Supplementary Material

Stixel World

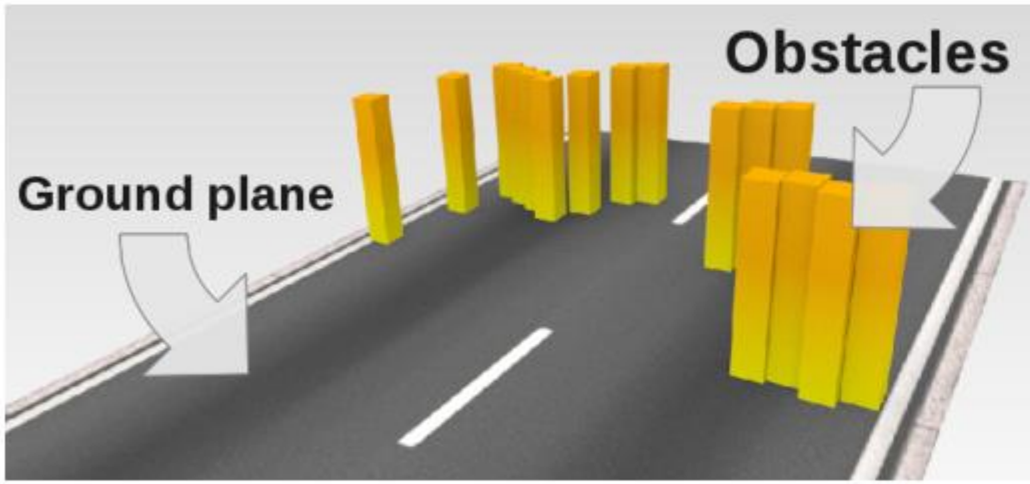
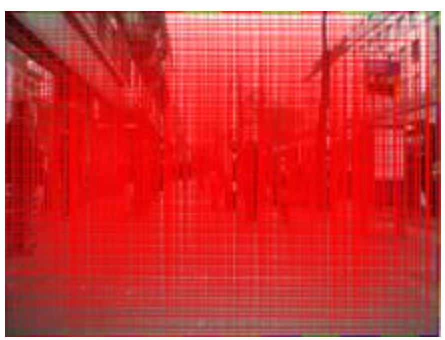


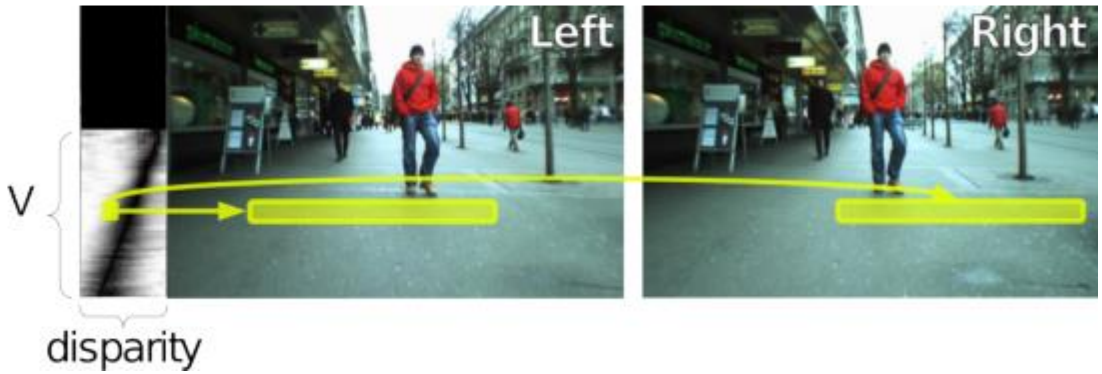
Figure 1: The stixel world is composed of the ground plane and vertical sticks describing the obstacles.



效果

Stixel – Ground Estimation

原理：
地面(平面)在v-disparity空间为直线



Stixel – Distance Estimation

原理：优化方程

- 数据项：是行人+脚下是地面
- 平滑项

(a) Object evidence (b) Ground evidence

Figure 3: The object and ground costs are computed by matching pixels in the left and right images. White dots on the image indicates object-ground boundary candidates, based on horizontal gradient maxima.

Stixel – Distance Estimation

$$d_s^*(u) = \underset{d(u)}{\operatorname{argmin}} \sum_u c_s\left(u, d(u)\right) + \sum_{u_a, u_b} s_s\left(d(u_a), d(u_b)\right) \tag{1}$$

where u_a, u_b are neighbours ($|u_a - u_b| = 1$).

数据项

The cost c_s is the result of summing two costs: $c_o(u, d)$ ("object cost"), the cost of a vertical object being present, and $c_g(u, d)$ ("ground cost"), the cost of a supporting ground being present (see figure 3).

$$c_s(u, d) = c_o(u, d) + c_g(u, d) \tag{2}$$

$$\begin{aligned} c_o(u, d) &= \sum_{v = v(\check{h}_o, d)}^{v(d)} c_m(u, v, d) \quad , \\ c_g(u, d) &= \sum_{v = v(d)}^{|V|} c_m(u, v, f_{ground}(v)) \end{aligned} \tag{3}$$

where $|V|$ indicates the number of rows in the image and the smallest v is at the top of the image.

平滑项

$$s_s\left(d_a, d_b\right) = \begin{cases} \infty & \text{if} \quad d_a < d_b - 1 \\ c_o\left(u_a, d_a\right) & \text{if} \quad d_a = d_b - 1 \\ 0 & \text{if} \quad d_a > d_b - 1 \end{cases} \tag{4}$$

where $d_a = d(u_a)$, $d_b = d(u_b)$, and u_a is one pixel to the left of u_b. The case $s_s = \infty$ ensures that no stixel distance estimate will violate the occlusion constraint.

Feature Response Estimation

N models, 1 image scale
(a) Naive approach

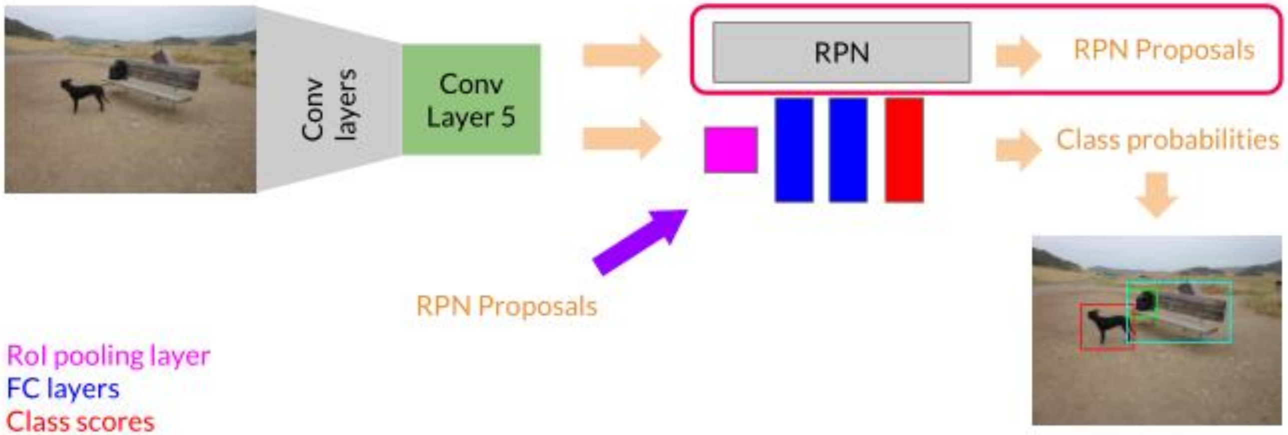
1 model, N image scales
(b) Traditional approach

1 model, N/K image scales
(c) FPDW approach

N/K models, 1 image scale
(d) Our approach

Figure 2: Different approaches to detecting pedestrians at multiple scales.

Faster-rcnn



i = anchor index in minibatch

Log loss

Ground truth objectness label

Smooth L1 loss

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*).$$

True box coordinates

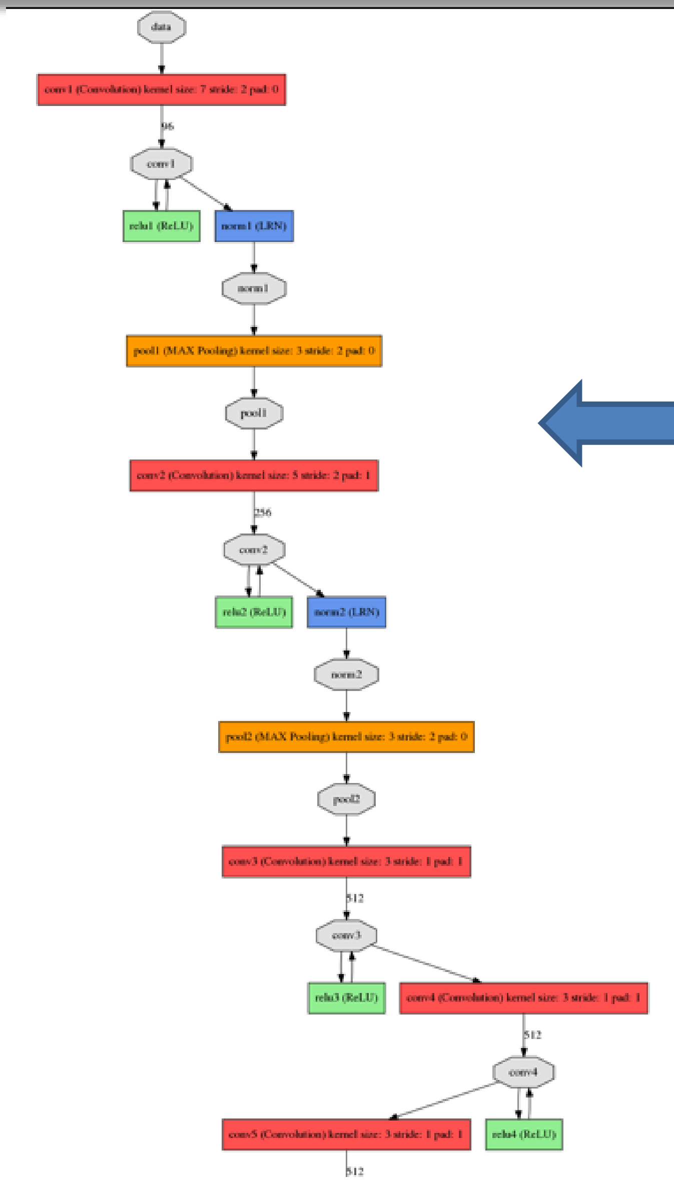
Coordinates of the predicted bounding box for anchor i

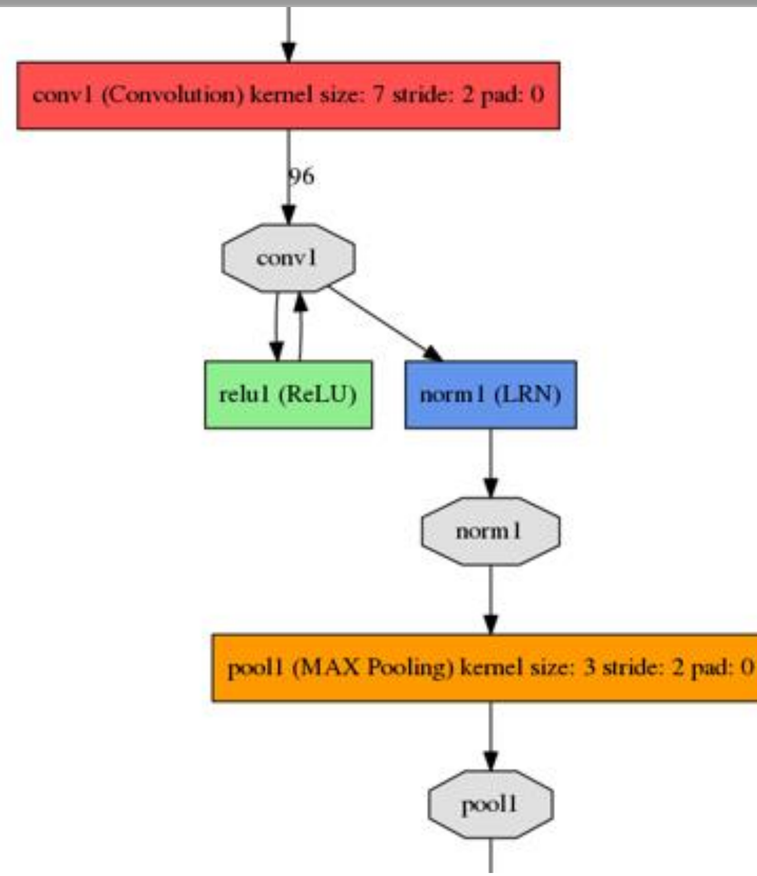
Predicted probability of being an object for anchor i

In practice λ= 10, so that both terms are roughly equally balanced

N_{cls} = Number of anchors in minibatch (~ 256)
N_{reg} = Number of anchor locations (~ 2400)

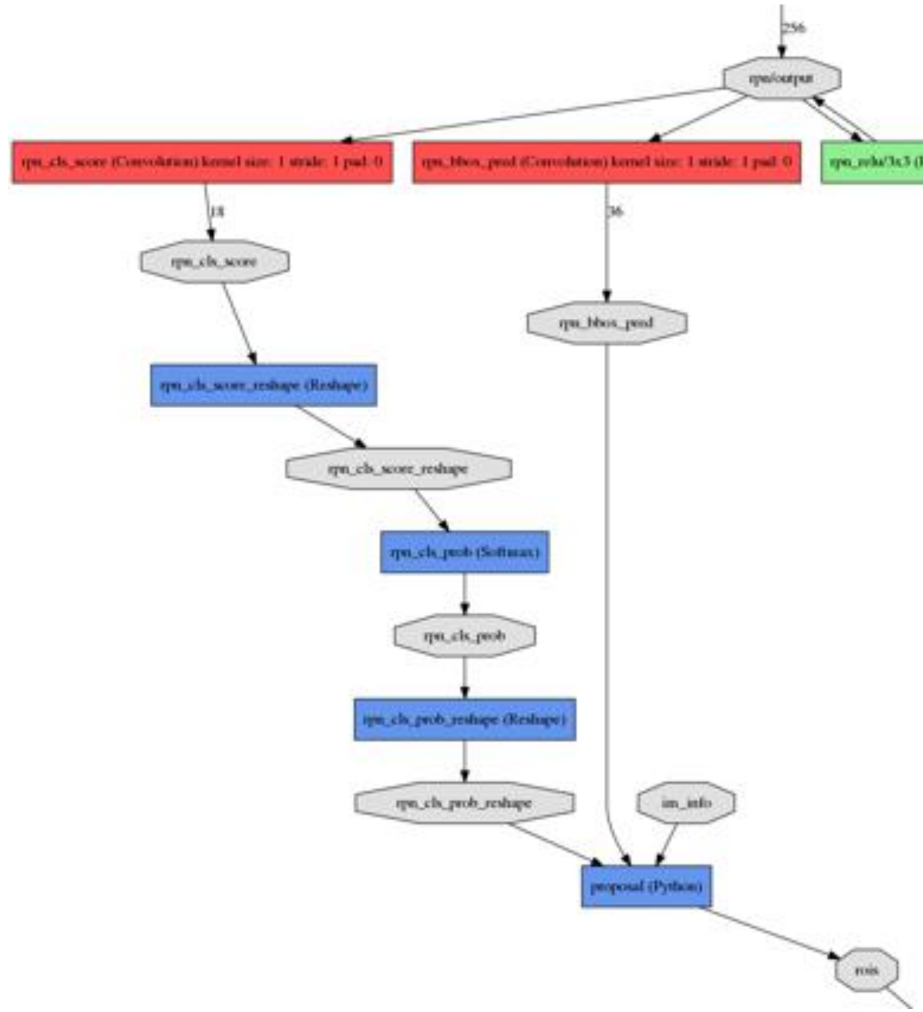
Faster-rcnn





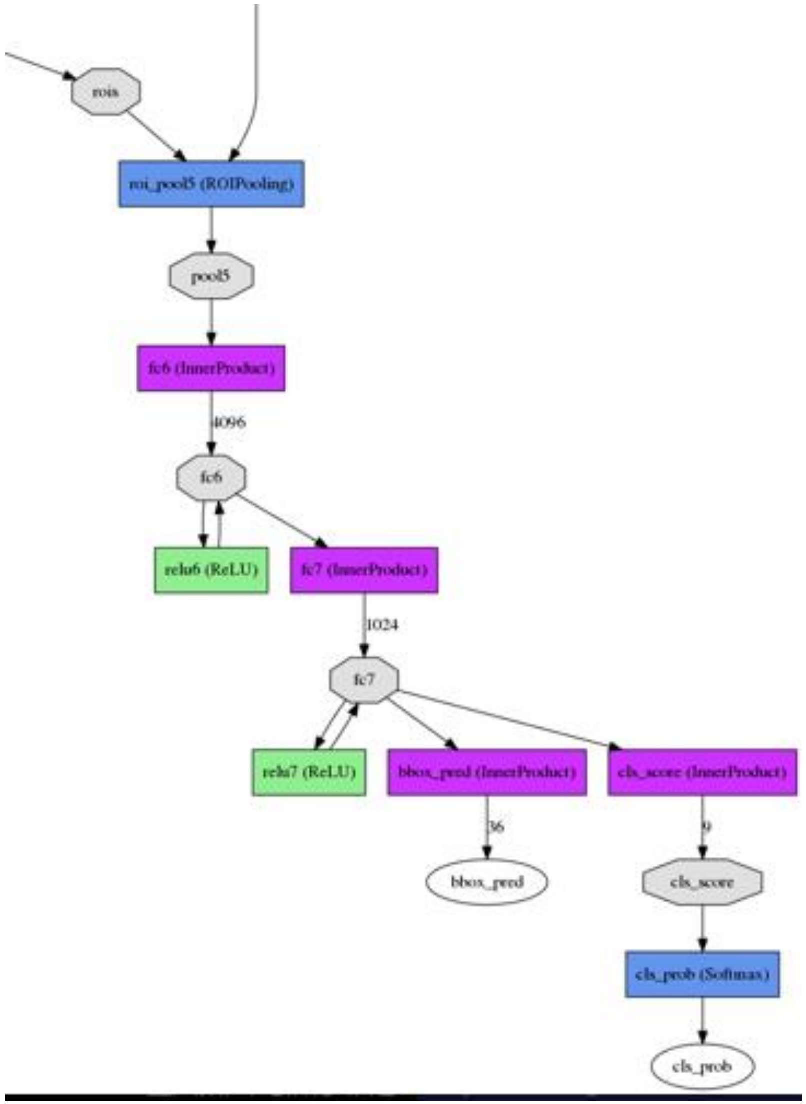
VGG net
特征提取

Faster-rcnn



RPN
ROI proposal

Faster-rcnn



Detection

Faster-rcnn Problem

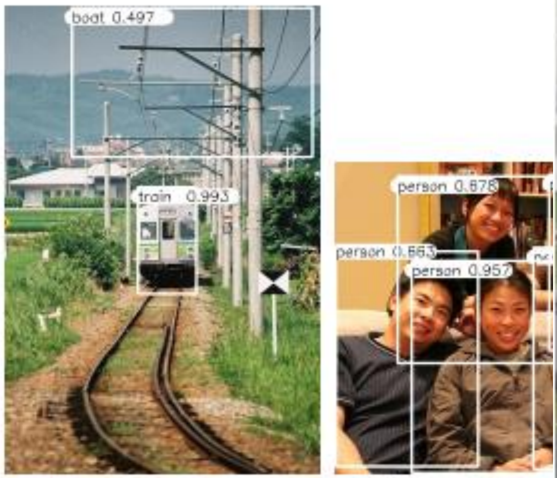
- 没有真正解决尺度问题
- 没有考虑场景语义
- 泛目标→具体目标
- 没能学习到行人/人脸的**hierarchical**的特征
  - 如果原始数据集就是人脸又会好一些，但是没有那么大的数据集
  - 训练数据与测试数据不同
    - 场景不同
    - 训练数据中缺乏某一类数据

Scale-invariance

- 没有真正解决尺度问题：
  - RPN网络中有6种anchor，bbox regression一定范围内修正bbox大小和位置
  - 不足以解决尺度问题
  - 尤其是训练数据尺度不够丰富
  - 不能仅靠训练数据多样性，和数据增强手段
- 分形网络
  - 通过多个pooling层的组合，能形成丰富的尺度

Context/Semantic info

- 没有考虑场景语义



(a) Boat and Train (b) Partial Occlus

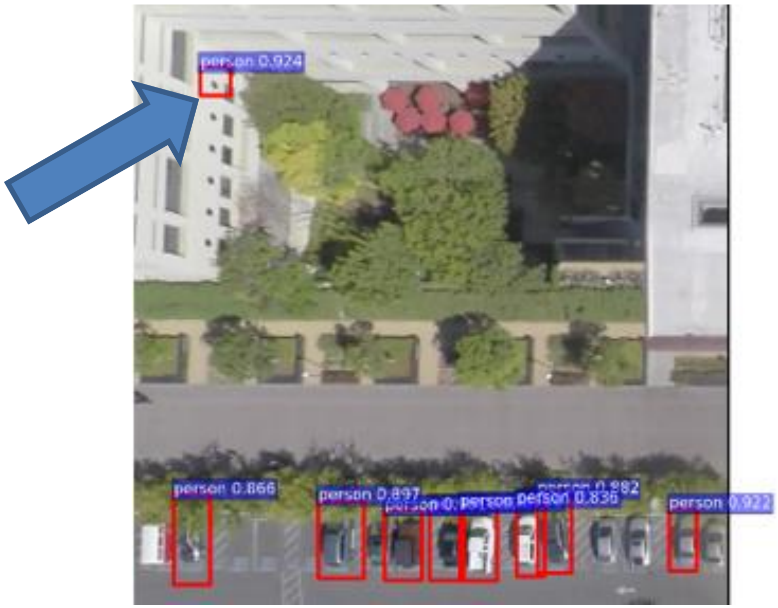
(b) Aeroplanes in air

Fig. 2. Image-level contextual information

Fig. 1. Object-level contextual information

Context/Semantic info

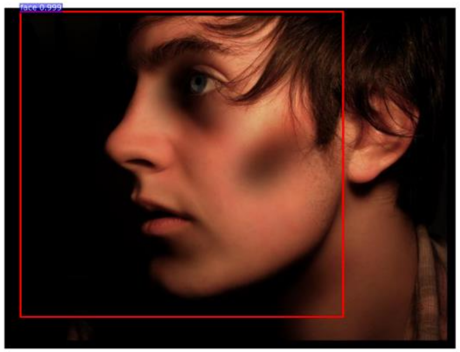
- 没有考虑场景语义

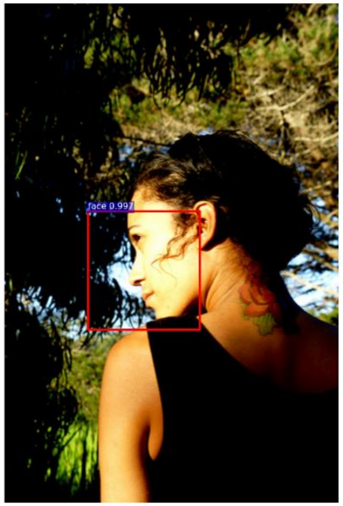


Apply to face

On Training data

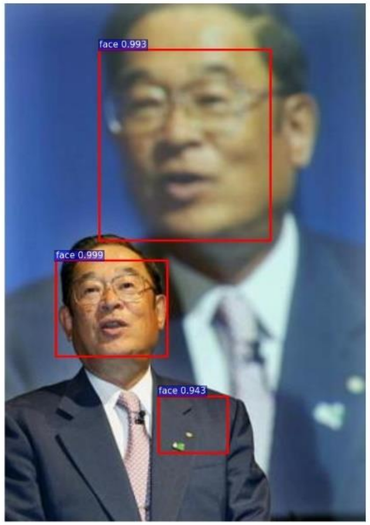
- Train(finetune) On AFLW
- Test On AFLW:
  - Bbox和标注风格有关
  - 效果不错

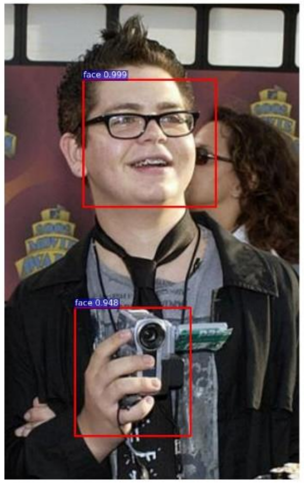


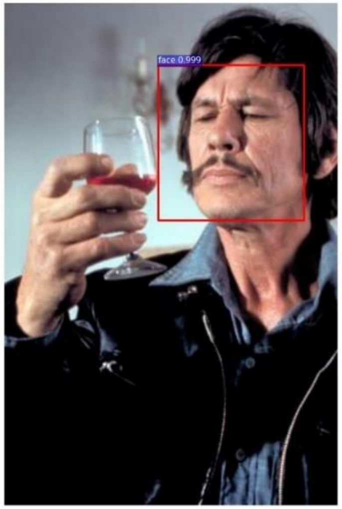


Apply to face

换一个数据集(FDDB)测试

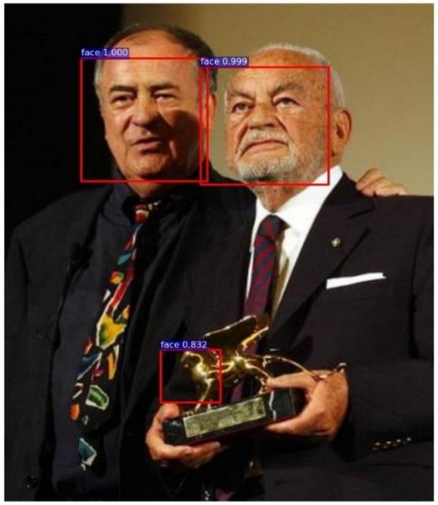






Apply to face

换一个数据集(FDDB)测试

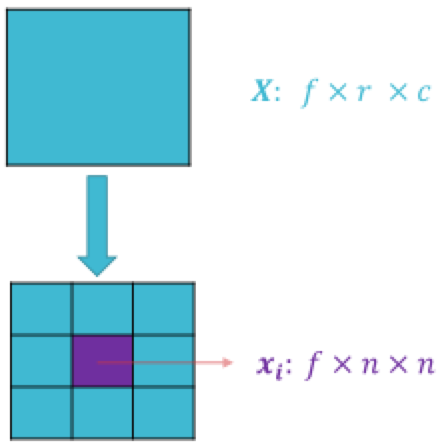


- Test On FDDB
- 误识别原因
  - 本身是做泛目标的
  - 自学习显著特征

Learn Face Structure

Grid Loss: Detecting Occluded Faces.(ECCV2016)

Loss函数考虑局部加全局，能使网络学习到人脸的具体部分，但是还是没有考虑部分之间的内在关系



- Balance part detectors with holistic detector

$$l(\theta) = \max\left(0, 1 - y(\mathbf{w}^T \mathbf{x} + b)\right) + \lambda \cdot \sum_{i=1}^{N} \max\left(0, m - y \cdot (\mathbf{w}_i^T \mathbf{f}_i + b_i)\right),$$

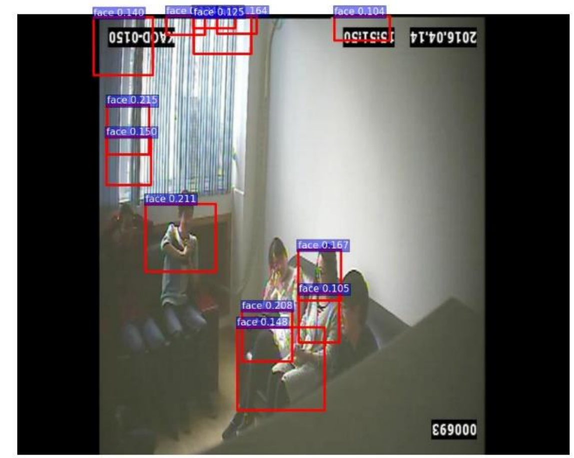
$$X = \{x_1, x_2, \ldots, x_N\}, \quad N = \left\lfloor \frac{r}{n} \right\rfloor \cdot \left\lfloor \frac{c}{n} \right\rfloor$$

$$\mathbf{w} = [\mathbf{w}_1, \mathbf{w}_1, \ldots, \mathbf{w}_1], b = \sum_i b_i$$

- The number of additional parameters compared to a regular classification layer is **N -1**

Apply to Anthor Data

Test On Bad Surveillance Data



- 主要是训练数据和测试数据不是一个风格的问题
  - 场景
  - 分别率
- 在数据量不足时，指定位置，让GAN生成人脸
  - 感觉目前生成数据不会很好
    - Resolution & size
    - Natural img

Stereo可以考虑的网络

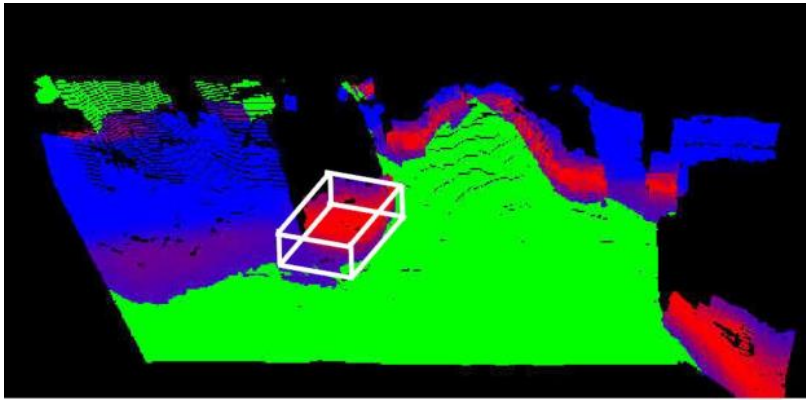
- 1、 Object Proposal

类似于Stixel World中的优化表达式

$$\mathbf{y}^* = \operatorname*{argmin}_{\mathbf{y}} E_{pc}(\mathbf{x},\mathbf{y}) + E_{fs}(\mathbf{x},\mathbf{y}) + E_{ht}(\mathbf{x},\mathbf{y}) + E_{ht-contr}(\mathbf{x},\mathbf{y})$$

物体脚下是底面

物体与周围3D空间有对比



里面有一个物体

物体高度是已知的

Stereo可以考虑的网络

- 2、 Object Detection
- Context info. 把bbox放大一点点，输入网络实现的
- HHA又包含了Stereo信息
- End-to-end