and high stretch
- 

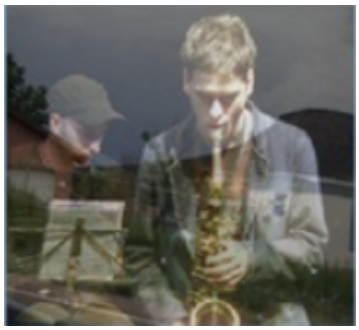
- Classification uses sigmoid activation to support multiple classification of targets
- 

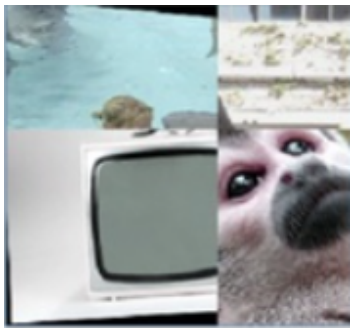
- What losses need to be calculated?
- 1. anchor frame where the real frame falls (positive example)
  - Positioning loss: the "deformation" coefficient tx ty tw t returns to the value that can be transformed into a real frame

        - Objectness loss: Return the objectness of the prediction box to 1
        - Classification loss: the classification probability is regressed to the one-hot encoding of the true box category
    - 2. There is no anchor frame where the real frame falls (negative example)
        - Objectness loss: return the objectness of the prediction box to 0
    - 3. Only the anchor frame where the real frame falls is given a chance to return to the real frame? The sample is skewed? (non-positive, non-negative)
        - It is not a matched anchor box but its corresponding prediction box and any real box IoU are not less than ignore threshold (0.7), no loss is calculated
- NMS
    - Group by class
    - Sort by score
    - Calculate IOU according to x, y, w, h
- drawbacks:
    - Low recall
    - Location precision need to be improved
    - Poor detection of close target groups \ small object \ occluded object
-

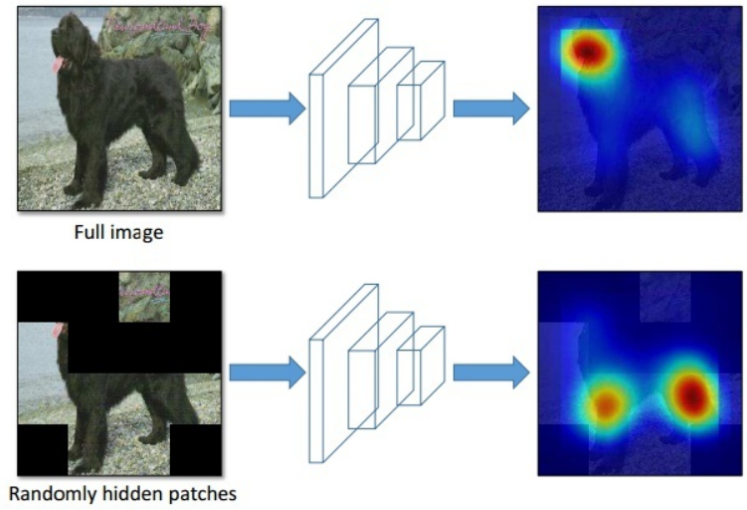


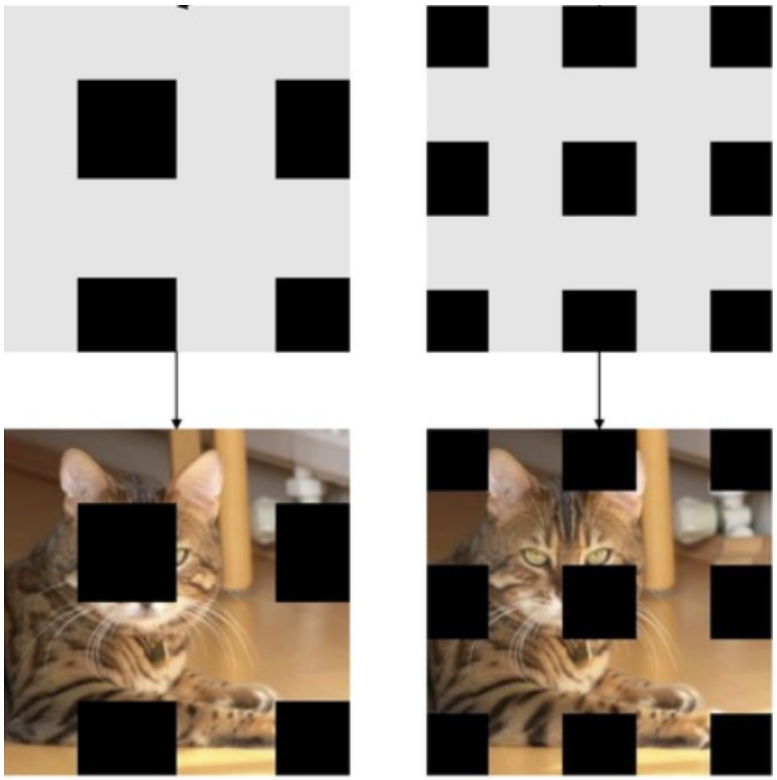
- YOLOV4
    - Data augment
        - Mixup
            - Convex overlay of image pairs and their labels
            -



- CutMix
  - Pasting a clipped portion of another image into an enhanced image The clipping of an image forces the model to learn to make predictions based on a large number of features
  -
- Mosaic
  - combined four training images into one image in a certain proportion, which enabled the model to learn to recognize objects in a smaller range and also helped significantly reduce the need for batch-size
  - 
- Blur
  -
- Crop, Rotation, Flip, Hue,Saturation,Exposure, Aspect.

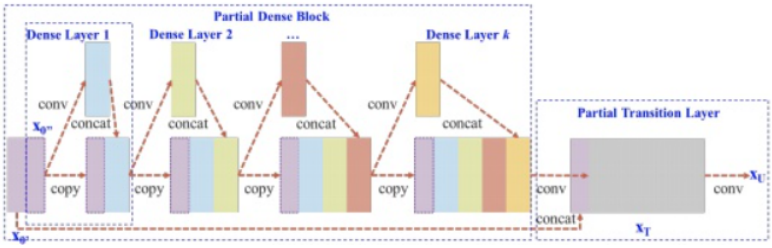
- Random Erase
  - Replace the region of the image with a random value or the average pixel value of the training set
  -
- Cutout
  - Use the clipping block Mask only for input to the first layer of CNN
  -
- Hide and Seek
  - The image is divided into a grid composed of SxS image patches, and some patches are hidden randomly according to the probability Settings, so that the model can learn the appearance of the whole object instead of a single piece, for example, it does not rely on the animal's face alone for recognition
  -



- Grid Mask
  - Hiding areas of the image in the grid also serves to make the model learn the entire composition of the object
  - 
- Self-Adversarial Training
  - Self-adversarial Training is a data enhancement technique that can resist counter attacks to a certain extent.CNN calculates Loss, and then changes the picture information through back propagation to form the illusion that there is no target on the picture, and then carries out normal target detection on the modified image.It should be noted that there is no need to change the network weights during the backpropagation of SAT.
  - Using adversarial generation can improve the weak links in the learning decision boundary and improve the robustness of the model.Therefore, this

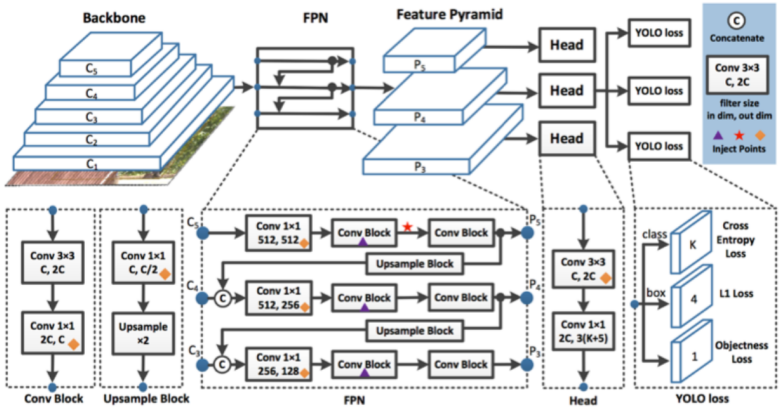
data enhancement method is used by more and more object detection frameworks.

- Label Smooth

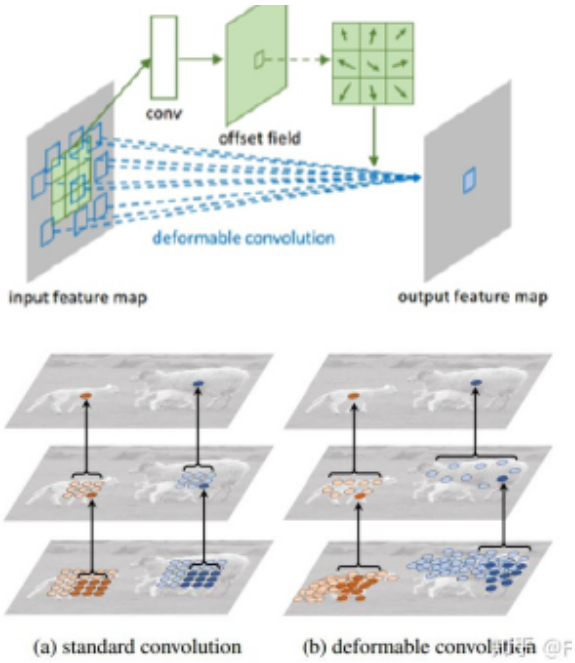
- BackBone: CSPDarknet
  - CSPNet solves the problem of repeated gradient information in network optimization in Backbone of other large convolutional neural network framework. It integrates the gradient changes into the feature graph from beginning to end, thus reducing the number of model parameters and FLOPS, which not only ensures the inference speed and accuracy, but also reduces the model size
  - 

- Neck: PANNet 、 SPP
  - PANET is based on the Mask R-CNN and FPN frameworks, while strengthening information transmission.The feature extractor of the network adopts a new FPN structure with enhanced bottom-up path, which improves the propagation of low-level features.Each stage of the third path takes the feature maps of the previous stage as input and processes them with a 3x3 convolution layer.The outputs are added via horizontal connections to the same phase feature diagrams of the top-down pathway that provide information for the next phase.Meanwhile, Adaptive feature pooling is used to restore the damaged information paths between each candidate area and all feature levels, and aggregate each candidate area on each feature level to avoid arbitrary allocation.
- activation function: Leaky Relu + Mish
- Loss: CIOU Loss \ DIOU_nms

- YOLOV5
  - data augment
    - Zoom, color space adjustment and Mosaic enhancement
    - auto learning bounding box anchors
  - BackBone: CSPDarknet
  - Neck: PANNet
  - Activation function: Leaky Relu + Sigmoid
  - Loss: GIOU Loss \ weighted-nms
- PPYOLO
  -



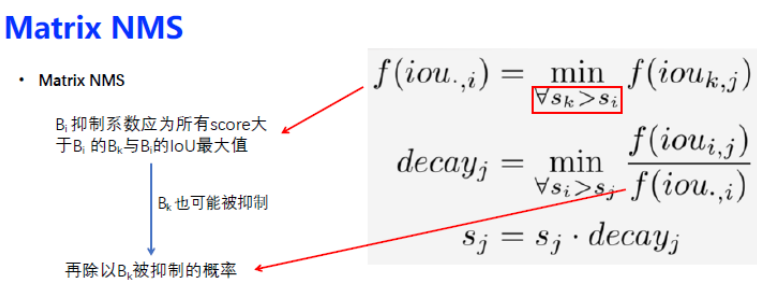
-
  - **数据增强**
    - Image Mixup
    - AutoAugment
    - Image Mosaic
  - **数据采样/正负样本采样**
    - OHEM
    - Class aware Sampling
  - **网络结构**
    - 更优的backbone
    - Deformable Conv
    - Coord Conv
    - SPP
    - DropBlock
  - **学习策略**
    - IoU/GIoU/DIoU/CIoU Loss
    - IoU Aware
    - Label smooth
    - 调整loss权重
    - Grid Sensitive
  - **训练策略**
    - 调整batch size/learning rate/迭代次数等
    - Synchronize batch norm
    - EMA
  - **后处理**
    - Soft NMS
    - Matrix NMS
- Image Mixup
  - Improve the generalization ability of the network to spatial disturbances
  - To a certain extent, increase the proportion of positive samples and increase the recall rate
- Label Smooth
  - One-hot hard tags will learn logit to -∞ or +∞
  - If too much is too much, it is easy to overfit and reduce the classification accuracy
  - A small amount of disturbance of the hard label becomes a soft label
- Synchronized Batch Norm
  - For multi-card parallel scenarios, the batch norm of the general framework is implemented as the independent statistical mean and method of each card
  - Synchronize mean and variance between cards, ‘increased batch size, mean variance is more reasonable, convergence speed and convergence effect are better
- Deformable Conv
  -



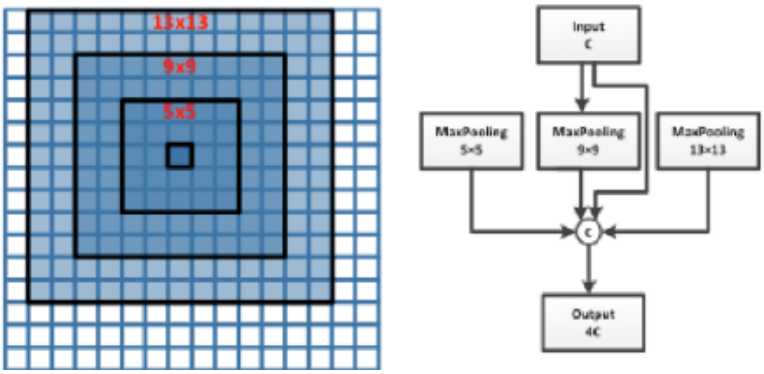
- Optimization point
  - Is it possible to learn the shape of the convolution kernel while learning the weights?
- Optimization
  - Learn an offset to represent the offset of the convolution kernel (v1)
  - Learn another mask to indicate which points should be calculated (v2)

- Drop Block
  - Optimization point
    - Dropout is a common method commonly used in deep learning to reduce overfitting and improve the generalization ability of the network. Is it suitable for detection
    - DropBlock drops an area instead of scattered points each time, which is more suitable for the target detection task of "detecting connected domains"
  - Calculation method
    - Replace random DROP discrete points with random DROP blocks
- Exponential Moving Average
  - $$\begin{aligned} \mathrm{EMA}_0 &= 0 \\ \mathrm{EMA}_t &= \mathrm{decay} * \mathrm{EMA}_{t-1} + (1 - \mathrm{decay}) * \theta_t \end{aligned}$$
  - In the model training process, the historical moving average of the parameters is used to make the training process more smooth and stable
- IoU Aware
  - Optimization point

        - The score of the accurate detection frame is not high, but it is filtered out by NMS
        - score = objectmess * classification Add the positioning accuracy index to the score
    - Calculation
        - Add a channel to learn IoU, that is, the number of output channels increases from B*(5+C) to B*(6+C)
        - When predicting, the predicted value of the objecness channel is updated as follows
        - $$S_{det} = p_i^{\alpha} IoU_i^{(1-\alpha)}$$
- Grid Sensitive
    - When the center coordinates of the real frame fall on the edge of the grid, it will tend to learn from logits to ∞ (same as abel smooth)
    - Grid Sensitive is introduced in the author of YOLOv4, when calculating the center coordinates, an offset and scaling are added
- Matrix NMS
    - NMS directly suppresses prediction boxes with low scores, and false suppression will occur for close objects of the same category.

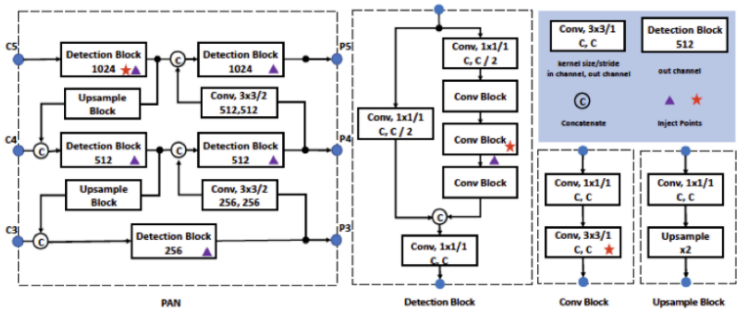


- Coord Conv
    - Convolution only perceives the receptive field corresponding to the convolution kernel, and is it useful to perceive global coordinates?
    - Add two channels, input x, y coordinate information
- SPP
    - multi-scale pooling to extract feature



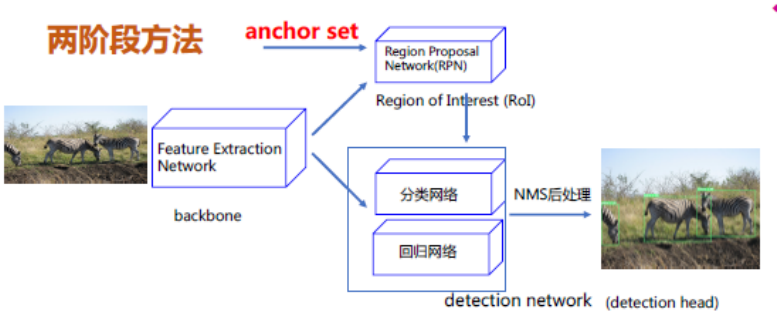
- PP-YOLO优化
  - 骨干网络：ResNet50vd-DCN
  - 训练策略：Larger Batch Size + EMA + DropBlock
  - 损失函数：IoU Loss + IoU Aware
  - 后处理：Grid Sensitive + Matrix NMS
  - 特征提取：SPP + Coord Conv
  - 其他：SSLD pretrain + 2x scheduler

- PPYOLOV2
  - Neck: PAN + mish

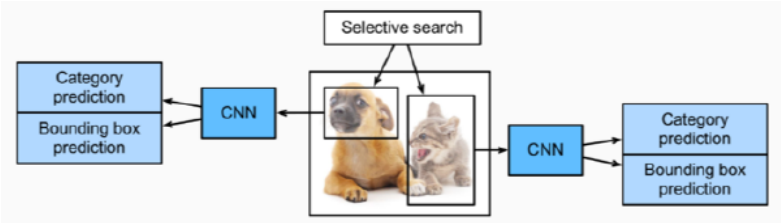


  - IoU Aware Branch

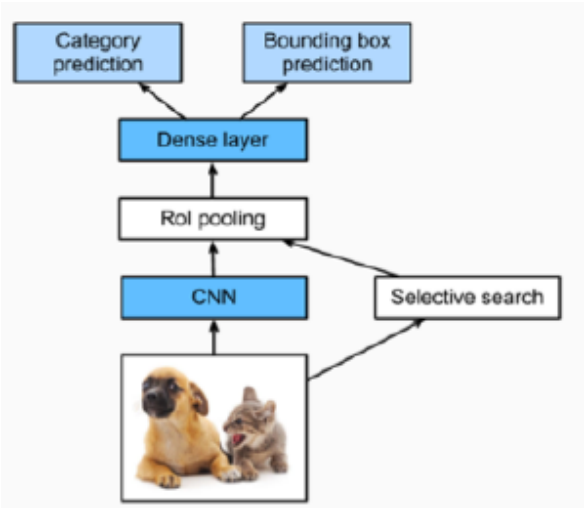
- Two stage

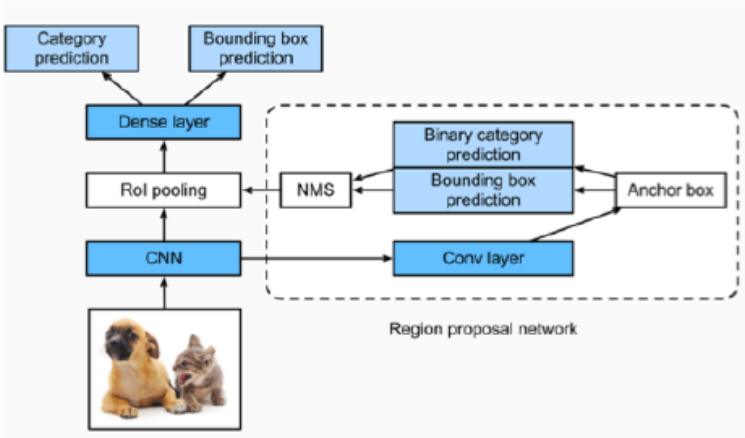


- Use anchors to classify the foreground and background
- Use candidate boxes for further regression and classification
- RCNN: ROI + Classification and location

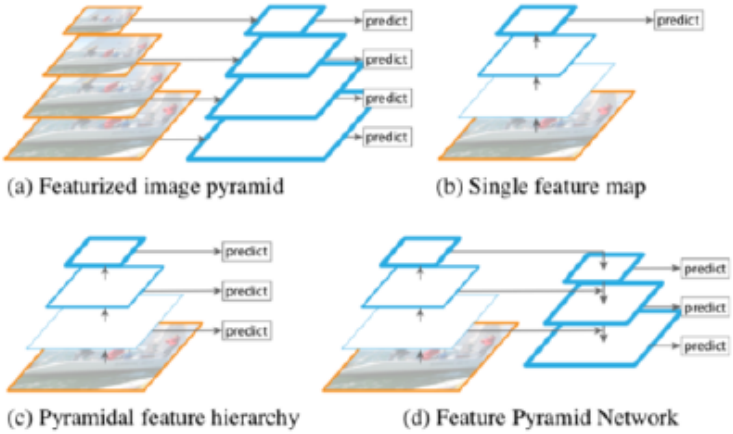


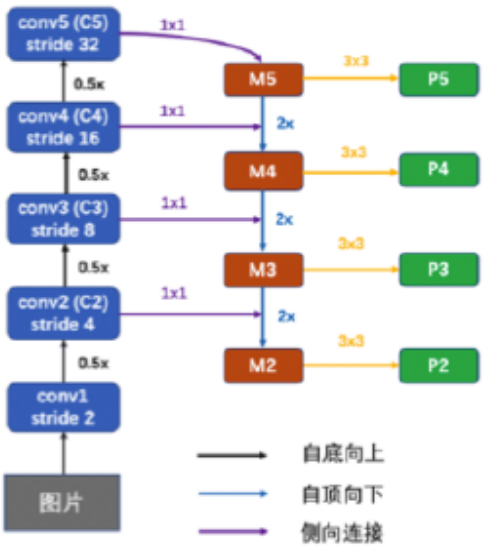
RCNN

- drawbacks
  - All ROI -> CNN train -> massive calculation + time-consuming
  - selective search -> bad quality
  - Feature extraction and support vector machine classifiers are separately trained in modules, Systematic optimization is not combined
- Fast-RCNN: simplify R-CNN
  - 

  - resize -> ROI pooling to deal with the problem to repetetive feature extraction
  - put classification and location together
  - svm + regressor -> softmax+regressor
  - drawbacks:
    - still using selective search -> time-consuming
- Faster-RCNN: RPN to replace selective search
  - 

- RPN to learn ROI
- increase accuracy and decrease time

- FPN: increase the model ability of detecting multi-scale objects
  - (a) Featurized image pyramid: Scale the input image to different scales and use multiple models for prediction
  - (b) Single feature map: Only use the features of the last layer as the input for the subsequent part of the detection model
  - (c) Pyramidal feature hierarchy: predict each level separately
  - (d) Feature Pyramid Network (FPN): fusion and classification prediction of features of different layers
  - 

  - 

  - The input is the output of each layer of the backbone network; the features are up-sampled, and then added to the features of the previous layer to obtain the output

of each layer of the FPN structure. The FPN structure and the backbone network are independent of each other.

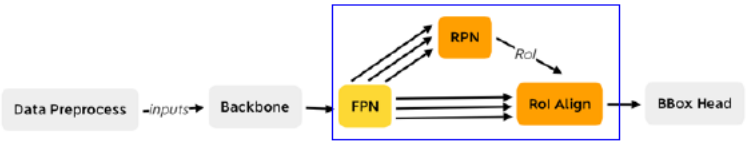
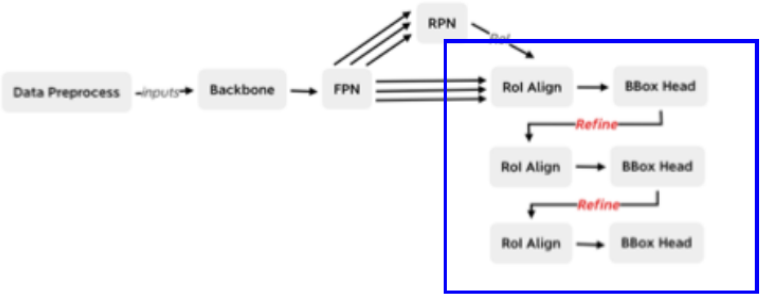
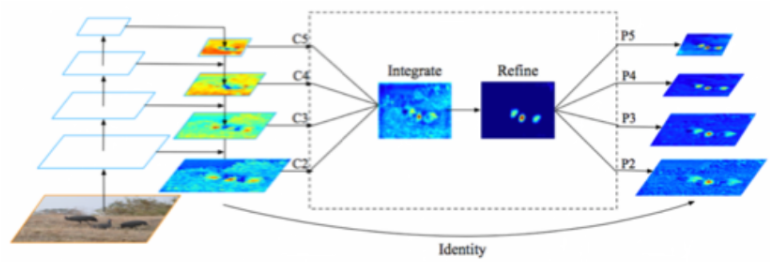
- 

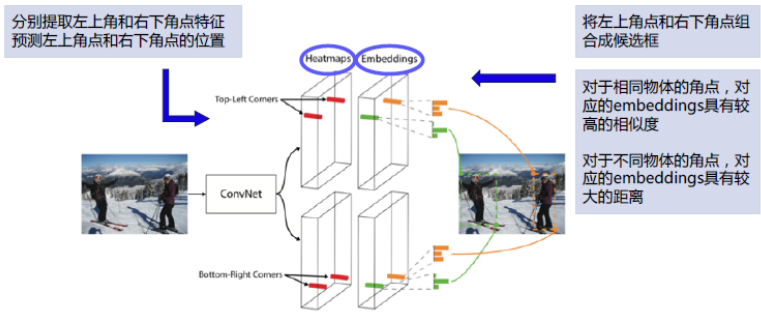
图 Faster R-CNN + FPN网络结构

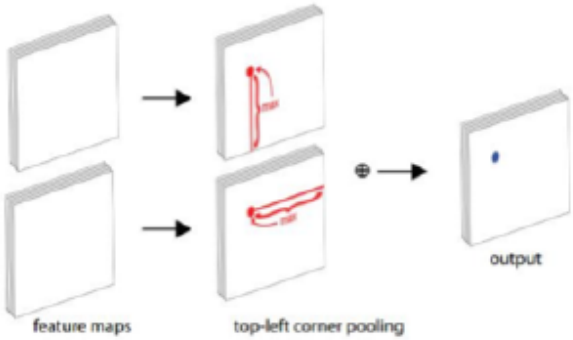
- Cascade R-CNN: improvement of Faster-RCNN -> Fine-tune the box three times, each time the offset of the BBox head and the RoI decoding are used as the RoI input for the next stage
  - 

  - The performance of the trainer is the best when the threshold of RoI itself is close to the threshold used for training by the trainer. -> Introduce multiple heads to adjust RoI
- Libra R-CNN
  - Imbalance problems in detection:
    - Feature level >How can the extracted features of different levels be really fully utilized?> FPN feature fusion
      - 

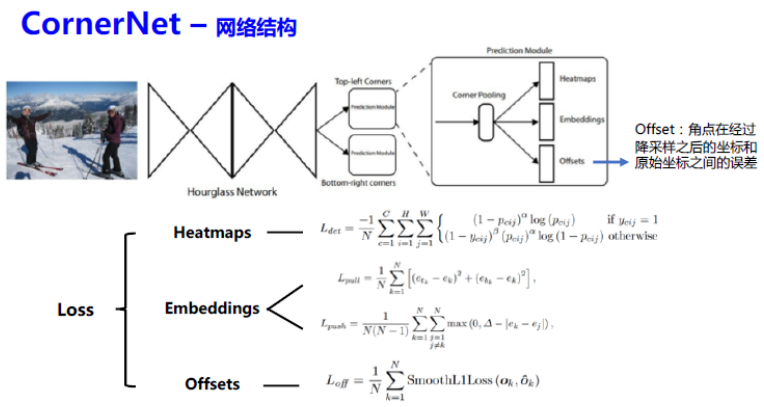
        - Rescale
          - Unify the feature maps of different levels to the C4 layer by difference or down-sampling
        - Integrate
          - Combine the collected features
        - Refine
          - Use non-local structure to further strengthen the fusion feature

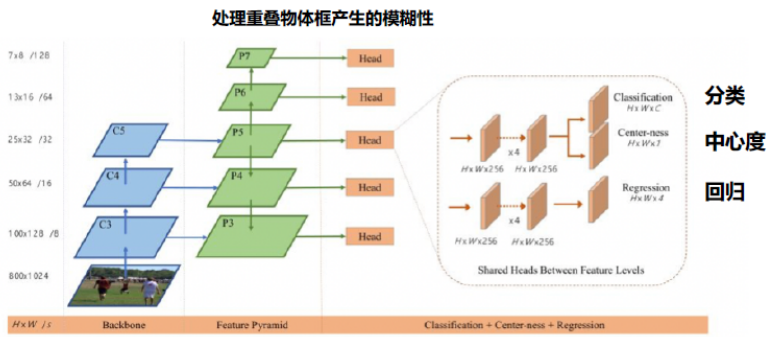
            - Strengthen
                - Sum the optimized features with the original features at different levels
    - Sample level> Is the sampled candidate area representative?> Sampling strategy in RPN
        - Positive Sample
            - Calculate the number of samples required for each category
            - Sample each category separately
        - Negative Sample
            - Divide the negative sample into two parts according to IoU
            - Samples larger than the threshold are divided into buckets according to IoU, and the number of samples that should fall in each bucket is calculated, and finally a negative sample with uniform IoU distribution is obtained.
            - Random sampling of samples below the threshold
    - Objective level >Can the currently designed loss function guide the target detector to better converge?> Loss
        - A sample with a larger Loss is considered a difficult sample, and a sample with a smaller Loss is an easy sample
        - The gradient corresponding to the difficult sample in Smooth L1 is larger than the gradient of the easy sample, resulting in an imbalance in the learning ability of different samples
        - In Balance L1, the gradient at the boundary between difficult samples and easy samples is smoother
- Anchor drawbacks
    - manually design
        - how many? how big? what is the ratio of width and height?
        - Hyperparameter
        - too many? how to deal with imbalance data?
- Anchor Free
    - base on center area: FCOS \ CenterNet \TTFNet
    - base on multiple key points Joint expression: CornerNet\Repoints
    - CornerNet
        - use a pair of corner point



- find corner: Reduce penalty + Corner Pooling
  - 

  - 1. Do max pooling from top to bottom
  - 2. Do max pooling from right to left
  - 3. Then merge (add)
- Desired result:
  - The similarity of one-diagonal points of the same target is high
  - The distance between a diagonal point of different targets is large
- Embeddings?
  - Encode each point on the input feature map, and the encoded vector obtained is embeddings, which represents the position information of the point.
  - CornerNet uses one-dimensional embeddings features, that is, the output dimension of the Embeddings module is (N, 1, H, W)
  - The embeddings distance between the upper left corner point and the lower right corner point is taken as the similarity of the two corner points.
-

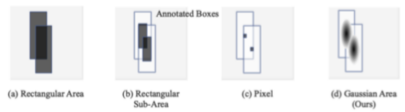


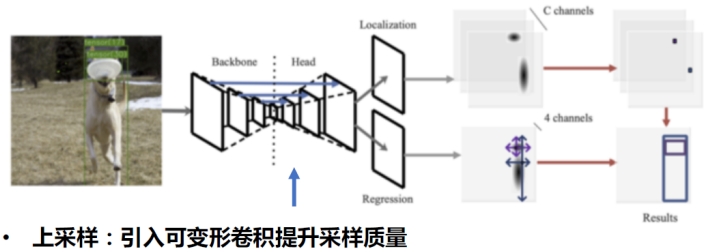
- CornerNet Squeeze
  - Reduce the processing of each pixel
  - Combine the design ideas of SqueezeNet and MobileNet
- FCOS
  - 

  - Loss : Focal loss(classification) + IoU loss(regression)
  - Core idea:
    - The distance from the detection point and the point to the four sides of the target frame
  - Classification + regression branch
    - The points in the real box are all positive samples
    - Determine the level by the maximum distance from the point to the four sides
  - Centrality branch;
    - Remove low-quality prediction boxes that are far from the center point
    - The farther the pixel point is from the center point, the smaller the center degree
- CenterNet
  - The detection problem is transformed into a key point problem, and the anchor is removed
  - There is only one positive sample per location, and NMS is no longer used
    - extract the max on heatmap
  - Directly predict the size of the target frame

    - Can be extended to other tasks
    - time-consuming
- TTFNet
    - increase LR
    - decrease data preprocessing
    - accelerate convergence speed
    - improve the quality of surveillance signals
    -

	CenterNet	TTFNet
中心点定位	中心点周围采用半径为r的高斯核产生负样本监督信号	中心点周围采用椭圆高斯核产生负样本监督信号，椭圆的长径短径与真实框大小成比例
尺寸回归	仅在中心点处产生真实框大小的监督信息	中心点周围采用椭圆高斯核产生采样区域，区域内的点均产生回归分支监督信息
Offset	弥补降采样带来的坐标误差	由于回归分支预测了点到真实框四边距离，去除offset分支





- **上采样：引入可变形卷积提升采样质量**
- **定位分支：采用Focal loss**
- **回归分支：采用GIoU loss，配合采样权重**

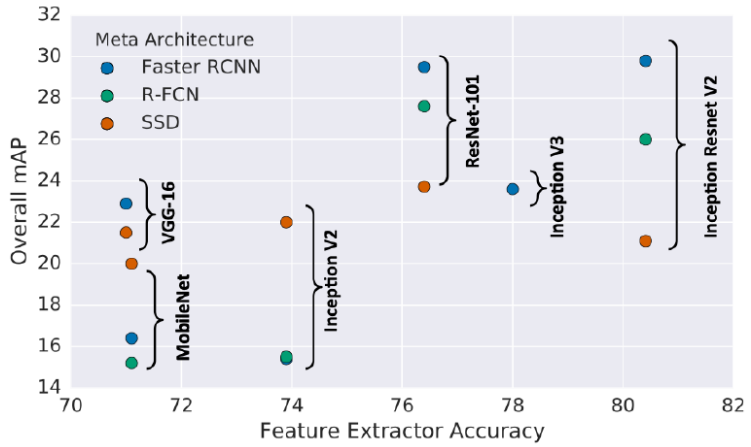
$$W_{ij} = \begin{cases} \log(a_m) \times \dfrac{G_m(i,j)}{\sum_{(x,y)\in A_m} G_m(x,y)} & (i,j) \in A_m \\ 0 & (i,j) \notin A \end{cases}$$

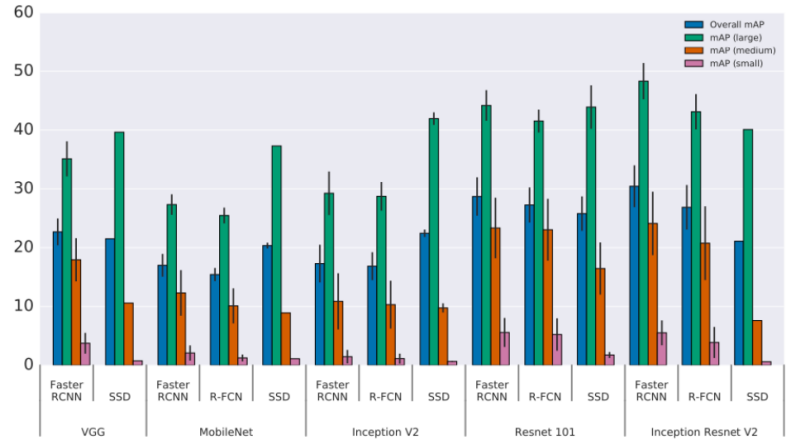
-

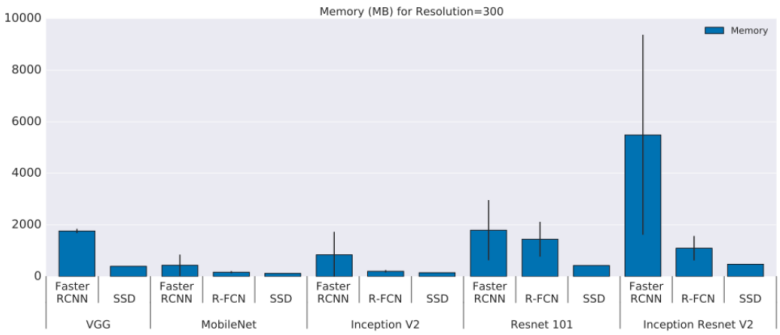
模型	检测机制	优点	缺点	使用范围
CornerNet	一对角点	corner pooling, embeddings 提高模型精度	角点分组难度和计算量较大 对边缘敏感，忽略内部信息	多目标检测
FCOS	点+点到四条边距离	多尺度预测解决目标模糊性 设计复杂度低	多尺度监督信息需要增强 中心度可解释性需要增强	可用于实例分割
CenterNet	一个中心点+长宽	简单高效，没有NMS后处理	只使用中心点回归，可获得监督信息较少	可用于2D, 3D目标检测及人体姿态识别

- Compare

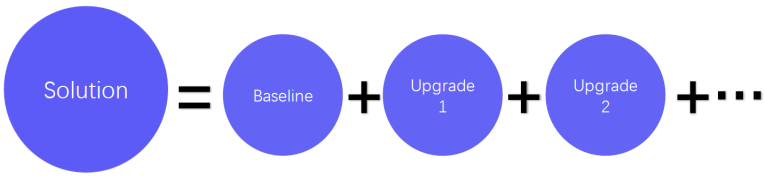
	Anchor-Based单阶段	Anchor-Based两阶段	Anchor-Free
网络结构	简单	复杂	简单
精度	优	更优	较优
预测速度	快	稍慢	快
超参数	较多	多	相对少
扩展性	一般	一般	较好

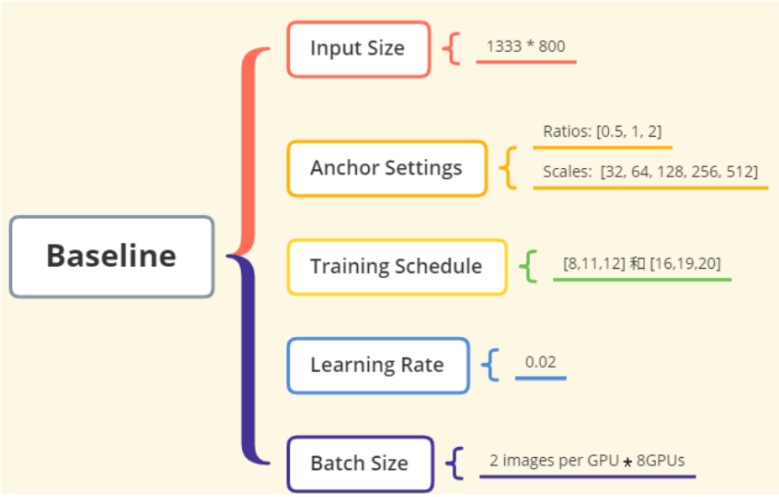






- Solution Pipeline





- Flow
  - 1. Analysis and evaluation indicators
    - The indicator used in the general detection field is mAP, what needs to be understood
    - The result will increase mAP, what result will decrease mAP.
  - 2. Visualize the results of the Baseline
    - Take a look at the errors of the Baseline.
  - 3. Classify errors
    - Under what circumstances will these problems occur.
  - 4. Research paper
    - To find solutions to these problems.
- Input: Image, Patches, Image Pyramid
- Basic Network
  - detection algorithm: one-stage\two-stage\Anchor Free
  - Backbone Network: ResNet \ DarkNet
  - Neck: FPN, PANet, Bi-FPN
  - Head:
    - Dense Prediction: RPN, YOLO, SSD, RetinaNet, FCOS
    - Sparse Prediction: Faster R-CNN, R-FCN
  - Loss Function: FocalLoss\ IoU Loss\ GIOU Loss
- Network Improvement
  - training tricks
    - Data augment
    - LR adjust strategy
    - Hyperparameter training
  - Model deploy
    - Compress

        - Prune
        - Quantify
        - Distilling
- Upgrade example
    - Small goal problem -> Adjust the network size
        - The resolution of the single picture is very high, and the screen size is a quarter of the original picture, 4096*3000. Under high resolution, the size of the defect is very small.
        - A lower size will miss many targets, and a higher size will increase the amount of calculation.
    - Multi-scale problem -> FPN
        - There is a big difference in the size of product defects. Some scratches are 10mm in length, and small scratches are 0.3mm. How to balance the size is a problem.
        - Fusion of feature maps of different layers, which can effectively enhance the performance of the model at multiple scales