Introduction of Scene Text Detection

李鑫

2020/12/01

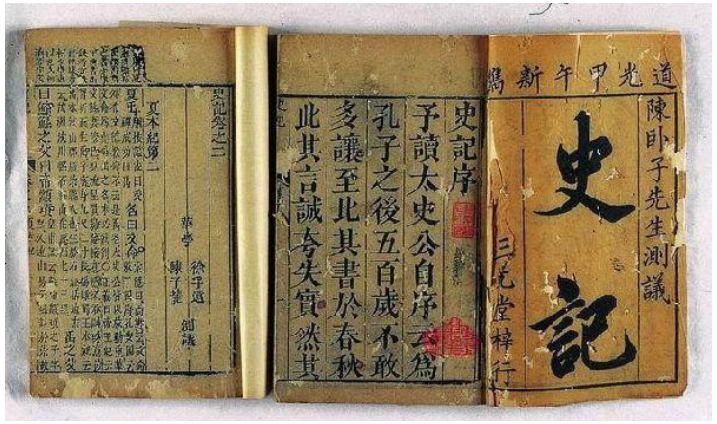
Outline

- Background and Introduction
- Datasets and Related Knowledge
- Conventional Methods
- Deep Learning Methods
- Conclusion and Outlook

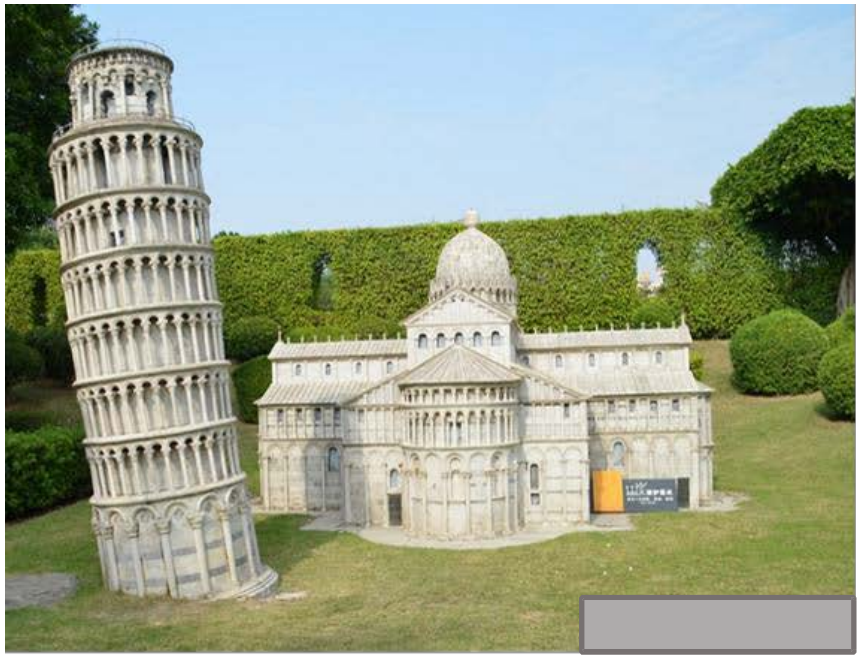
Text as a Carrier of High Level Semantics

Text is an invention of humankind that

- carries rich and precise high level semantics
- conveys human thoughts and emotions



Text as a Cue in Visual Recognition



(a)

(b)

Text as a Cue in Visual Recognition

Text is complementary to other visual cues, such as contour, color and texture

(a)

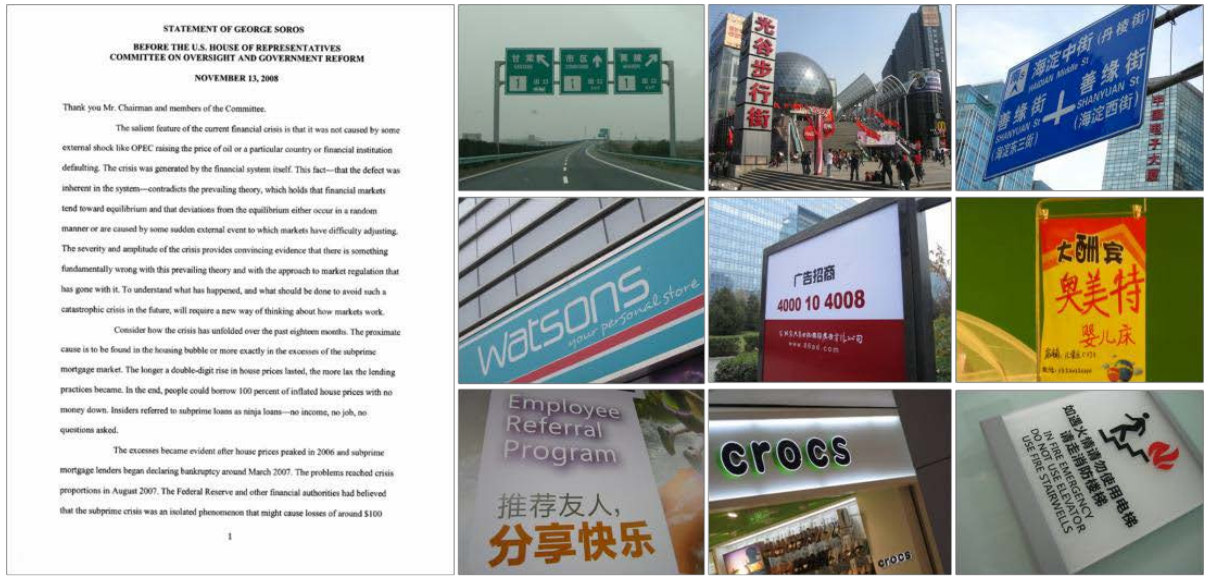
(b)

Problem Definition

Scene text detection is the process of predicting the presence of text and localizing each instance (if any), usually at word or line level, in natural scenes

Challenges

Traditional OCR *vs.* Scene Text Detection and Recognition



- clean background vs. cluttered background
- regular font vs. various fonts
- plain layout vs. complex layouts
- monotone color vs. different colors

Challenges

Diversity of scene text:

different colors, scales, orientations, fonts, languages…

Challenges

Complexity of background:

elements like signs, fences, bricks, and grasses are virtually indistinguishable from true text

Challenges

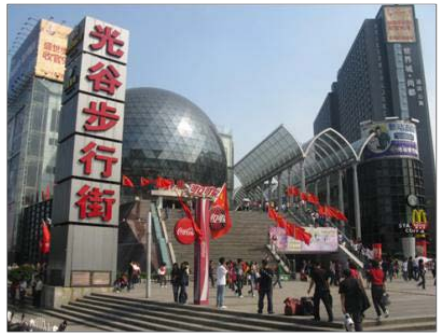
Various interference factors:

noise, blur, non-uniform illumination, low resolution, partial occlusion…

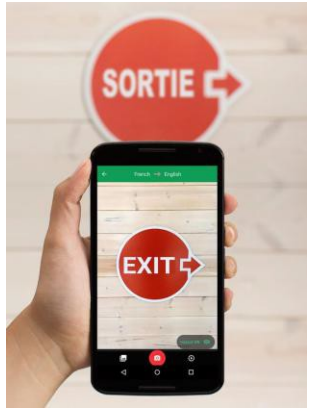
Applications

Card Recognition

Product Search



Geo-location



Instant Translation

Self-driving Car

Industry Automation

Outline

- Background and Introduction
- Datasets and Related Knowledge
- Conventional Methods
- Deep Learning Methods
- Conclusion and Outlook

ICDAR 2013

- 485 images containing text in a variety of colors and fonts on different backgrounds
- mostly horizontal text

http://rrc.cvc.uab.es/?ch=2&com=introduction

ICDAR 2015

- 1500 images in total, with text instances of different orientations
- incidental scene text: without the user having taken any specific prior action to cause its appearance or improve its positioning / quality in the frame
- only English text

http://rrc.cvc.uab.es/?ch=4&com=introduction

MSRA-TD500

- 500 images in total, with text instances of different orientations
- both Chinese and English text
- adopted by IAPR as official dataset

http://www.iapr-tc11.org/mediawiki/index.php/MSRA_Text_Detection_500_Database_(MSRA-TD500)

ICDAR 2015

Ranking Table

Description Paper Source Code

Date	Method	Recall	Precision	Hmean
2020-07-31	TextFuseNet	90.56%	93.96%	92.23%
2020-01-22	TH	89.46%	94.03%	91.69%
2019-08-13	JDAI	90.85%	92.50%	91.67%
2019-03-11	Alibaba-PAI V2	89.41%	93.32%	91.32%
2019-08-08	Eleme-AI	89.31%	93.03%	91.13%
2018-09-07	Sogou_MM	90.03%	92.21%	91.11%
2018-07-03	Baidu VIS v2	88.11%	94.04%	90.98%
2019-05-14	ArtDet	88.40%	92.91%	90.60%
2018-01-31	Alibaba-PAI	87.34%	93.84%	90.47%
2018-01-22	FOTS	87.92%	91.85%	89.84%
2018-11-15	Pixel-Anchor(Multiscale)	86.95%	92.28%	89.54%
2018-03-05	HoText_v1	83.58%	96.34%	89.51%
2019-08-07	CM-CV&AR	87.53%	91.49%	89.47%
2020-08-12	RRPN++ (single scale)	87.19%	91.84%	89.45%

- very popular benchmark
- with text instances of different orientations
- Only English text

http://rrc.cvc.uab.es/?ch=4&com=introduction

SCUT-CTW1500

- First dataset about Curved text
- 10k text annotations in 1,500 images (1000 for training and 500 for testing).
- facilitate a new research direction for the scene text community

Total-Text

- 1555 images with different text orientations: Horizontal, Multi-Oriented, and Curved
- facilitate a new research direction for the scene text community

https://github.com/cs-chan/Total-Text-Dataset

MLT-2017

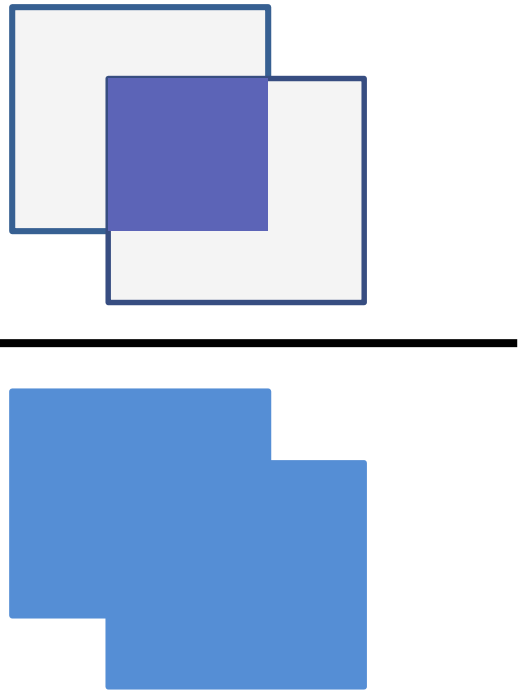
- multilingual dataset, 9 languages: Chinese, Japanese, Korean, English, French, Arabic, Italian, German and Indian
- for text detection, script identification and recognition

http://rrc.cvc.uab.es/?ch=8&com=introduction

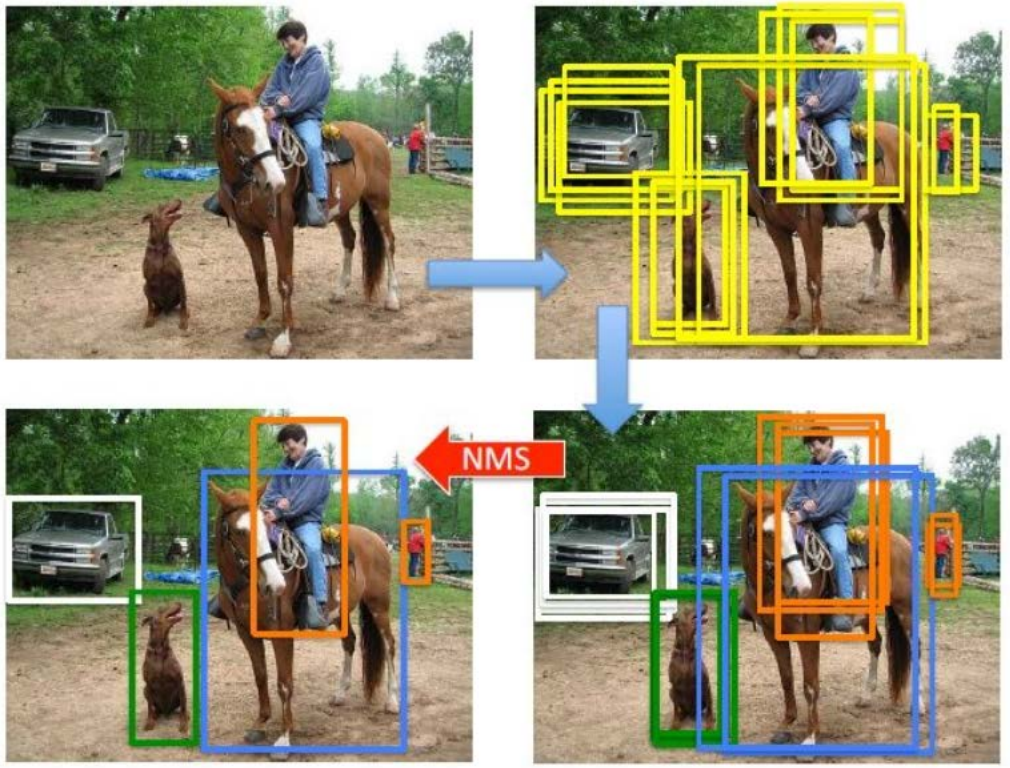
IOU & NMS

$$IOU = \frac{DetectionResult \bigcap GroundTruth}{DetectionResult \bigcup GroundTruth}$$

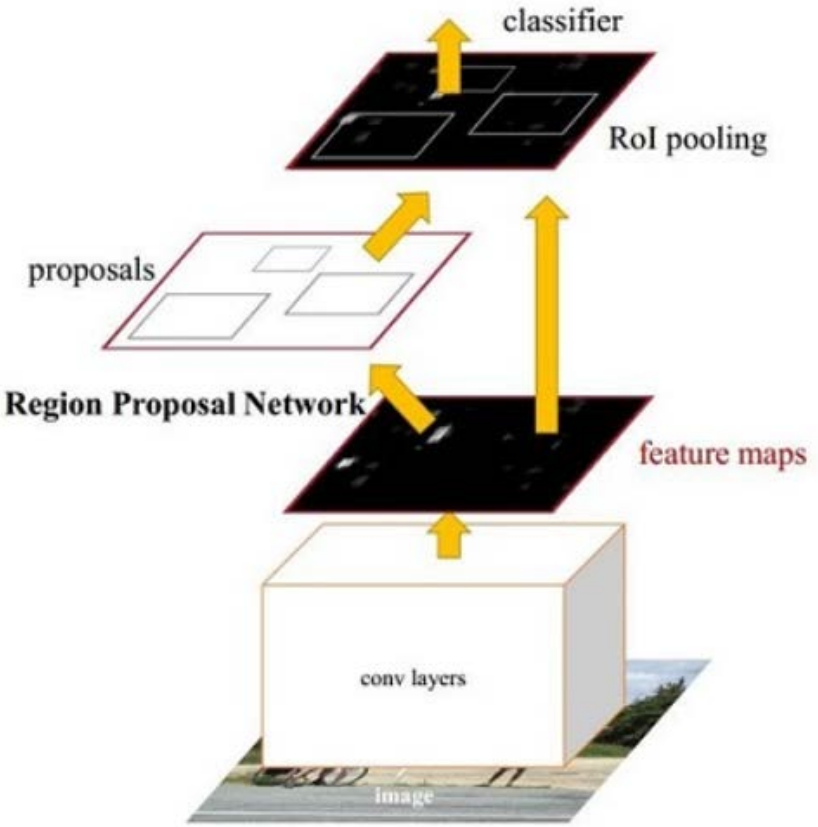
IoU=

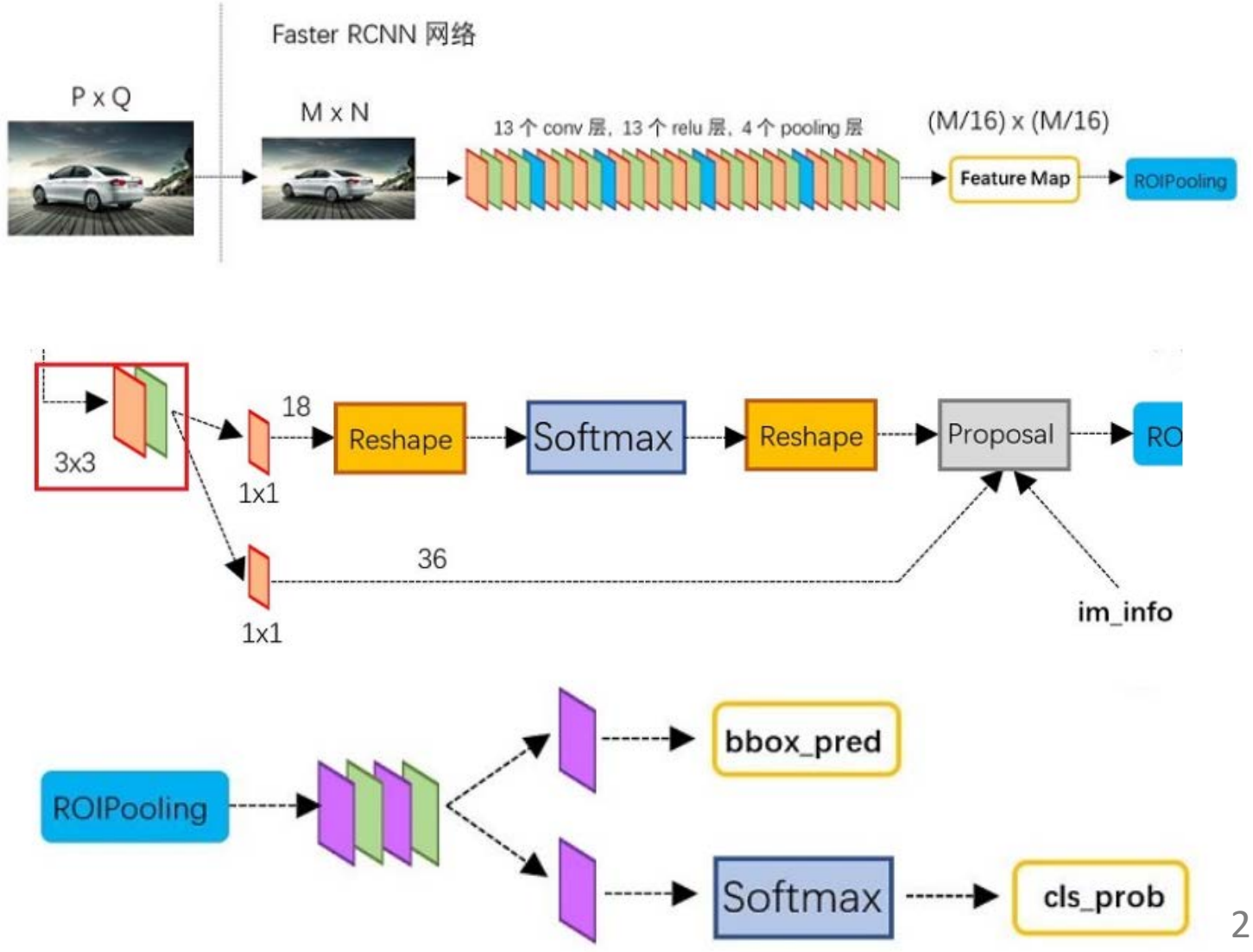


Non-Maximum Suppression

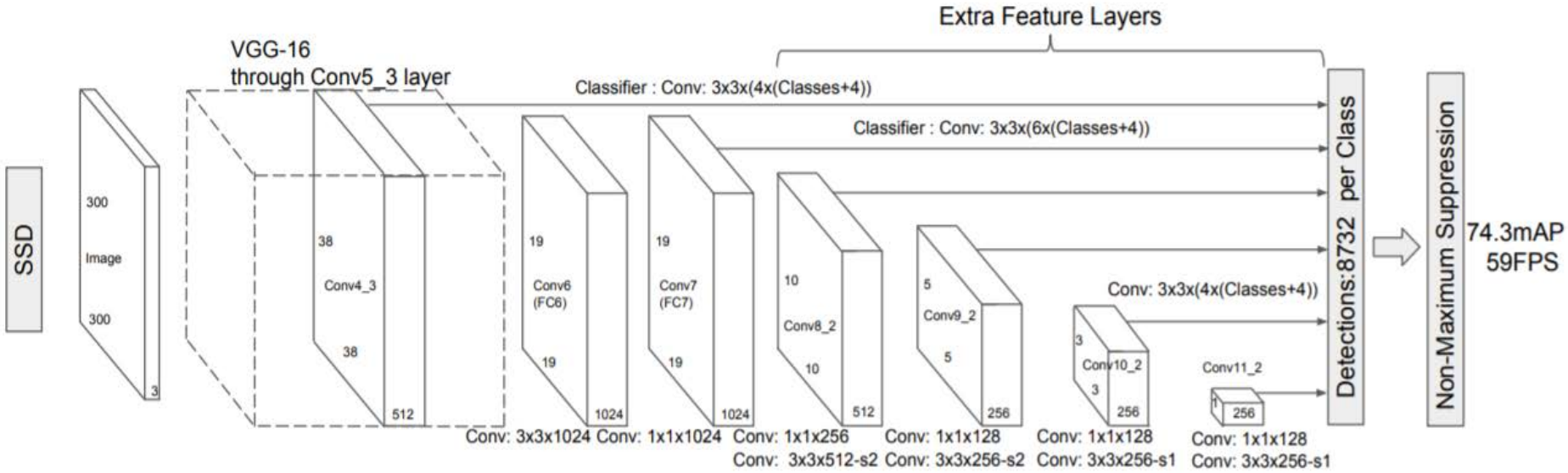


Faster RCNN


classifier
RoI pooling
proposals
Region Proposal Network
feature maps
conv layers
image


Faster RCNN 网络
P x Q
M x N
13 个 conv 层，13 个 relu 层，4 个 pooling 层
(M/16) x (M/16)
Feature Map
ROIPooling
3x3
1x1
18
Reshape
Softmax
Reshape
Proposal
RO
1x1
36
im_info
ROIPooling
bbox_pred
Softmax
cls_prob

SSD

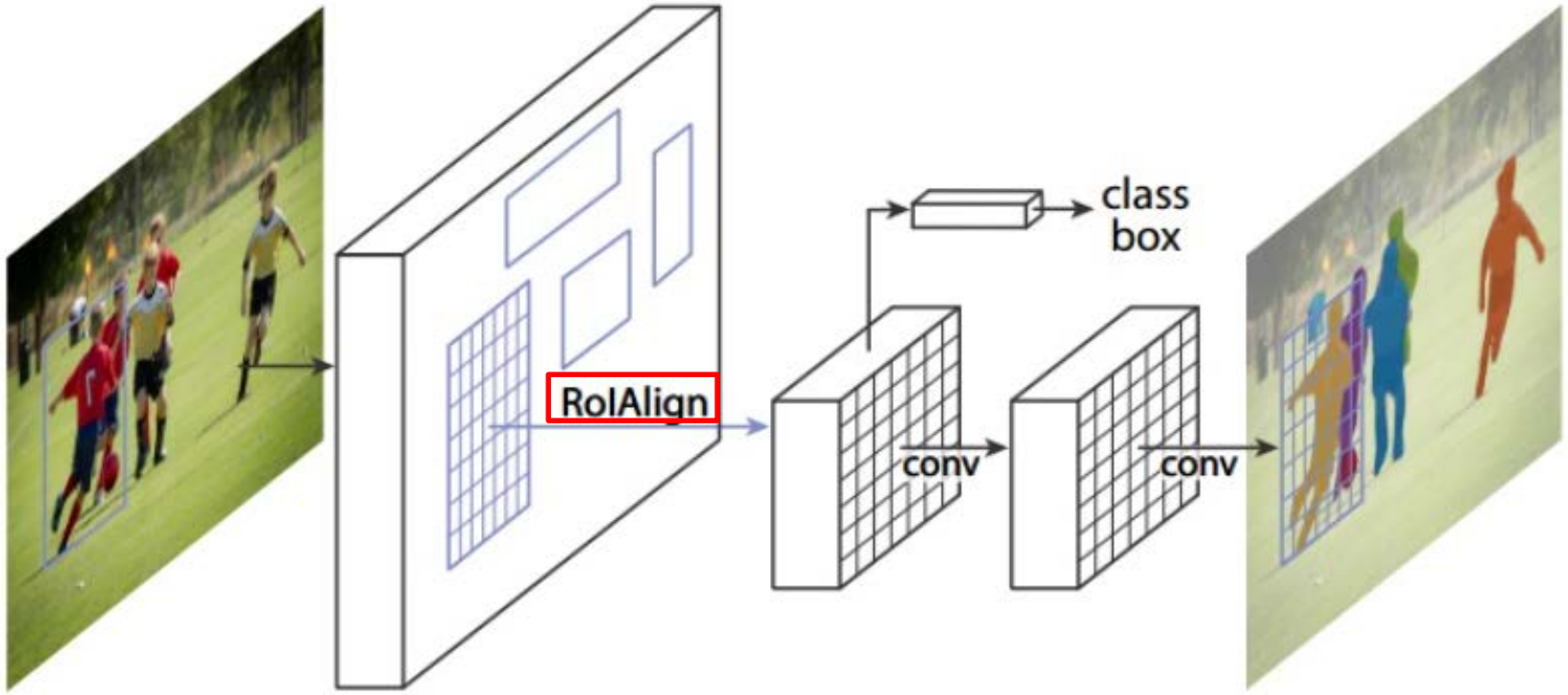


- Multi-scale feature map
- Default boxes (anchor)

Feature map	Anchor number	Feature map size
Conv4_3	4 anchor	(38,38)
Conv7	6 anchor	(19,19)
Conv8_2	6 anchor	(10,10)
Conv9_2	6 anchor	(5,5)
Conv10_2	4 anchor	(3,3)
Conv11_2	4 anchor	(1,1)

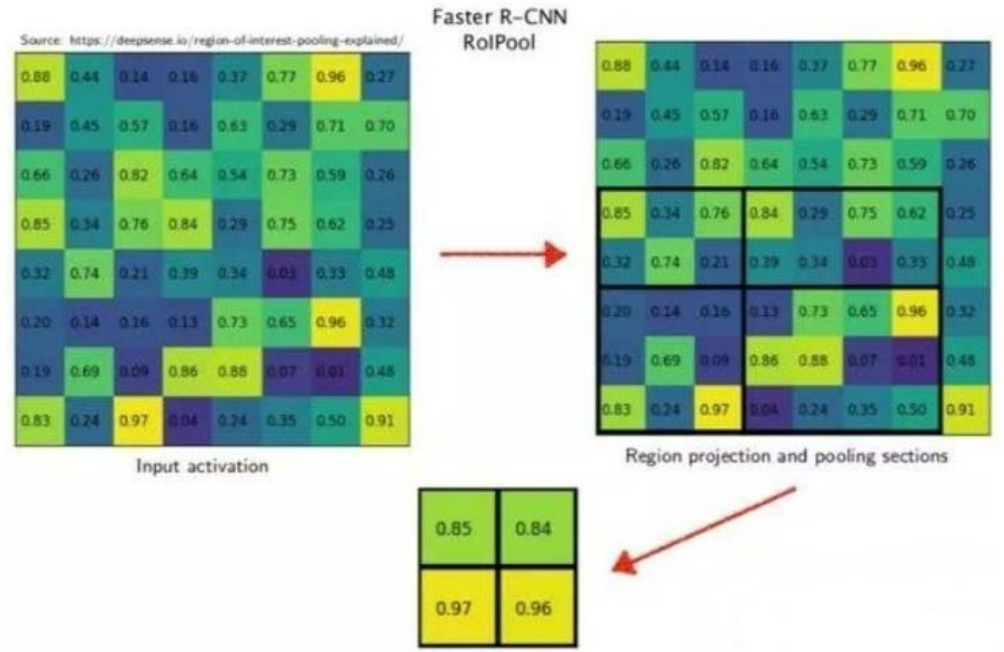
Liu et al., SSD: a single shot detector ECCV 2016

Mask RCNN

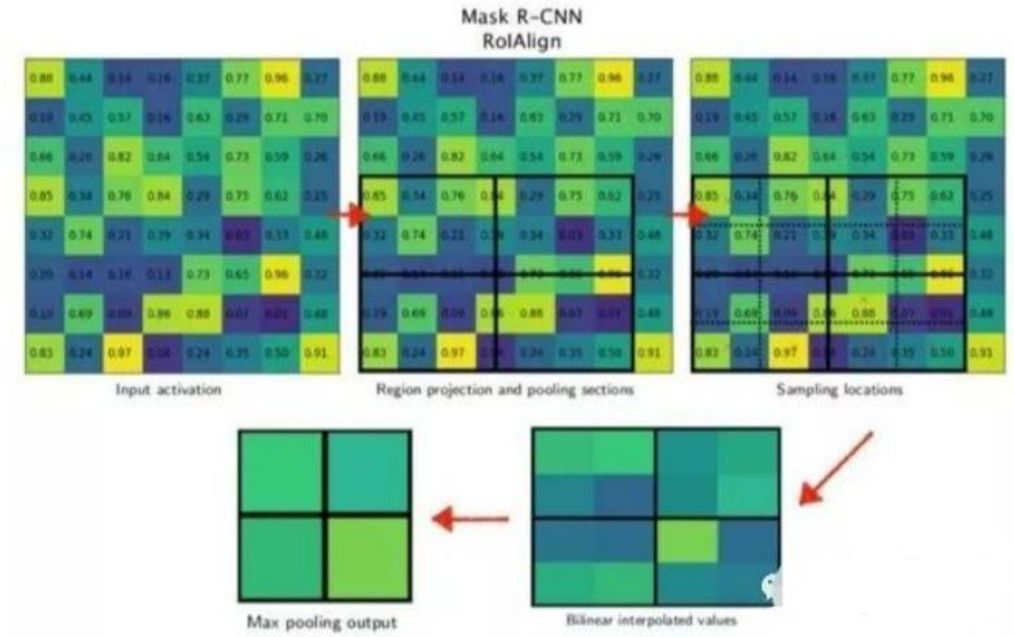


- RoI pooling→RoI Align
- Add a mask branch
- Introduce the FPN

ROI Pooling & ROI Align



ROI Pooling

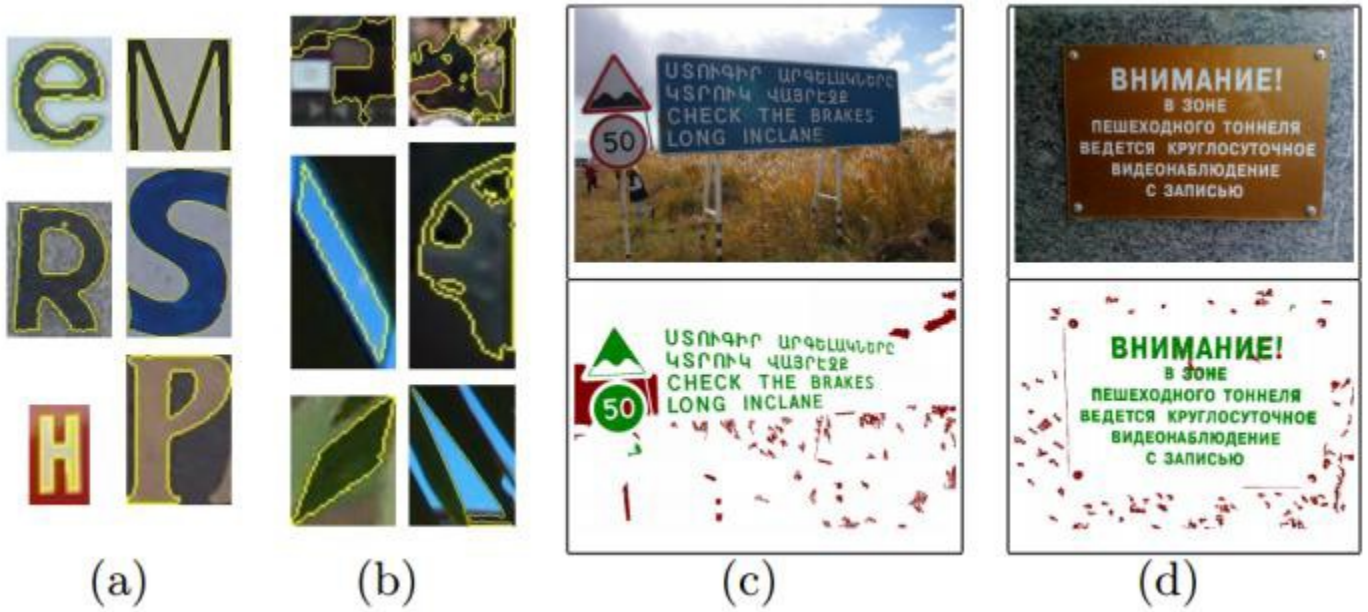


ROI Align

Outline

- Background and Introduction
- Datasets and Related Knowledge
- Conventional Methods
- Deep Learning Methods
- Conclusion and Outlook

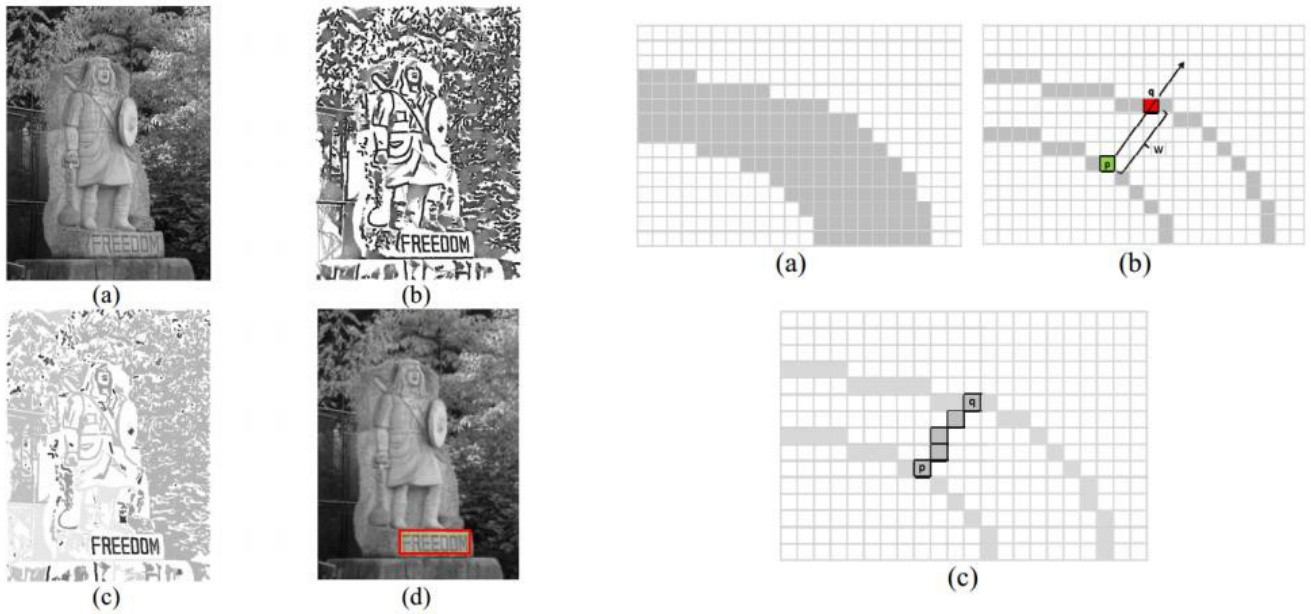
MSER



- extract character candidates using MSER (Maximally Stable Extremal Regions), assuming similar color within each character
- robust, fast to compute, independent of scale
- limitation: can only handle horizontal text, due to features and linking strategy

Neumann and Matas et al. A method for text localization and recognition in real-world images. ACCV, 2010.

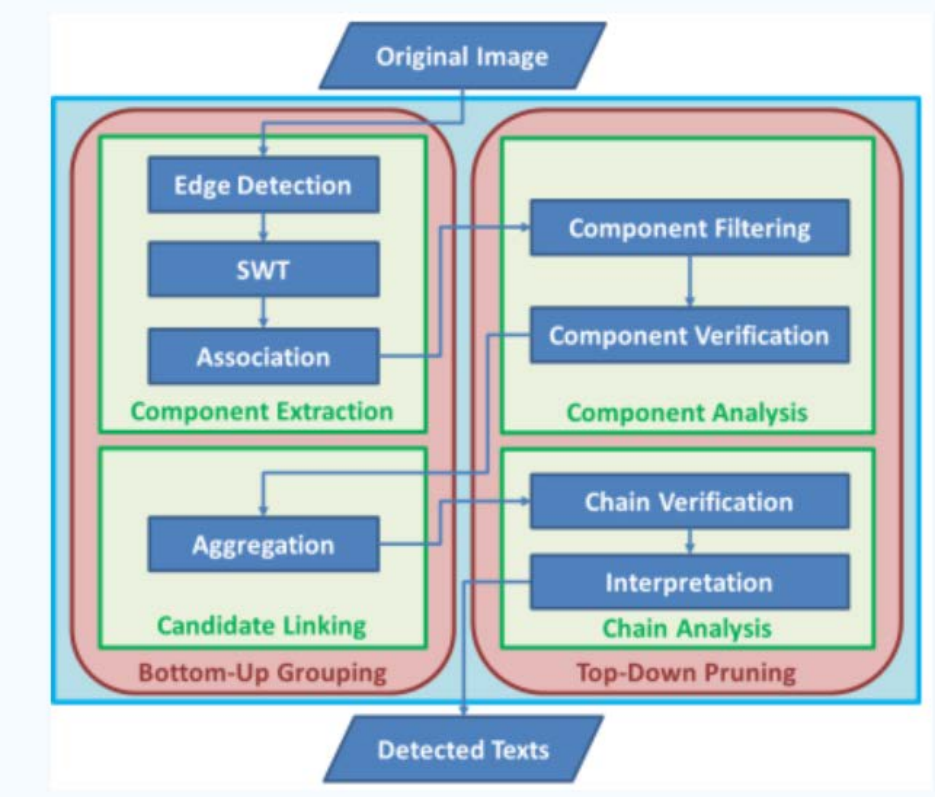
SWT

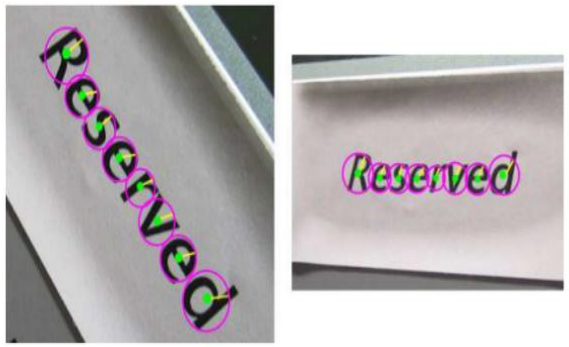


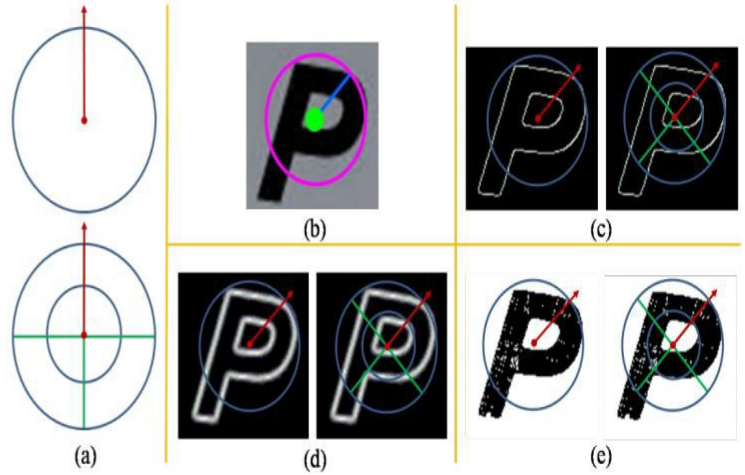
- extract character candidates with SWT (Stroke Width Transform), assuming consistent stroke width within each character
- robust, fast to compute, independent of scale
- limitation: can only handle horizontal text, due to features and linking strategy

Epshtein et al. Detecting Text in Natural Scenes with Stroke Width Transform. CVPR, 2010.

SWT







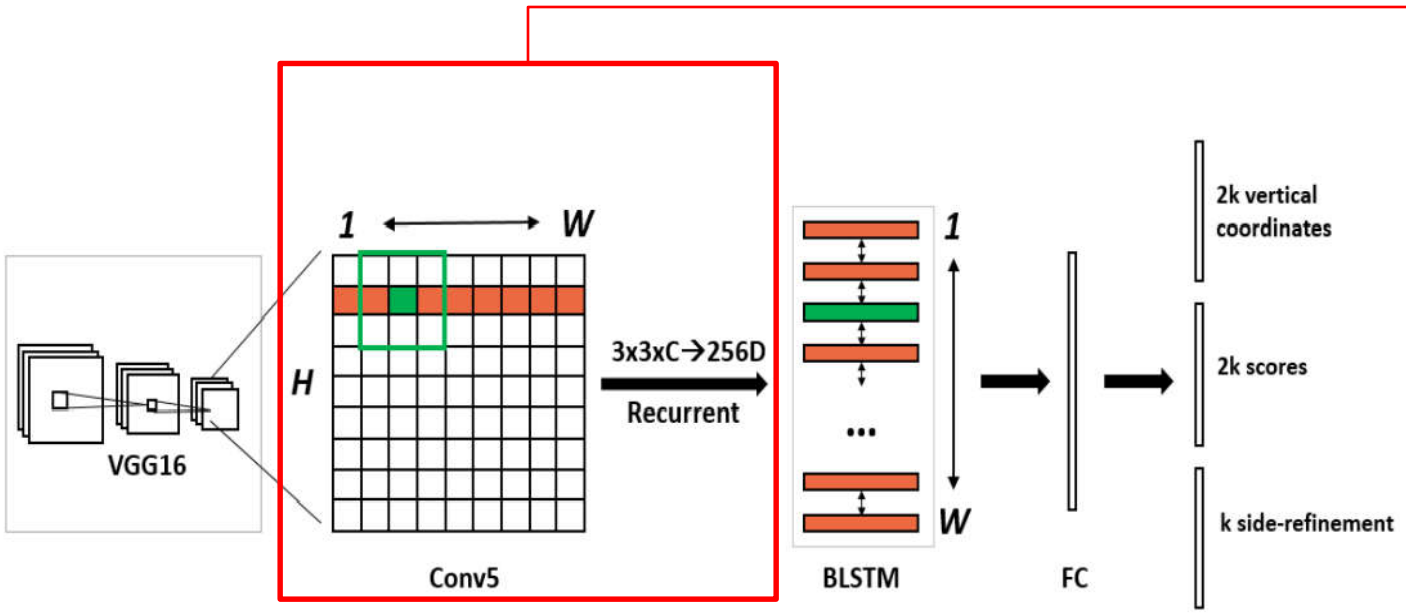
- adopt SWT to hunt character candidates
- design rotation-invariant features that facilitate multi-oriented text detection
- propose a new dataset (MSRA-TD500) that contains text instances of different directions

Yao et al., Detecting texts of arbitrary orientations in natural images. CVPR, 2012.

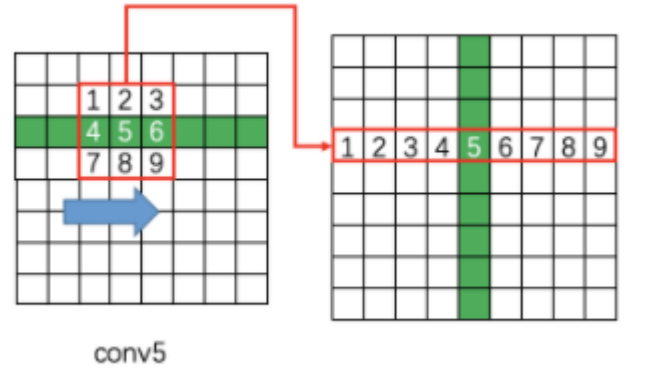
Outline

- Background and Introduction
- Datasets and Related Knowledge
- Conventional Methods
- Deep Learning Methods
  - Regression
  - Segmentation
  - Hybrid Methods
- Conclusion and Outlook

CTPN



$N \times C \times W \times H$ N:image numbers



Reshape: $N \times 9C \times H \times W \rightarrow (NH) \times W \times 9C$

- Dense sliding windows on feature maps to extract a feature vector of every location.
- BLSTM to capture the sequential context information.
- Fully-connected layer simultaneously predicts text/non-text scores, y-axis coordinates and side-refinement offsets of k anchors.

Tian et al., Detecting Text in Natural Image with Connectionist Text Proposal Network ECCV, 2016

CTPN

1. Fine-scale Proposals

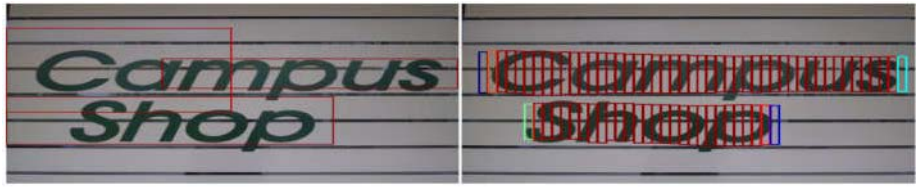
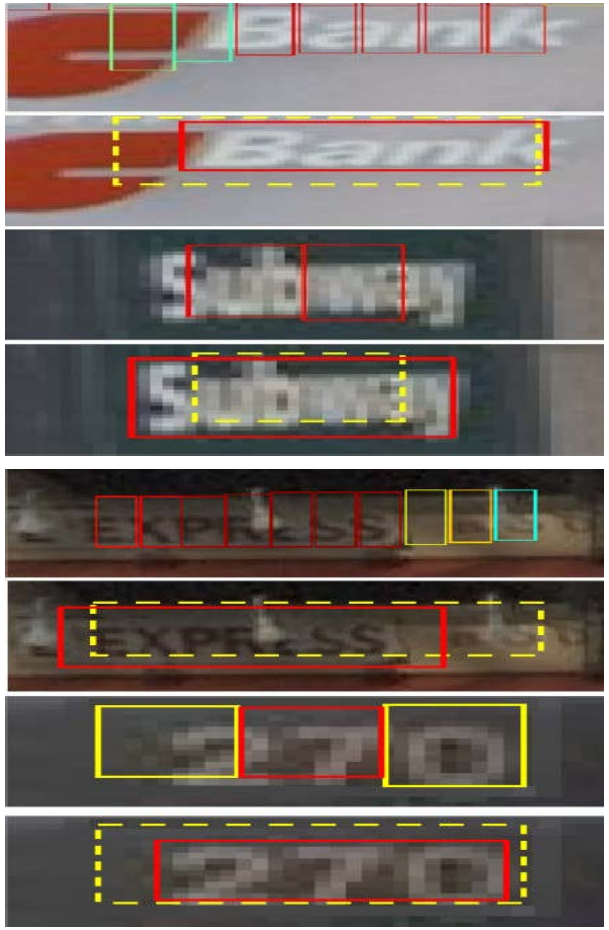


Fig. 2: **Left**: RPN proposals. **Right**: Fine-scale text proposals.

2. Recurrent Connectionist Text Proposals

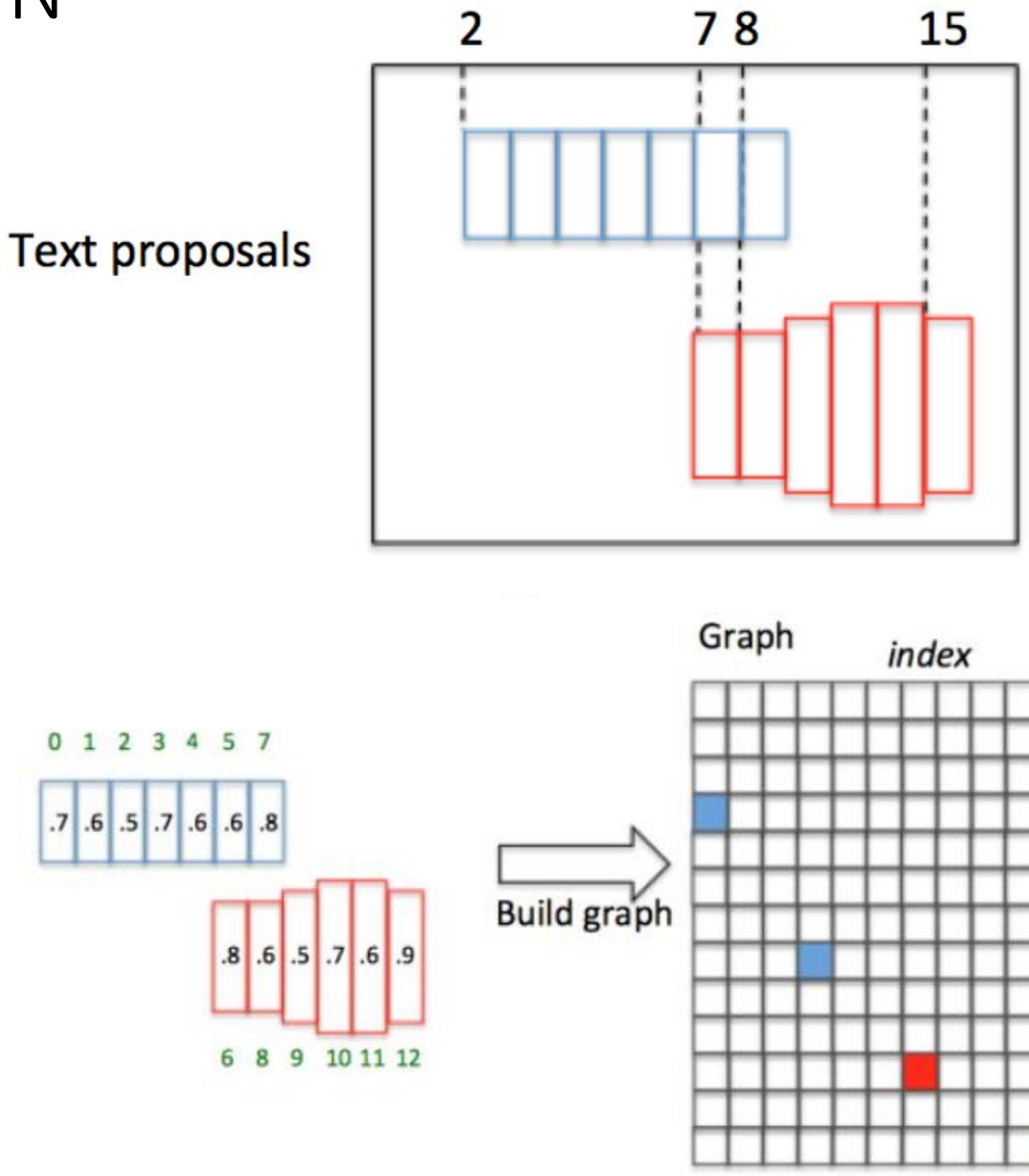
Fig. 3: **Top**: CTPN without RNN. **Bottom**: CTPN with RNN connection.

3. Side-refinement



Tian et al., Detecting Text in Natural Image with Connectionist Text Proposal Network ECCV, 2016

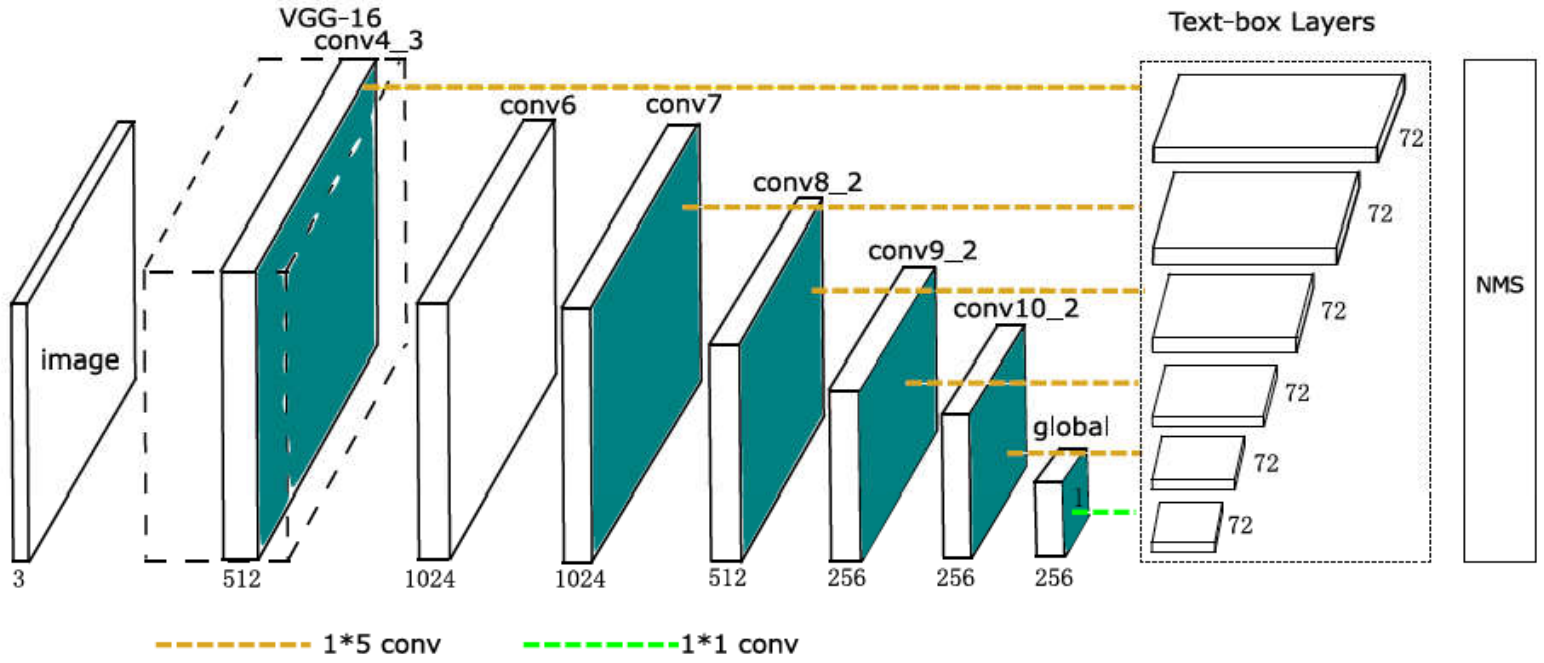
CTPN



1. Sorted by x axis
2. Calculate each anchor box and generate a $\mathrm{pair}(\mathrm{box}_i, \mathrm{box}_j)$
3. Build a graph

Tian et al., Detecting Text in Natural Image with Connectionist Text Proposal Network ECCV, 2016

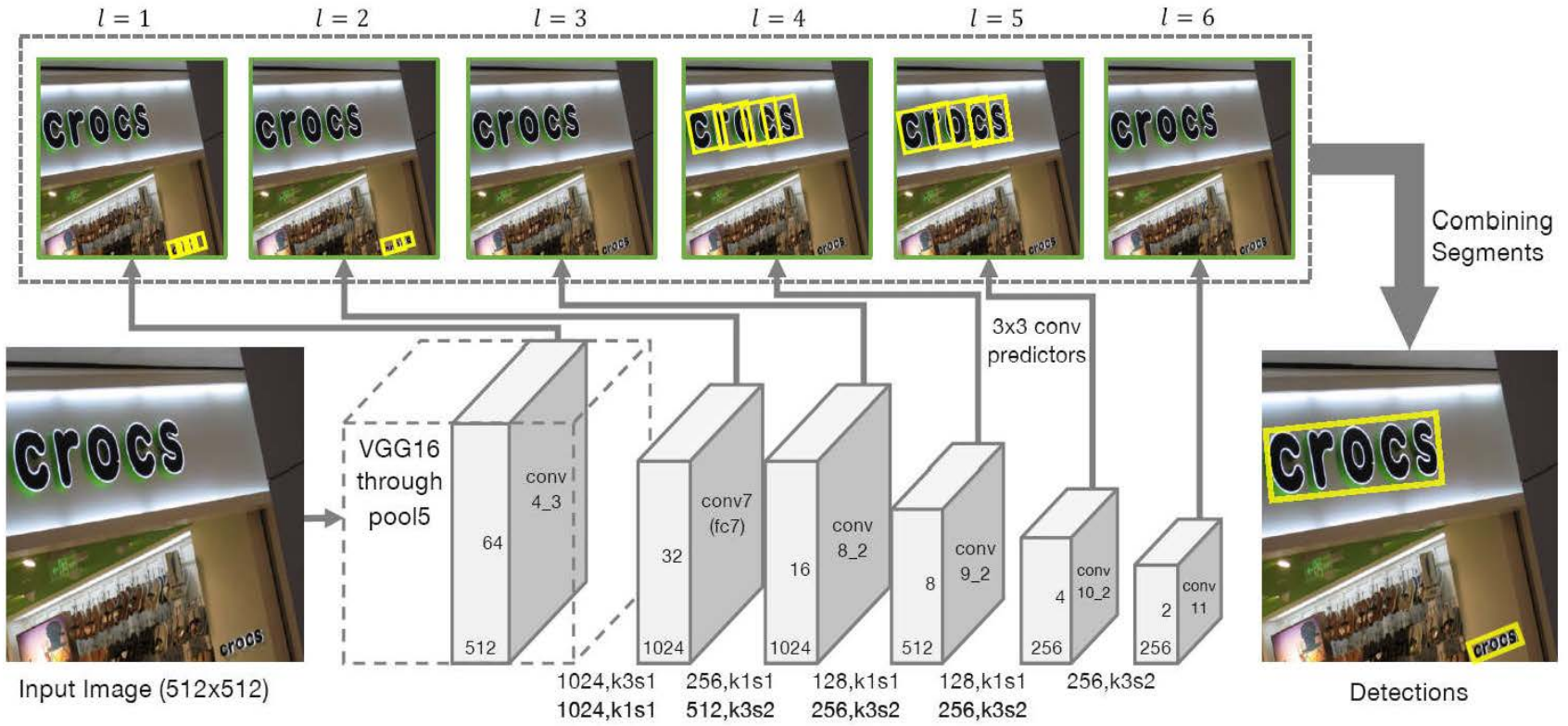
TextBoxes



- Fully convolutional network based on SSD.
- On every map location, a text-box layer predicts a 72-d vector(text presence scores (2-d) and offsets (4-d) for 12 default boxes)
- Special designed default boxes

Liao et al., TextBoxes: A fast text detector with a single deep neural network AAAI,2017.

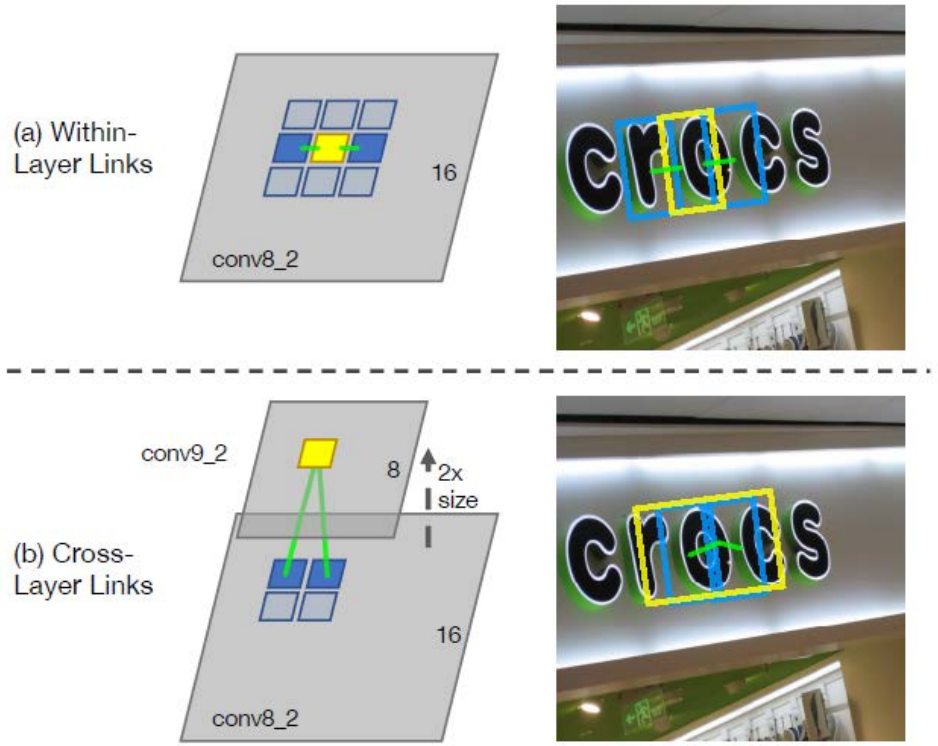
SegLink



- Fully convolutional network inspired by SSD
- Multi-stage outputs for segments and their links
- Solve the problem of CNN receptive field for long texts

Shi et al., Detecting Oriented Text in Natural Images by Linking Segments CVPR 2017.

SegLink



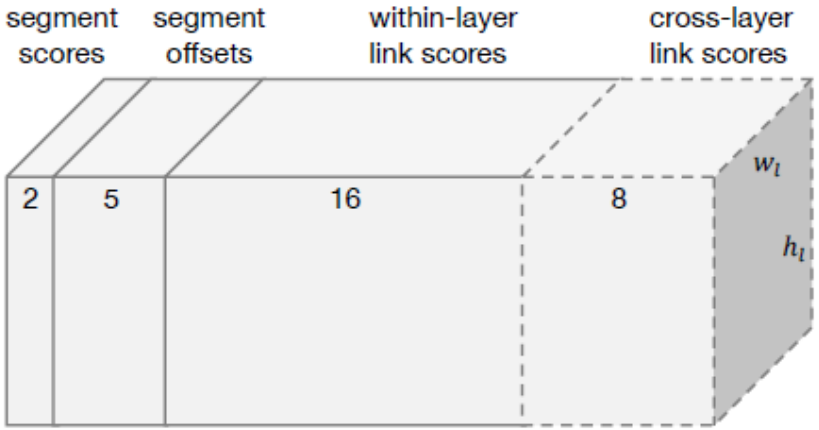
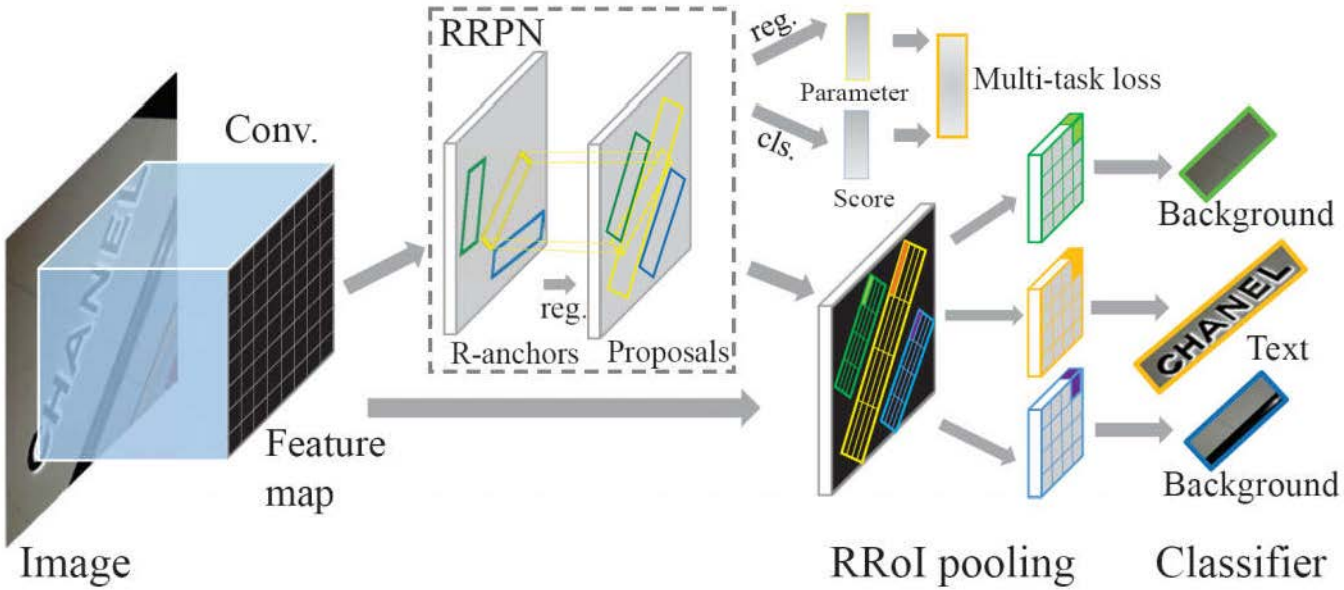


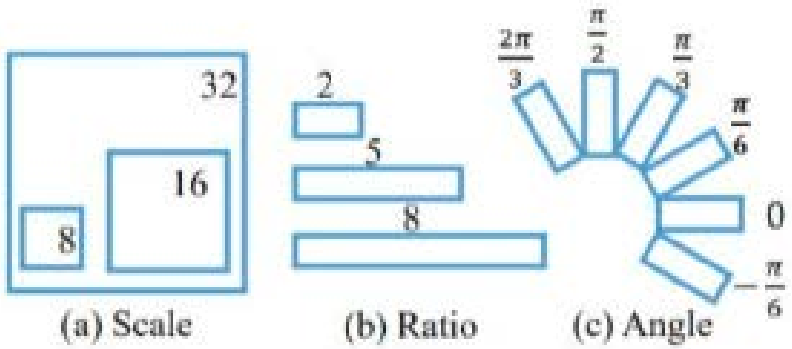
Figure 4. Output channels of a convolutional predictor. The block shows a $w_l \times h_l$ map of depth 31. The predictor of $l = 1$ does not output the channels for corss-layer links.

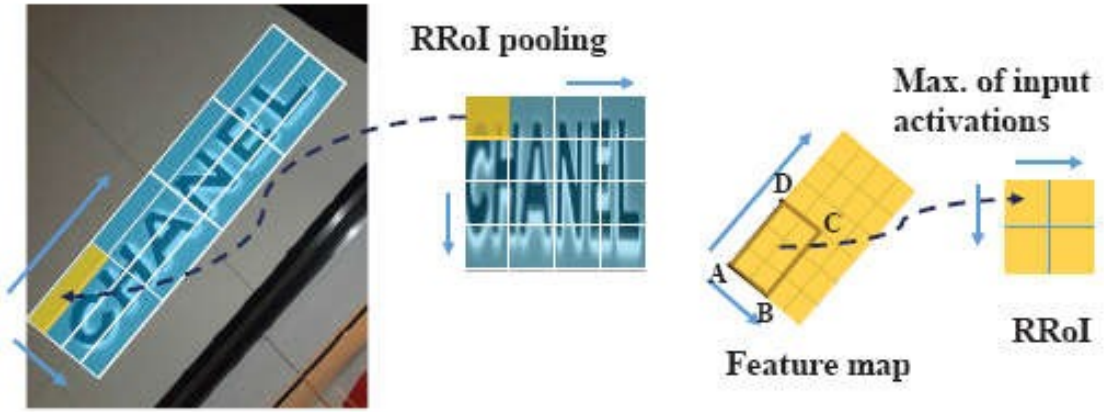
Shi et al., Detecting Oriented Text in Natural Images by Linking Segments CVPR 2017.

RRPN



- Use the architecture of Faster-rcnn
- RPN ->Rotated RPN
- RoI pooling ->Rotated RoI pooling

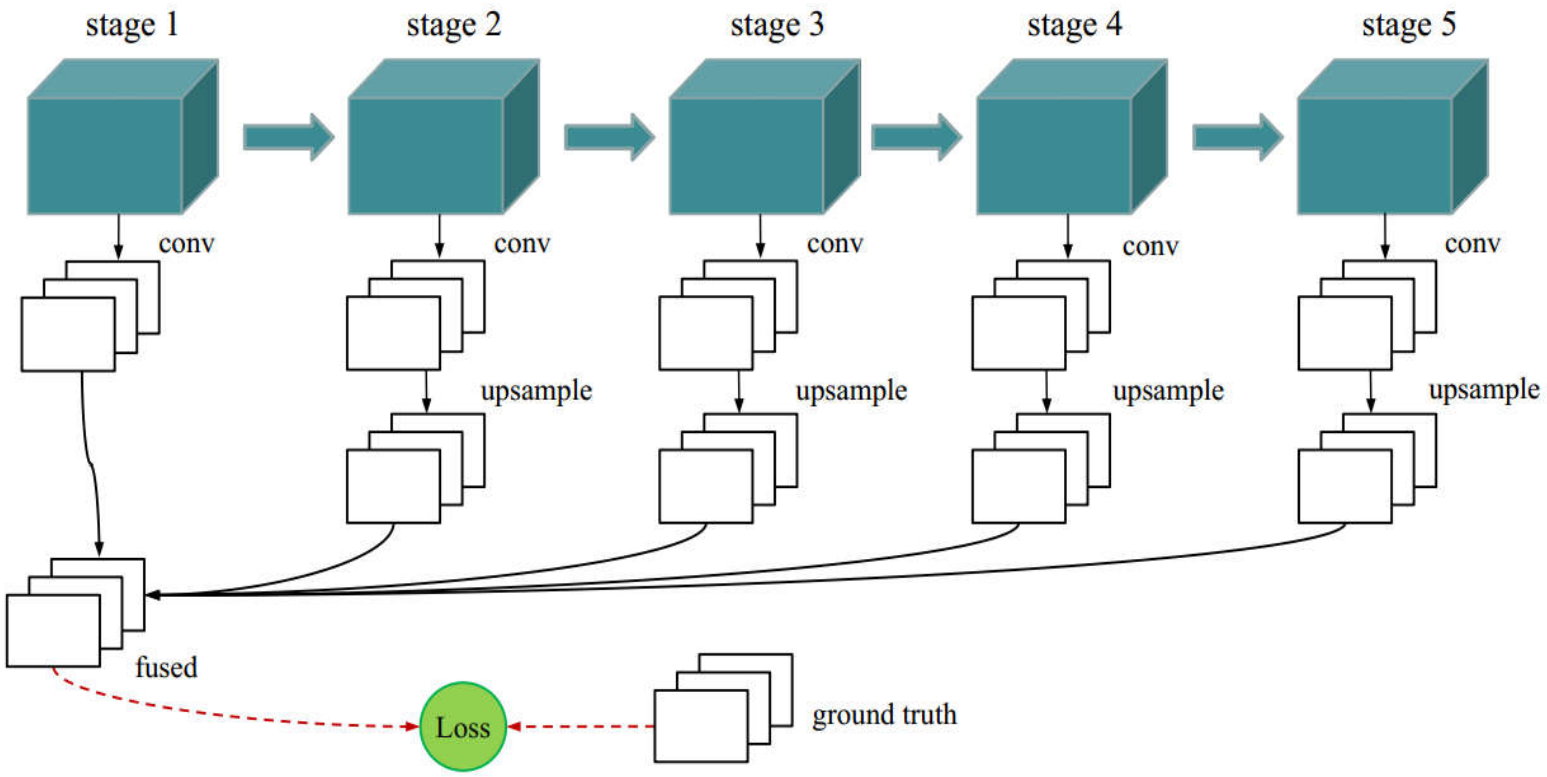




Jianqi Ma et al., Arbitrary-Oriented Scene Text Detection via Rotation Proposals ACM MM, 2018.

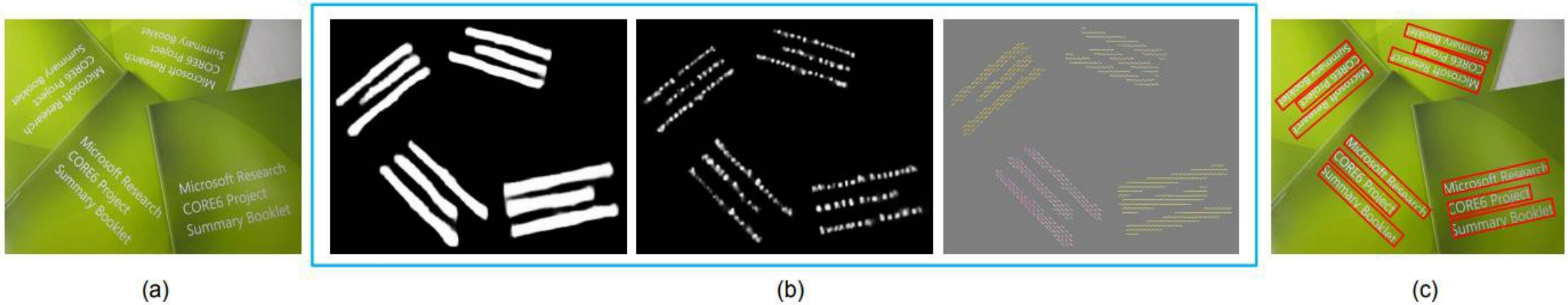
Outline

- Background and Introduction
- Conventional Methods
- Deep Learning Methods
    - object detection
    - Segmentation
    - Hybrid Methods
- Datasets and Competitions
- Conclusion and Outlook



- FCN based network.
- Multi task. Text region, individual characters and their relationship are estimated simultaneously.

Yao et al., Scene Text Detection Via Holistic, Multi-Channel Prediction arXiv:1606.09002, 2016



- holistic, pixel-wise predictions: text region map, character map and linking orientation map detections are formed using these three maps

- can simultaneously handle horizontal, multi-oriented and curved text in real-world natural images

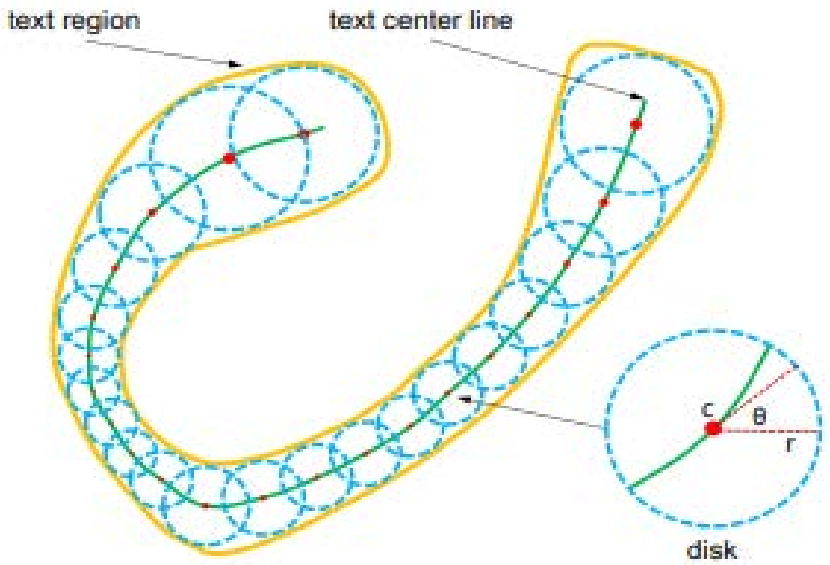
Yao et al., Scene Text Detection Via Holistic, Multi-Channel Prediction arXiv:1606.09002, 2016

TextSnake

- A flexible and general representation for scene text of arbitrary shapes;
- Able to effectively and precisely describe the geometric properties,such as location,scale,and bending of curved text,while the other representations struggle

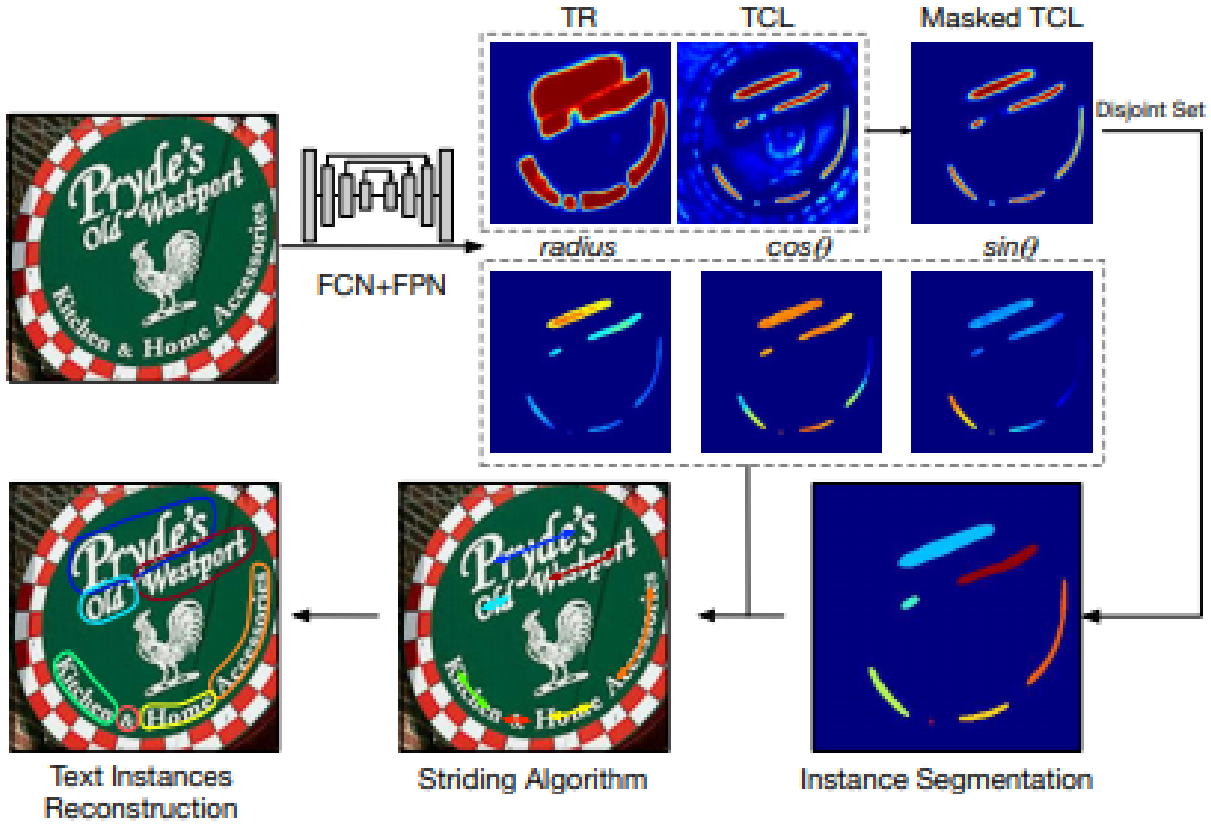
Long S , Ruan J, Zhang W, et al. TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes ECCV 2018.

TextSnake



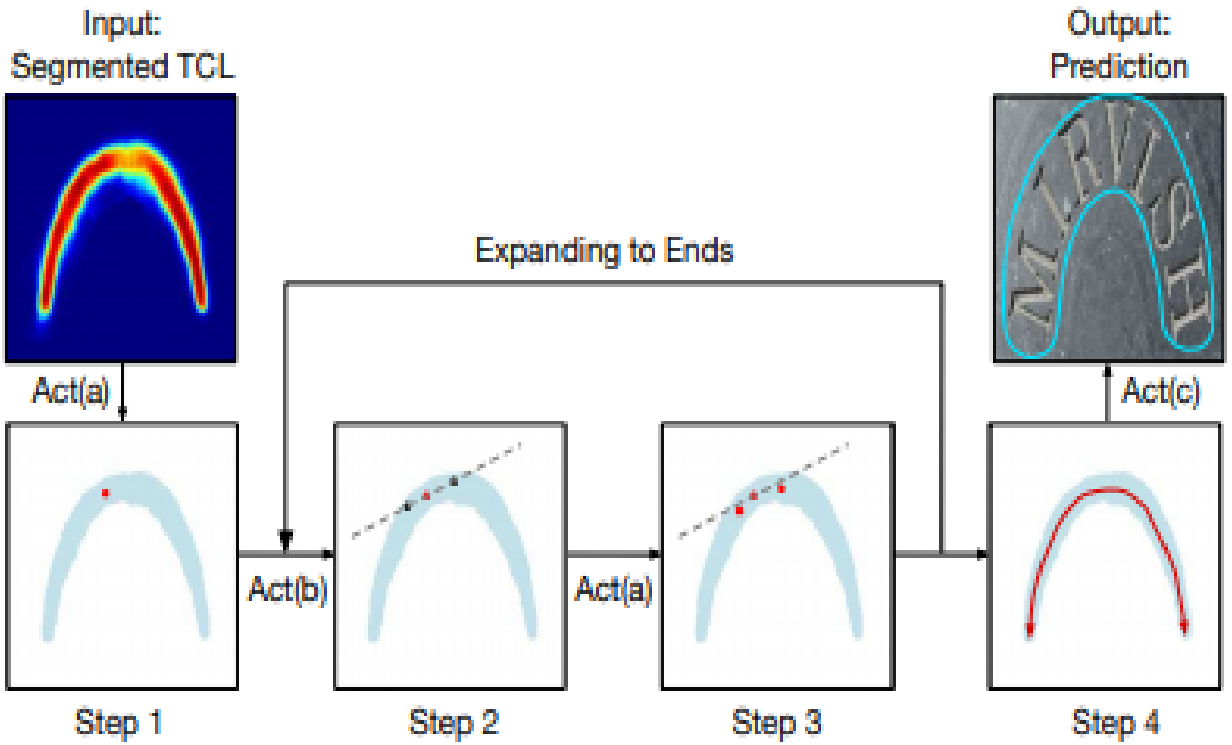
$$S(t) = \{D_0, D_1, \cdots, D_i, \cdots, D_n\}$$

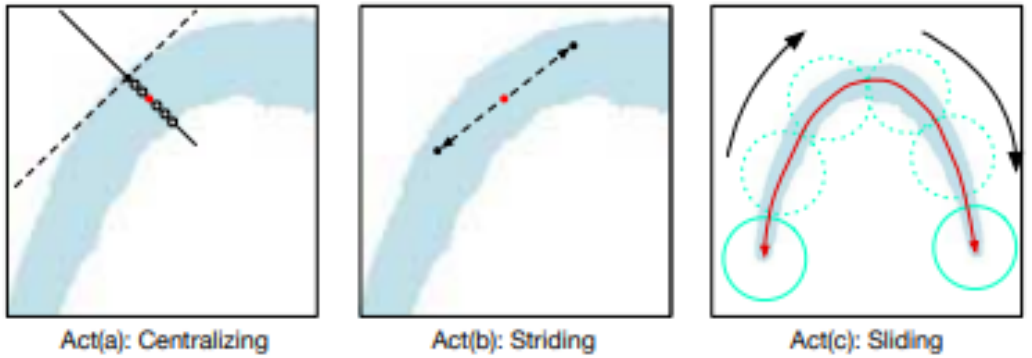
$$D = (c, r, \theta)$$



Long S , Ruan J, Zhang W, et al. TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes ECCV 2018.

TextSnake





$$L = L_{cls} + L_{reg}$$

$$L_{cls} = \lambda_1 L_{tr} + \lambda_2 L_{tcl}$$

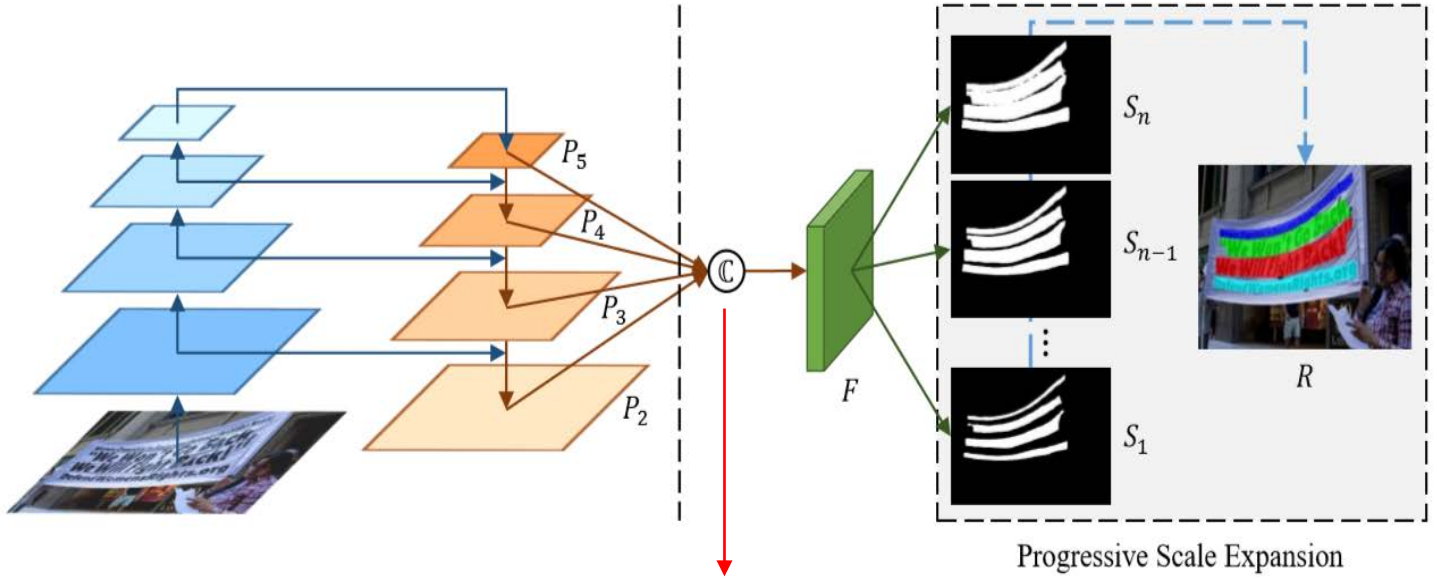
$$L_{reg} = \lambda_3 L_r + \lambda_4 L_{sin} + \lambda_5 L_{cos}$$

Long S , Ruan J, Zhang W, et al. TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes ECCV 2018.

PSENet

Text instances with irregular shapes

Fail to separate text areas with very adjacent edges

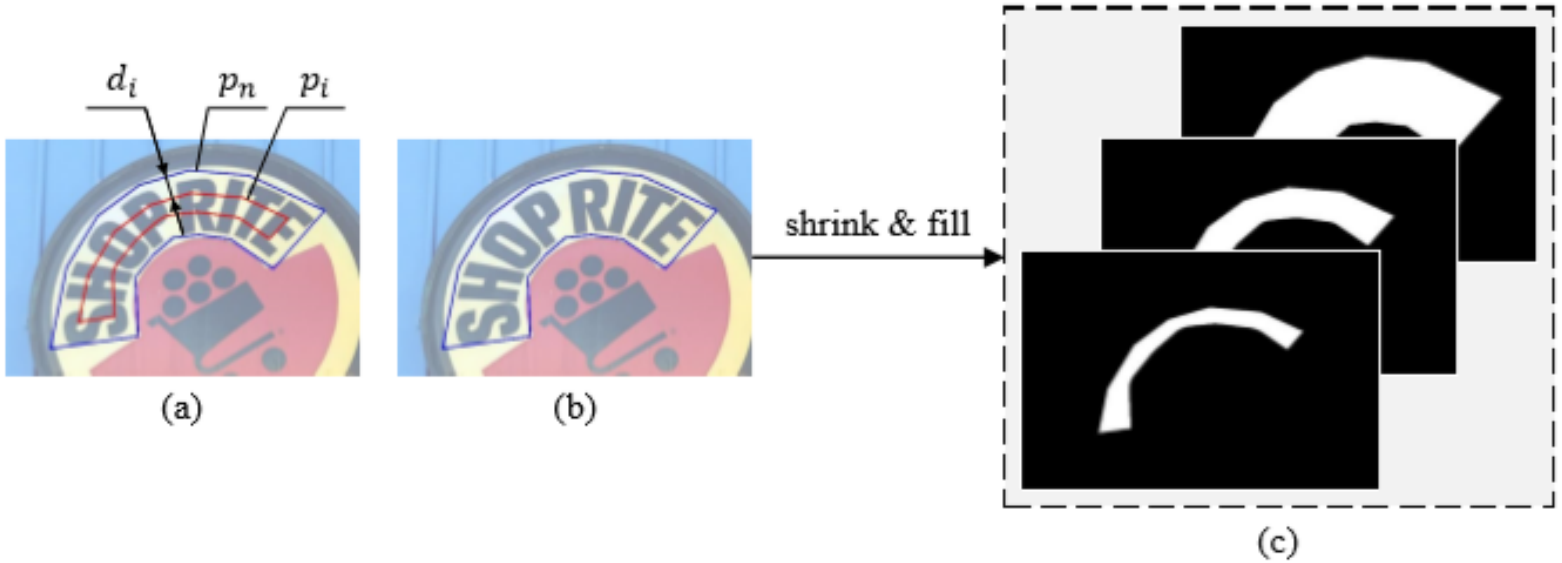


$$F = \mathbb{C}(P_2, P_3, P_4, P_5)$$
$$= P_2 \parallel \mathrm{Up}_{\times 2}(P_3) \parallel \mathrm{Up}_{\times 4}(P_4) \parallel \mathrm{Up}_{\times 8}(P_5)$$

Wenhai Wang et al. Shape Robust Text Detection with Progressive Scale Expansion Network CVPR 2019

PSENet

Label Generation

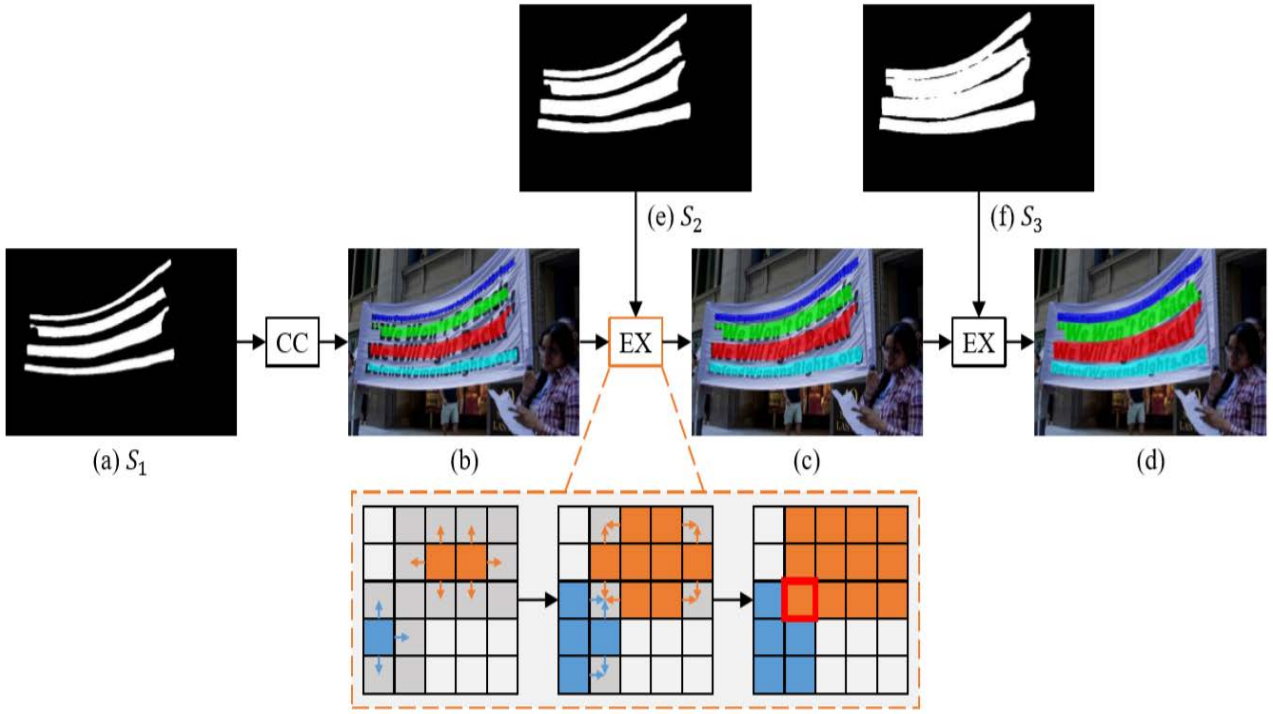


$$d_i = \frac{\text{Area}(p_n) \times (1 - r_i^2)}{\text{Perimeter}(p_n)}$$

$$r_i = 1 - \frac{(1 - m) \times (n - i)}{n - 1}$$

Wenhai Wang et al. Shape Robust Text Detection with Progressive Scale Expansion Network CVPR 2019

PSENet



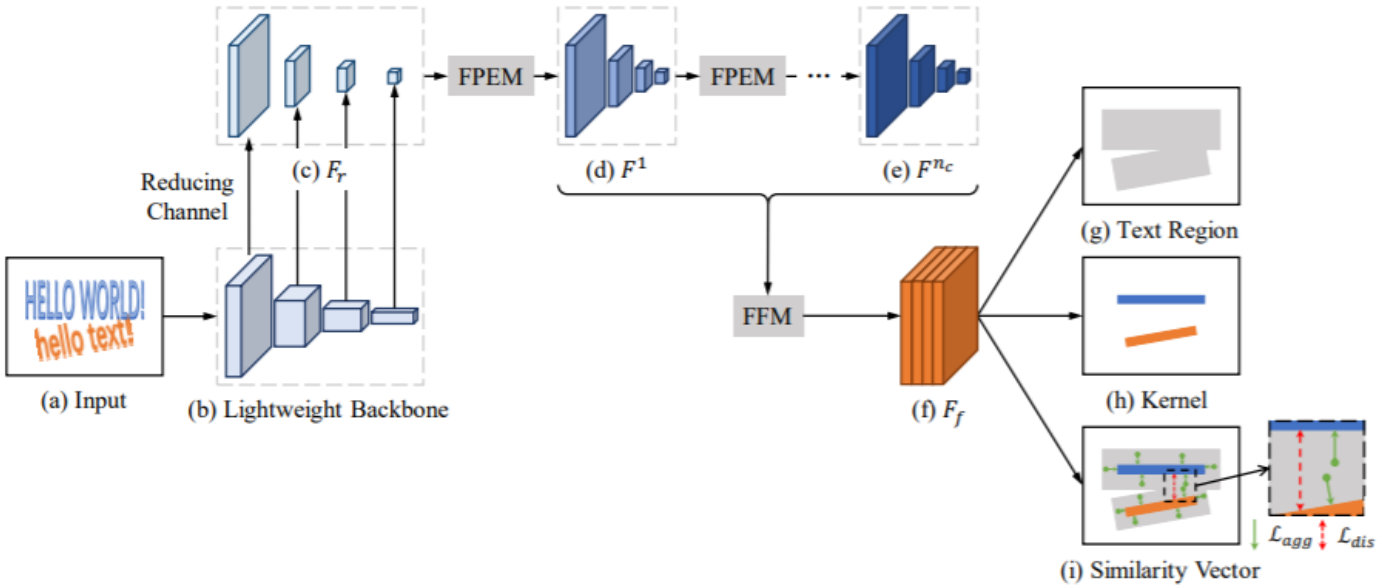
Algorithm 1 Scale Expansion Algorithm

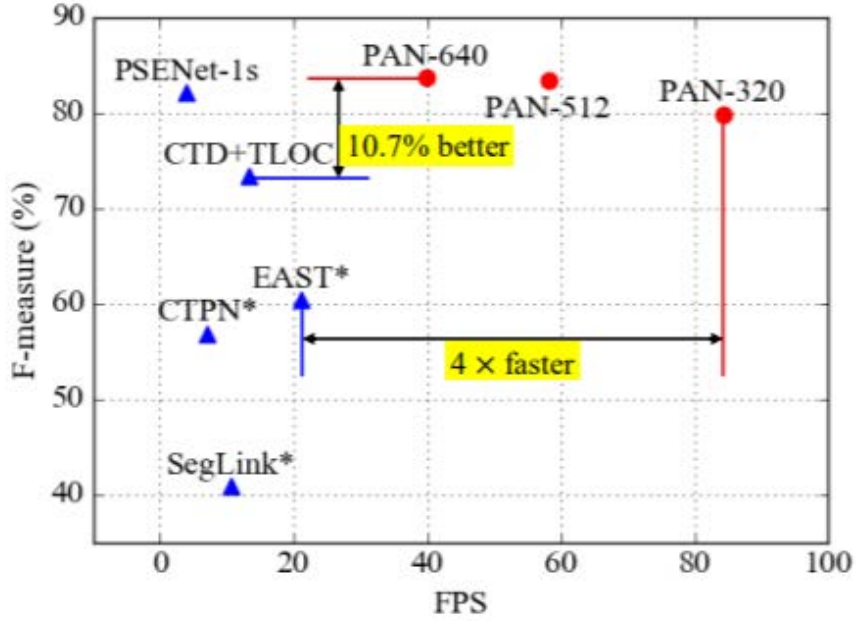
Require: Kernels: C, Segmentation Result: S_i
Ensure: Scale Expanded Kernels: E

function EXPANSION(C, S_i)
  $T \leftarrow \emptyset$; $P \leftarrow \emptyset$; $Q \leftarrow \emptyset$
  for each $c_i \in C$ **do**
    $T \leftarrow T \cup \{(p, label) \mid (p, label) \in c_i\}$
    $P \leftarrow P \cup \{p \mid (p, label) \in c_i\}$
    Enqueue(Q, c_i) // push all the elements in c_i into Q
  end for
  while $Q \neq \emptyset$ **do**
    $(p, label) \leftarrow$ **Dequeue**(Q) // pop the first element of Q
    if $\exists q \in$ **Neighbor**(p) and $q \notin P$ and $S_i[q] =$ True **then**
      $T \leftarrow T \cup \{(q, label)\}$; $P \leftarrow P \cup \{q\}$
      Enqueue$(Q, (q, label))$ // push the element $(q, label)$ into Q
    end if
  end while
  $E =$ **GroupByLabel**(T)
  return E
end function

Wenhai Wang et al. Shape Robust Text Detection with Progressive Scale Expansion Network CVPR 2019

PANet

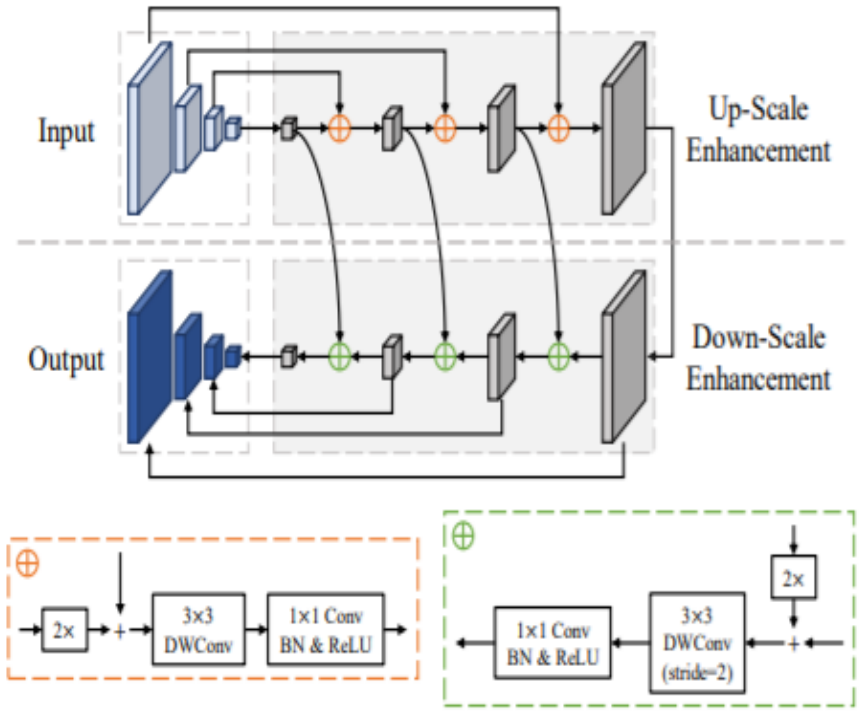




- The trade-off between speed and accuracy
- To model the arbitrary-shaped text instance

Wenhai Wang et al. Efficient and AccurateArbitrary-Shaped Text Detection with Pixel Aggregation Network ICCV 2019

PANet



FPEM

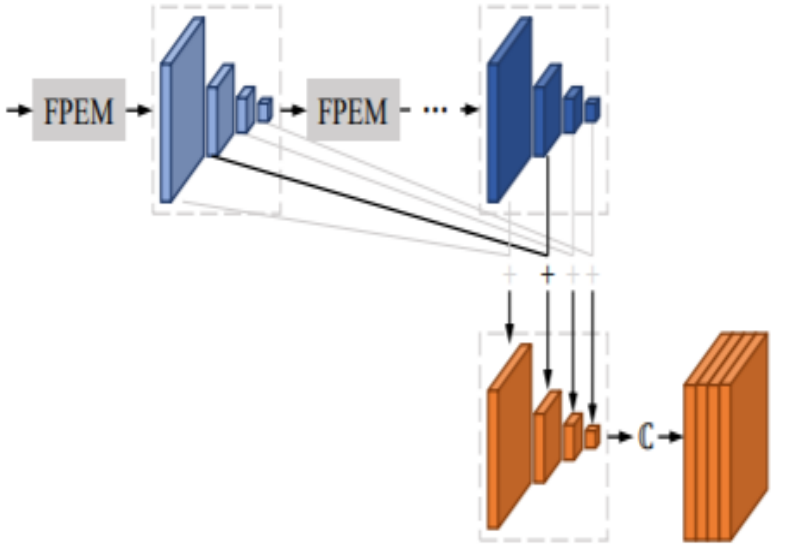
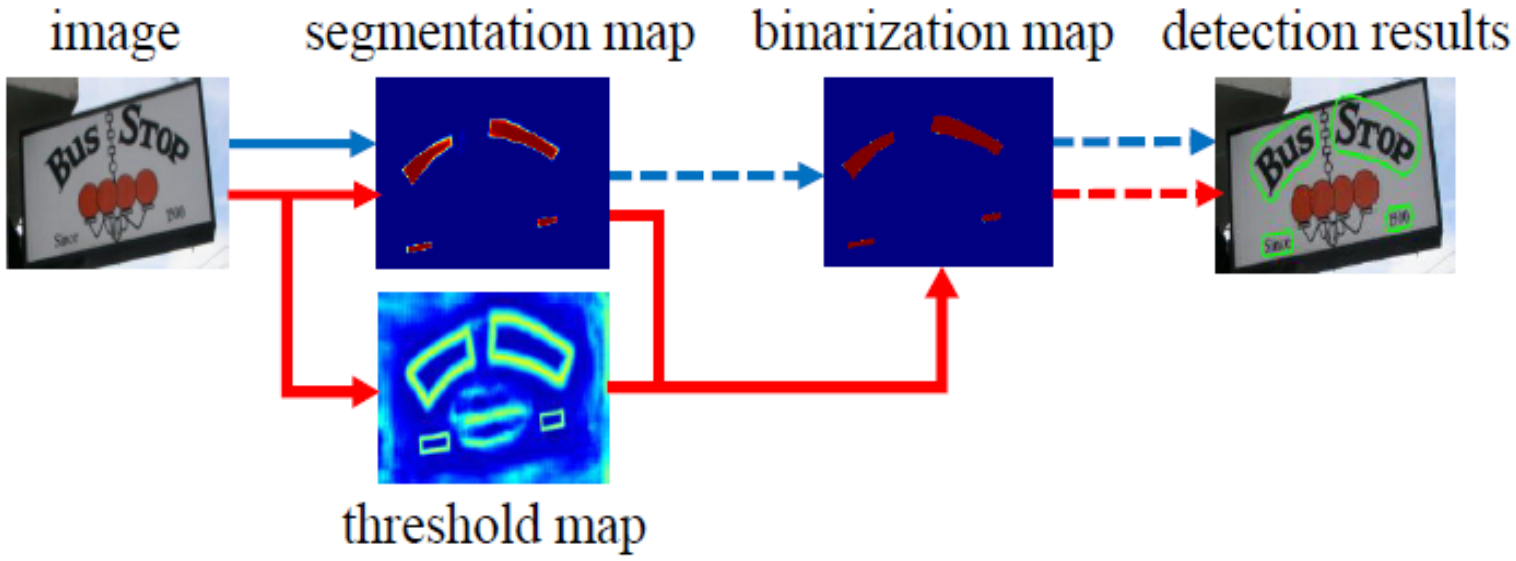


Figure 5. The detail of FFM. "+" is element-wise addition. "C" is the operation of upsampling and concatenating.

FFM

Wenhai Wang et al. Efficient and AccurateArbitrary-Shaped Text Detection with Pixel Aggregation Network ICCV 2019

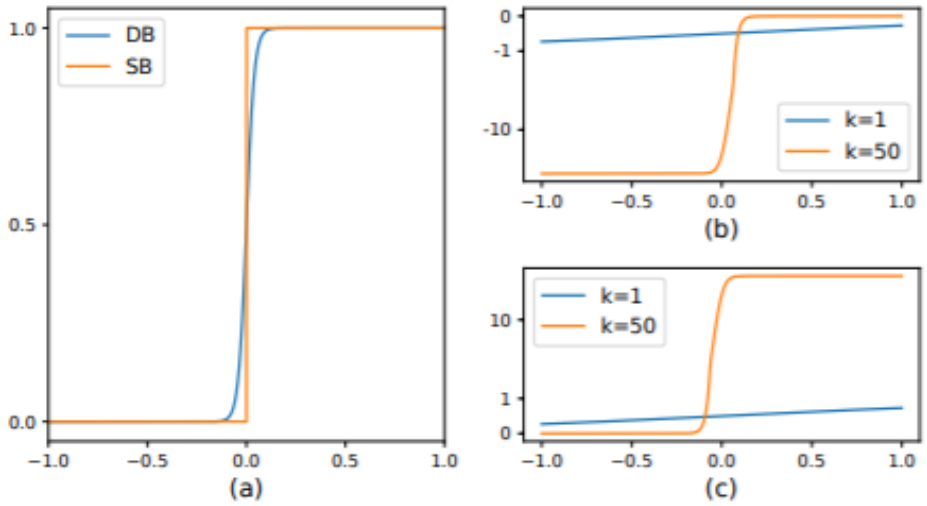
DBNet



- Differentiable binarization
- Adaptive threshold
- Simplified yet effective post-processing

Liao, Minghui, et al. "Real-Time Scene Text Detection with Differentiable Binarization." AAAI. 2020.

DBNet

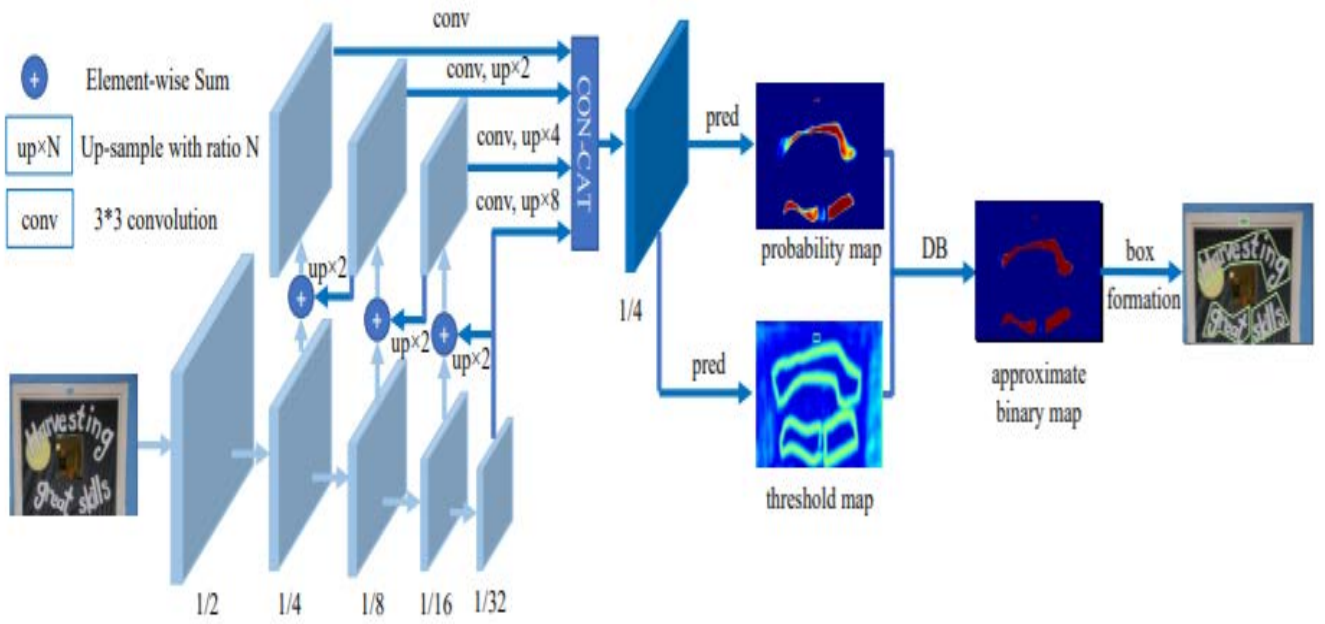


Standard binarization

$$B_{i,j} = \begin{cases} 1 & \text{if } P_{i,j} >= t, \\ 0 & \text{otherwise.} \end{cases}$$

Differentiable binarization

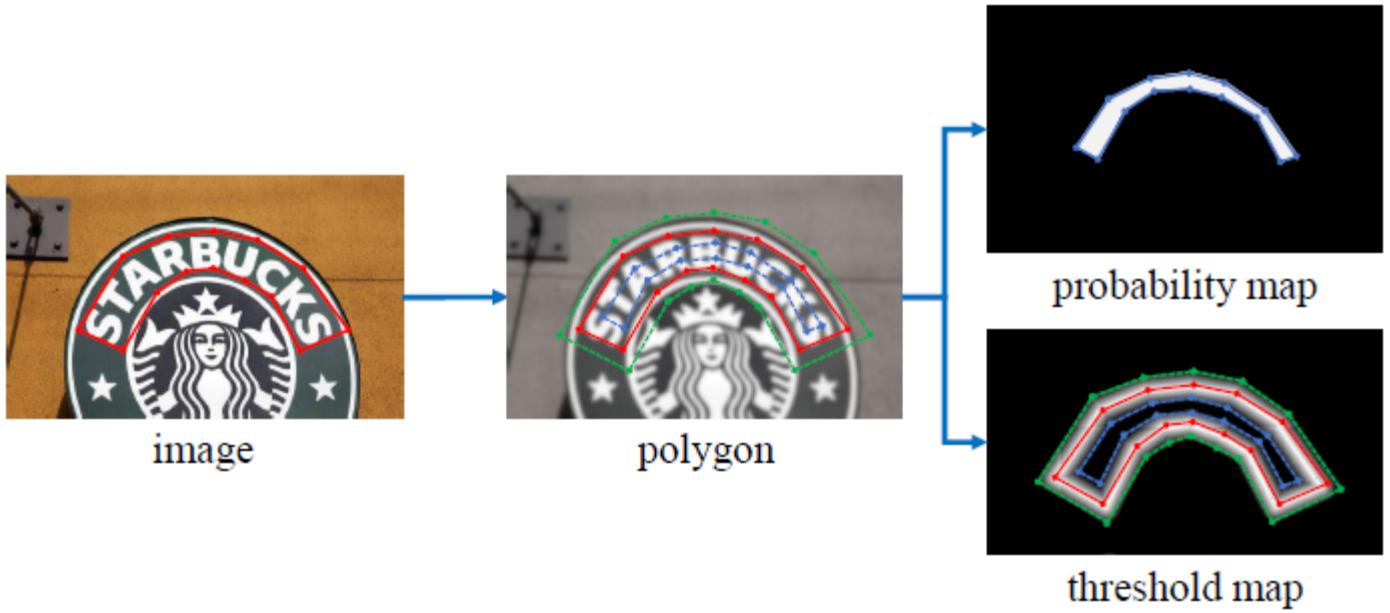
$$\hat{B}_{i,j} = \frac{1}{1 + e^{-k(P_{i,j} - T_{i,j})}}$$



Liao, Minghui, et al. "Real-Time Scene Text Detection with Differentiable Binarization." AAAI. 2020.

DBNet

Label Generation



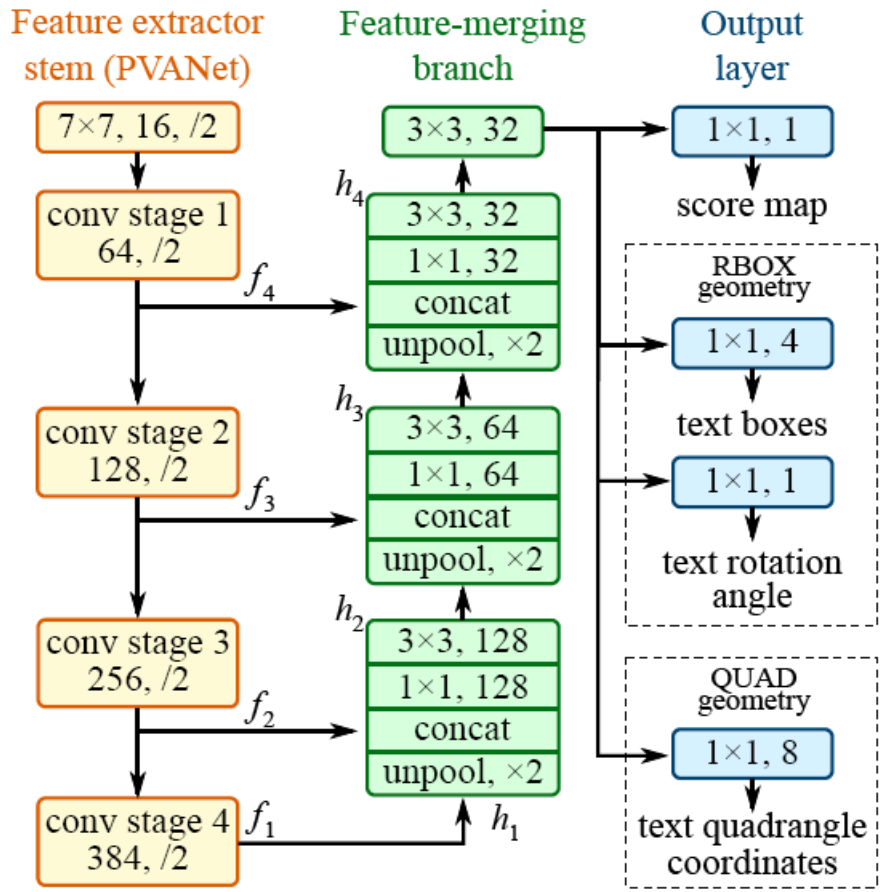
- Shrink polygon
- Fill the dilated with positive label
- Fill the gap area with gradual value

Liao, Minghui, et al. "Real-Time Scene Text Detection with Differentiable Binarization." AAAI. 2020.

Outline

- Background and Introduction
- Datasets and Related Knowledge
- Conventional Methods
- Deep Learning Methods
  - Regression
  - Segmentation
  - Hybrid Methods
- Conclusion and Outlook

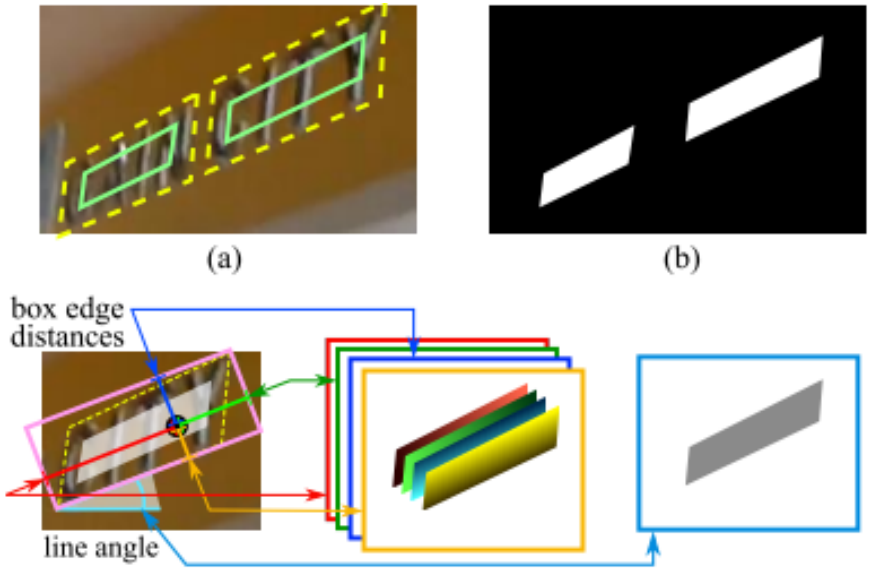
EAST

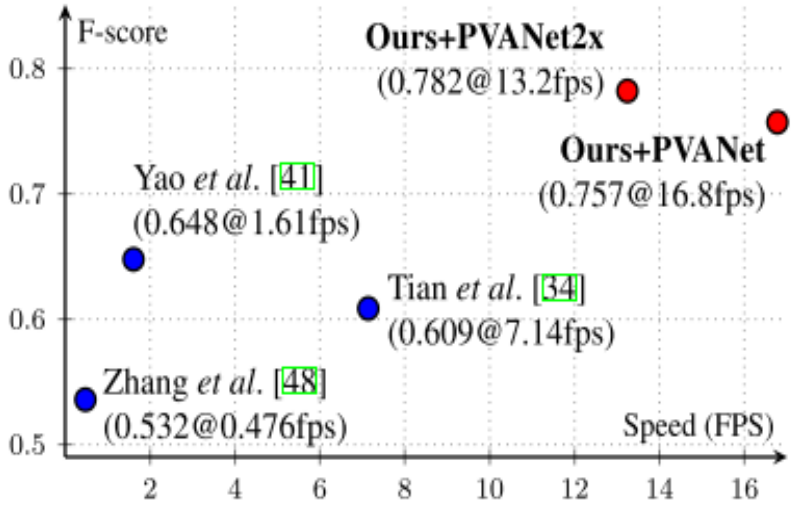


- PVANet(faster than VGG16)
- Multi-channel :
  - Score map
  - Rotated bounding boxes
  - Quadrangle bounding boxes

Zhou et al., EAST: An Efficient and Accurate Scene Text Detector CVPR 2017

EAST





- main idea: predict location, scale and orientation of text with a single model and multiple loss functions (multi-task training)
- advantages:
  - (a). accuracy: allow for end-to-end training and optimization
  - (b). efficiency: remove redundant stages and processings

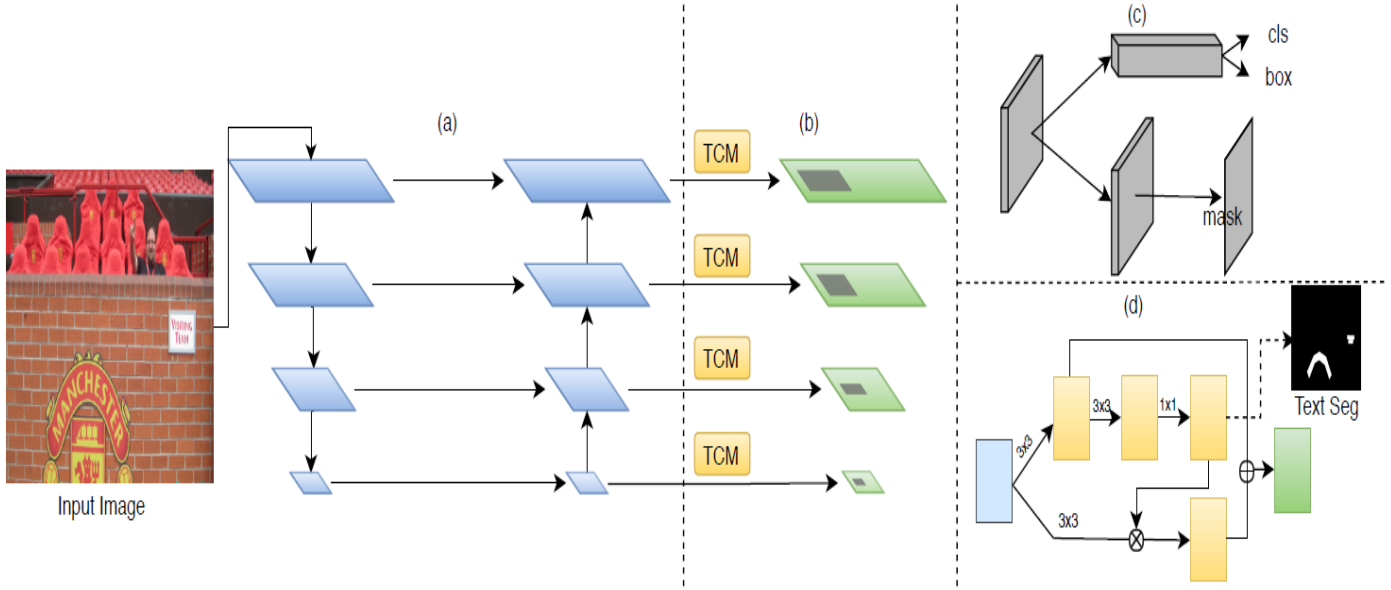
Zhou et al., EAST: An Efficient and Accurate Scene Text Detector CVPR 2017

SPCNet

- Lack of context information clues.
- Inaccurate classification score.

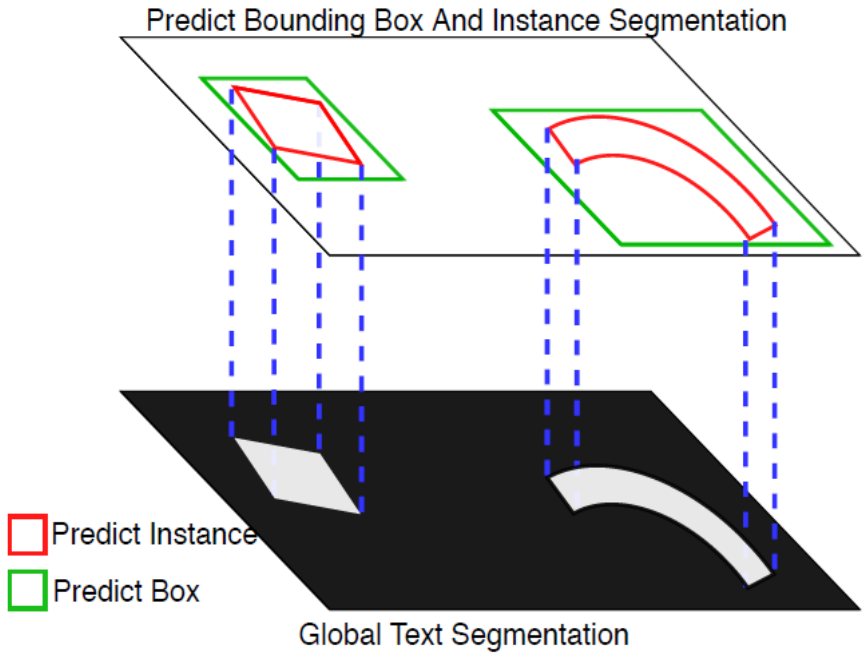
Enze Xie, et al. Scene Text Detection with Supervised Pyramid Context Network. In AAAI 2019.

SPCNet



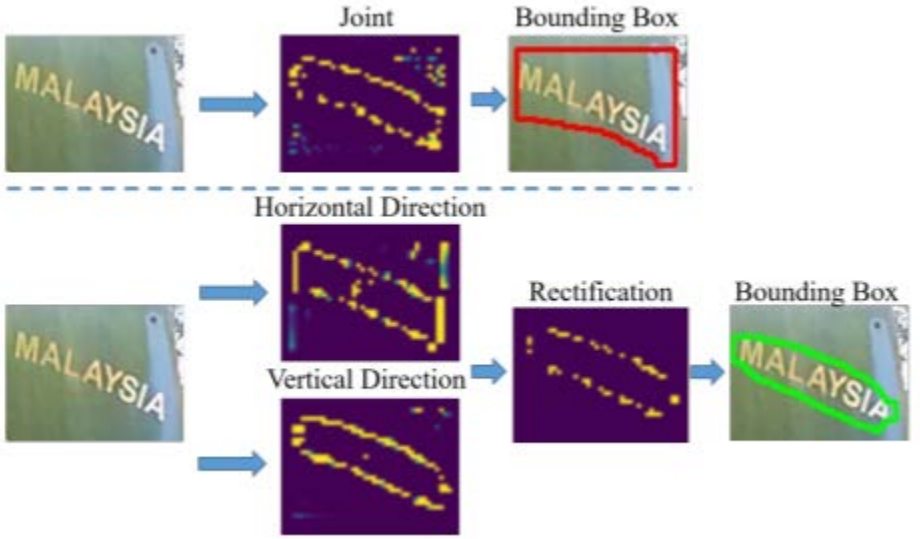
- Based Mask RCNN
- Text context (TCM)
- A new re-score mechanism

Re-score mechanism



Enze Xie, et al. Scene Text Detection with Supervised Pyramid Context Network. In AAAI 2019.

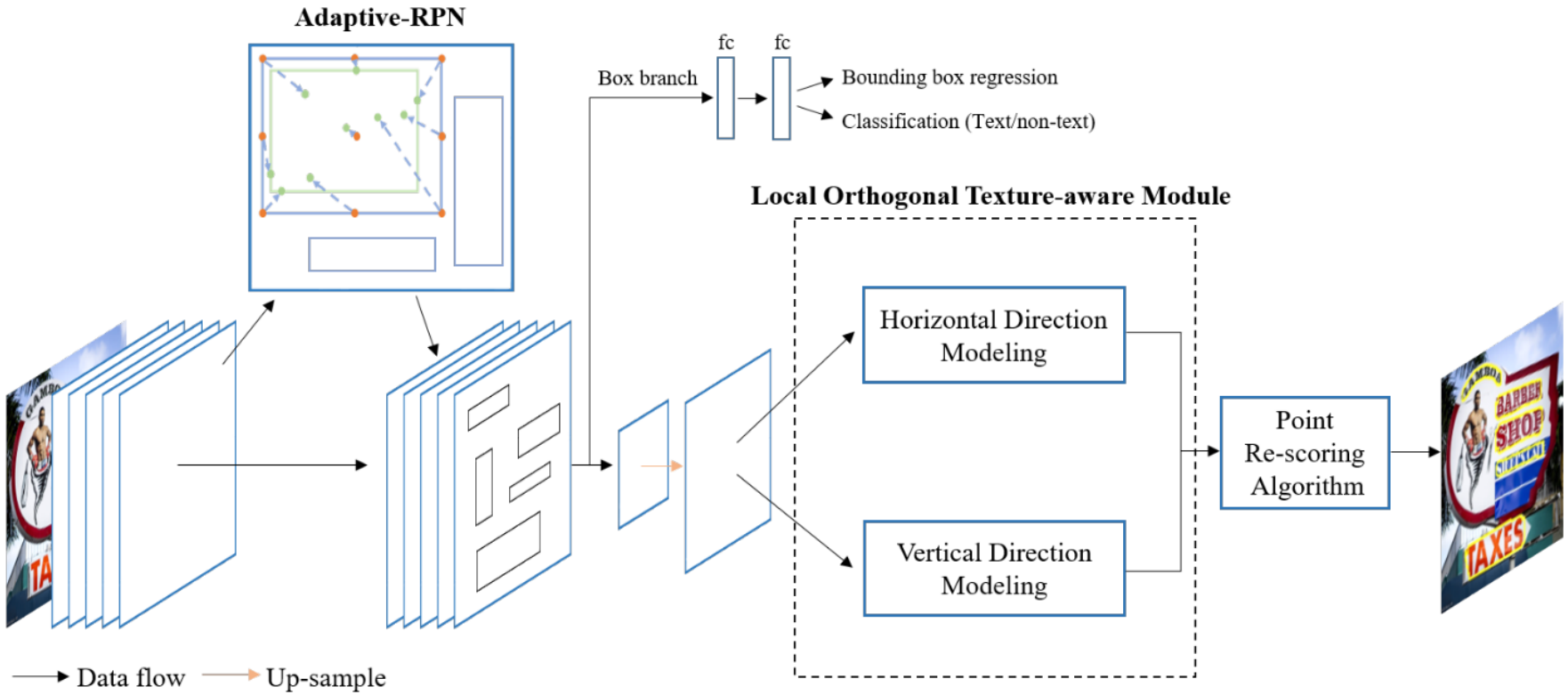
ContourNet



- Existing methods suffer from false positives in their text representations;
- The large scale variance of scene texts makes it hard for network to learn samples.

Yuxin Wang, et al. ContourNet: Taking a Further Step toward Accurate Arbitrary-shaped Scene Text Detection. CVPR 2020.

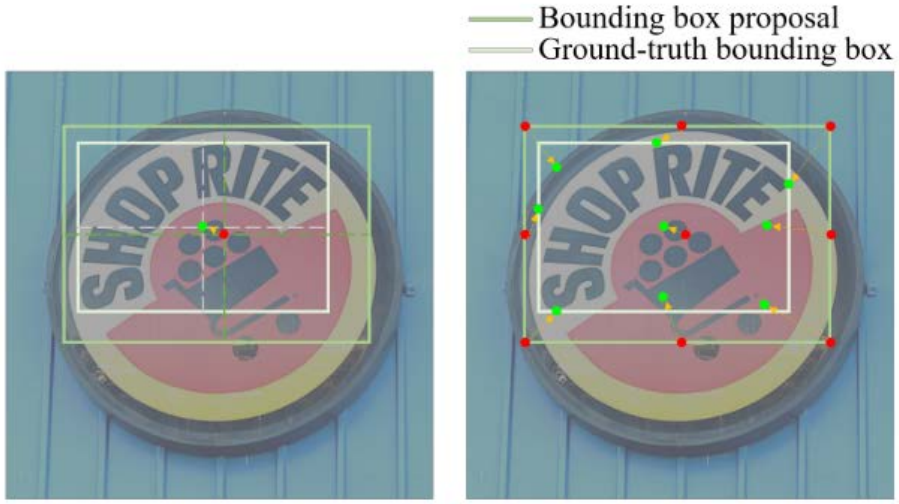
ContourNet

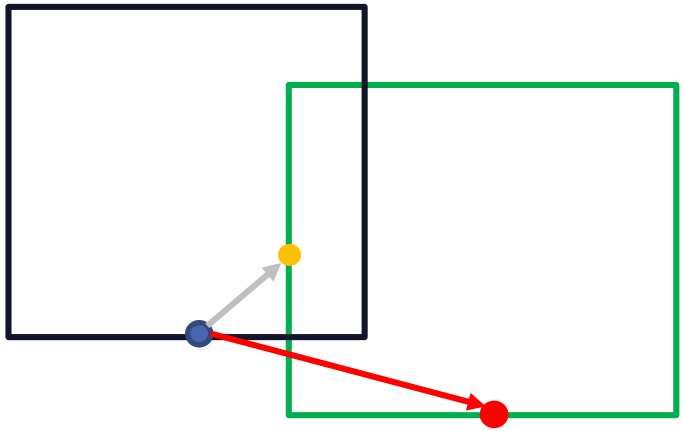


- Based Mask RCNN
- Proposed the Adaptive-RPN
- Local orthogonal Texture-aware module

Yuxin Wang, et al. ContourNet: Taking a Further Step toward Accurate Arbitrary-shaped Scene Text Detection. CVPR 2020.

ContourNet





RPN

Predict 4-d regression vector $\{\Delta x, \Delta y, \Delta w, \Delta h\}$
refine current bounding box proposal $B_c = \{x_c, y_c, w_c, h_c\}$

$$B_t = \{x_c + w_c \Delta x_c, y_c + h_c \Delta y_c, w_c e^{\Delta w_c}, h_c e^{\Delta h_c}\}$$

Adaptive RPN

Use a set of pre-defined points P $P = \{(x_l, y_l)\}_{l=1}^{n}$

$$R = \{x_r, y_r\}_{r=1}^{n} = \{(x_l + w_c \Delta x_l, y_l + h_c \Delta y_l)\}_{l=1}^{n}$$

$$\begin{aligned} Proposal &= \{x_{tl}, y_{tl}, x_{rb}, y_{rb}\} \\ &= \{min\{x_r\}_{r=1}^{n}, min\{y_r\}_{r=1}^{n}, \\ &\quad max\{x_r\}_{r=1}^{n}, max\{y_r\}_{r=1}^{n}\} \end{aligned}$$

Yuxin Wang, et al. ContourNet: Taking a Further Step toward Accurate Arbitrary-shaped Scene Text Detection. CVPR 2020.

Outline

- Background and Introduction
- Datasets and Related Knowledge
- Conventional Methods
- Deep Learning Methods
- Conclusion and Outlook

Conclusion and Outlook

- Evolution path
  - Pre-deep-learning era [1914-2013]: conventional techniques and features
    - MSER [Neumann *et al.*, 2010;]
    - SWT [Epshtein *et al.*, 2010; Yao *et al.*, 2012]
    - HOG [Wang *et al.*, 2011]
    - CRF [Mishra *et al.*, 2011]
  - Transition period [2013-2015]: mixture of conventional techniques/features and deep models/features
    - HOG+DNN [Bissacco *et al.*, 2013]
    - MSER+CNN [Huang *et al.*, 2014; Zhang *et al.*, 2015]
    - HOG+LSTM [Su *et al.*, 2014]
  - Deep learning era [2015-now]: "pure" deep models/features
    - CNN [Gupta *et al.*, 2016]
    - RNN [Ghosh *et al.*, 2016]
    - FCN [Yao *et al.*, 2016; Zhou *et al.*, 2017]
    - Faster-RCNN [Busta *et al.*, 2017]

https://en.wikipedia.org/wiki/Optical_character_recognition

Conclusion and Outlook

- Grand challenges remain
  - Diversity of text: language, font, scale, orientation, arrangement, etc.
  - Complexity of background: virtually indistinguishable elements (signs, fences, bricks and grasses, etc.)
  - Interferences: noise, blur, distortion, low resolution, nonuniform illumination, partial occlusion, etc.

Conclusion and Outlook

- Future Trends
  - Stronger models (accuracy, efficiency, interpretability)
  - Data synthesis
  - Muiti-oriented text
  - Curved text
  - Muiti-language text

Thank You!