A Introduction to Text Detection Methods

liuxuebo@sensetime.com

ICDAR

ICDAR 2015 Task2:Focused Scene Text

img_1.png

```
22 249 113 286 "The"
142 249 287 286 "Photo"
326 245 620 297 "Specialists"
```

GT_1.txt

img_2.png

```
158 128 411 181 "Footpath"
443 128 501 169 "To"
64 200 363 243 "Colchester"
394 199 487 239 "and"
72 271 382 312 "Greenstead"
```

GT_2.txt

Deteval			
Method	Recall	Precision	Hmean
RRPN-4	87.31 %	95.19 %	91.08 %
MSRA_v1	88.58 %	93.67 %	91.06 %
Baidu IDL	87.11 %	92.83 %	89.88 %
XvBaoBao	85.37 %	91.46 %	88.31 %
CAS_HotEye	86.65 %	89.99 %	88.29 %
CTPN	82.98 %	92.98 %	87.69 %
SCUT-HCII	84.22 %	90.66 %	87.32 %
TextConv+Wor...	81.64 %	93.40 %	87.13 %
mser+rtree	88.02 %	85.96 %	86.97 %
TextConv+Wor...	81.02 %	93.38 %	86.76 %
SenseTime	84.22 %	88.84 %	86.47 %

ICDAR

ICDAR 2015 Task4:Incidental Scene Text

img_1.jpg

```
1001,249,1062,255,1062,270,1001,270,###
980,37,1041,27,1042,53,981,63,South
981,80,1074,75,1078,130,985,134,L14
994,185,1055,182,1056,205,994,202,#14-01
1000,217,1060,219,1060,241,999,238,#14-03
1043,26,1115,16,1115,43,1044,53,Tower
1058,187,1077,187,1077,206,1059,206,###
...
```

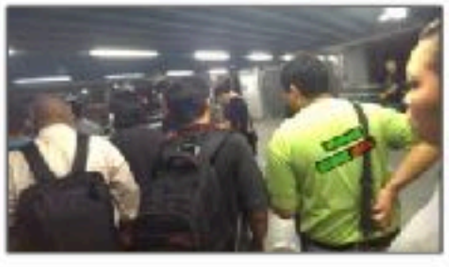
gt_img_1.txt

img_2.jpg


```
341,129,399,117,395,153,337,165,East
405,113,479,97,479,133,403,147,West
491,92,561,75,559,110,487,132,Line
338,248,433,242,423,319,328,325,Way
444,235,539,228,534,296,439,303,out
```

gt_img_2.txt

img_3.jpg



```
931,400,1040,348,1052,374,943,426,LEADERSHIP
913,460,1005,417,1016,444,924,487,THROUGH
1009,416,1076,382,1087,407,1020,442,###
```

gt_img_3.txt

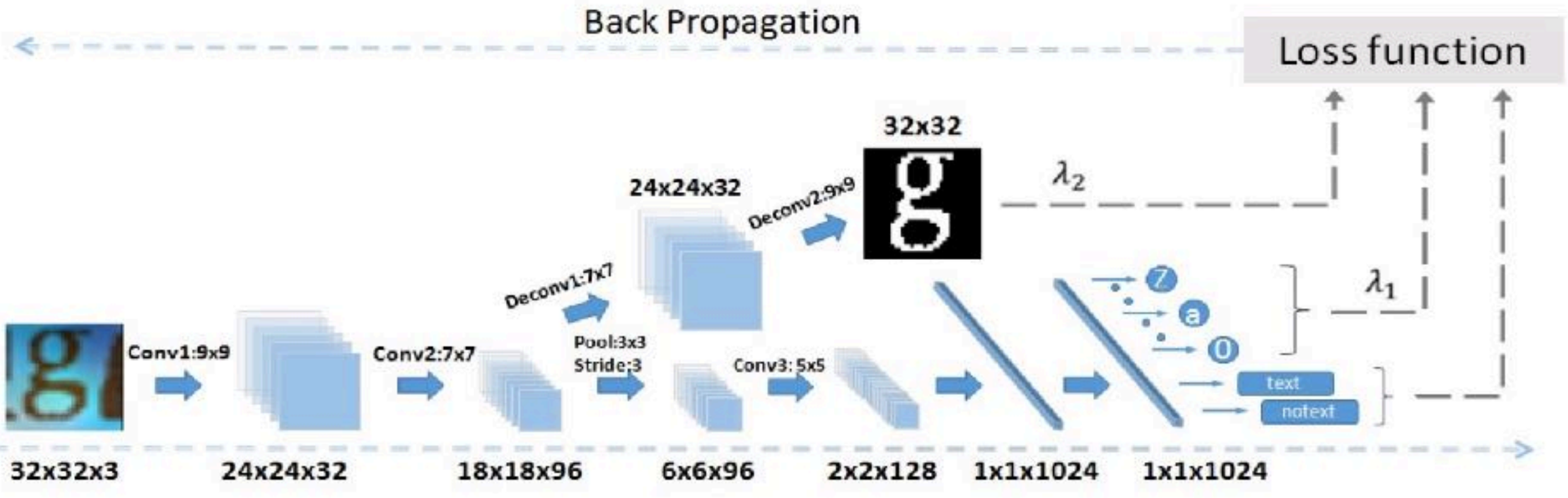
Method	Recall	Precision	Hmean
RRPN-4	77.13 %	83.52 %	80.20 %
MSRA_v1	74.10 %	85.22 %	79.27 %
RRPN-3	73.23 %	82.17 %	77.44 %
SRC-B-Machine...	69.86 %	86.11 %	77.14 %
SCUT_DMP_v2	76.89 %	77.00 %	76.95 %
Baidu IDL v2	72.75 %	77.41 %	75.01 %
HUST_Oriente...	74.15 %	69.03 %	71.49 %
SCUT_DMPNet	68.22 %	73.23 %	70.64 %
RRPN-2	72.65 %	68.53 %	70.53 %
Baidu IDL	68.22 %	71.53 %	69.84 %
Megvii-Image++	56.96 %	72.40 %	63.76 %
CTPN	51.56 %	74.22 %	60.85 %
RRPN-1	72.56 %	45.42 %	55.87 %
MCLAB_FCN	43.09 %	70.81 %	53.58 %

A Introduction to Text Detection Methods

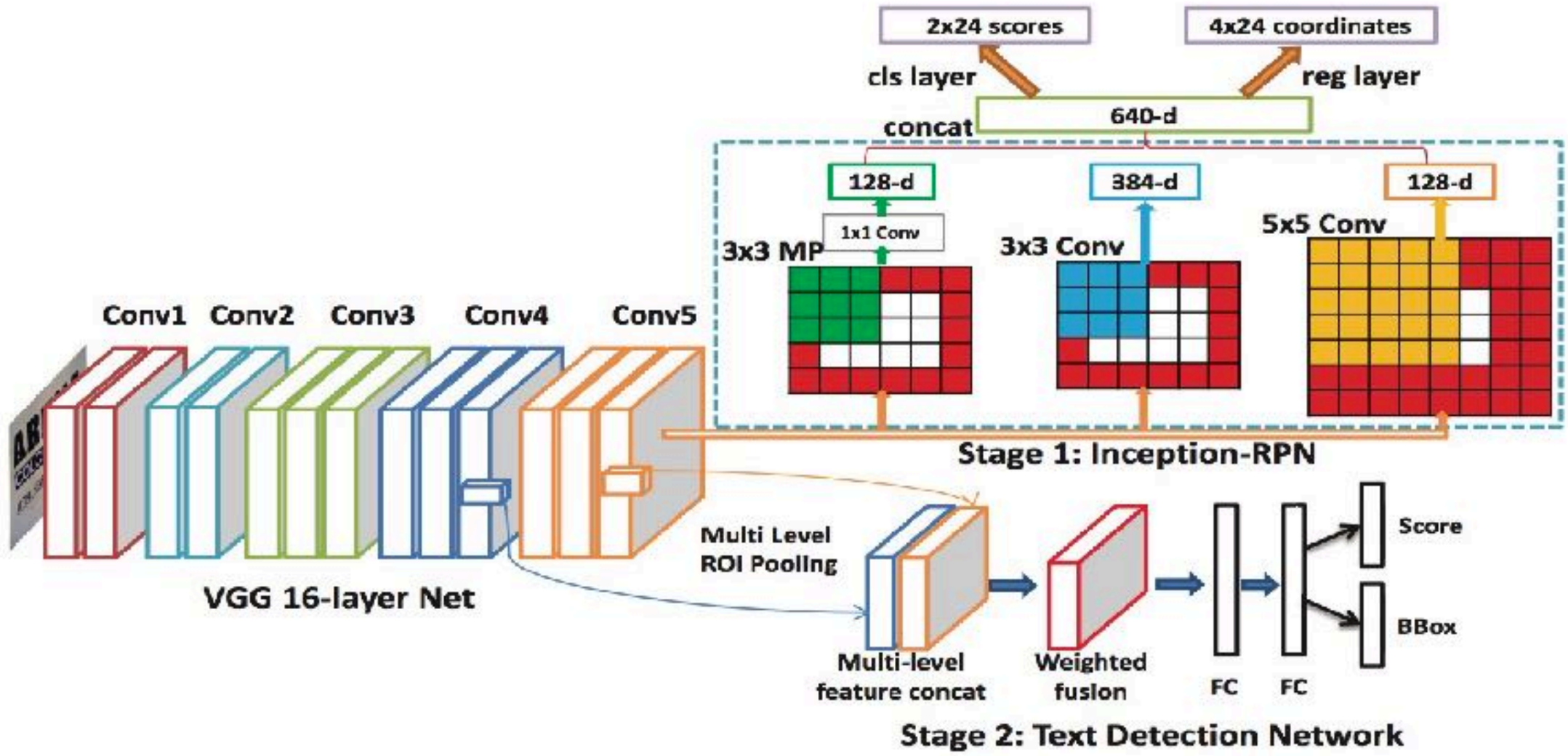
- SWT/MSER + CNN
  - Text-Attentional Convolutional Neural Network for Scene Text Detection
- RPN based (正负样本选择, 文字倾斜)
  - DeepText: A Unified Framework for Text Proposal Generation and Text Detection in Natural Images
  - TextBoxes: A Fast Text Detector with a Single Deep Neural Network
  - Detecting Text in Natural Image with Connectionist Text Proposal Network
- FCN based (heat map -> bounding box)
  - Accurate Text Localization in Natural Image with Cascaded Convolutional Text Network
  - Scene Text Detection via Holistic, Multi-Channel Prediction

Text-Attentional Convolutional Neural Network for Scene Text Detection

MSER获取character-level的proposal,输入CNN进行筛选
Result: ICDAR 2013: 82%



DeepText: A Unified Framework for Text Proposal Generation and Text Detection in Natural Images



DeepText: A Unified Framework for Text Proposal Generation and Text Detection in Natural Images

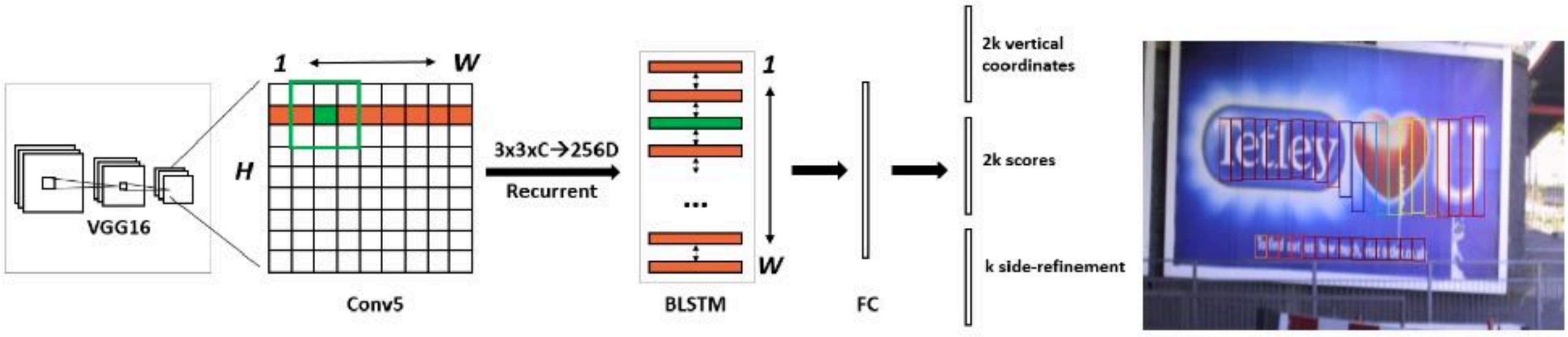
- ATC(ambiguous text category):为了解决文字检测中IOU小的proposal也可能是正样本的问题,新增加ambiguous text,即IOU在0.2-0.5之间的样本
- MLRP(multi-level region-of-interest pooling):在conv4和conv5都进行ROI pooling,将得到的feature concat,更好地handle multi-scale的text
- Iterative bounding box voting:将迭代不同次数的模型预测的bbox放在一起进行NMS,提高recall

Result: ICDAR 2013: 85%

Detecting Text in Natural Image with Connectionist Text Proposal Network

- 新的正负样本选取方法:anchor固定宽度,计算IOU时不考虑x方向,只考虑y方向的overlap,预测时只预测y方向的坐标以及x方向与边缘的距离
- 为了更好利用x方向的信息,图片经过CNN后每一行过一个LSTM,扩大x方向的receptive field

Result: ICDAR 2013: 87.7%



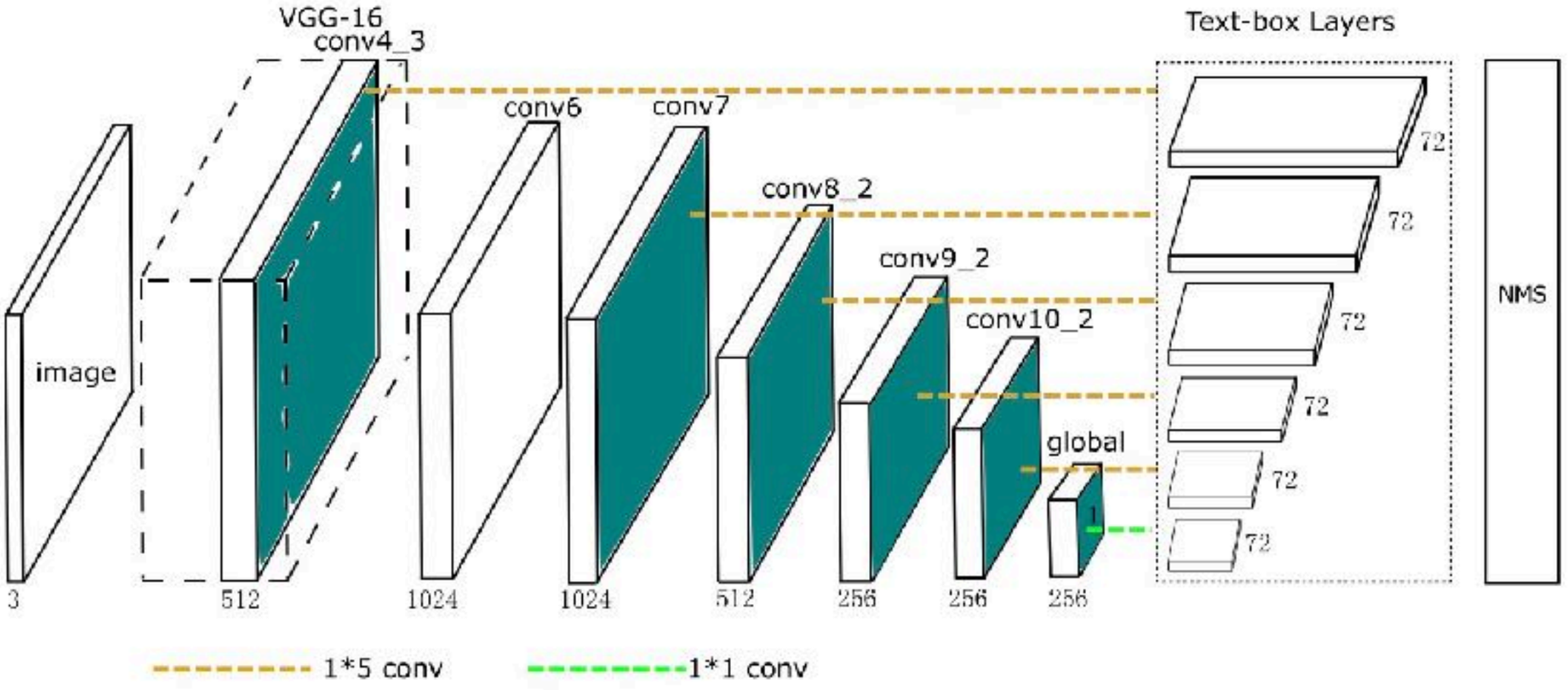
(a)

(b)

TextBoxes: A Fast Text Detector with a Single Deep Neural Network

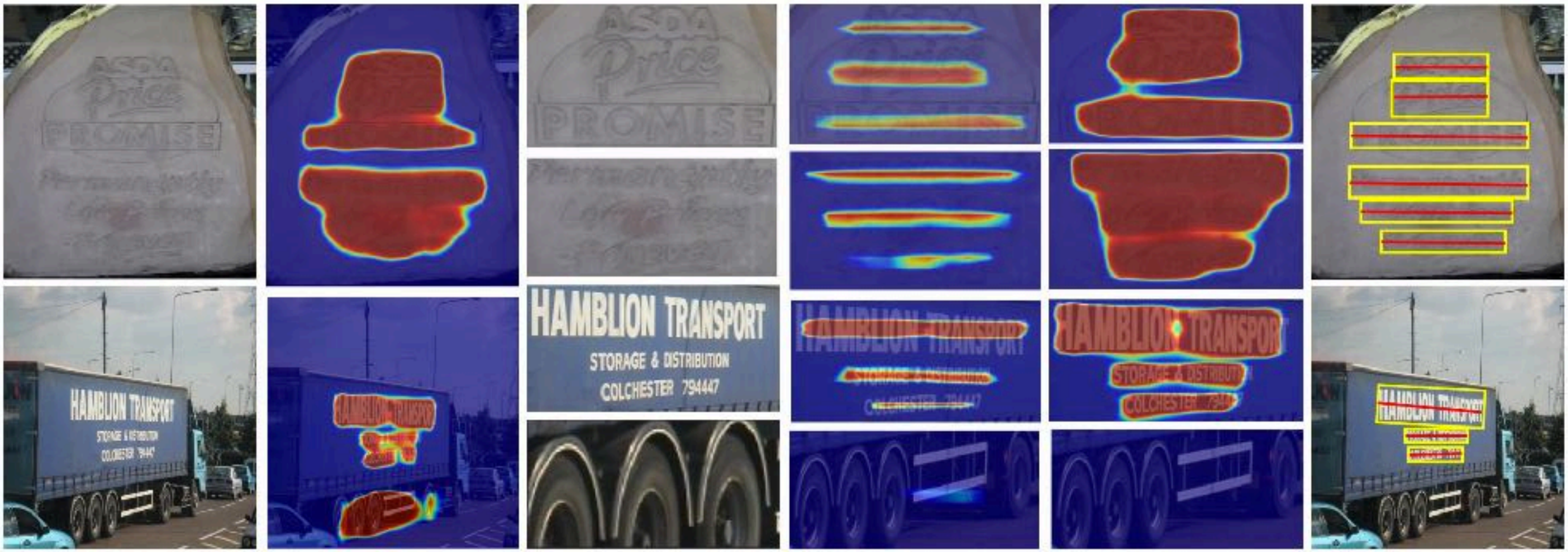
- 结构类似SSD
- 对预测结果进行文字识别,过滤掉score低的

Result: ICDAR 2013: 86%(输入多个scale的图片) 只输入300x300图片时81%,用时0.09s(TITAN X)



Accurate Text Localization in Natural Image with Cascaded Convolutional Text Network

两个CNN,第一个输入resize后的图片(更快),预测文字区域,第二个输入文字区域,预测每行文字中心位置和精准文字区域

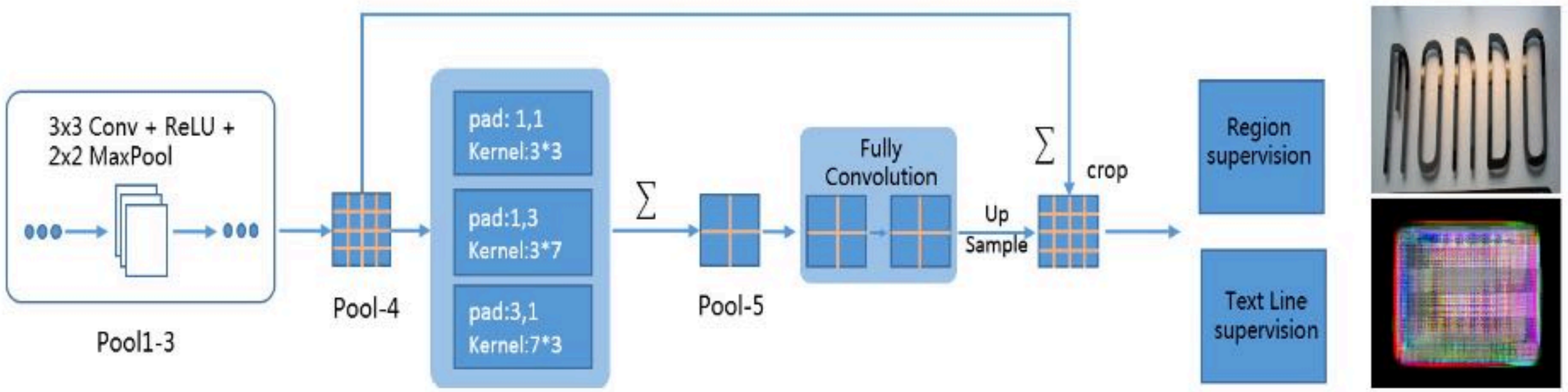


(a) Image (b) TR-Heatmap (c) Text Region (d) CL-Heatmap (e) TX-Heatmap (f) Text Line

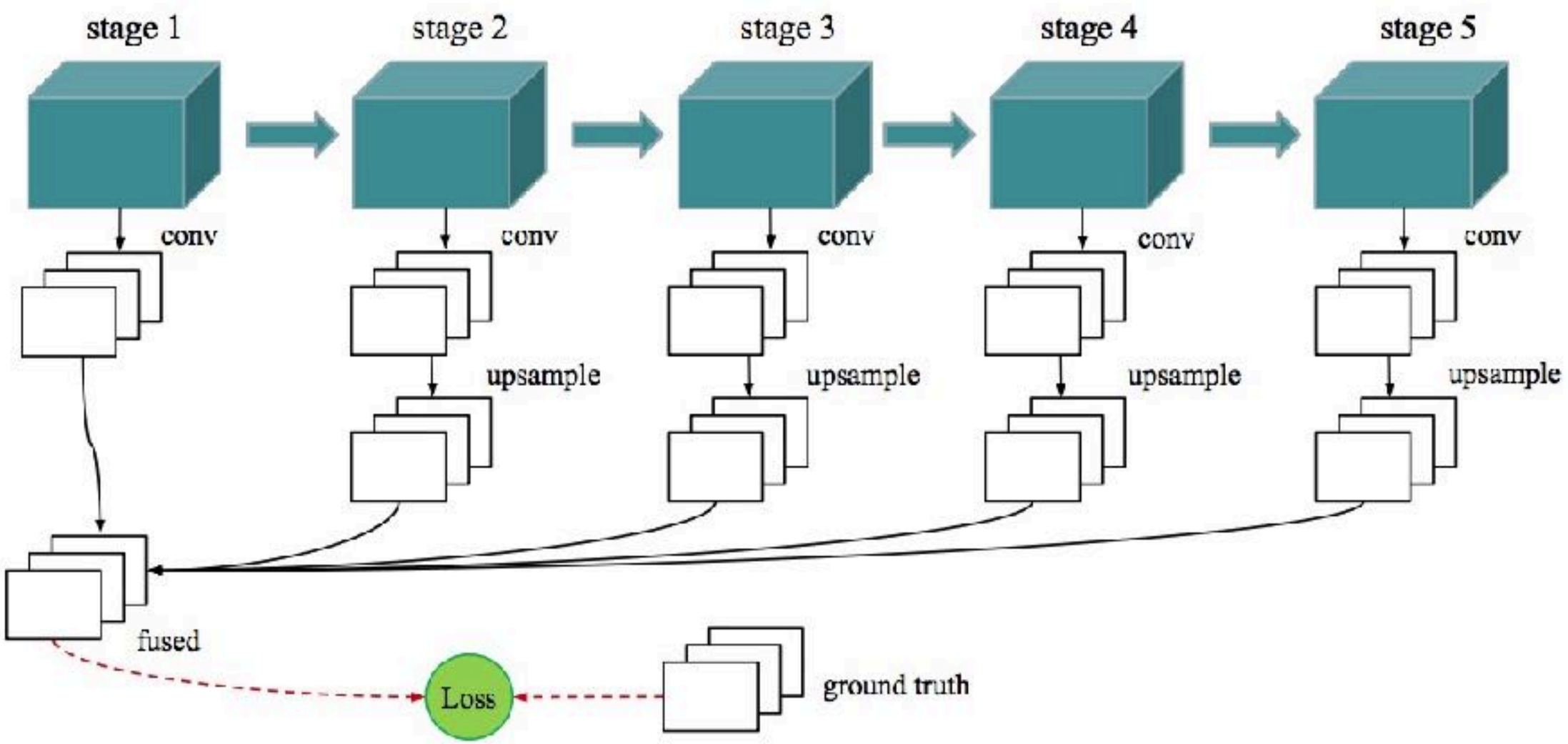
Accurate Text Localization in Natural Image with Cascaded Convolutional Text Network

- pool4和pool5之间有三个卷积层,为了更好地处理细长文字,kennel size分别为3x3,3x7,7x3
- text line中心为1,0.25H和0.75H为0的高斯分布

Result: ICDAR 2013: 86%



Scene Text Detection via Holistic, Multi-Channel Prediction

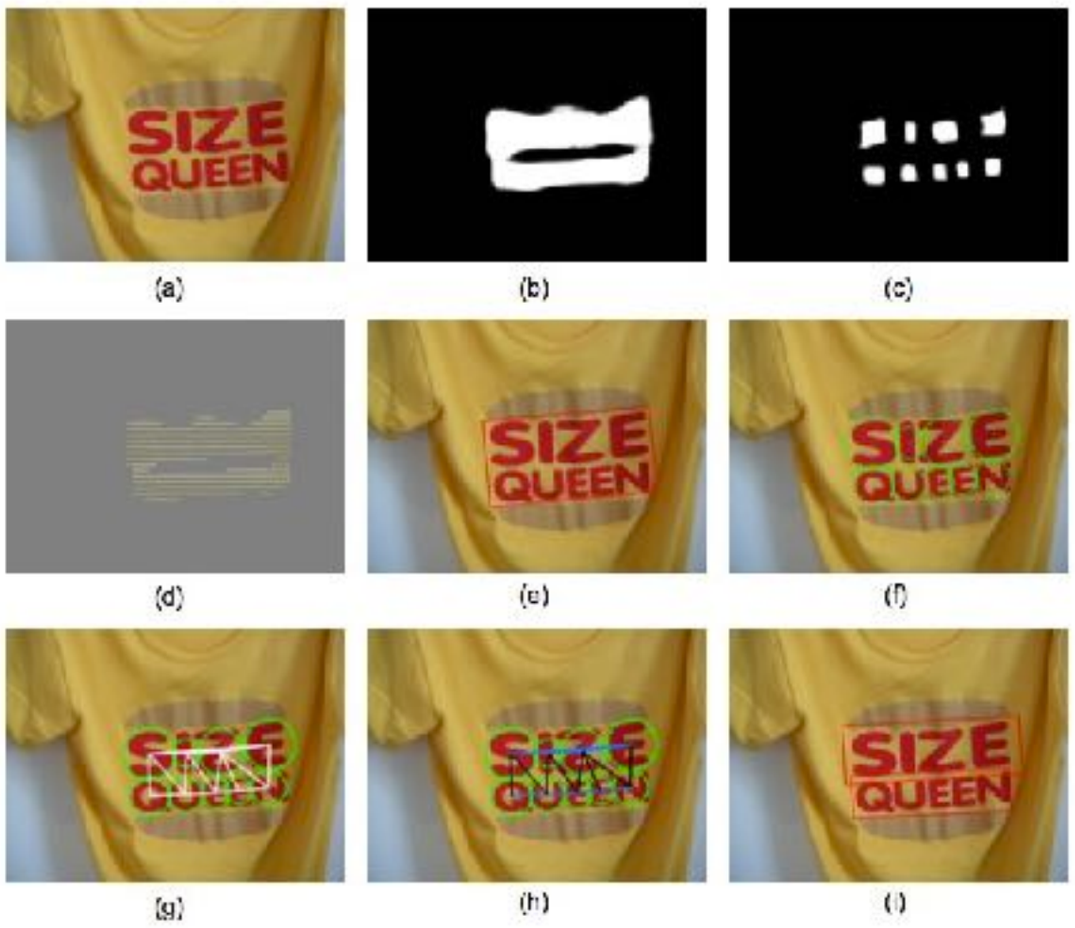


Scene Text Detection via Holistic, Multi-Channel Prediction

预测map of text region, map of character, map of linking orientation,为避免多个字符重合,每个字符的ground truth位置缩小一半,文字方向为(-2/π, 2/π),归一化到(0,1),文字方向的loss function为:

$$\Delta_o(\hat{\Theta}, \Theta; R, \mathbf{W}, \mathbf{w}) = \sum_{j=1}^{|R|} R_j(\sin(\pi|\hat{\Theta}_j - \Theta_j|)).$$

Scene Text Detection via Holistic, Multi-Channel Prediction



根据map of text region, map of character分别得到文字区域的位置和每个字符的位置,把每个字符当做一个点,使用delaunay triangulation算法将这些点连起来,delaunay triangulation算法用线段连接这些点产生尽可能多的三角形,这样得到一个图,图的每条边权重计算如下:

$$s(i,j) = \frac{2a(i,j)o(i,j)}{a(i,j)+o(i,j)}, \quad a(i,j) = \exp(-\frac{d^2(i,j)}{2D^2}), \quad o(i,j) = cos(\Lambda(\phi(i,j)-\psi(i,j)))$$

Scene Text Detection via Holistic, Multi-Channel Prediction

得到这个图之后,先用最大生成树算法得到这个图的生成树,然后确定text line的个数k和每个text line包含的字符:

$$S_{vm} = \sum_{i=1}^{K} \frac{\lambda_{i1}}{\lambda_{i2}},$$

where K is the number of clusters (text lines), λ_{i1} and λ_{i2} are the largest and second largest eigenvalues of the covariance matrix C_i. C_i is computed using the coordinates of the centers of characters the in the ith cluster. The optimal segmentation of text lines is achieved, when the value of function S_{vm} reaches its maximum.

Result: ICDAR 2013: 84.3% ICDAR 2015: 64.7%

THANK YOU!