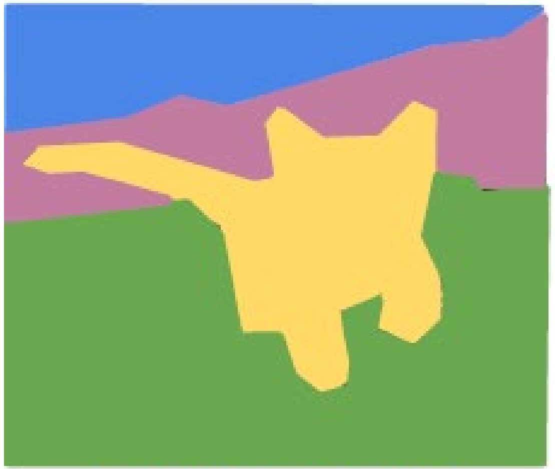
目标检测与分割

Object Detection and Segmentation

胡浩基
浙江大学信息与电子工程学院
haoji_hu@zju.edu.cn

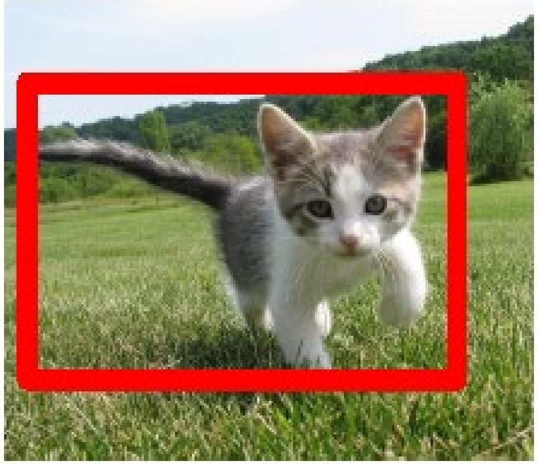
目标检测与分割

语义分割（Semantic Segmentation）



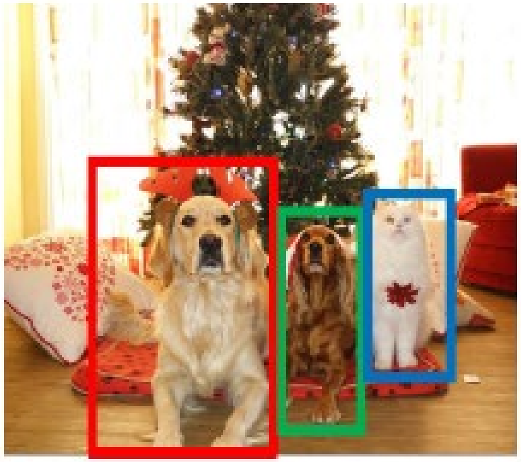
GRASS, CAT, TREE, SKY

目标定位与识别（Classification and Localization）



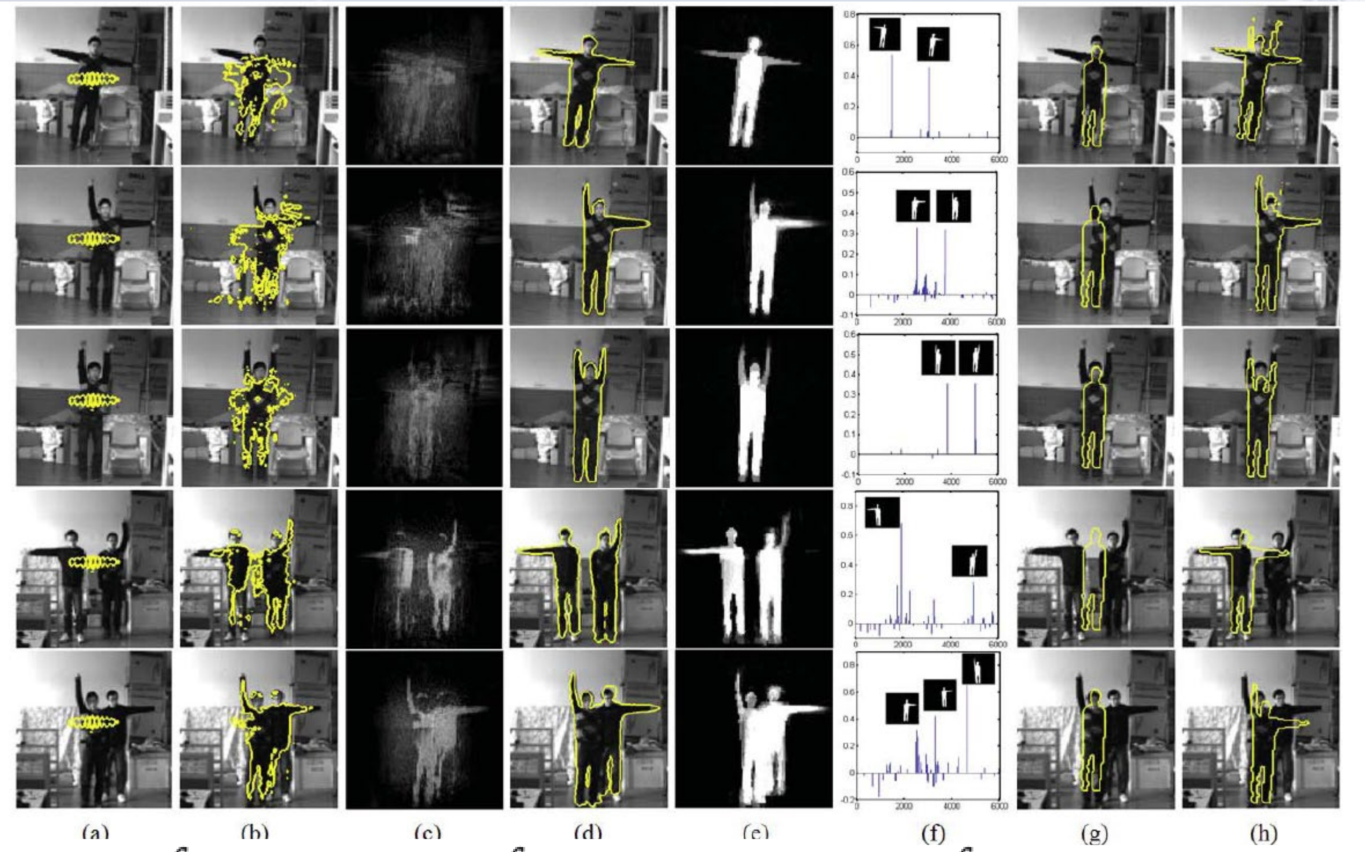
CAT

目标检测（Object Detection）



DOG, DOG, CAT

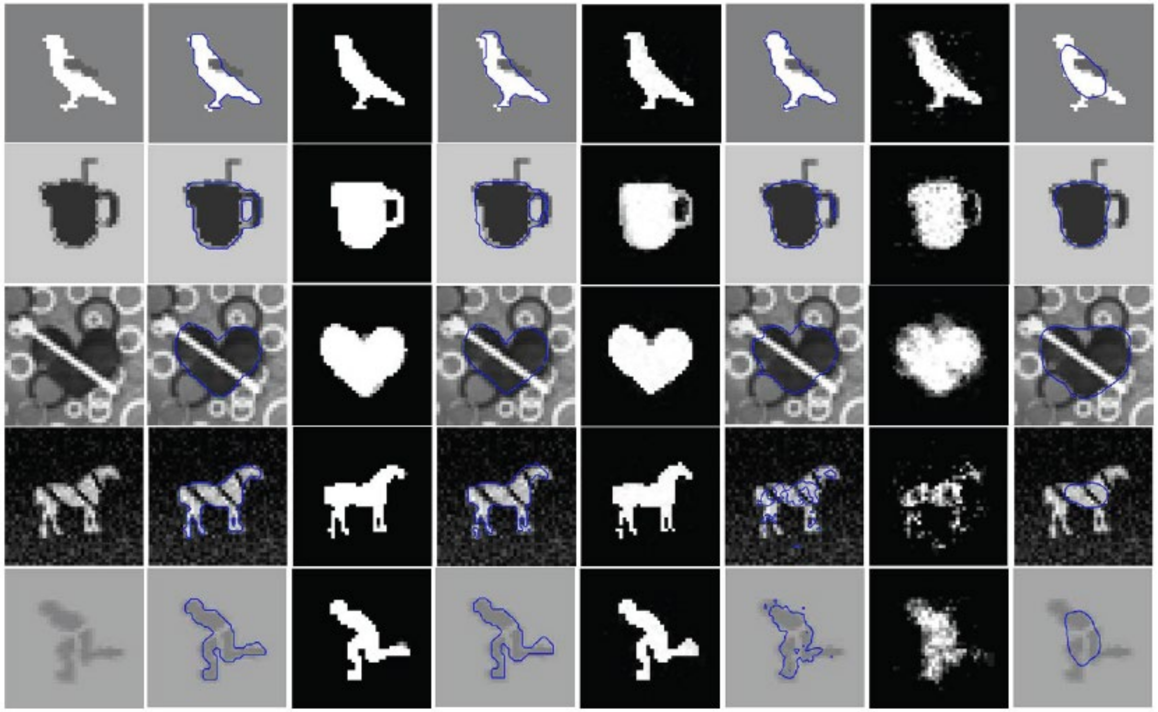
基于形状稀疏的图像分割

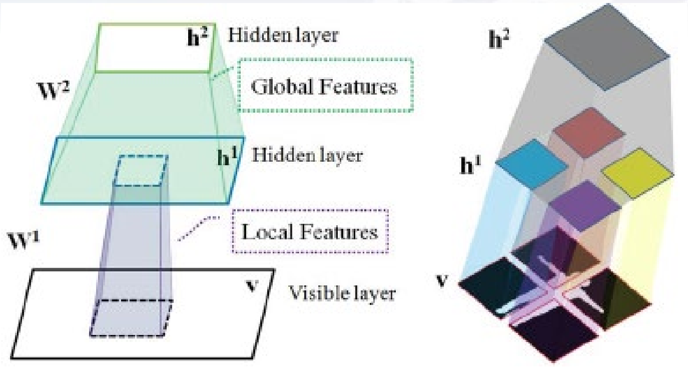


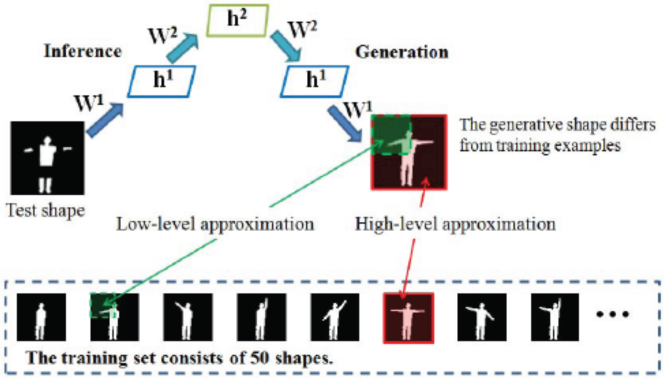
$$E_i(\mathbf{q}) = \int_\Omega \mathbf{r}_o(x)\mathbf{q}(x)dx + \int_\Omega \mathbf{r}_b(x)(1-\mathbf{q}(x))dx + \int_\Omega \mathbf{r}_e(x)|\nabla \mathbf{q}(x)|dx$$

Fei Chen, Huimin Yu, Roland Hu: Shape Sparse Representation for Joint Object Classification and Segmentation. IEEE Transactions on Image Processing 22(3): 992-1004 (2013)

深度学习的图像分割方法

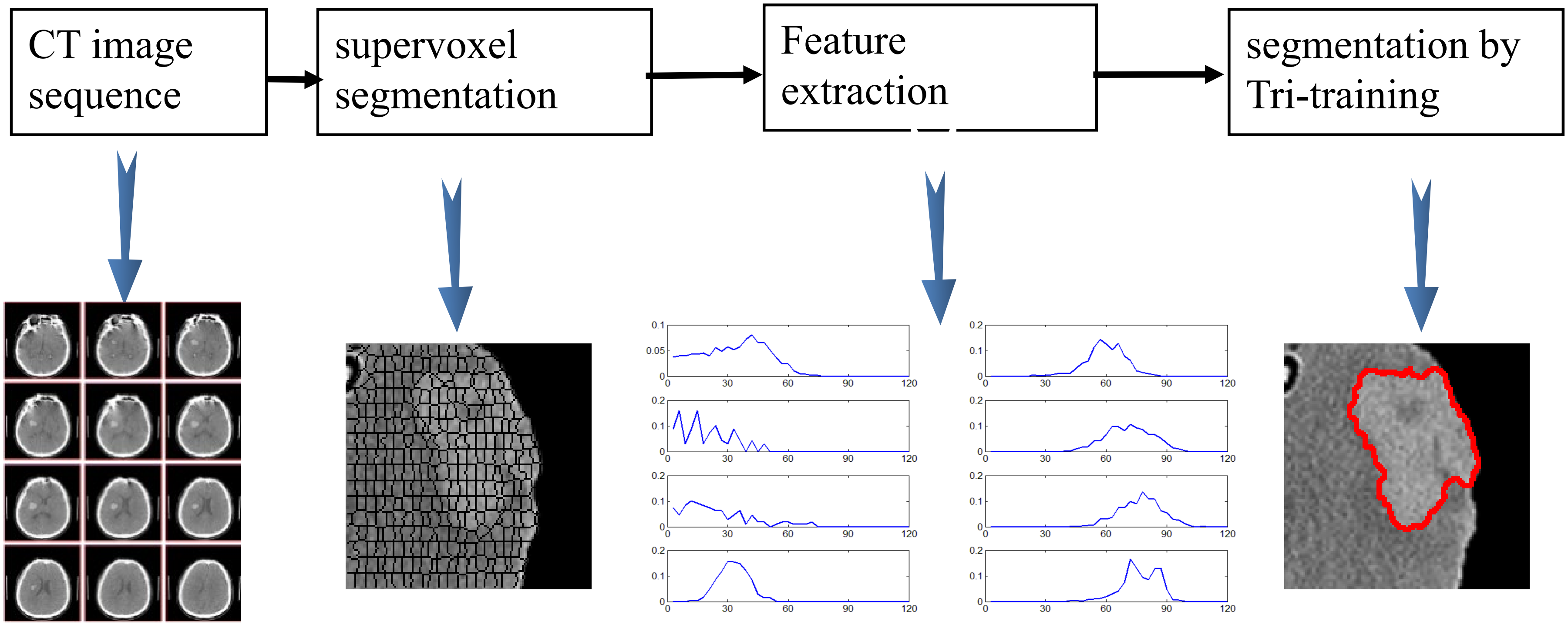





Inference
W²
h²
W²
Generation
h¹
h¹
W¹
W¹
The generative shape differs from training examples
Test shape
Low-level approximation
High-level approximation
The training set consists of 50 shapes.

Fei Chen, Huimin Yu, Roland Hu, Xunxun Zeng: Deep learning shape priors for object segmentation. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, CVPR 2013.

目标检测与分割



CT image sequence

supervoxel segmentation

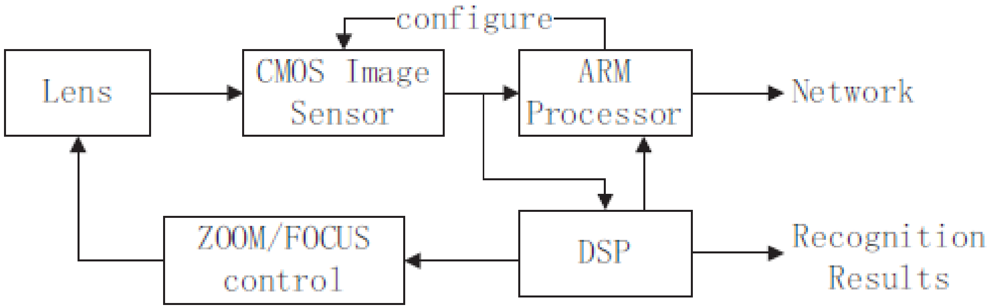
Feature extraction

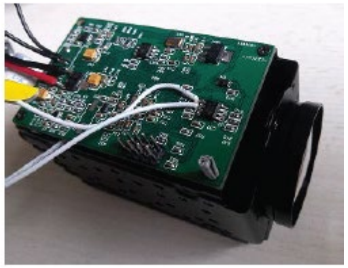
segmentation by Tri-training

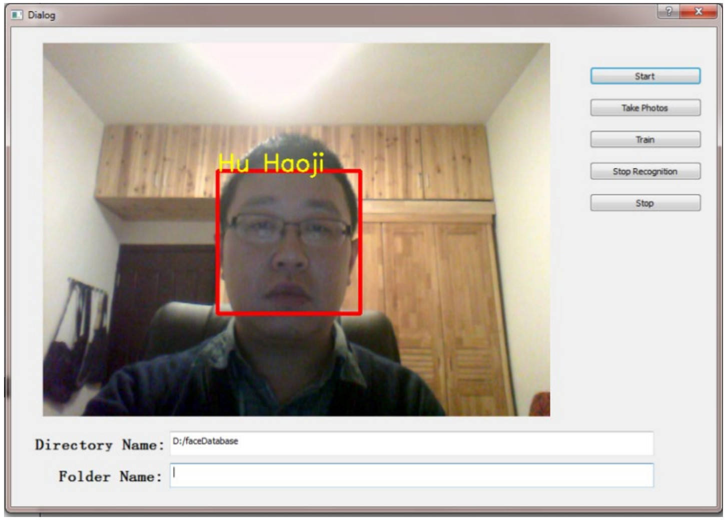
Mingjie Sun, Roland Hu, Huimin Yu, Bao Zhao, Huipeng Ren, Tri-Training for Semi-Supervised Segmentation of Intracranial Hemorrhages on Brain CT Images, VCIP 2016 (submitted).

目标检测与分割

- 人脸检测与识别硬件系统

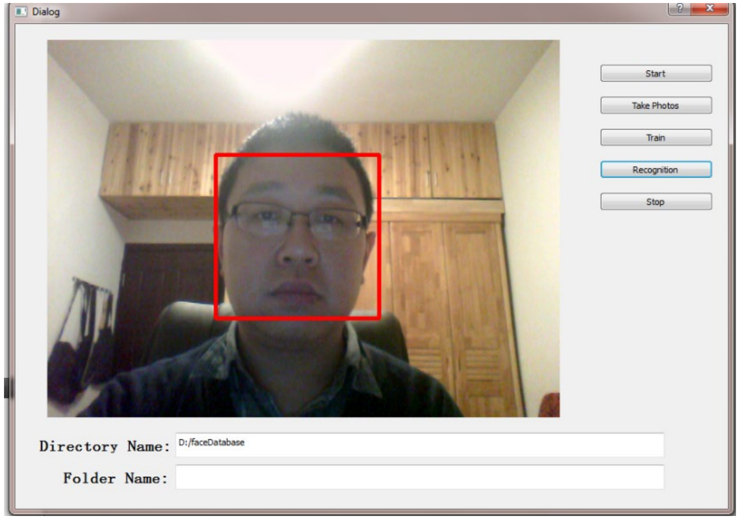






Directory Name: D:/faceDatabase

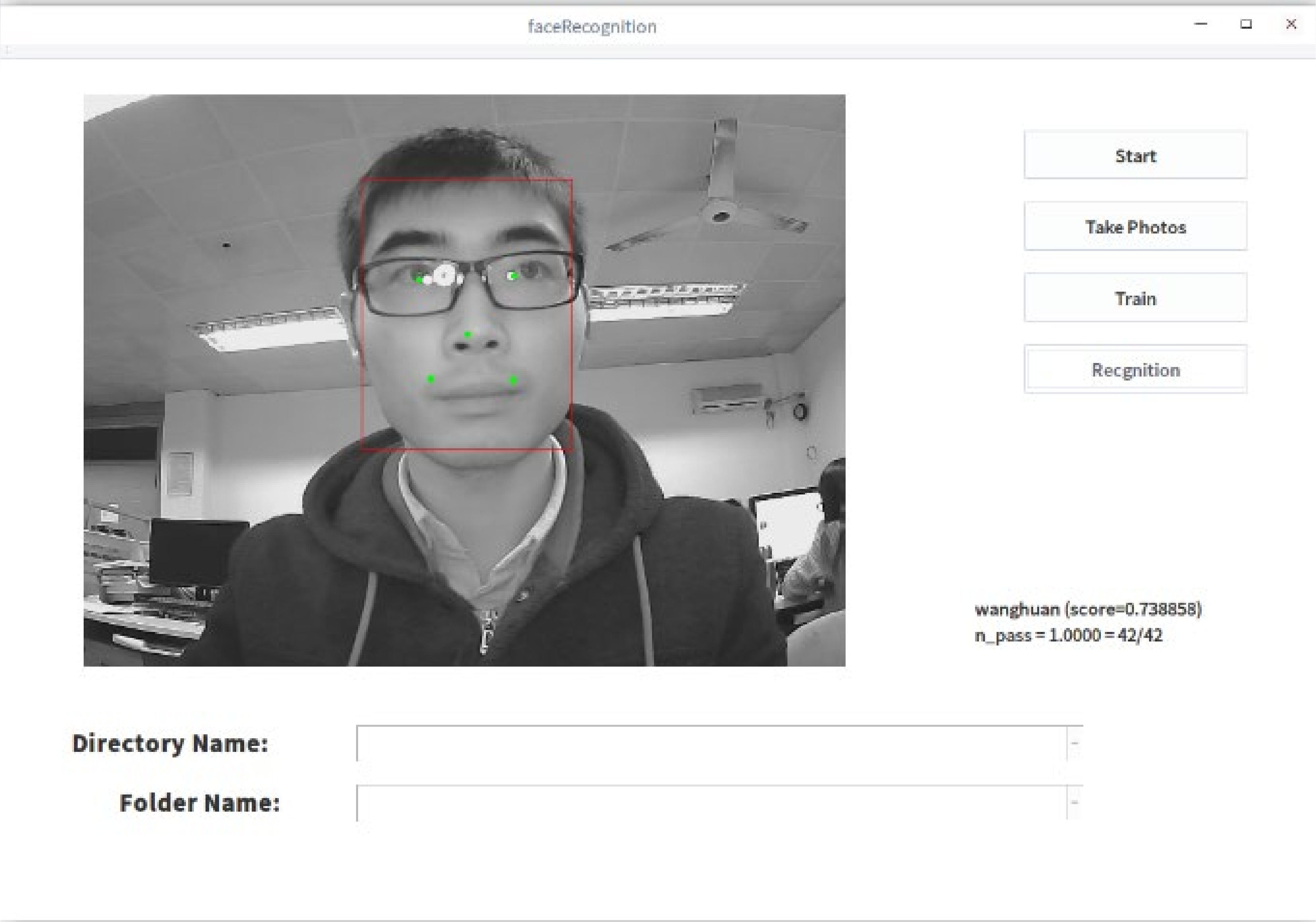
Folder Name:



Directory Name: D:/faceDatabase

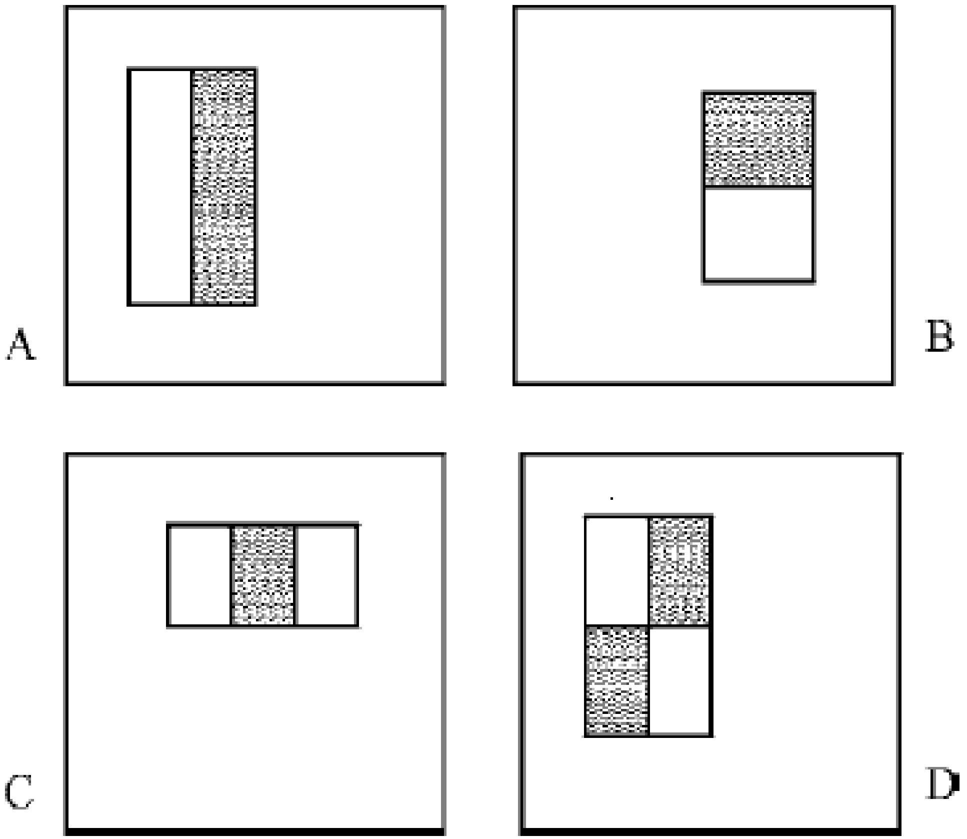
Folder Name:

目标检测与分割



```
wanghuan (score=0.738858)
n_pass = 1.0000 = 42/42
```

目标检测（AdaBoost）



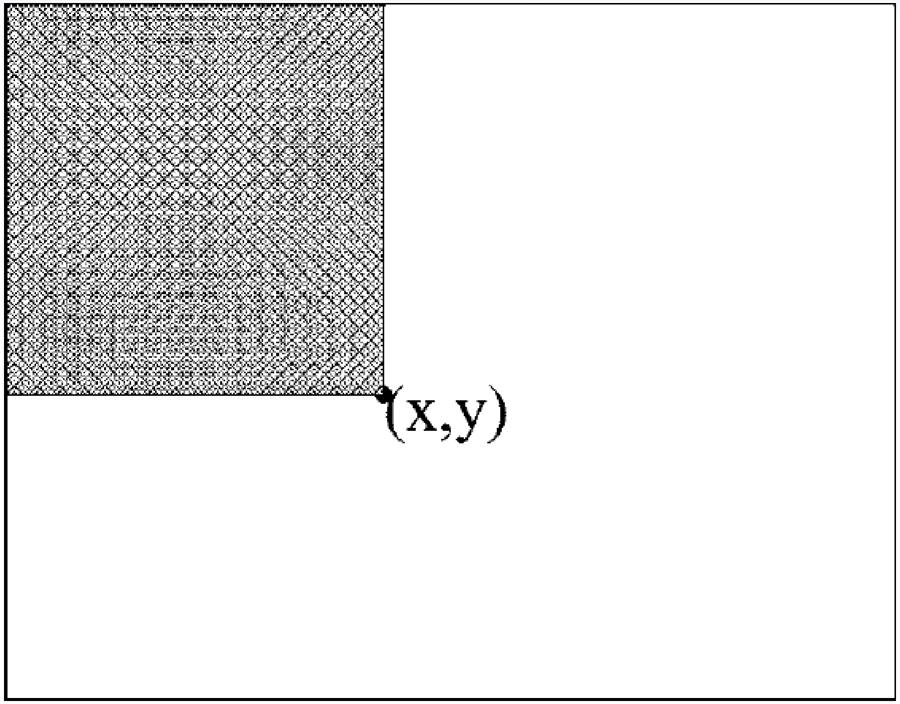
Haar-like Features

（1）白色区域的像素值，减去黑色区域的像素值。
（2）每一个FEATURE所有区域长度和宽度一致。
（3）FEATURE可以在整幅图上平移，只需要满足（1）和（2）即可。
（4）可以取左图这四种形式。

对于一个24*24的图像，所有的Haar-like feature 个数为20万左右。

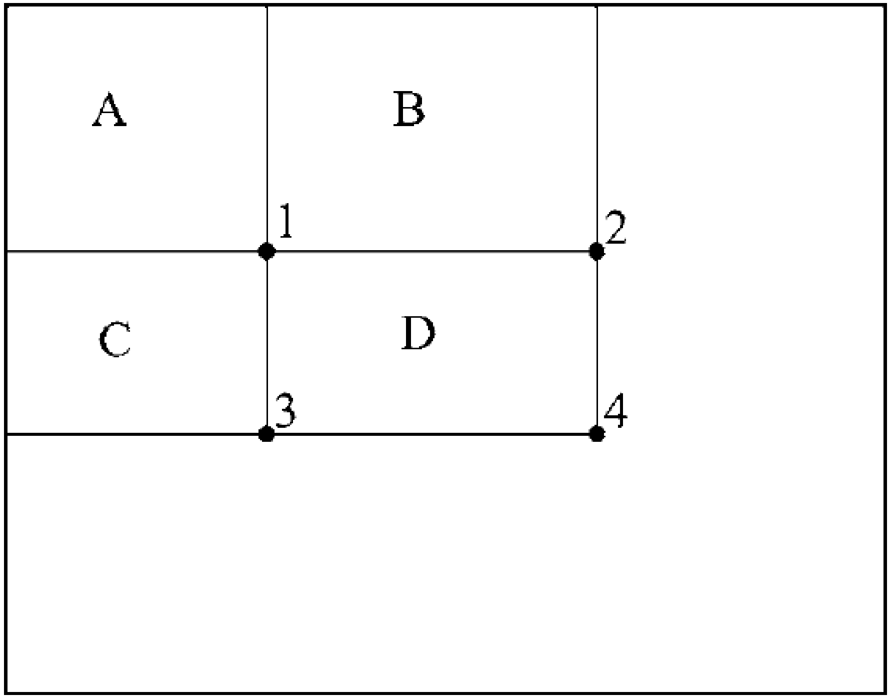
Viola and Jone, Robust Real-Time Face Detection, International Journal of Computer Vision 57(2), 137–154, 2004

目标检测（AdaBoost）



Integral(x,y) = $\sum_{u \le x} \sum_{v \le y} Image(u, v)$

目标检测（AdaBoost）



Image(D) =Integral(4)+Integral(1)-Integral(3)-Integral(2)

目标检测（AdaBoost）

分类器构造：取一些人脸（6000张左右）和一些非人脸（7万张）作为训练样本。

$$h(x, f, p, \theta) = \begin{cases} 1 & \text{if } pf(x) < p\theta \\ 0 & \text{otherwise} \end{cases}$$

x: 图像。

f: 某一个Haar-like feature。

θ: 阈值。

p: POLARITY，只能取+1或-1。

对某个特定的 f,求最佳的 θ, p 取值，使h在训练样本上识别率最高。

目标检测（AdaBoost）

分类器构造：取一些人脸（6000张左右）和一些非人脸（7万张）作为训练样本。

$$h(x, f, p, \theta) = \begin{cases} 1 & \text{if } pf(x) < p\theta \\ 0 & \text{otherwise} \end{cases}$$

x: 图像。

f: 某一个Haar-like feature。

θ: 阈值。

p: POLARITY， 只能取+1或-1。

对某个特定的 f, 求最佳的 θ, p 取值， 使h在训练样本上识别率最高。

ADABOOST算法流程

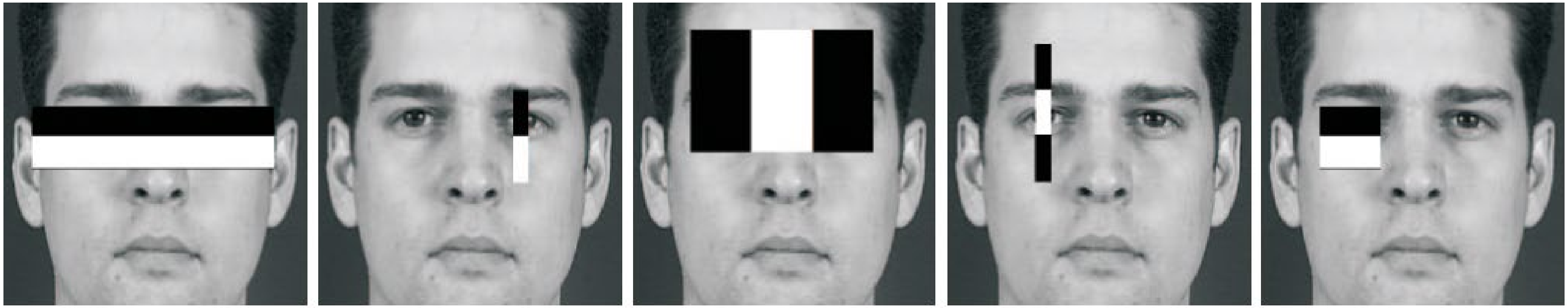
1. 首先在数据集D中选取正确率最高的特征，用F1表示。
2. 将数据集D分为两类，{F1分对的数据}和{F1分错的数据}。
3. 以较大概率取F1分错的数据，以较小概率去F1分对的数据，形成新的集合D2。
4. 在D2中选取正确率最高的特征，用F2表示。

ADABOOST算法流程

5. 将D分为：　{F1、F2都分对的数据}，{F1分对而F2分错的数据，以及F1分错而F2分对的数据}，{F1，F2都分错的数据}。
6. 以最大概率取{F1，F2都分错的数据}，以次大概率取{F1分对而F2分错的数据，以及F1分错而F2分对的数据}，　以最小概率取{F1、F2都分对的数据}，得到数据集D3.
7. 在D3中选取正确率最高的特征，用F3表示。循环，以此类推。
8. 　用各个特征的线性组合构建分类器。

目标检测（AdaBoost）

- 最能表征人脸的Haar 特征：

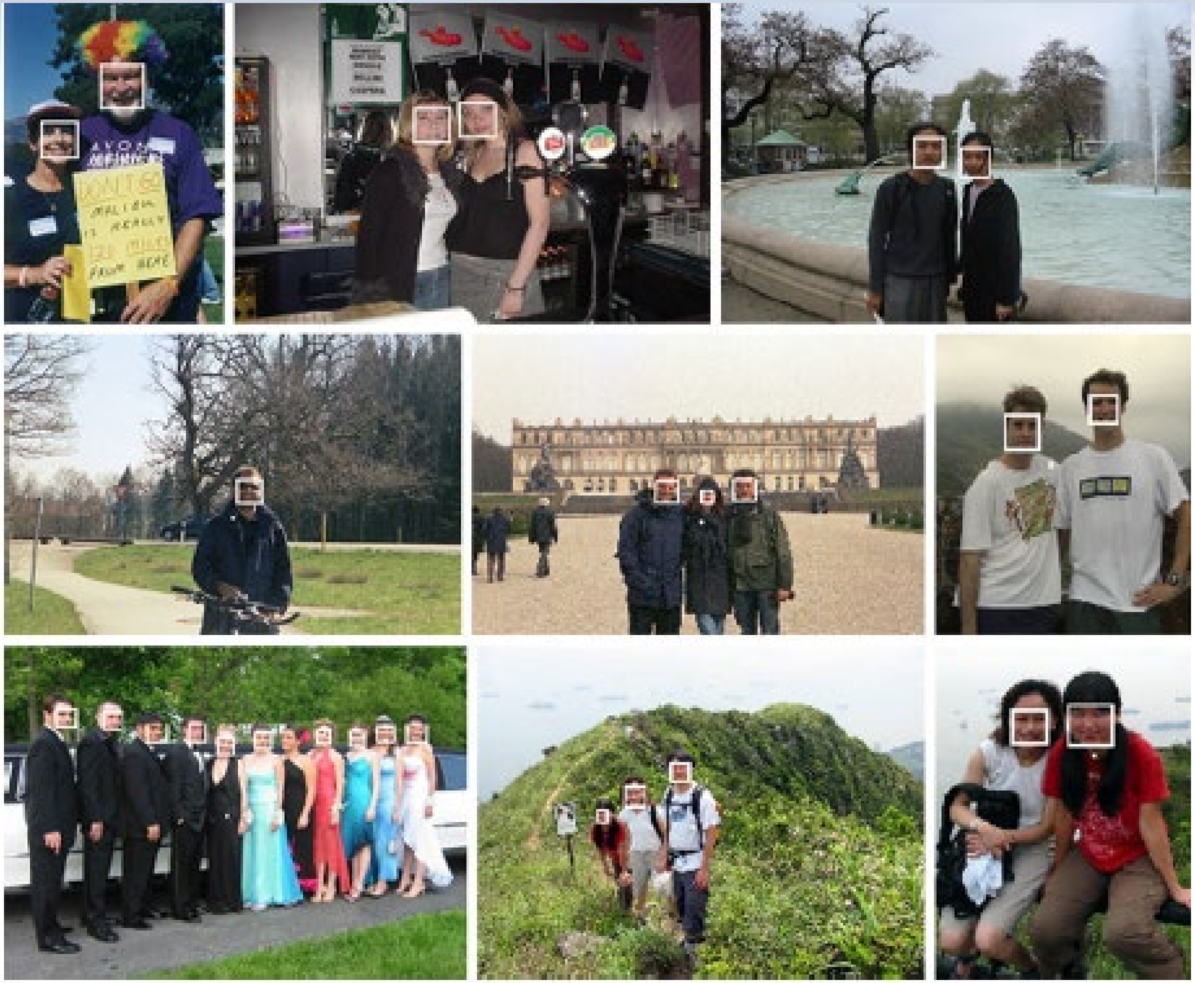


目标检测（AdaBoost）

AdaBoost人脸检测流程：

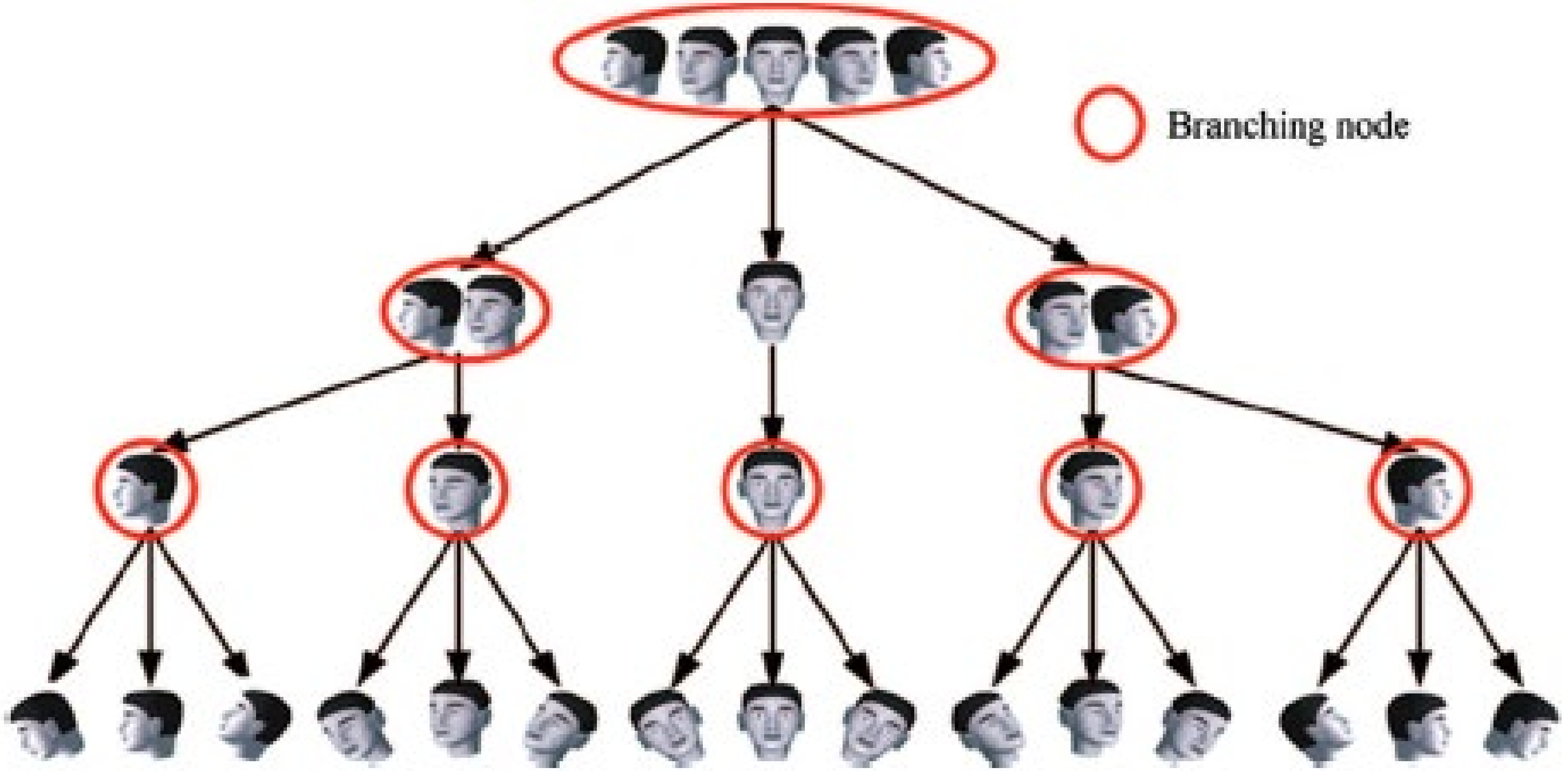
1. 在图像中，对每一个24*24的格子遍历使用分类器，如果是人脸，则输出。
2. 将图像缩小，长宽同时除以1.2，在用分类器遍历每一个24*24的格子。如果是人脸，将该处位置坐标乘以1.2， 等比例放大到原图。
3. 重复2，直到图像长或宽小于24个像素为止。

目标检测（AdaBoost）



检测效果

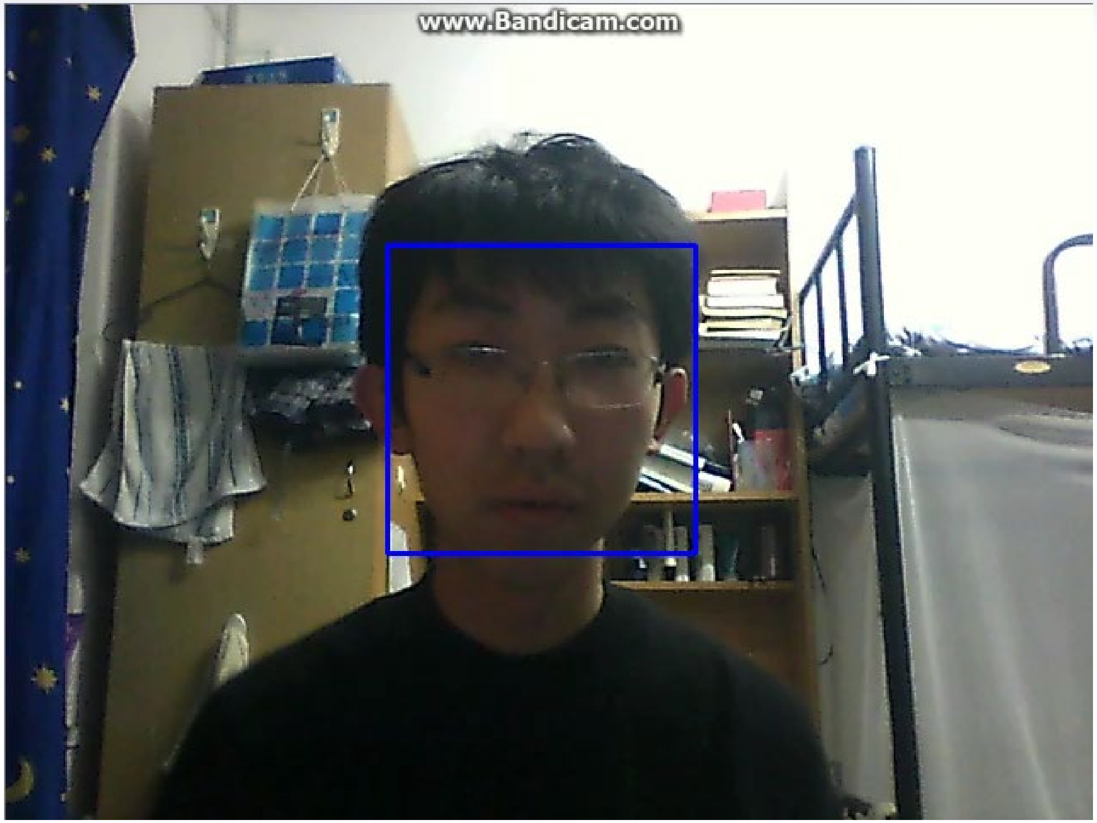
目标检测（AdaBoost）



多个姿态人脸都可以构建AdaBoost分类器，树状级联后可以获得多姿态人脸检测器

Huang et al. High-performance rotation invariant Multiview face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(4), 671-686 (2007)

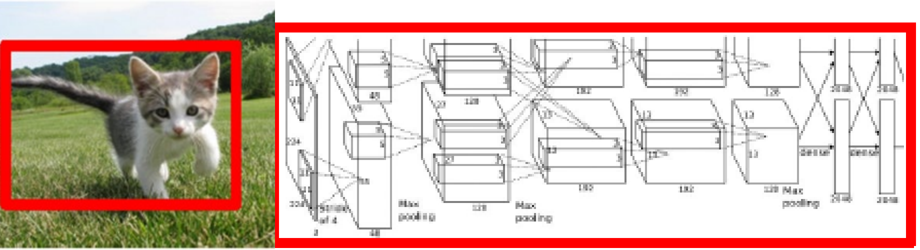
目标检测（AdaBoost）



AdaBoost 人脸检测展示：OPENCV中的例程程序

Classification and Localization

Classification + Localization



CAT

Often pretrained on ImageNet
(Transfer learning)

Vector:
4096

Fully Connected:
4096 to 1000

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

Correct label:
Cat

Softmax Loss

+ → **Loss**

Fully Connected:
4096 to 4

Box Coordinates
(x, y, w, h)

L2 Loss

Correct box:
(x', y', w', h')

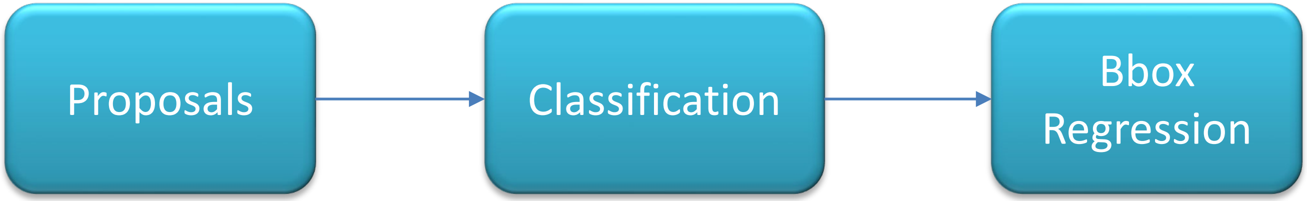
Treat localization as a regression problem!

目标检测（RCNN）

如何将卷积神经网络（CNN）用在目标检测上？

主要问题:

用大大小小的方框遍历所有图像不现实，如何快速挑出可能有物体的区域（Region of Interest, ROI）。我们需要一个计算量不那么大的算法，提出ROI的候选区域（Region of Proposals, or Proposals）



目标检测（RCNN）

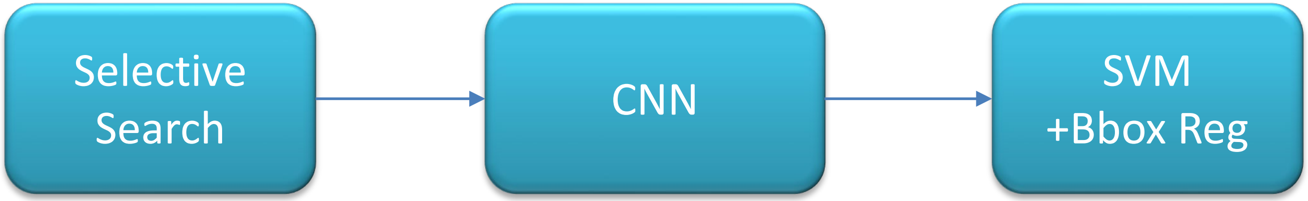
2014-R-CNN, a naïve deep detection model

Basic Ideas:

1. Use selective search to generate proposals
2. Scale and resize proposals to fit the CNN
3. SVM for final decisions

Main Problems:

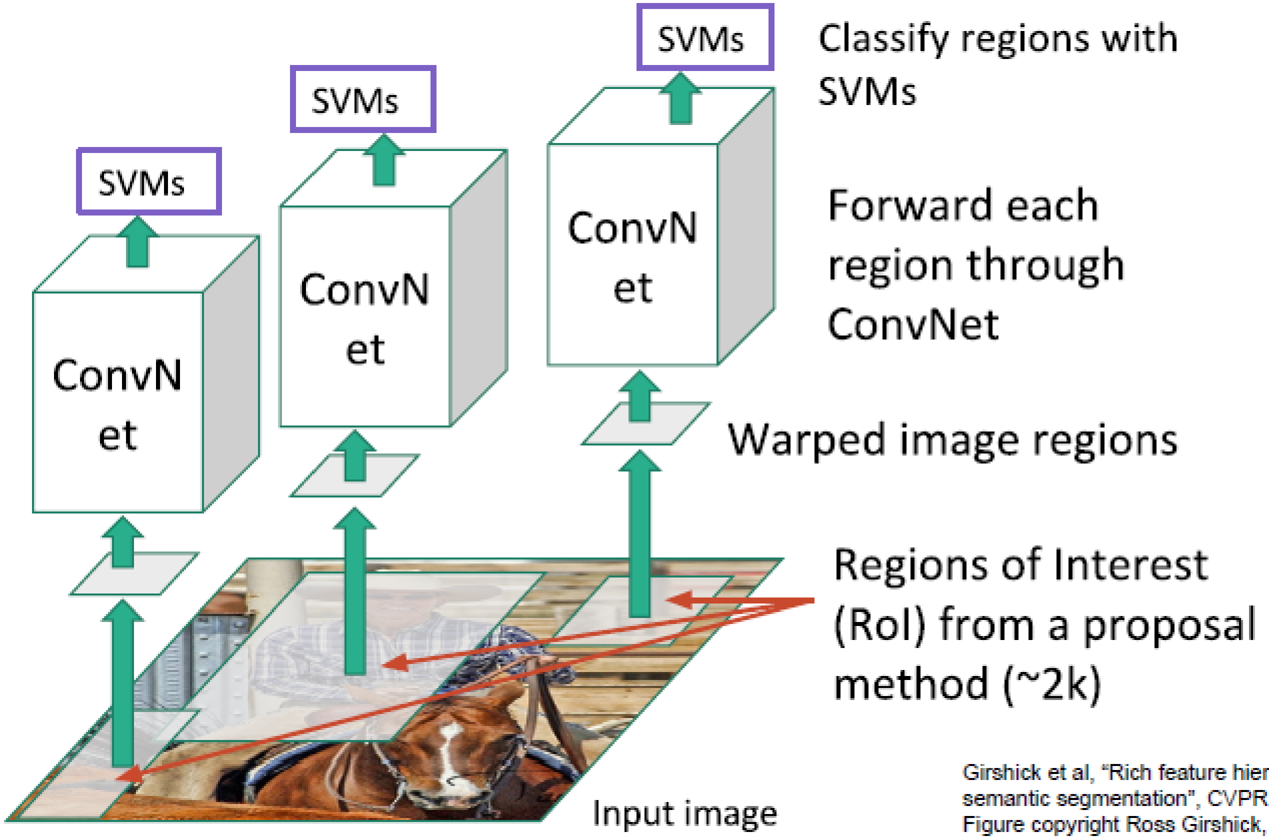
1. High cost to perform Selective Search (~5s per image)
2. Too many passes to CNN (~2000 proposals per image)
3. Lead to unacceptable test time (~50s per image)
4. High space cost to train SVM (millions of 1024-d features)



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

目标检测（RCNN）

R-CNN



Classify regions with SVMs

Forward each region through ConvNet

Warped image regions

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

Input image

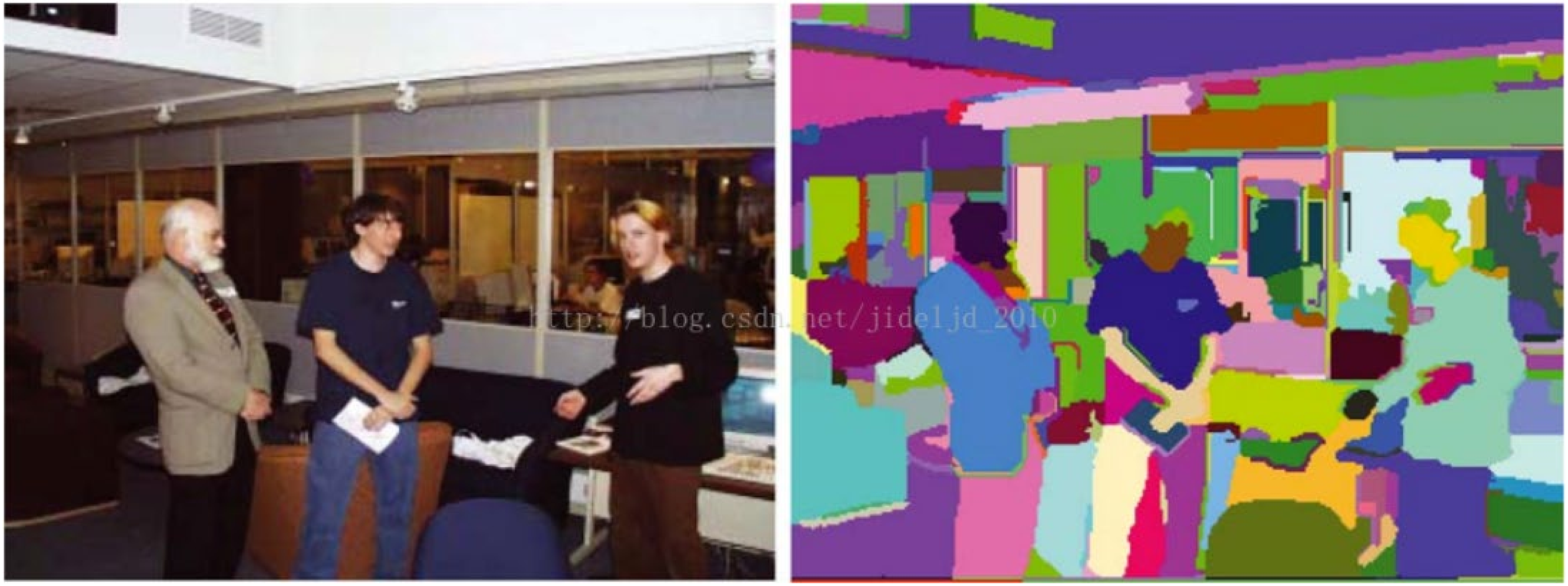
Girshick et al, “Rich feature hierarchies for accurate object detection and semantic segmentation”, CVPR 2014.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

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

目标检测（RCNN）

Region Proposals (Selective Search, SS)

给定一张图片，首先使用 Efficient Graph-BasedImage Segmentation 算法，将图片进行过分割 (Over-Segmentation)

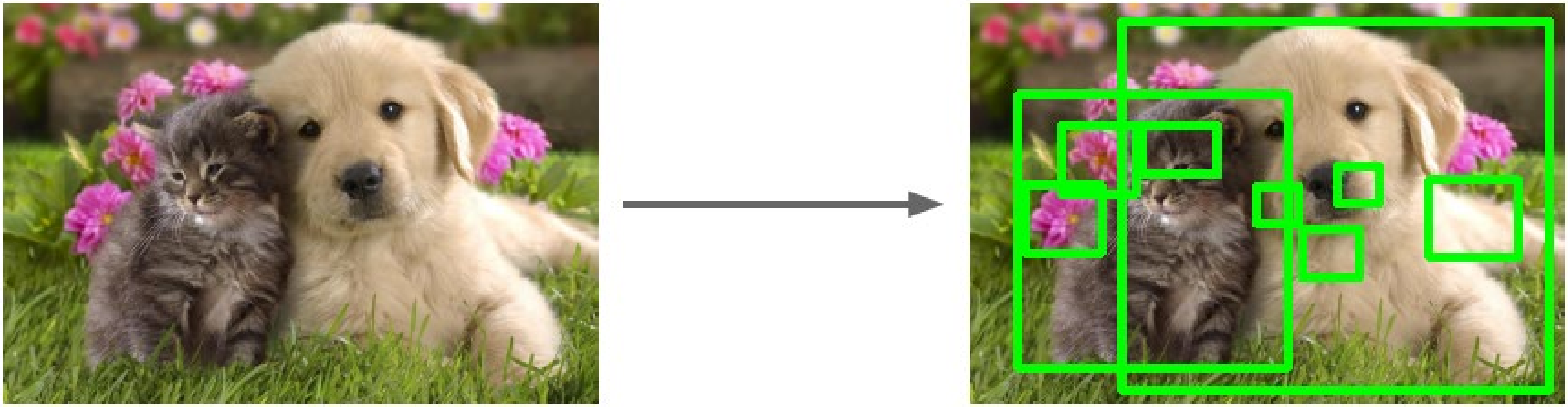


如图所示，过分割后的每个region非常小，以此为基础，对相邻的region进行相似度判断并融合，形成不同尺度下的region。每个region对应一个bounding

目标检测（RCNN）

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



Alexe et al, “Measuring the objectness of image windows”, TPAMI 2012
Uijlings et al, “Selective Search for Object Recognition”, IJCV 2013
Cheng et al, “BING: Binarized normed gradients for objectness estimation at 300fps”, CVPR 2014
Zitnick and Dollar, “Edge boxes: Locating object proposals from edges”, ECCV 2014

目标检测（RCNN）

2015-fast R-CNN, ROI pooling

Basic Ideas:

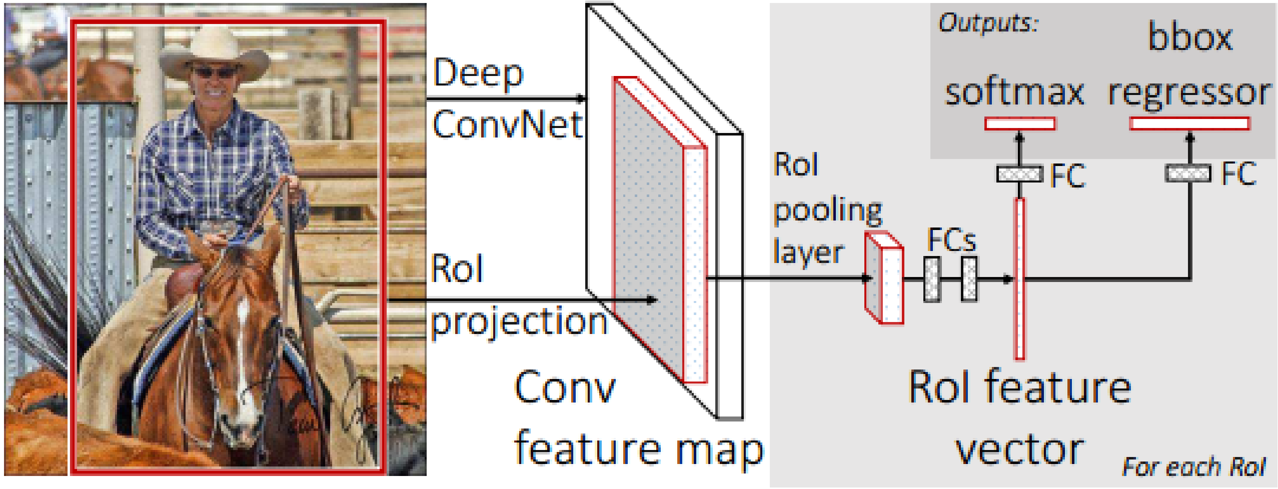
Reduce the computation redundancy caused by overlaps

Main Contributions:

1. ROI pooling layer
2. Replace SVM with softmax inside CNN
3. Use SVD to accelerate fully connected layer

Main Problems:

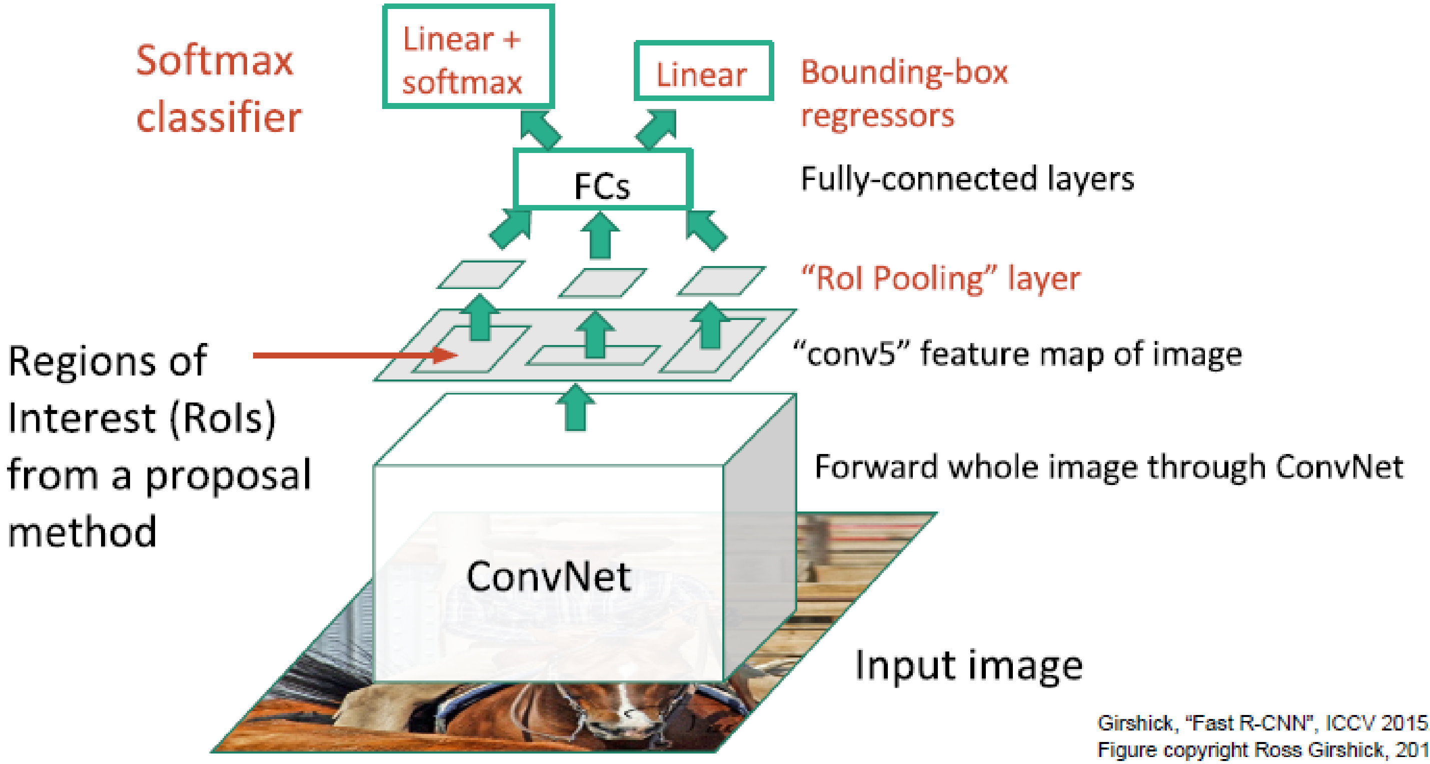
1. SS costs too much time (~2s for a fast version)



Girshick, Ross. "Fast r-cnn." CVPR. 2015.

目标检测（RCNN）

Fast R-CNN

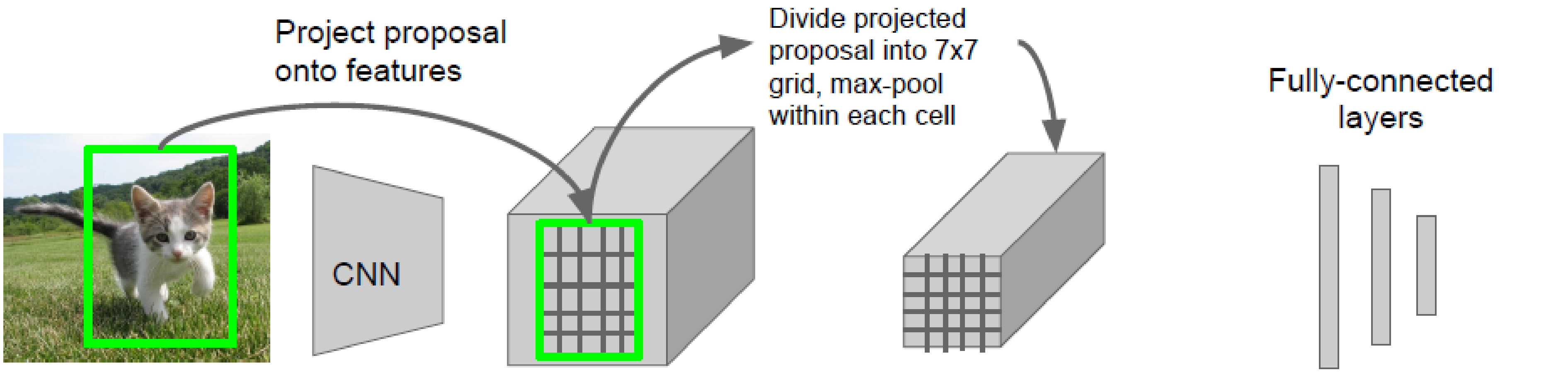


Girshick, “Fast R-CNN”, ICCV 2015.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Girshick, Ross. "Fast r-cnn." CVPR. 2015.

目标检测（RCNN）

ROI Pooling



Project proposal onto features

Divide projected proposal into 7x7 grid, max-pool within each cell

Fully-connected layers

Hi-res input image:
3 x 640 x 480
with region
proposal

Hi-res conv features:
512 x 20 x 15;

Projected region
proposal is e.g.
512 x 18 x 8
(varies per proposal)

RoI conv features:
512 x 7 x 7
for region proposal

Fully-connected layers expect
low-res conv features:
512 x 7 x 7

Girshick, “Fast R-CNN”, ICCV 2015.

Girshick, Ross. "Fast r-cnn." CVPR. 2015.

目标检测（RCNN）

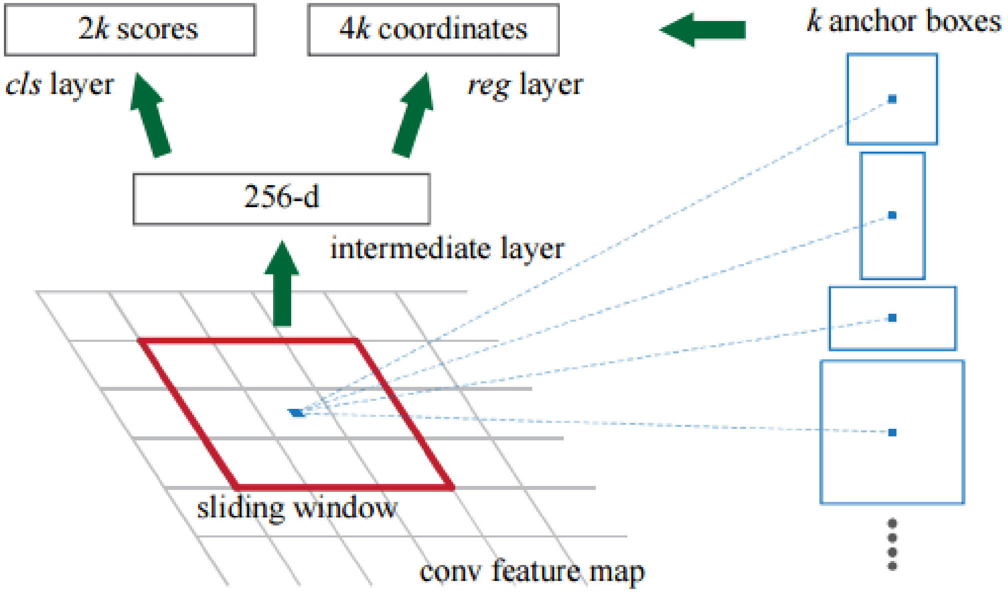
2015-faster R-CNN, RPN

Basic Ideas:

Reduce the time of generating region proposals

Main Contributions:

1. Region Proposal Network (RPN)
2. An end to end model finally!

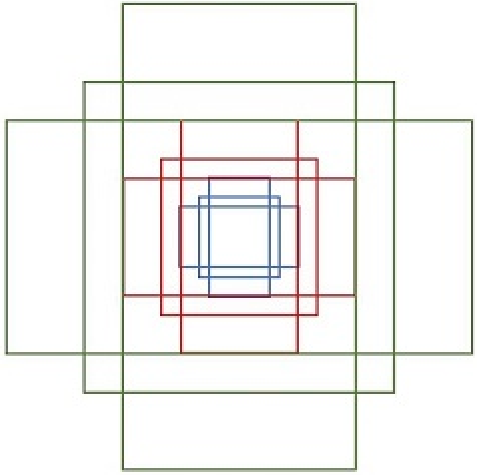


Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." NIPS. 2015.

目标检测（RCNN）

2015-faster R-CNN, RPN

对于特征图某个固定点，ANCHOR 生成9个矩形，共有3种形状，长宽比为大约为：width:height = [1:1, 1:2, 2:1]三种，实际上通过anchors就引入了检测中常用到的多尺度方法。



把任意大小的输入图像reshape成800x600（即图2中的M=800，N=600）。再回头来看anchors的大小，anchors中长宽1:2中最大为352x704，长宽2:1中最大736x384，基本是cover了800x600的各个尺度和形状。

目标检测（RCNN）

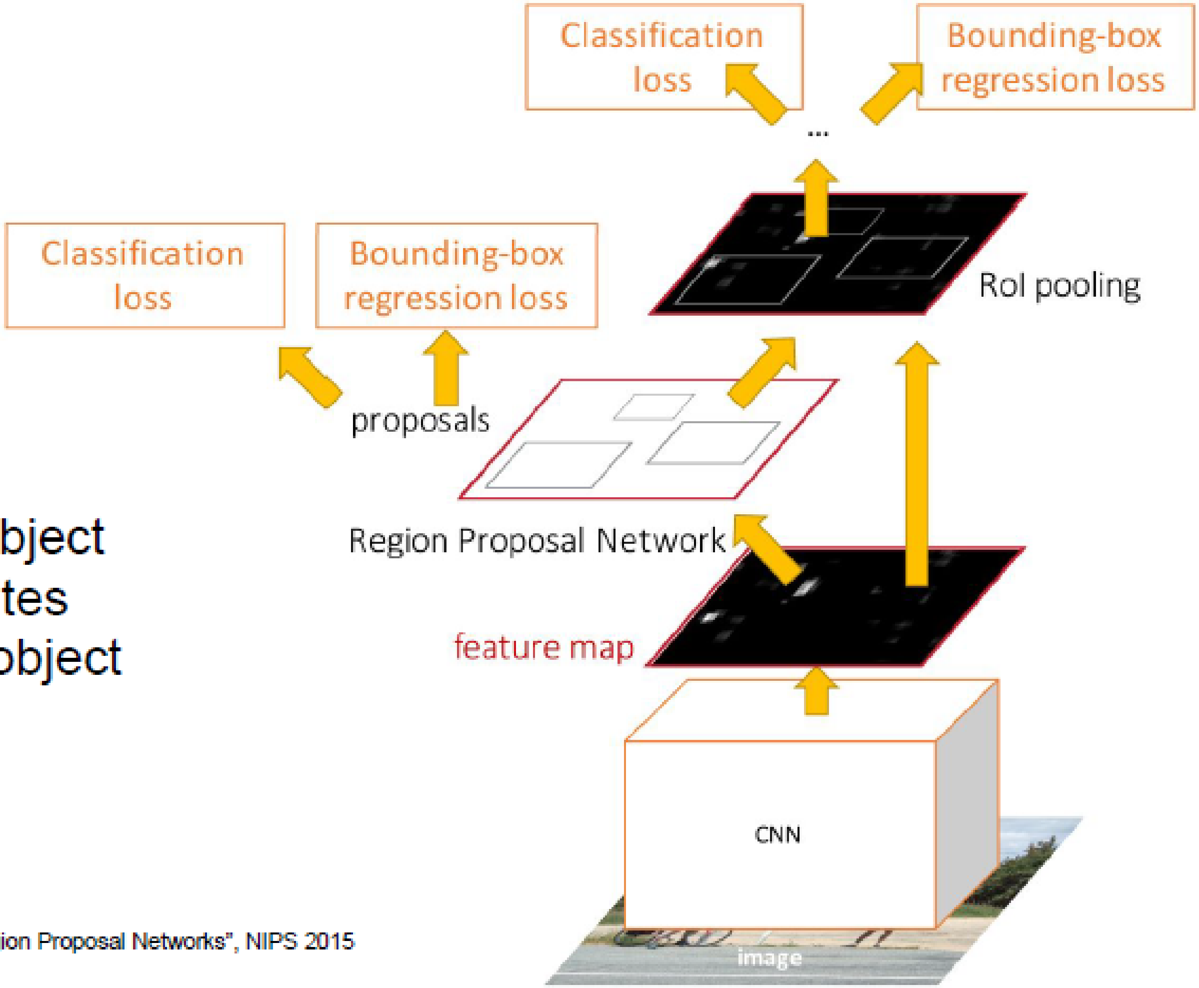
Faster R-CNN:

Make CNN do proposals!

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

Jointly train with 4 losses:

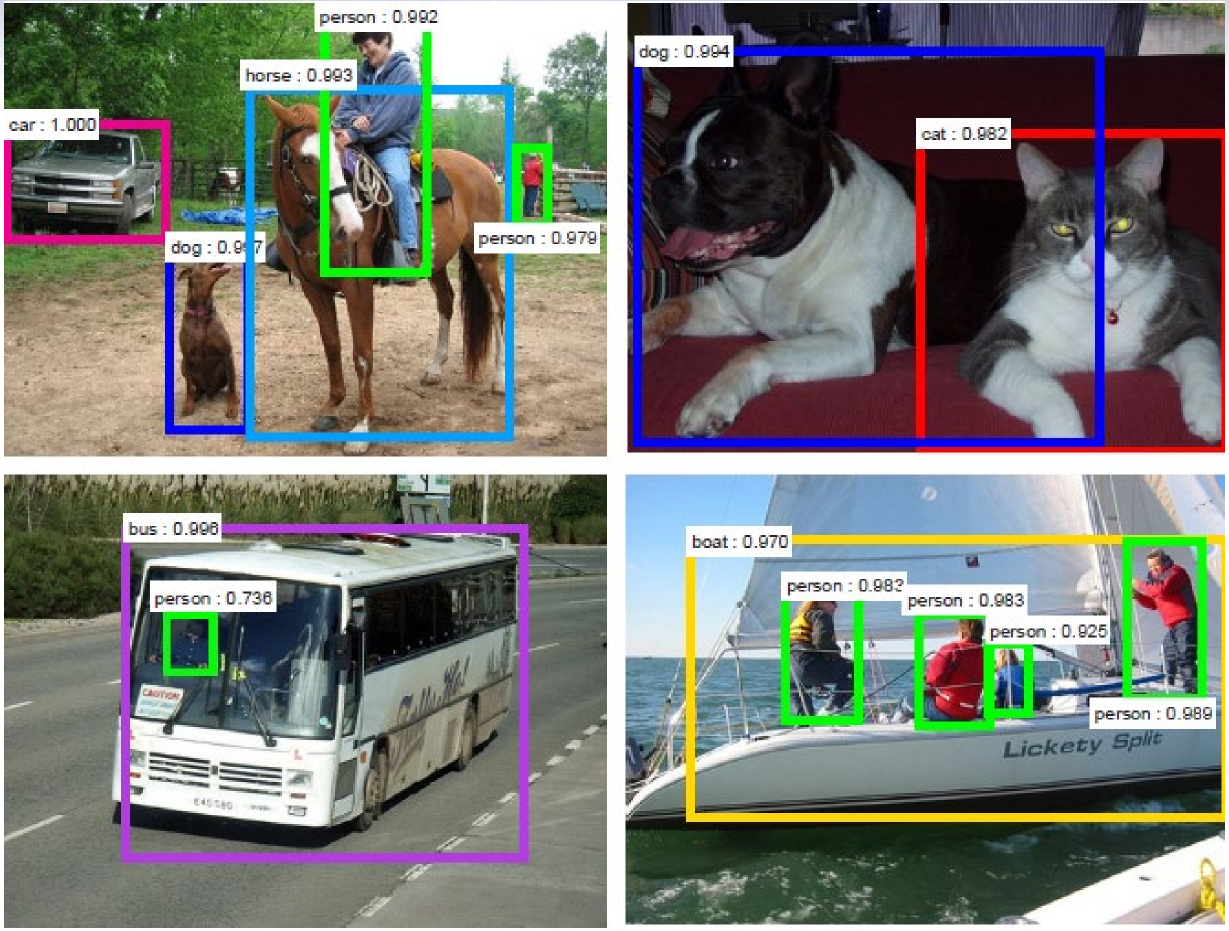
1. RPN classify object / not object
2. RPN regress box coordinates
3. Final classification score (object classes)
4. Final box coordinates



Ren et al, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NIPS 2015
Figure copyright 2015, Ross Girshick; reproduced with permission

Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." NIPS. 2015.

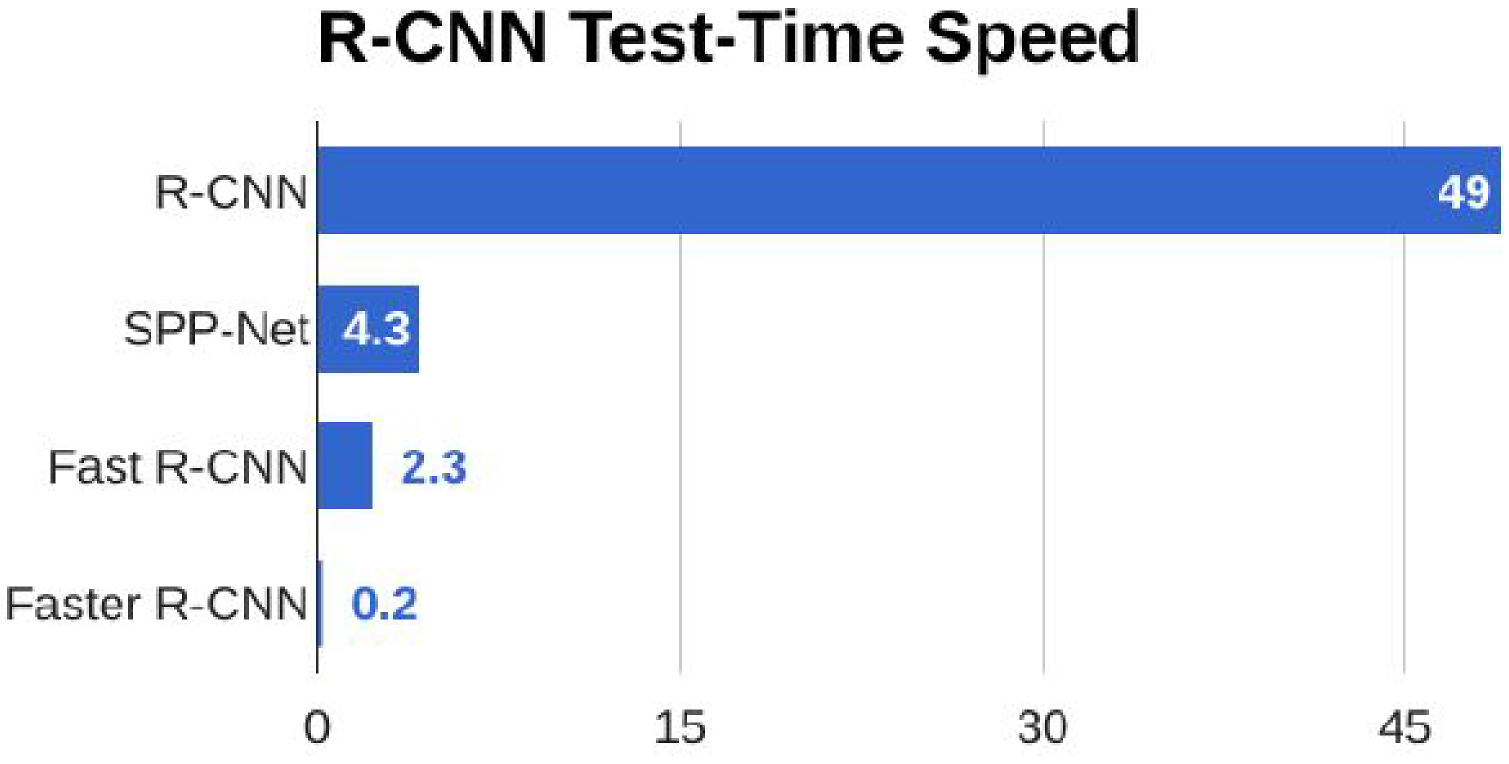
目标检测（RCNN）



Faster R-CNN检测结果

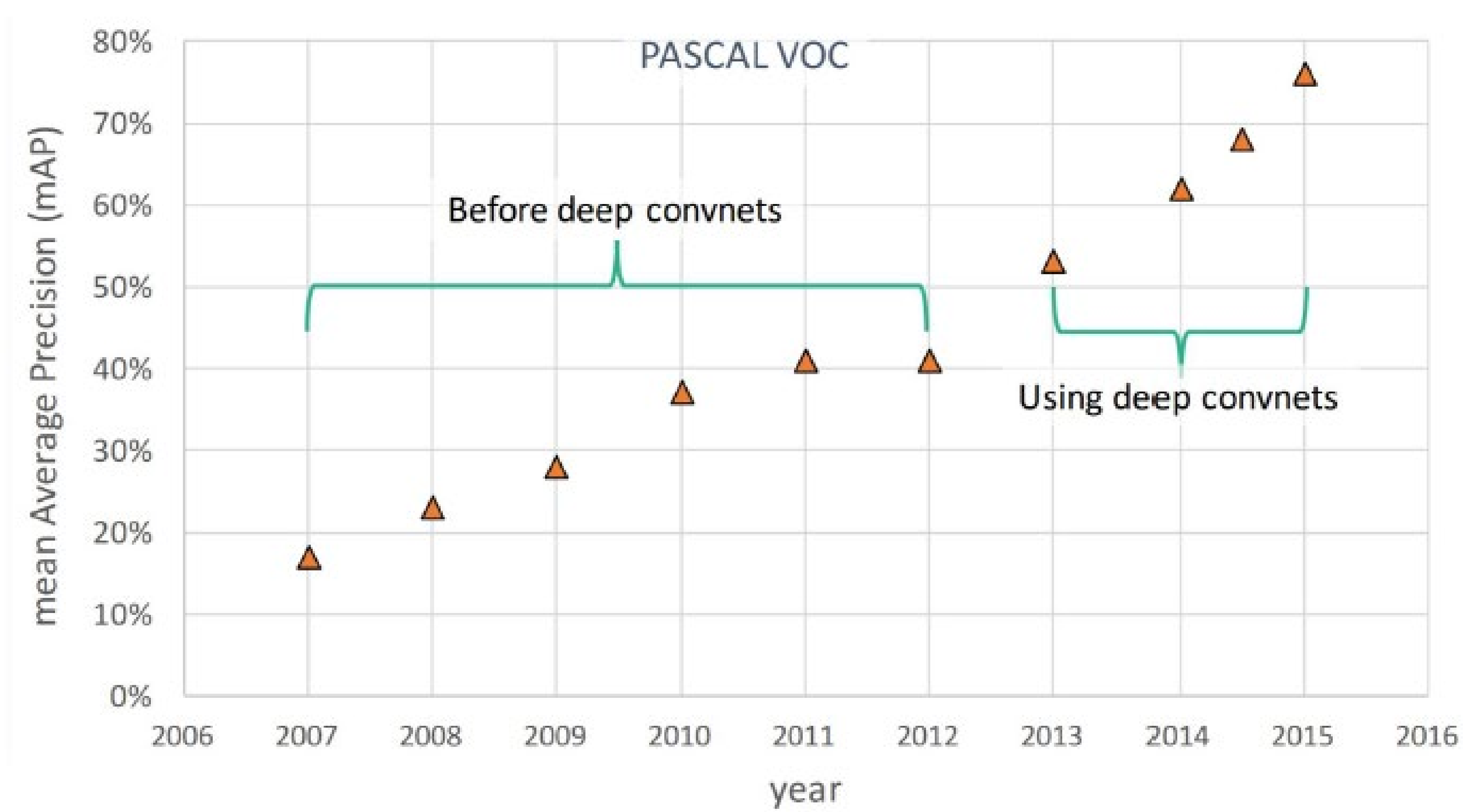
Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." NIPS. 2015.

目标检测（RCNN）



运行时间对比

目标检测（RCNN）



在PASCAL VOC上的性能对比

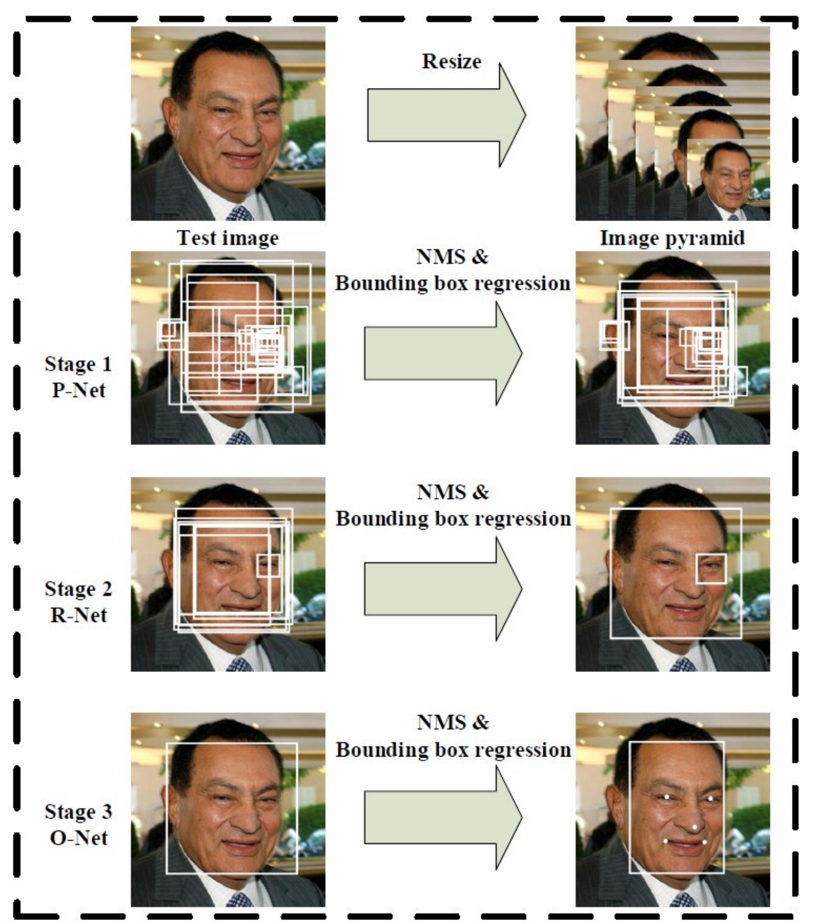
目标检测 – 以人脸检测为例

MTCNN

Multitask:

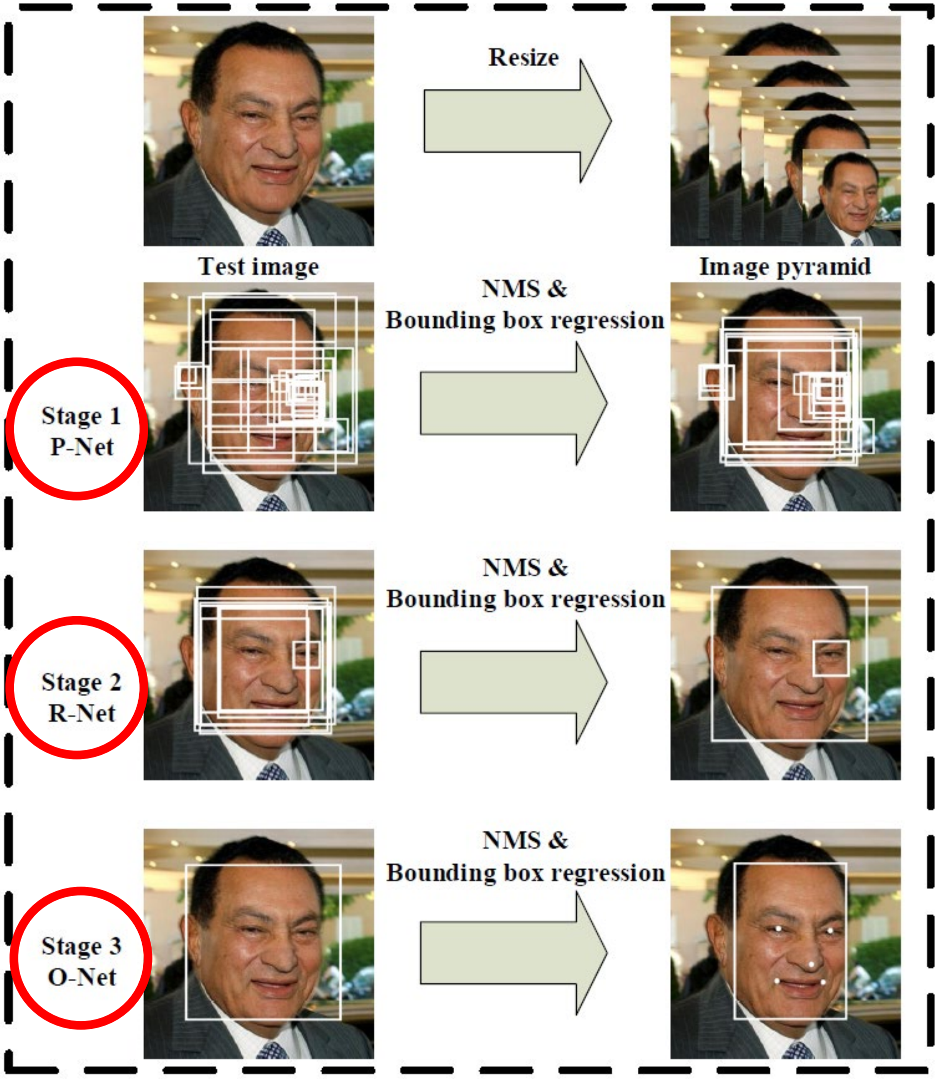
1 Face detection

2 Facial landmarks localization



Zhang K, Zhang Z, Li Z, et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks.

目标检测 – 以人脸检测为例

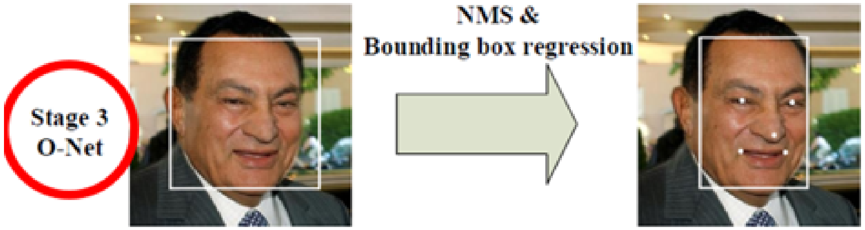


P-Net (Proposal Network)：
该网络主要是检测图中人脸，产生多个人脸候选框和回归向量，再用回归向量对候选窗口进行校准，最后通过非极大值抑制NMS来合并高度重叠的候选框。

R-Net (Refine Network)：
该网络同样输出候选框置信度（根据置信度削减候选框数量）和回归向量，通过边界框回归和NMS精调候选框的位置。

O-Net (Output Network)：
比R-Net层又多了一层卷积层，处理结果更加精细，作用和R-Net层作用一样（削减框数量同时精调回归框）。再者，该层对人脸区域进行了更多的监督，最后输出5个人脸关键点坐标。

目标检测 – 以人脸检测为例



以Onet的关键点训练为例
（实际mtcnn训练label应
有人脸框坐标）

训练输入

+

5个正确 landmarks 坐标（标定点）

运算结果

5个运算出来的坐标（运算点）

计算损失

-

如计算
运算点与标定点的
Euclidean loss

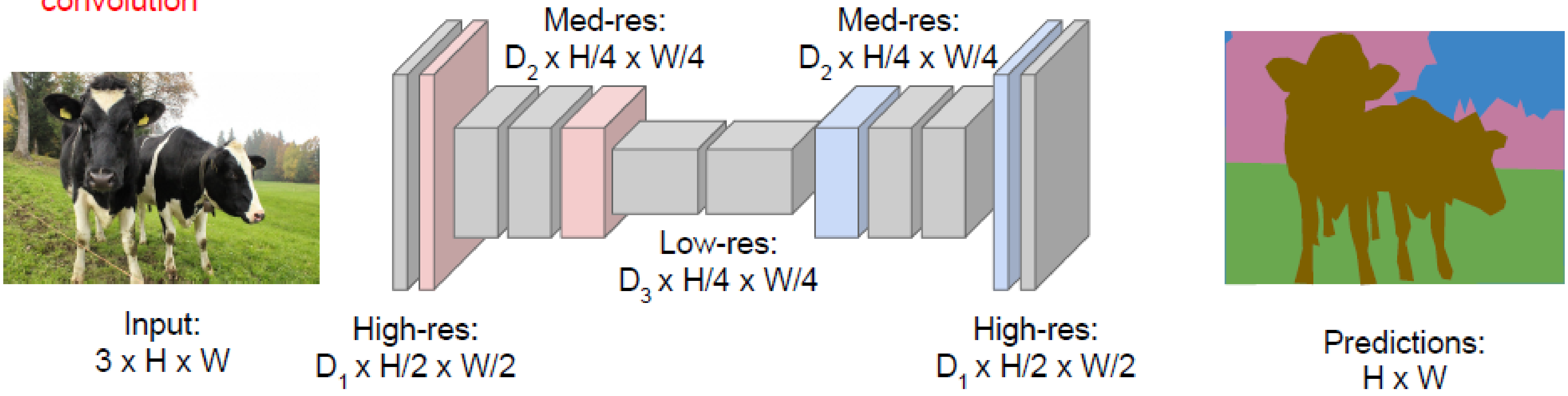
目标检测 – 语义分割

全卷积网络（Fully Convolutional Networks）

Downsampling:
Pooling, strided
convolution

Design network as a bunch of convolutional layers, with **downsampling** and **upsampling** inside the network!

Upsampling:
???

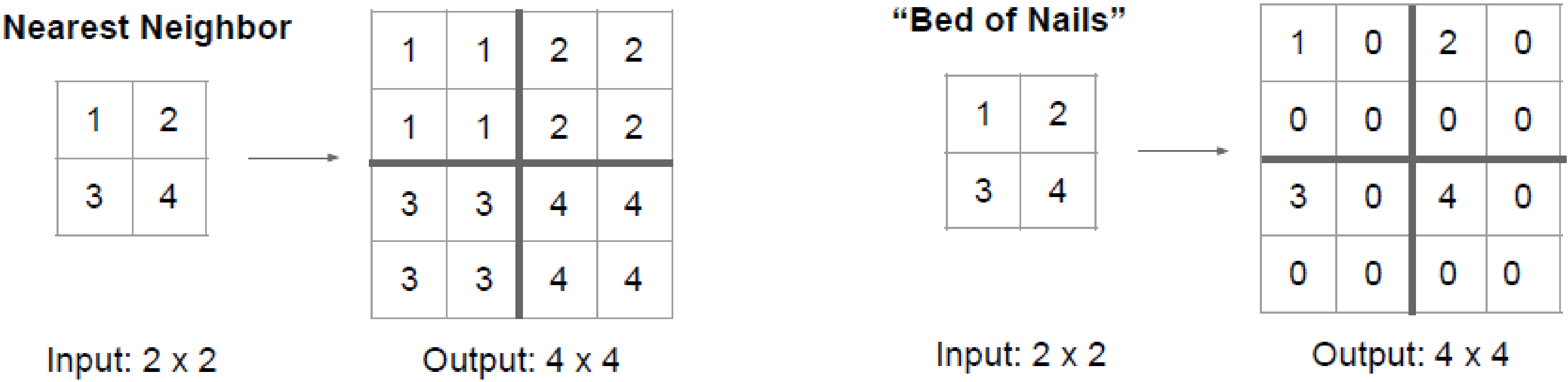


Long, Shelhamer and Darreli, Fully Convolutional Networks for Semantic Segmentation, CVPR 2015

目标检测 – 语义分割

Pooling 层的上采样（Upsampling）

（a） Average pooling



Nearest Neighbor

1	2
3	4

Input: 2 x 2

1	1	2	2
1	1	2	2
3	3	4	4
3	3	4	4

Output: 4 x 4

“Bed of Nails”

1	2
3	4

Input: 2 x 2

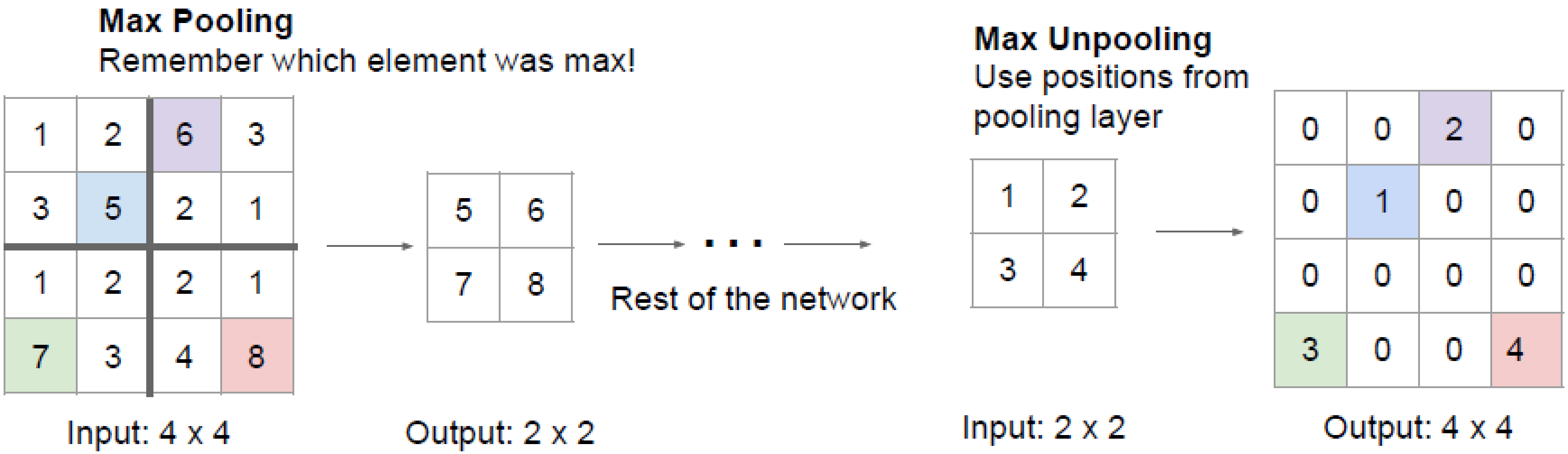
1	0	2	0
0	0	0	0
3	0	4	0
0	0	0	0

Output: 4 x 4

目标检测 – 语义分割

Pooling 层的上采样（Upsampling）

（b） Max pooling



Max Pooling

Remember which element was max!

1	2	6	3
3	5	2	1
1	2	2	1
7	3	4	8

Input: 4 x 4

5	6
7	8

Output: 2 x 2

. . .

Rest of the network

Max Unpooling

Use positions from pooling layer

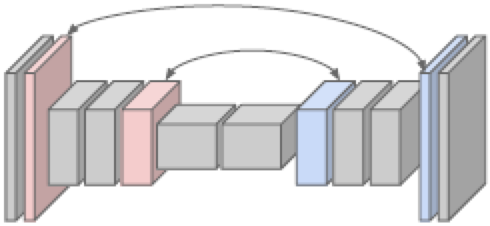
1	2
3	4

Input: 2 x 2

0	0	2	0
0	1	0	0
0	0	0	0
3	0	0	4

Output: 4 x 4

Corresponding pairs of downsampling and upsampling layers



目标检测 – 语义分割

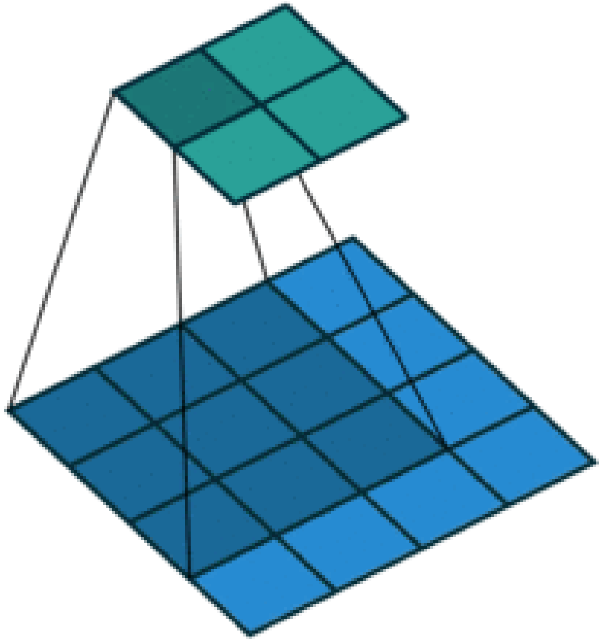
卷积层的上采样（Upsampling）

也叫反卷积（Deconvolution）或 转置卷积（Transpose Convolution）

目标检测 – 语义分割

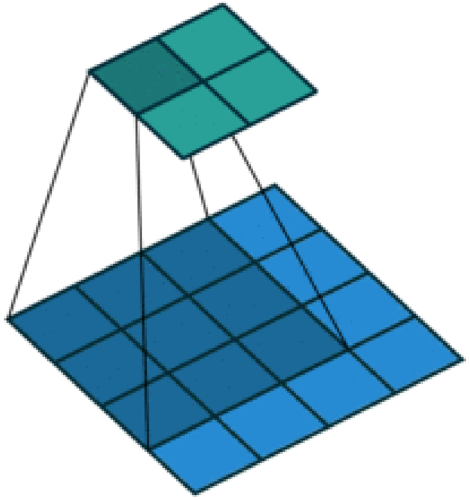
卷积层的上采样（Upsampling）

考虑如下一个卷积层，输入特征图4*4，卷积核3*3，步长1，卷积后获得特征图维度为2*2：



目标检测 – 语义分割

卷积流程



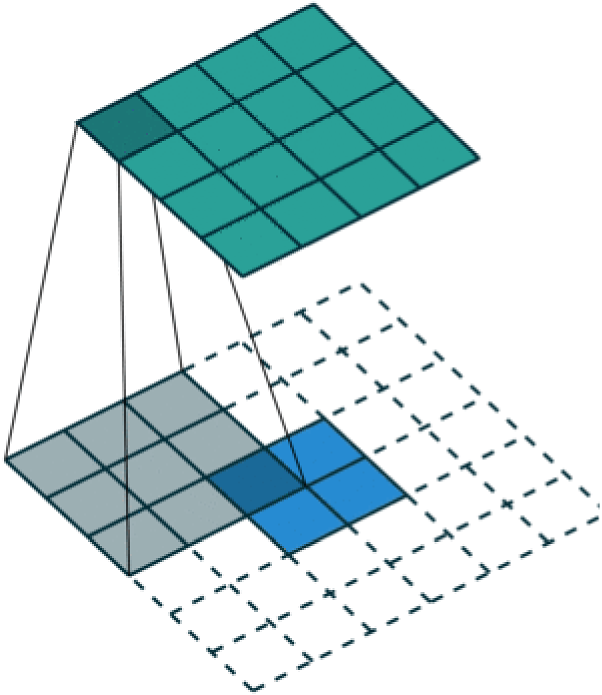
$$C=\begin{pmatrix} w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} & 0 & 0 & 0 & 0 & 0 \\ 0 & w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} & 0 \\ 0 & 0 & 0 & 0 & 0 & w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} \end{pmatrix}$$

$$X=[x\_1,x\_2,\ldots,x\_16]^{T}$$

$$Y=CX$$

目标检测 – 语义分割

反卷积流程

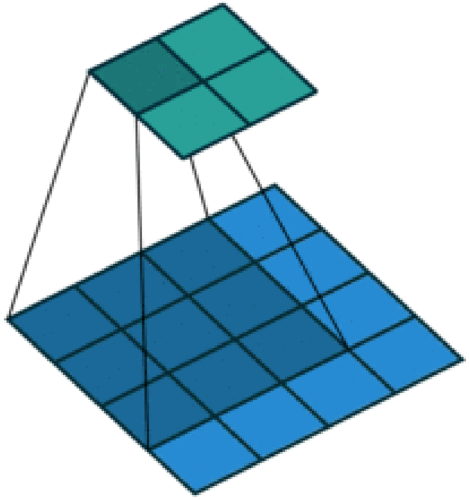


$$C = \begin{pmatrix} w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} & 0 & 0 & 0 & 0 & 0 \\ 0 & w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} & 0 \\ 0 & 0 & 0 & 0 & 0 & w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} \end{pmatrix}$$

$$X=[x\_1,x\_2,\dots,x\_16]^T$$

$$X=C^T Y$$

目标检测 – 语义分割



$$C=\begin{pmatrix} w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} & 0 & 0 & 0 & 0 & 0 \\ 0 & w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} & 0 \\ 0 & 0 & 0 & 0 & 0 & w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} \end{pmatrix}$$

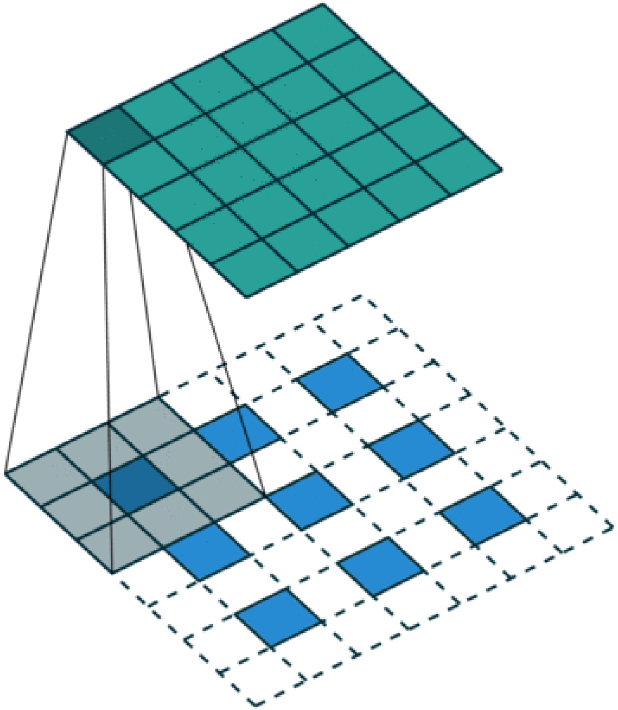
X=[x_1,x_2,…,x_16]^T

Y=CX

目标检测 – 语义分割

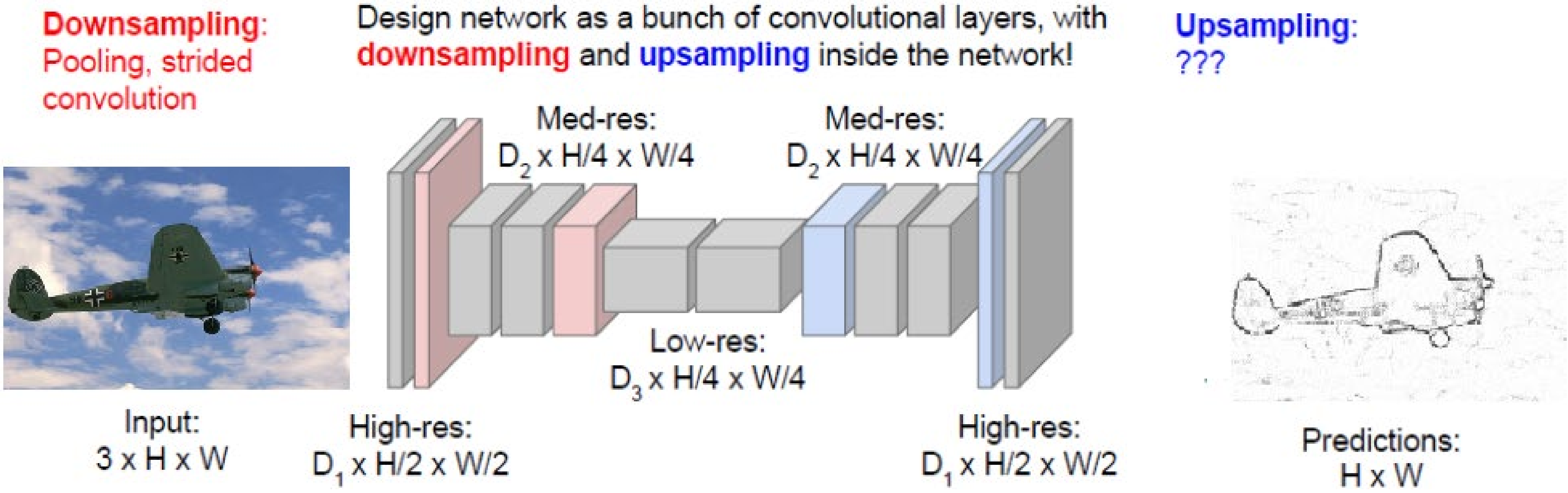
卷积层上采样另一个例子：

考虑一个卷积层，输入特征图5*5，卷积核3*3，步长2，补零1， 卷积后获得特征图维度为3*3，其反卷积示意图如下：



目标检测 – 语义分割

全卷积网络 – 边缘提取



Downsampling: Pooling, strided convolution

Design network as a bunch of convolutional layers, with downsampling and upsampling inside the network!

Upsampling: ???

Med-res: D_2 x H/4 x W/4

Med-res: D_2 x H/4 x W/4

Low-res: D_3 x H/4 x W/4

Input: 3 x H x W

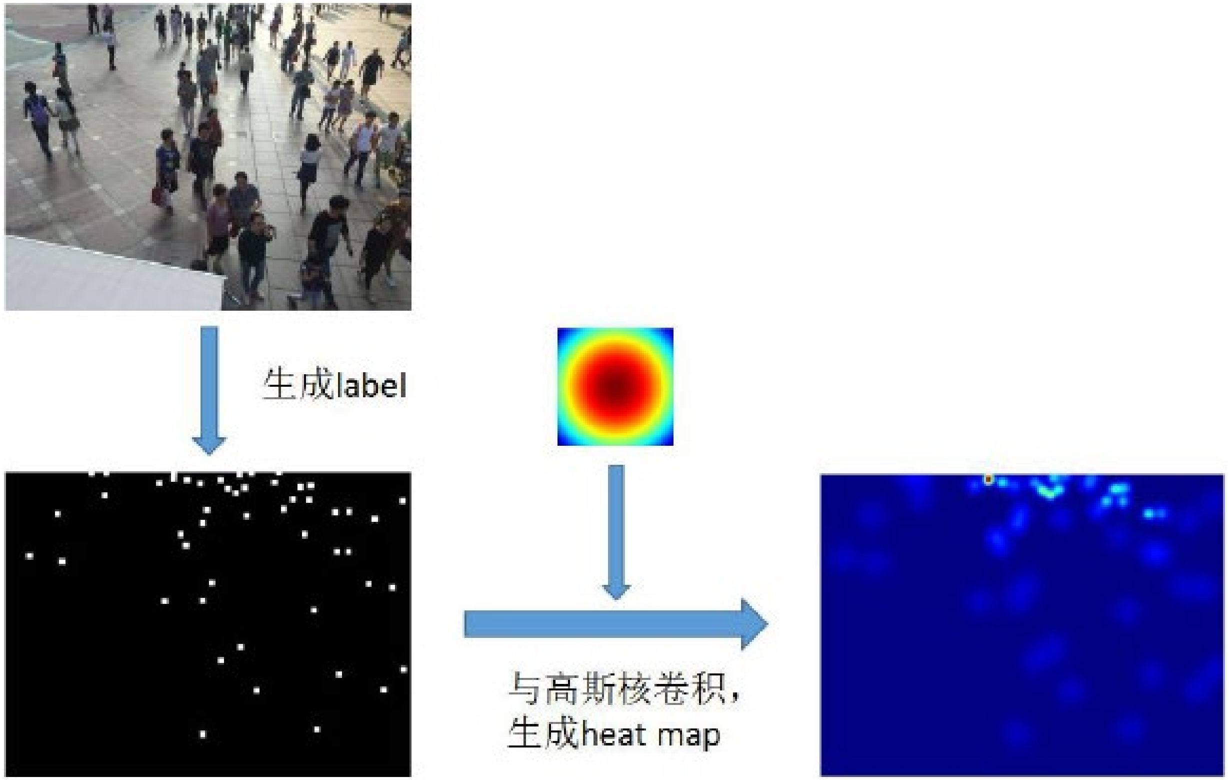
High-res: D_1 x H/2 x W/2

High-res: D_1 x H/2 x W/2

Predictions: H x W

目标检测 – 语义分割

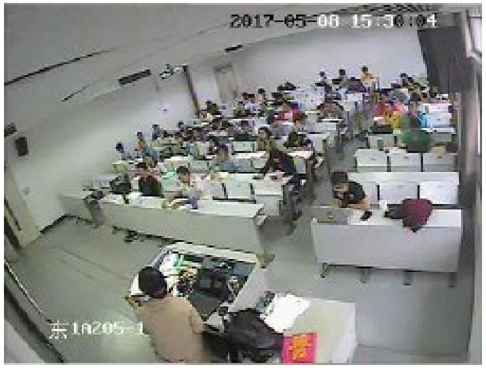
全卷积网络 – 视频场景人数估计



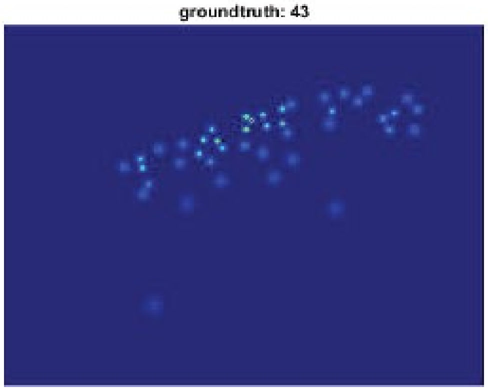
训练流程：**Heat map** 生成

目标检测 – 语义分割

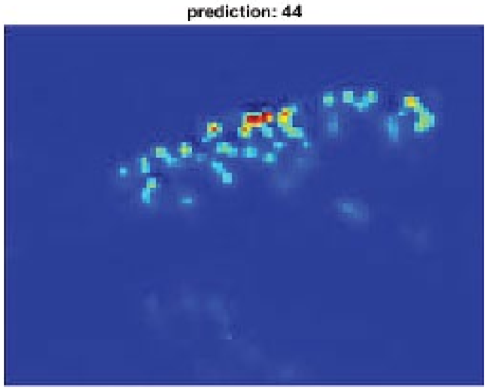
全卷积网络 – 视频场景人数估计



图像



Ground Truth = 43



Prediction= 44

测试流程：数据的Ground Truth 与 Prediction

目标检测 – 语义分割

全卷积网络 – 视频场景人数估计

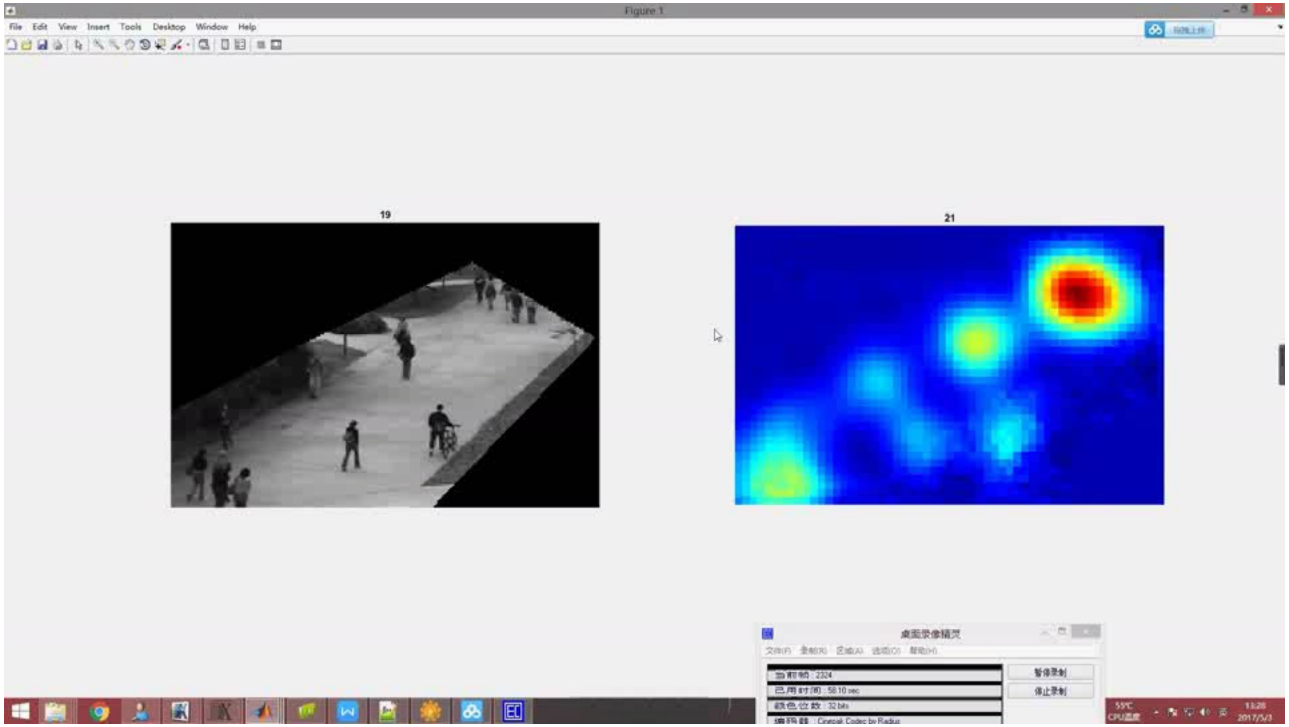


Figure 1
File Edit View Insert Tools Desktop Window Help
19
21
桌面录像精灵
文件(F) 录制(R) 区域(A) 选项(O) 帮助(H)
暂停录制
停止录制
55℃
CPU温度
13:28
2017/5/3

当前帧：2324

已用时间：58.10 sec

颜色位数：32 bits

编码器：Cinepak Codec by Radius

Thank you and comments are welcomed