Rotated BBox Detection

R2CNN
EAST
X-Line

周至公 2019 09 23

R^{2}CNN: Rotational Region CNN for Orientation Robust Scene Text Detection

Yingying Jiang, Xiangyu Zhu, Xiaobing Wang, Shuli Yang, Wei Li, Hua Wang, Pei Fu and Zhenbo Luo

Samsung R&D Institute China - Beijing
{yy.jiang, xiangyu.zhu, x0106.wang, shuli.yang, wei2016.li, hua00.wang, pei.fu, zb.luo}@samsung.com

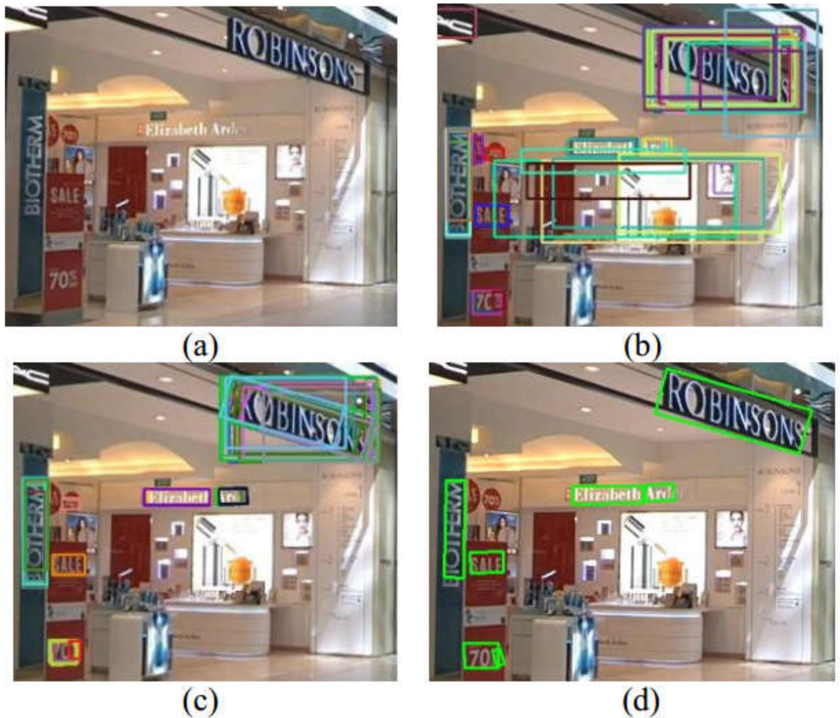
30 June 2017

Scene Text Detection

- Object detection: only one category
- Orientation Robust, inclined rectangle

(the angle target is not stable in some special points).

Use (x1, y1, x2, y2, h) to represent an inclined rectangle, (x1, y1) means the point at the left-top corner of the scene text. h means the height of the text.



(a) (b) (c) (d)

Fig. 1. The procedure of the proposed method R^{2}CNN. (a) Original input image; (b) text regions (axis-aligned bounding boxes) generated by RPN; (c) predicted axis-aligned boxes and inclined minimum area boxes (each inclined box is associated with an axis-aligned box, and the associated box pair is indicated by the same color); (d) detection result after inclined non-maximum suppression.

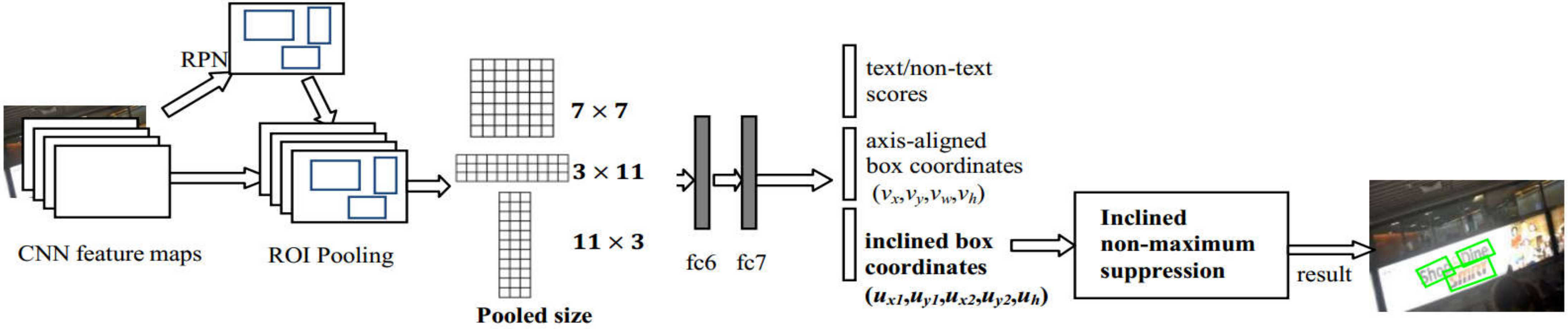

RPN
CNN feature maps
ROI Pooling
7 × 7
3 × 11
11 × 3
Pooled size
fc6
fc7
text/non-text scores
axis-aligned box coordinates (v_x, v_y, v_w, v_h)
inclined box coordinates $(u_{x1}, u_{y1}, u_{x2}, u_{y2}, u_h)$
Inclined non-maximum suppression
result

Fig.2. The network architecture of Rotational Region CNN (R^{2}CNN). The RPN is used for proposing axis-aligned bounding boxes that enclose the arbitrary-oriented texts. For each box generated by RPN, three ROIPoolings with different pooled sizes are performed and the pooled features are concatenated for predicting the text scores, axis-aligned box (v_x,v_y,v_w,v_h) and inclined minimum area box $(u_{x1},u_{y1},u_{x2},u_{y2},u_h)$. Then an inclined non-maximum suppression is conducted on the inclined boxes to get the final result.

Anchor: anchor scales **in Faster RCNN** are (8,16,32), in **R2CNN** (4,8,16) or (4,8,16,32). keep other settings of RPN the same as Faster. 小目标text
ROIPoolings: 7 x 7 , add 11 × 3 and 3 × 11 不同方向.
Data augmentation: rotate image at the following angles (-90, -75, -60, -45, -30, -15, 0, 15, 30, 45, 60, 75, 90)

$$\begin{aligned} L(p,t,v,v^*,u,u^*) = {} & L_{\text{cls}}(p,t) \\ & + \lambda_1 t \textstyle\sum_{i\in\{x,y,w,h\}} L_{reg}(v_i, v_i^*) \\ & + \lambda_2 t \textstyle\sum_{i\in\{x1,y1,x2,y2,h\}} L_{reg}(u_i, u_i^*) \end{aligned}$$

$$L_{reg}(w, w^*) = \text{smooth}_{\text{L1}}(w - w^*)$$

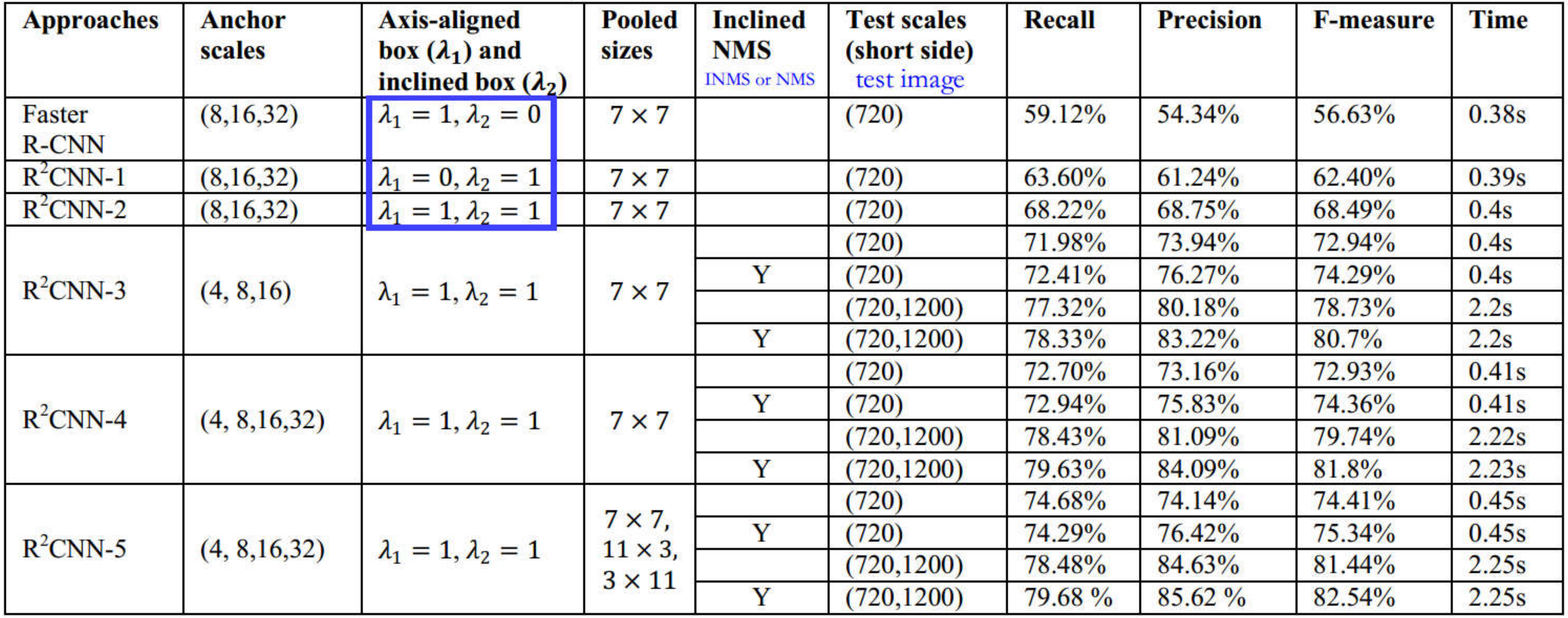
$$\text{smooth}_{L1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases}$$

倾斜NMS(INMS)

- 基本步骤(rbox代表旋转矩形框)
- 1.对输出的检测框rbox按照得分进行降序排序rbox_lists;
- 2.依次遍历上述的rbox_lists．具体的做法是：将当前遍历的rbox与剩余的rbox进行交集运算得到相应的相交点集合，并根据判断相交点集合组成的凸边形的面积，计算每两个rbox的IOU；对于大于设定阈值的rbox进行滤除，保留小于设定阈值的rbox;
- 3.得到最终的检测框

- 其他NMS变体： https://zhuanlan.zhihu.com/p/50126479

Table 1. Results of R^{2} CNN under different settings on ICDAR 2015.



Approaches	Anchor scales	Axis-aligned box (λ_1) and inclined box (λ_2)	Pooled sizes	Inclined NMS INMS or NMS	Test scales (short side) test image	Recall	Precision	F-measure	Time
Faster R-CNN	(8,16,32)	$\lambda_1 = 1, \lambda_2 = 0$	7×7		(720)	59.12%	54.34%	56.63%	0.38s
R^{2}CNN-1	(8,16,32)	$\lambda_1 = 0, \lambda_2 = 1$	7×7		(720)	63.60%	61.24%	62.40%	0.39s
R^{2}CNN-2	(8,16,32)	$\lambda_1 = 1, \lambda_2 = 1$	7×7		(720)	68.22%	68.75%	68.49%	0.4s
R^{2}CNN-3	(4, 8,16)	$\lambda_1 = 1, \lambda_2 = 1$	7×7		(720)	71.98%	73.94%	72.94%	0.4s
				Y	(720)	72.41%	76.27%	74.29%	0.4s
					(720,1200)	77.32%	80.18%	78.73%	2.2s
				Y	(720,1200)	78.33%	83.22%	80.7%	2.2s
R^{2}CNN-4	(4, 8,16,32)	$\lambda_1 = 1, \lambda_2 = 1$	7×7		(720)	72.70%	73.16%	72.93%	0.41s
				Y	(720)	72.94%	75.83%	74.36%	0.41s
					(720,1200)	78.43%	81.09%	79.74%	2.22s
				Y	(720,1200)	79.63%	84.09%	81.8%	2.23s
R^{2}CNN-5	(4, 8,16,32)	$\lambda_1 = 1, \lambda_2 = 1$	7×7, 11×3, 3×11		(720)	74.68%	74.14%	74.41%	0.45s
				Y	(720)	74.29%	76.42%	75.34%	0.45s
					(720,1200)	78.48%	84.63%	81.44%	2.25s
				Y	(720,1200)	79.68 %	85.62 %	82.54%	2.25s

- learning the additional axis-aligned box could help the detection of the inclined box.
- RPN is competent for generating text regions in the form of axis-aligned boxes for arbitrary-oriented texts
- small anchors could improve the scene text detection performance

EAST: An Efficient and Accurate Scene Text Detector

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang

Megvii Technology Inc., Beijing, China
{zxy, yaocong, wenhe, wangyuzhi, zsc, hwr, liangjiajun}@megvii.com

10 Jul 2017

End to End

此图凸显出结构simple

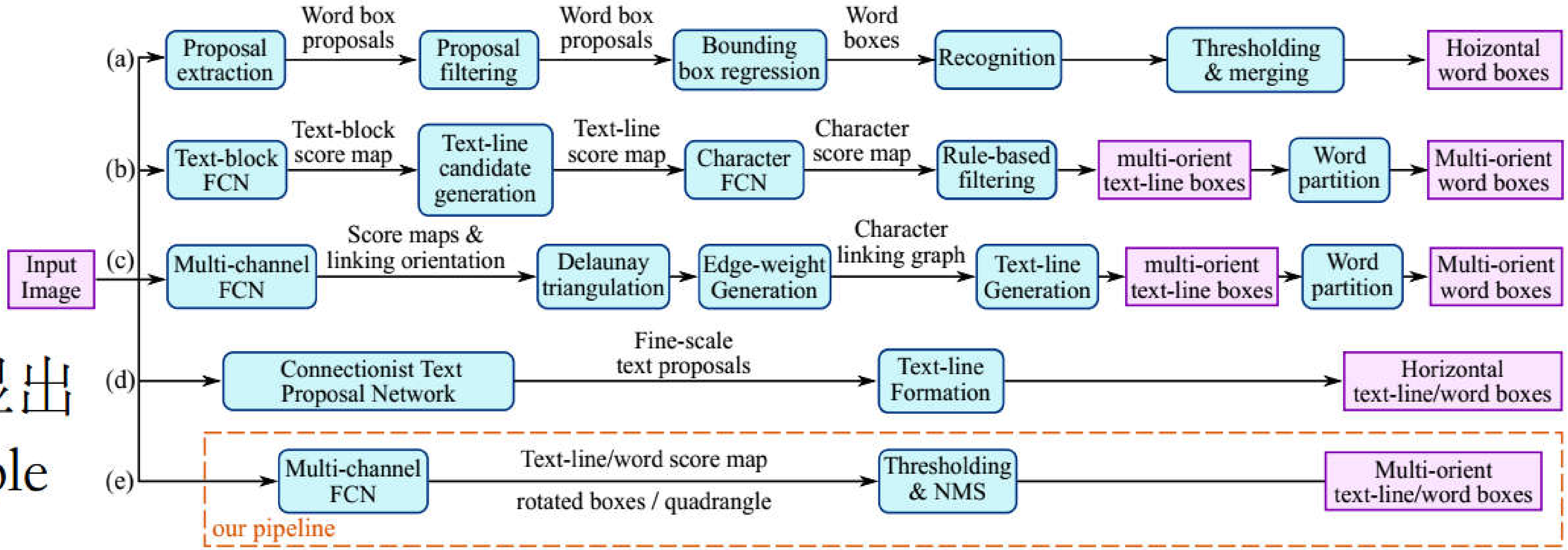
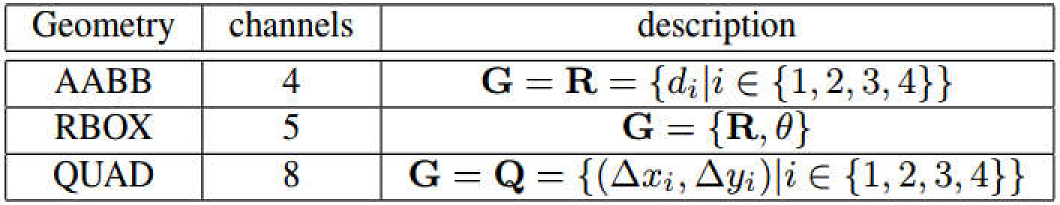


Figure 2. Comparison of pipelines of several recent works on scene text detection: (a) Horizontal word detection and recognition pipeline proposed by Jaderberg *et al.* [12]; (b) Multi-orient text detection pipeline proposed by Zhang *et al.* [48]; (c) Multi-orient text detection pipeline proposed by Yao *et al.* [41]; (d) Horizontal text detection using CTPN, proposed by Tian *et al.* [34]; (e) Our pipeline, which eliminates most intermediate steps, consists of only two stages and is much simpler than previous solutions.

Label Generation

Score map
先缩较长的一对边(平均长)，后缩短边
顶点pi向内侧移动0.3ri，ri为pi相连的较短边，
保守

box



Geometry	channels	description
AABB	4	$\mathbf{G} = \mathbf{R} = \{d_i \mid i \in \{1,2,3,4\}\}$
RBOX	5	$\mathbf{G} = \{\mathbf{R}, \theta\}$
QUAD	8	$\mathbf{G} = \mathbf{Q} = \{(\Delta x_i, \Delta y_i) \mid i \in \{1,2,3,4\}\}$

Table 1. Output geometry design

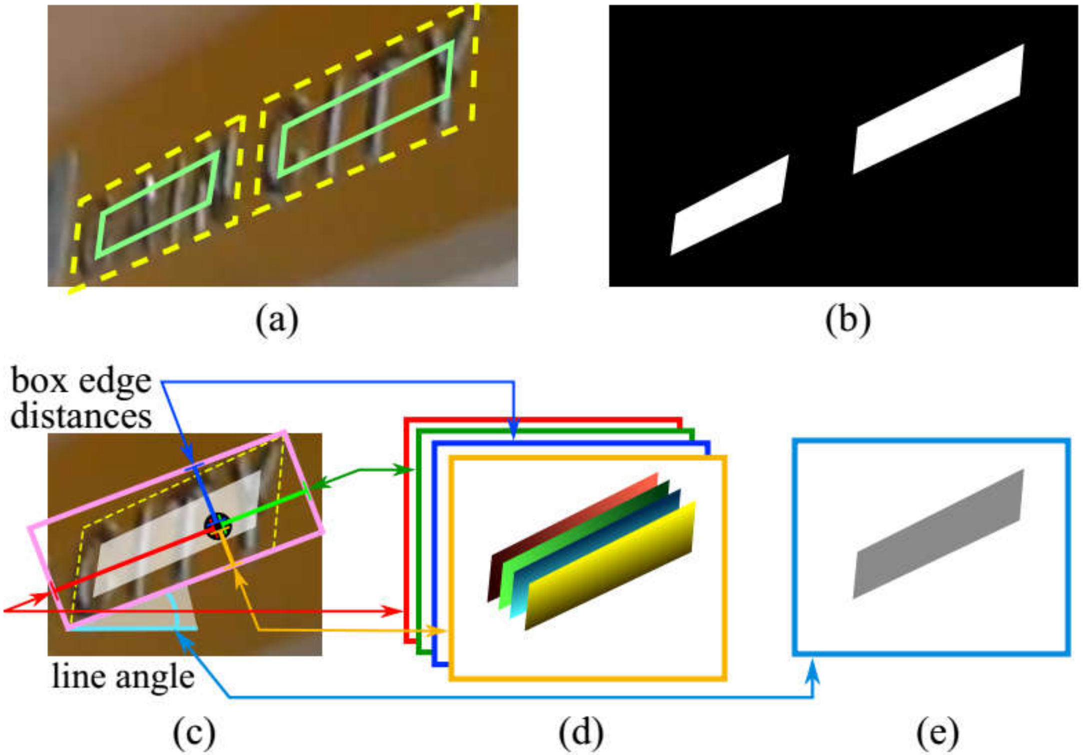


Figure 4. Label generation process: (a) Text quadrangle (yellow dashed) and the shrunk quadrangle (green solid); (b) Text score map; (c) RBOX geometry map generation; (d) 4 channels of distances of each pixel to rectangle boundaries; (e) Rotation angle.

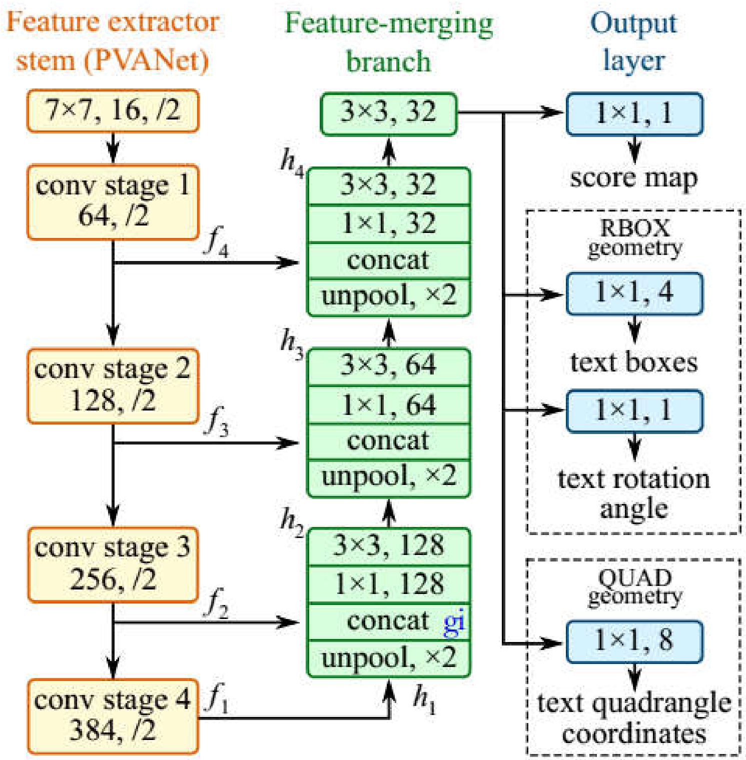


Figure 3. Structure of our text detection FCN.

images. The model is a fully-convolutional neural network adapted for text detection that outputs dense per-pixel predictions of words or text lines. This eliminates intermediate steps such as candidate proposal, text region formation and word partition. The post-processing steps only include thresholding and NMS on predicted geometric shapes. The detector is named as **EAST**, since it is an **E**fficient and **A**ccuracy **S**cene **T**ext detection pipeline.

Loss

L = Ls + λgLg losses for the **score map** and the **geometry**, set λg to 1

- **Loss for Score Map** balanced cross-entropy

$$L_s = \text{balanced-xent}(\hat{\mathbf{Y}}, \mathbf{Y}^*) = -\beta \mathbf{Y}^* \log \hat{\mathbf{Y}} - (1-\beta)(1-\mathbf{Y}^*)\log(1-\hat{\mathbf{Y}}) \quad (5)$$

$$\beta = 1 - \frac{\sum_{y^* \in \mathbf{Y}^*} y^*}{|\mathbf{Y}^*|}.$$

|Y*|表示的是所有的像素个数

- **Loss for Geometries** should be scale-invariant

RBOX *L*g = *L*AABB + *λθLθ*

$$L_\theta(\hat{\theta}, \theta^*) = 1 - \cos(\hat{\theta} - \theta^*).$$

$$L_{\text{AABB}} = -\log \text{IoU}(\hat{\mathbf{R}}, \mathbf{R}^*) = -\log \frac{|\hat{\mathbf{R}} \cap \mathbf{R}^*|}{|\hat{\mathbf{R}} \cup \mathbf{R}^*|}$$

$$w_{\mathbf{i}} = \min(\hat{d}_2, d_2^*) + \min(\hat{d}_4, d_4^*)$$

$$h_{\mathbf{i}} = \min(\hat{d}_1, d_1^*) + \min(\hat{d}_3, d_3^*)$$

$$|\hat{\mathbf{R}} \cup \mathbf{R}^*| = |\hat{\mathbf{R}}| + |\mathbf{R}^*| - |\hat{\mathbf{R}} \cap \mathbf{R}^*|.$$

QUAD

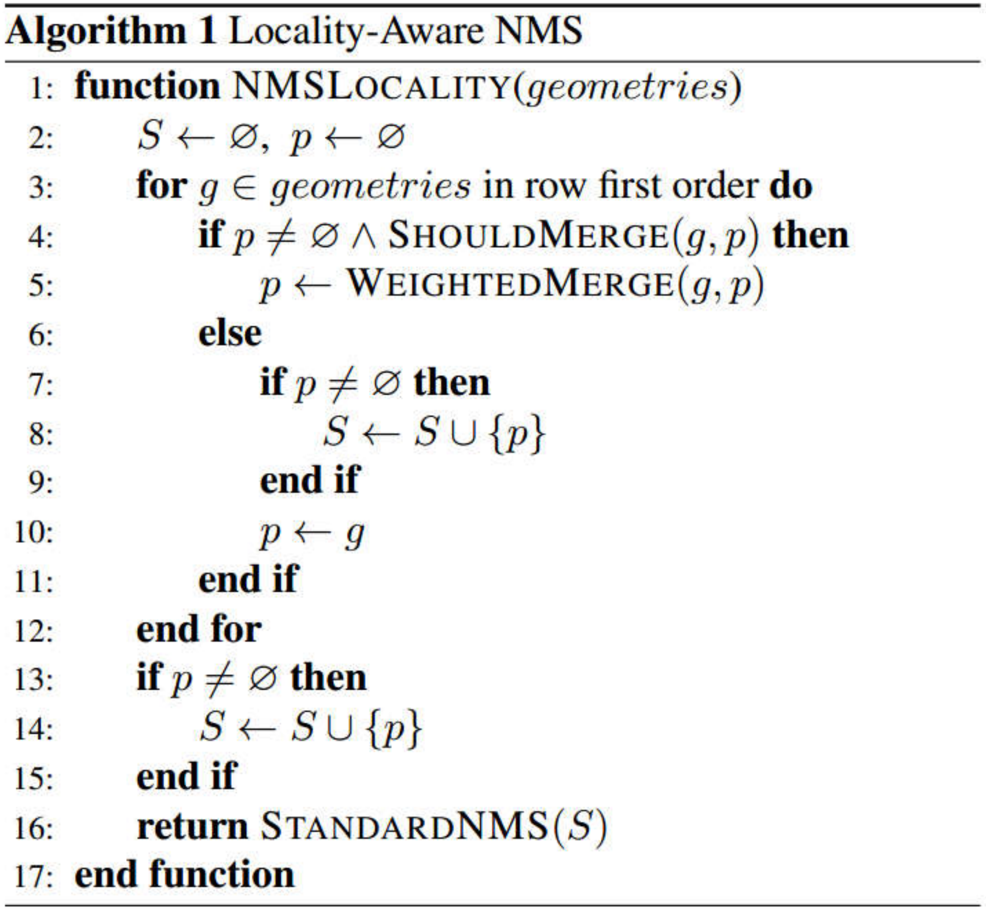
$$L_g = L_{\text{QUAD}}(\hat{\mathbf{Q}}, \mathbf{Q}^*) = \min_{\tilde{\mathbf{Q}} \in P_{\mathbf{Q}^*}} \sum_{\substack{c_i \in \mathrm{C}_{\mathbf{Q}}, \\ \tilde{c}_i \in \mathrm{C}_{\tilde{\mathbf{Q}}}}} \frac{\text{smoothed}_{L1}(c_i - \tilde{c}_i)}{8 \times N_{\mathbf{Q}^*}}$$

$$\mathrm{C}_{\mathbf{Q}} = \{x_1, y_1, x_2, y_2, \ldots, x_4, y_4\}$$

$$N_{\mathbf{Q}^*} = \min_{i=1}^{4} D(p_i, p_{(i \bmod 4)+1})$$

Locality-Aware NMS

- Merge the geometries row by row



Algorithm 1 Locality-Aware NMS

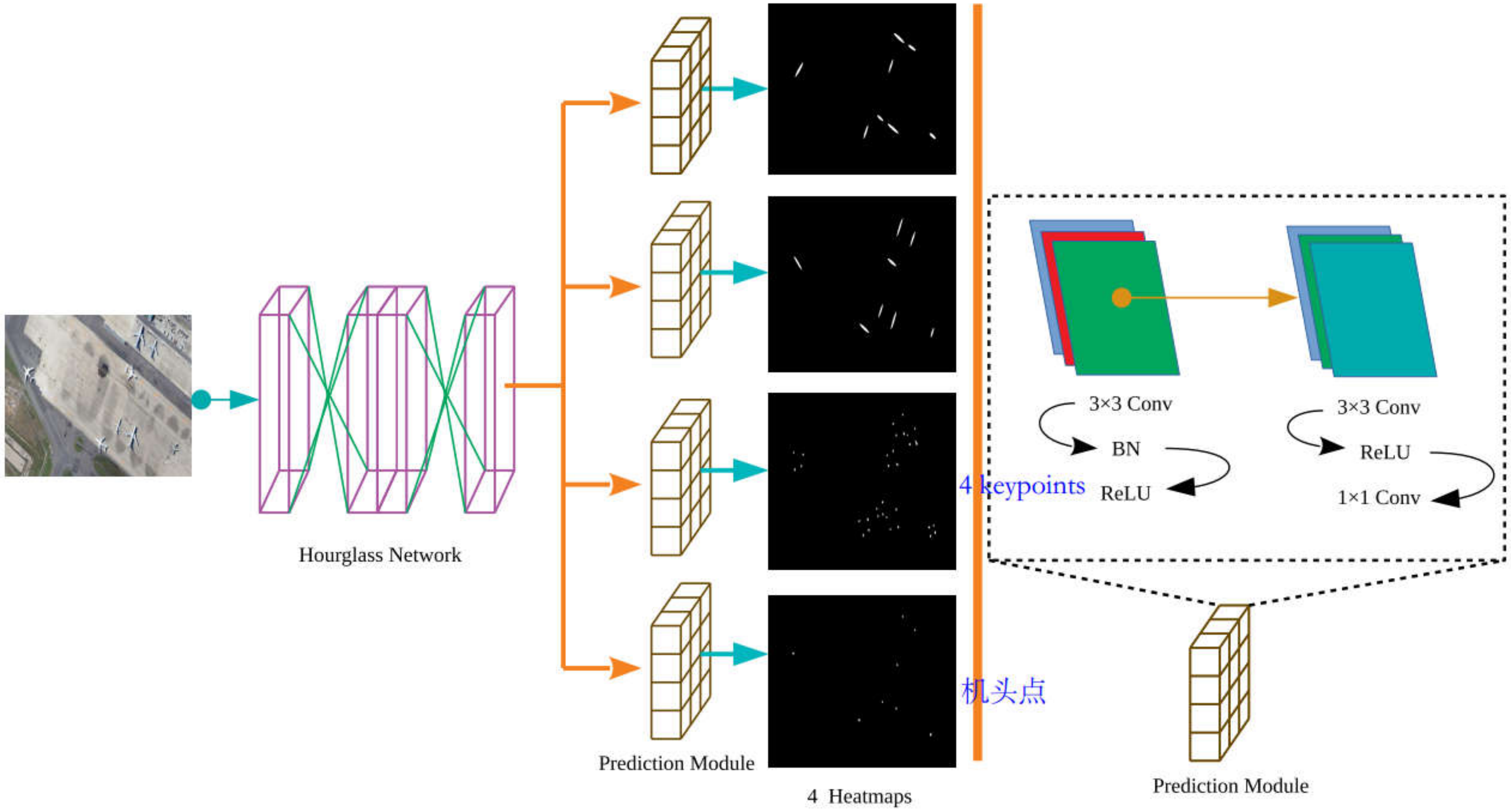
```
function NMSLocality($geometries$)
    $S \leftarrow \varnothing,\ p \leftarrow \varnothing$
    for $g \in geometries$ in row first order do
        if $p \neq \varnothing \wedge$ ShouldMerge$(g, p)$ then
            $p \leftarrow$ WeightedMerge$(g, p)$
        else
            if $p \neq \varnothing$ then
                $S \leftarrow S \cup \{p\}$
            end if
            $p \leftarrow g$
        end if
    end for
    if $p \neq \varnothing$ then
        $S \leftarrow S \cup \{p\}$
    end if
    return StandardNMS$(S)$
end function
```

X-LineNet: Detecting Aircraft in Remote Sensing Images by a pair of Intersecting Line Segments

Haoran Wei[a,b], Wang Bing[a,b], Zhang Yue[b]

[a] *University of Chinese Academy of Sciences, Beijing, China*
[b] *Institute of Electrics, Chinese Academy of Sciences, Beijing, China*

29 Jul 2019


Hourglass Network
Prediction Module
4 Heatmaps
4 keypoints
机头点
3×3 Conv
BN
ReLU
3×3 Conv
ReLU
1×1 Conv
Prediction Module