如何让计算机读懂图片

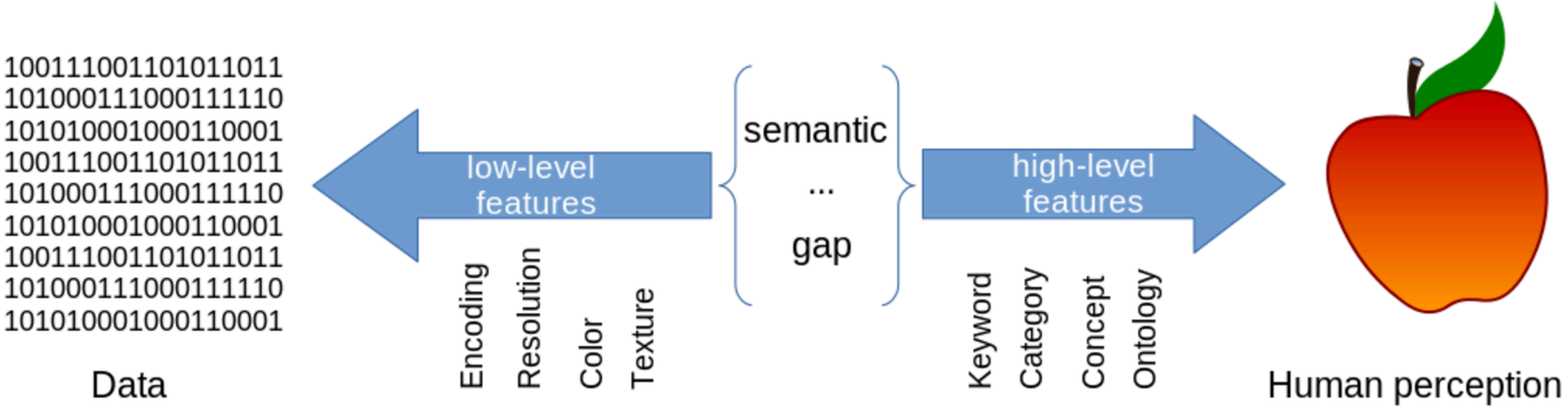
王佳军 · 饿了么

百年老店
周大福
CHOW TAI FOOK
医疗保险定点零售药店
20%

Weapons to Understand Images

- Object Detection
- Text Detection
- Text Recognition

Semantic Gap



100111001101011011
101000111000111110
101010001000110001
100111001101011011
101000111000111110
101010001000110001
100111001101011011
101000111000111110
101010001000110001

Data

low-level features

Encoding

Resolution

Color

Texture

semantic … gap

high-level features

Keyword

Category

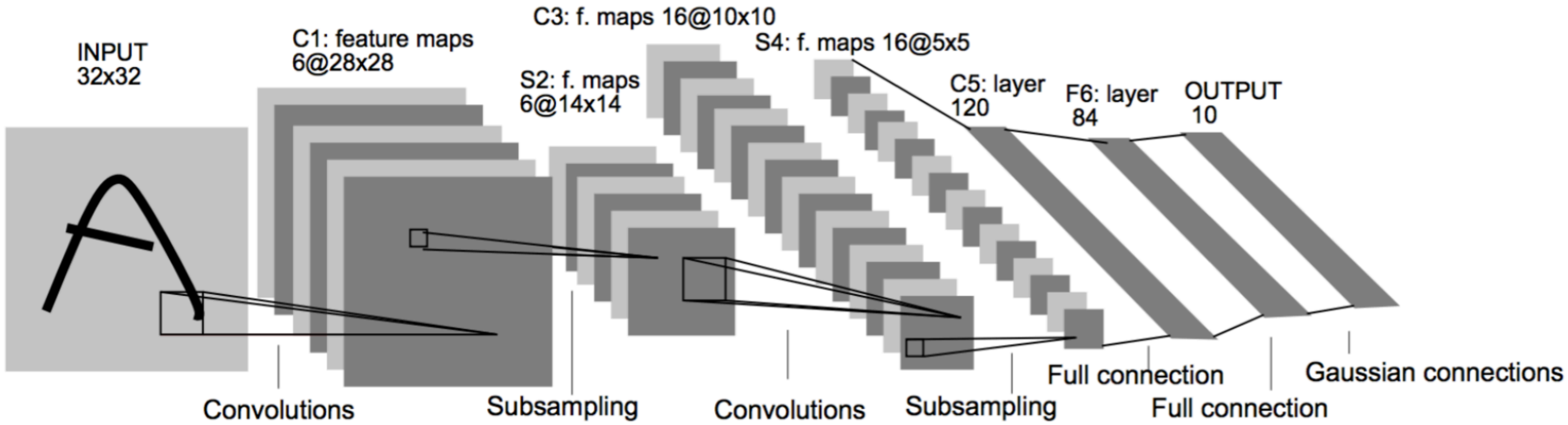
Concept

Ontology

Human perception

https://www.wikiwand.com/en/Semantic_gap

First Glimpse of Deep Learning



Foolproof image classifier

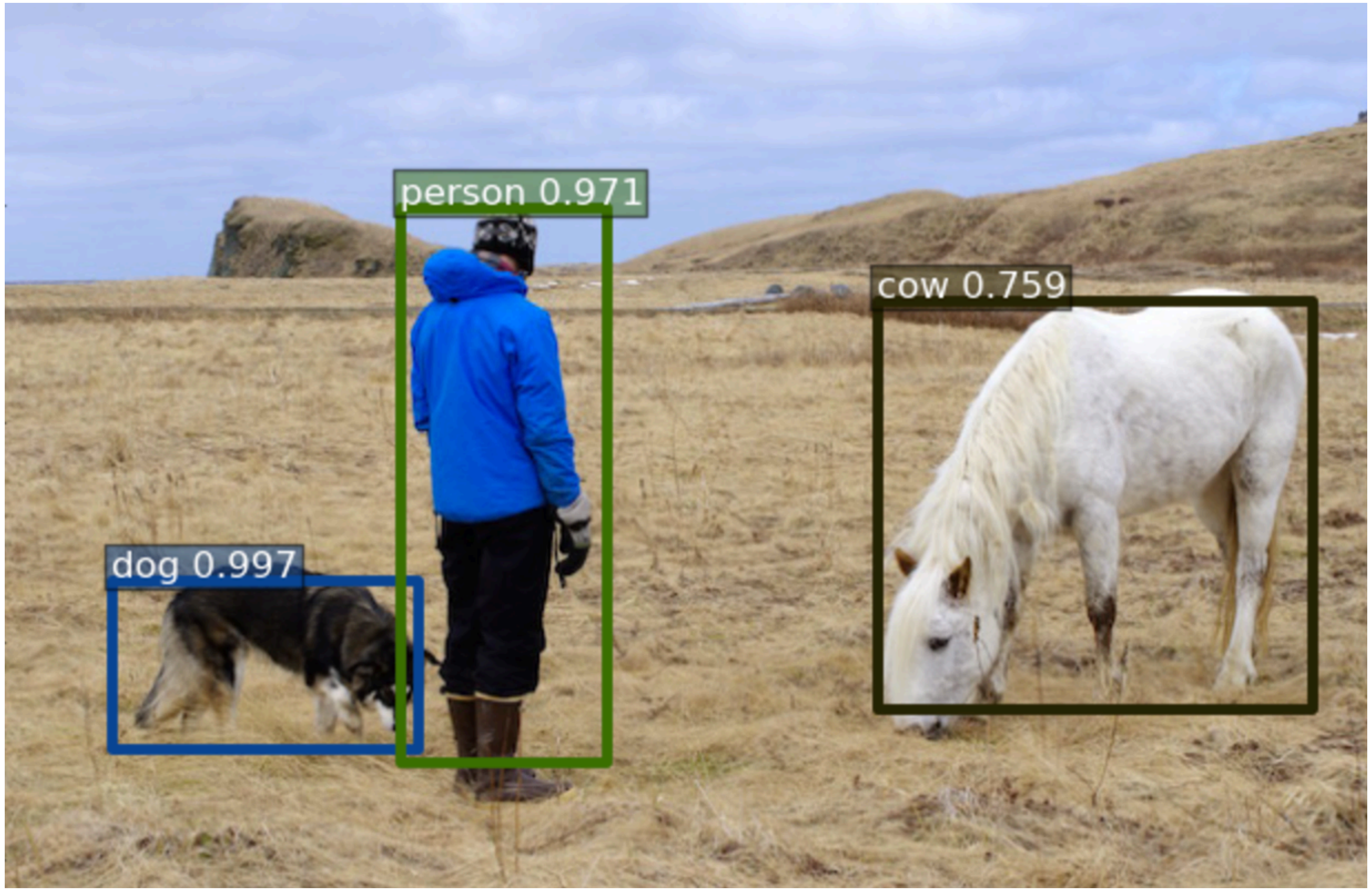
Weapons to Understand Images

- Object Detection

- Text Detection

- Text Recognition

What is Object Detection

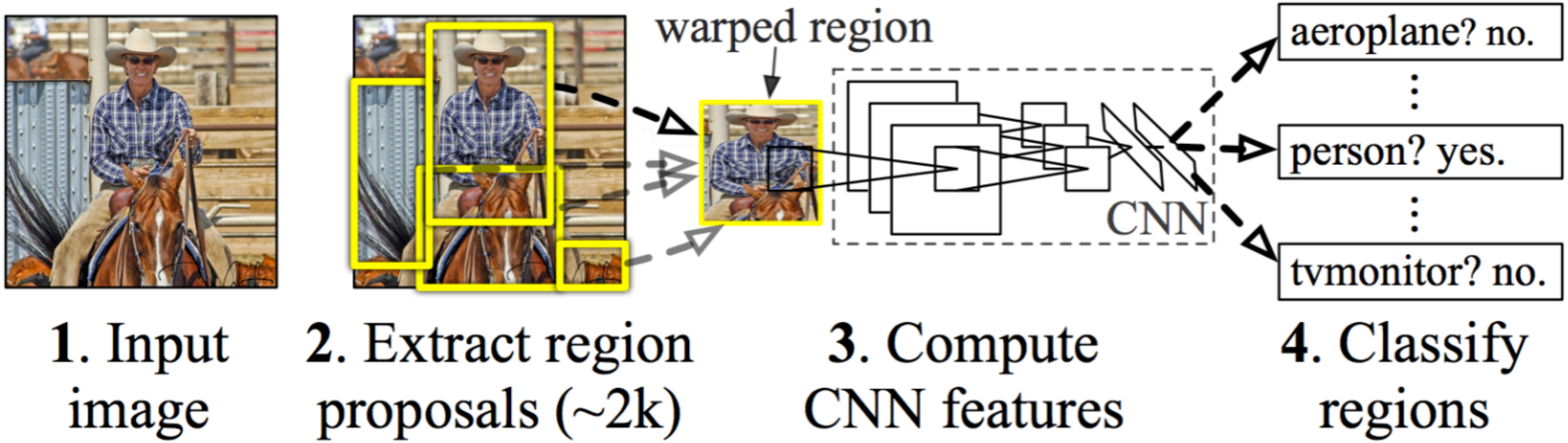


Scale

Ratio

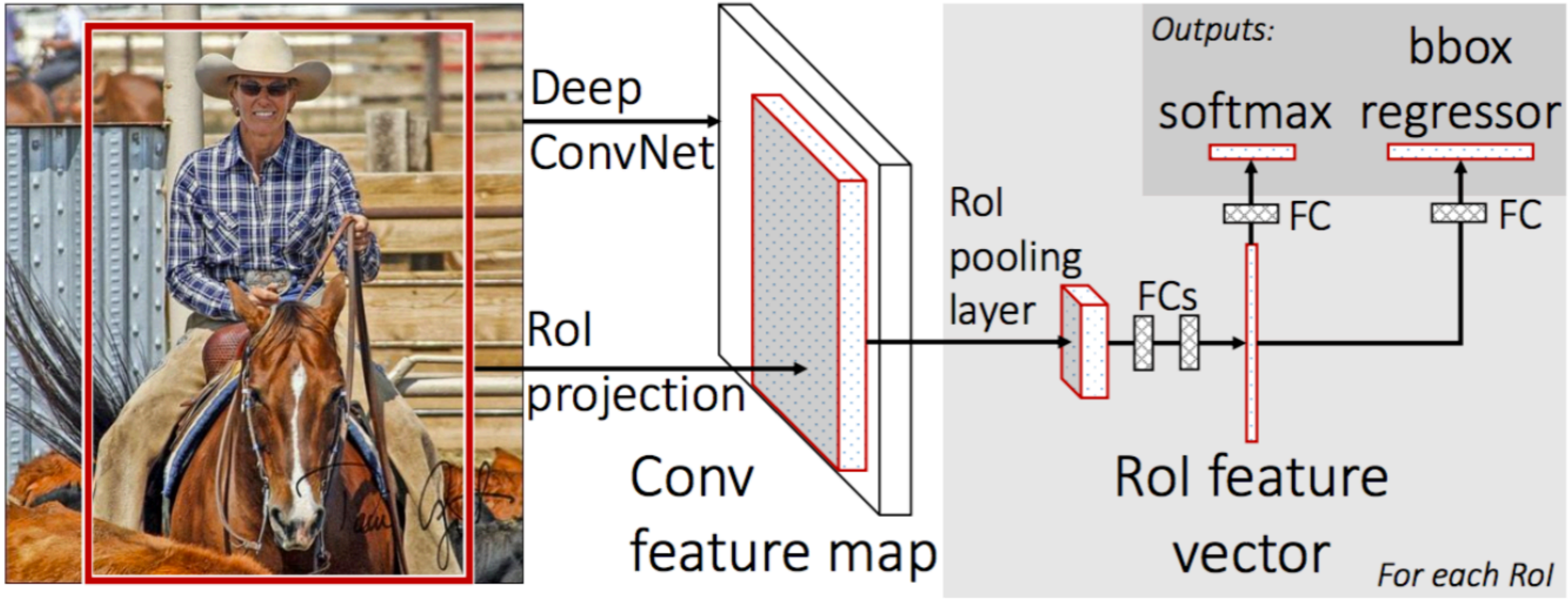
Location

Region Proposal Method



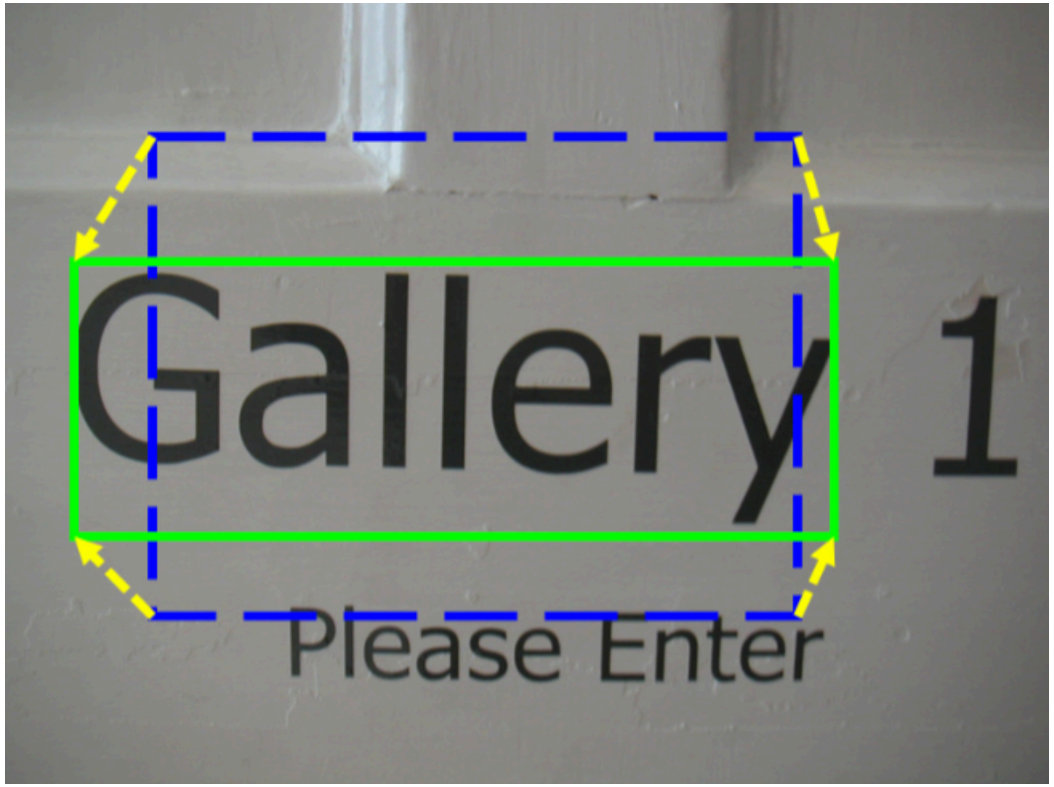
R-CNN (Girshick et al., 2014)

ROI Pooling

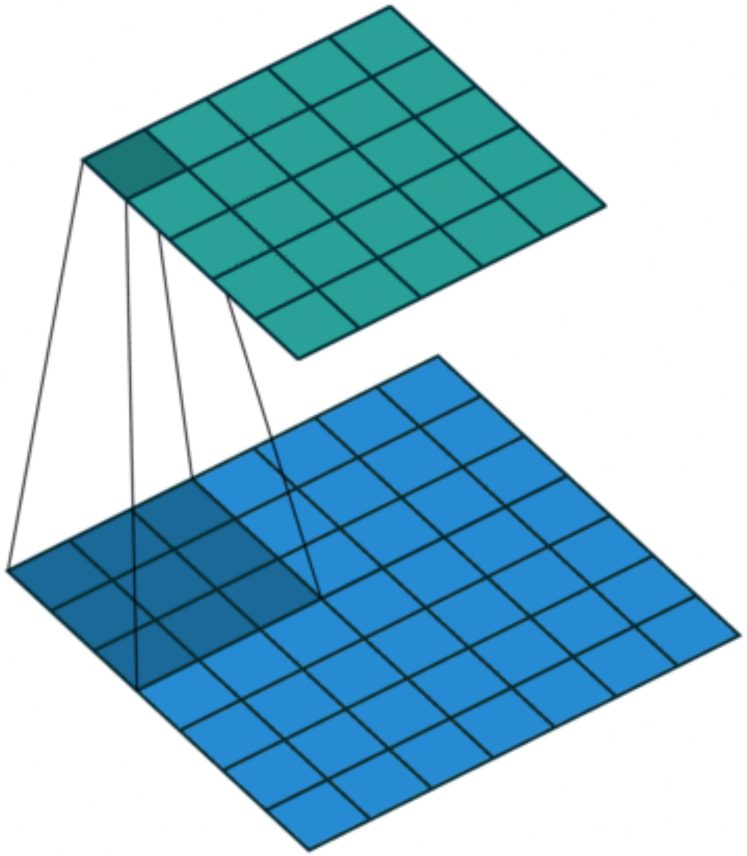


Fast R-CNN (Girshick, 2015)

Bounding Box Regression

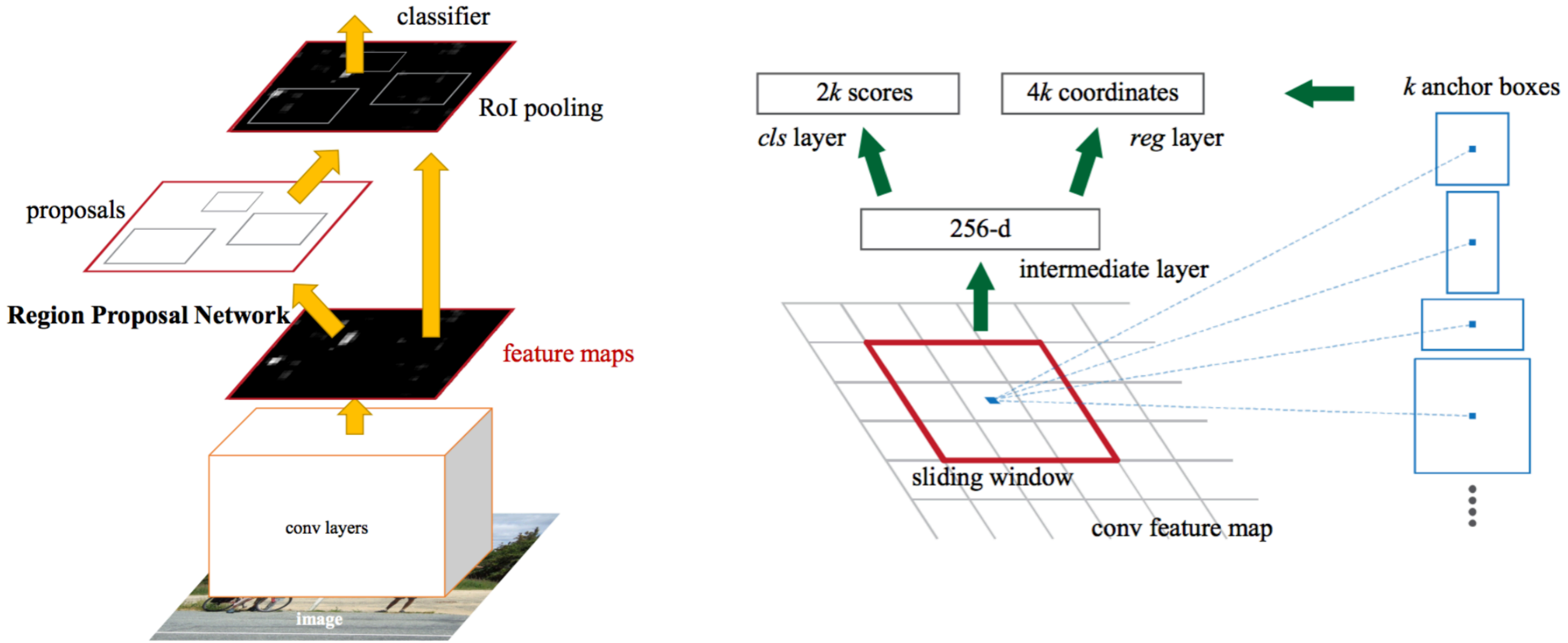


Convolution vs. Sliding Window



https://github.com/vdumoulin/conv_arithmetic

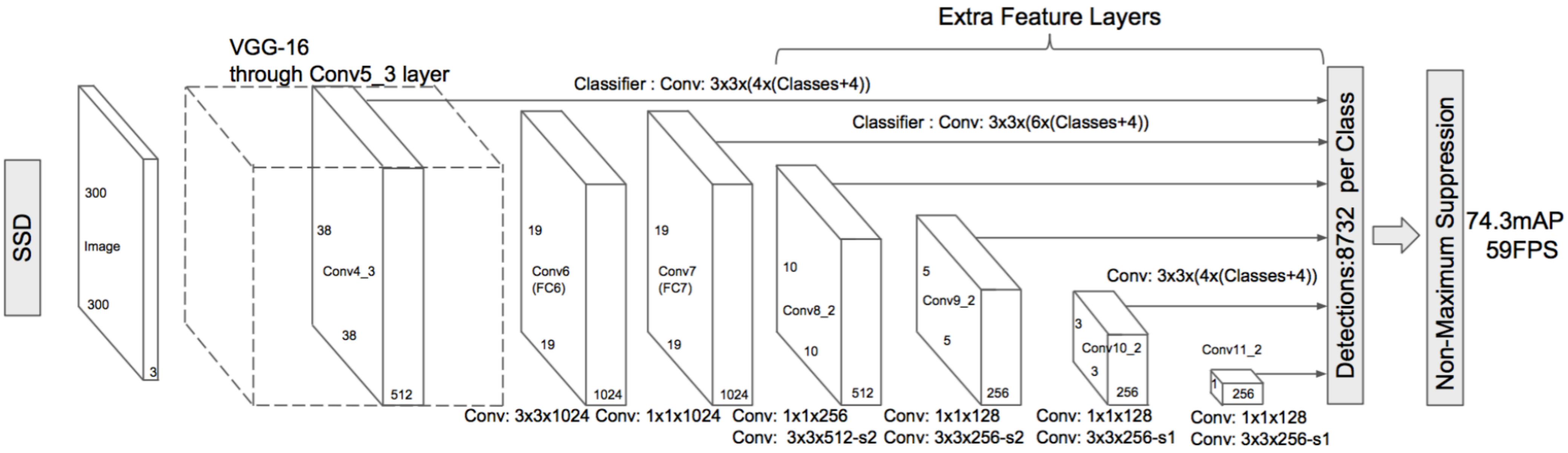
Region Proposal Network, Anchor Boxes



Faster R-CNN (Ren et al., 2015)

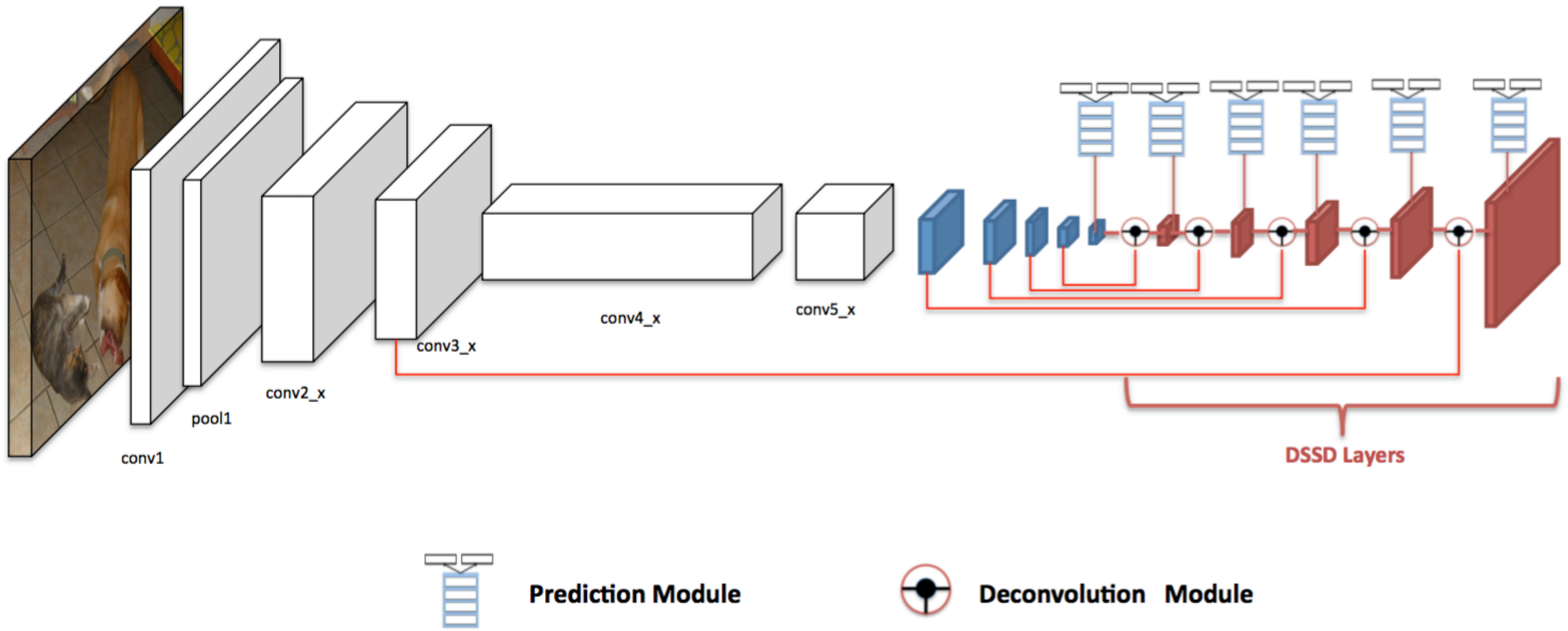
Dense Object Detection

$$kernel \rightarrow \#anchors \times (\#classes + 4)$$



SSD (Liu et al., 2016)

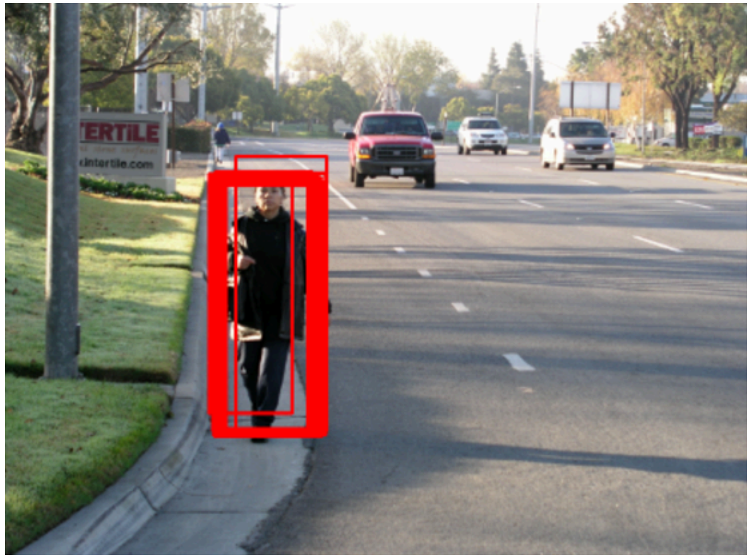
Hourglass Structure



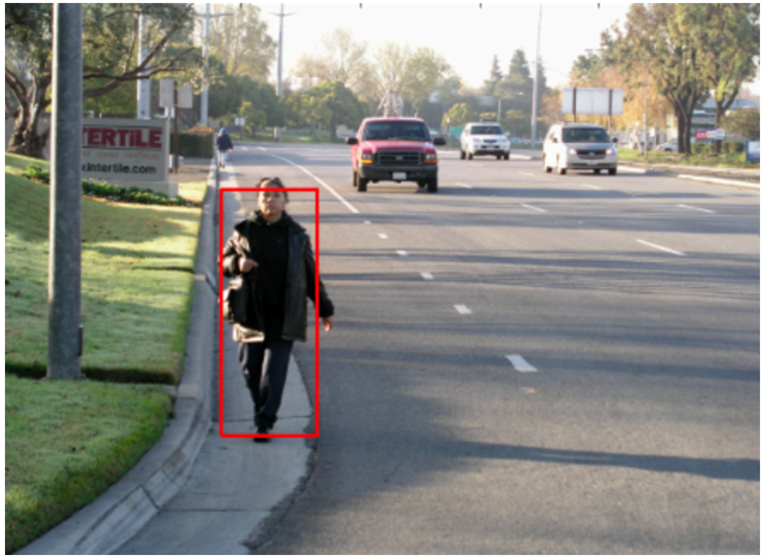
DSSD (Fu et al., 2017)

How to Do Inference

Network forward result



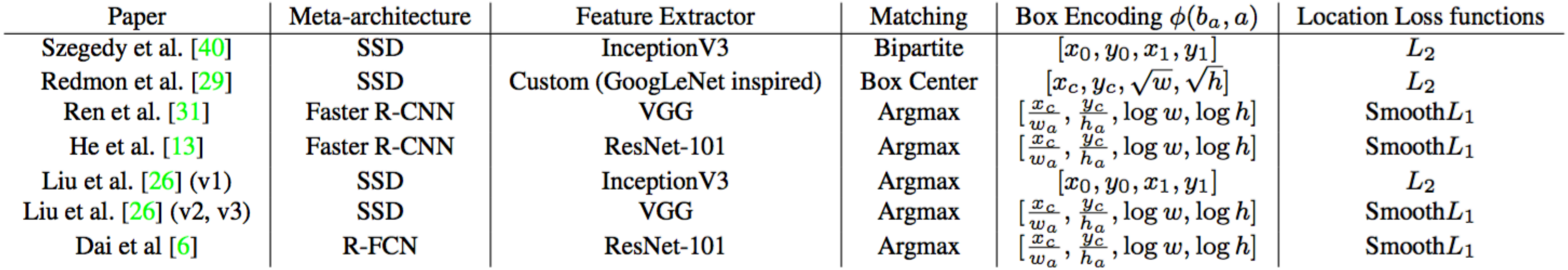
After Non-Maximum Suppression



One-Stage vs. Two-Stage Detector

One-Stage	**RCNN (Girshick et al., 2014)** **Fast RCNN (Girshick, 2015)** **Faster RCNN (Ren et al., 2015)** **R-RCN (Dai et al., 2016)**	**High performance**
Two-Stage	YOLO (Redmon et al., 2016) YOLO9000 (Redmon and Farhadi, 2016) SSD (Liu et al., 2016) DSSD (Fu et al., 2017)	Simple structure Fast speed

Things to be considered



Paper	Meta-architecture	Feature Extractor	Matching	Box Encoding $\phi(b_a, a)$	Location Loss functions
Szegedy et al. [40]	SSD	InceptionV3	Bipartite	$[x_0, y_0, x_1, y_1]$	L_2
Redmon et al. [29]	SSD	Custom (GoogLeNet inspired)	Box Center	$[x_c, y_c, \sqrt{w}, \sqrt{h}]$	L_2
Ren et al. [31]	Faster R-CNN	VGG	Argmax	$[\frac{x_c}{w_a}, \frac{y_c}{h_a}, \log w, \log h]$	SmoothL_1
He et al. [13]	Faster R-CNN	ResNet-101	Argmax	$[\frac{x_c}{w_a}, \frac{y_c}{h_a}, \log w, \log h]$	SmoothL_1
Liu et al. [26] (v1)	SSD	InceptionV3	Argmax	$[x_0, y_0, x_1, y_1]$	L_2
Liu et al. [26] (v2, v3)	SSD	VGG	Argmax	$[\frac{x_c}{w_a}, \frac{y_c}{h_a}, \log w, \log h]$	SmoothL_1
Dai et al [6]	R-FCN	ResNet-101	Argmax	$[\frac{x_c}{w_a}, \frac{y_c}{h_a}, \log w, \log h]$	SmoothL_1

(Huang et al., 2016)

Class Imbalance

- Online hard example mining (Shrivastava et al., 2016)
- Scale classification loss
  - By class frequency (Redmon et al., 2016)
  - By inferred probability (Lin et al., 2017)

Model Comparison (Huang et al., 2016)

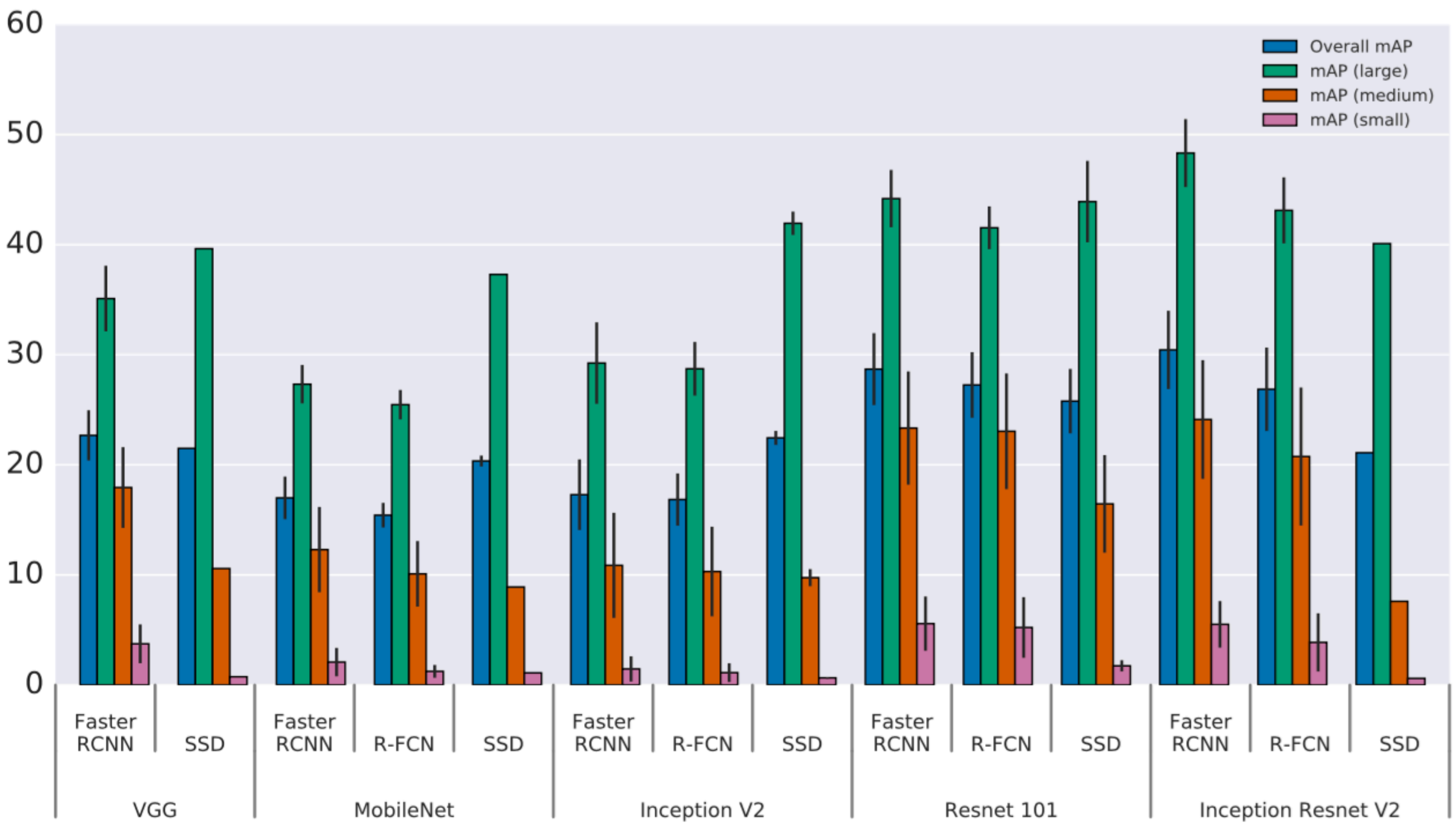


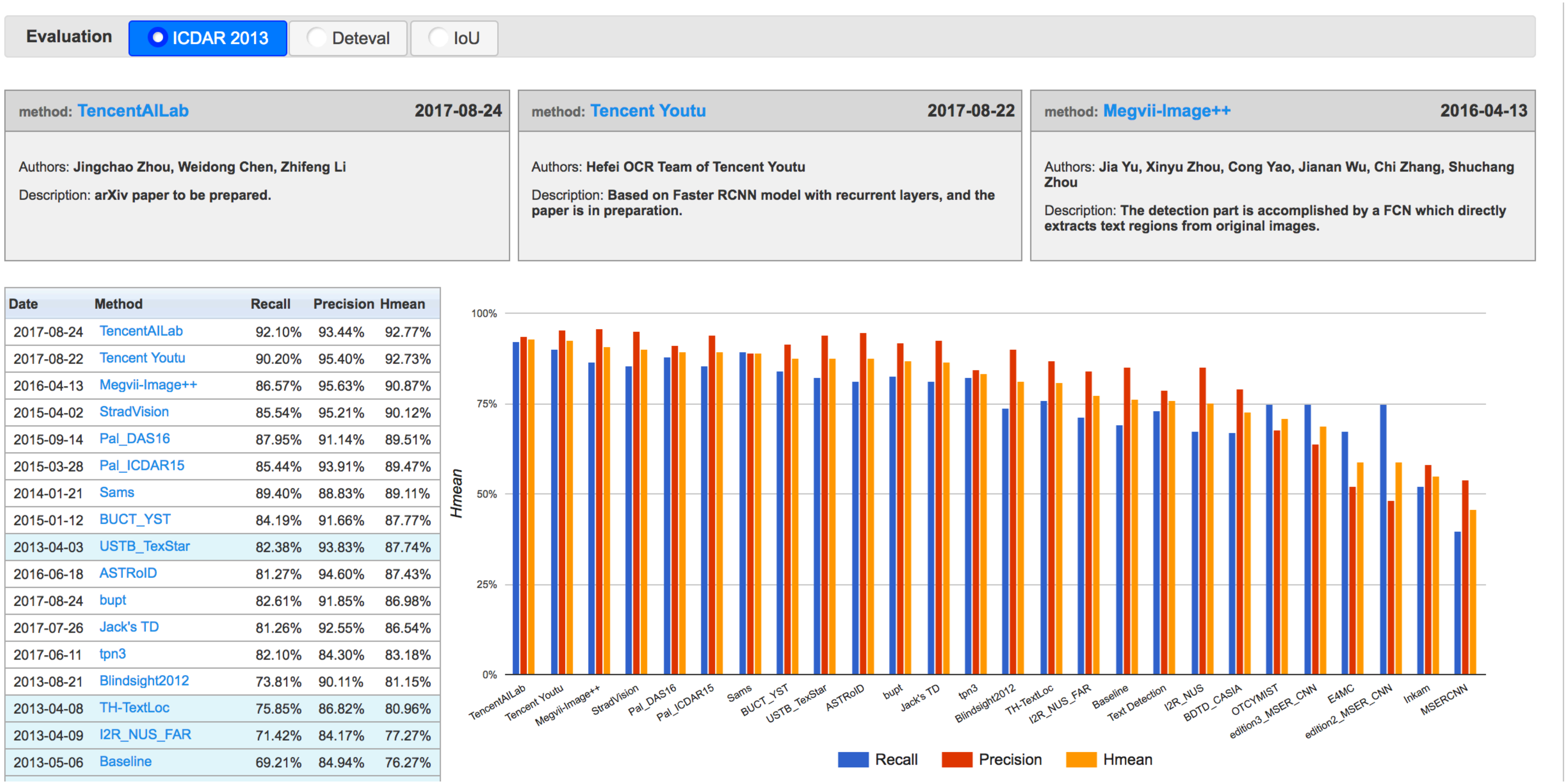
Figure 4: Accuracy stratified by object size, meta-architecture and feature extractor, We fix the image resolution to 300.

Ablation Study

Weapons to Understand Images

- Object Detection
- Text Detection
- Text Recognition

ICDAR Competition



Evaluation ● ICDAR 2013 ○ Deteval ○ IoU

method: TencentAILab	2017-08-24
Authors: Jingchao Zhou, Weidong Chen, Zhifeng Li	
Description: arXiv paper to be prepared.	

method: Tencent Youtu	2017-08-22
Authors: Hefei OCR Team of Tencent Youtu	
Description: Based on Faster RCNN model with recurrent layers, and the paper is in preparation.	

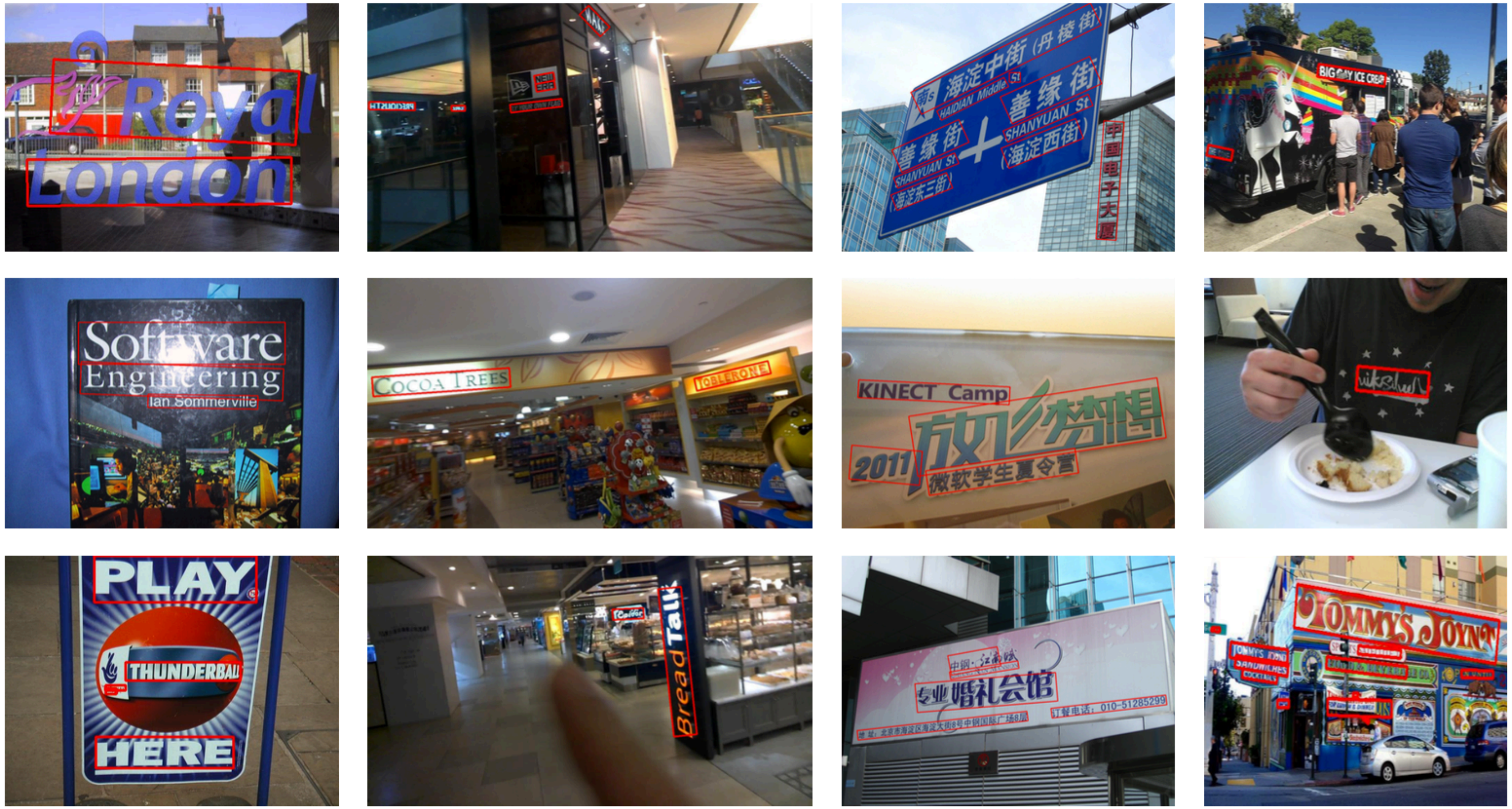
method: Megvii-Image++	2016-04-13
Authors: Jia Yu, Xinyu Zhou, Cong Yao, Jianan Wu, Chi Zhang, Shuchang Zhou	
Description: The detection part is accomplished by a FCN which directly extracts text regions from original images.	

Date	Method	Recall	Precision	Hmean
2017-08-24	TencentAILab	92.10%	93.44%	92.77%
2017-08-22	Tencent Youtu	90.20%	95.40%	92.73%
2016-04-13	Megvii-Image++	86.57%	95.63%	90.87%
2015-04-02	StradVision	85.54%	95.21%	90.12%
2015-09-14	Pal_DAS16	87.95%	91.14%	89.51%
2015-03-28	Pal_ICDAR15	85.44%	93.91%	89.47%
2014-01-21	Sams	89.40%	88.83%	89.11%
2015-01-12	BUCT_YST	84.19%	91.66%	87.77%
2013-04-03	USTB_TexStar	82.38%	93.83%	87.74%
2016-06-18	ASTRoID	81.27%	94.60%	87.43%
2017-08-24	bupt	82.61%	91.85%	86.98%
2017-07-26	Jack's TD	81.26%	92.55%	86.54%
2017-06-11	tpn3	82.10%	84.30%	83.18%
2013-08-21	Blindsight2012	73.81%	90.11%	81.15%
2013-04-08	TH-TextLoc	75.85%	86.82%	80.96%
2013-04-09	I2R_NUS_FAR	71.42%	84.17%	77.27%
2013-05-06	Baseline	69.21%	84.94%	76.27%

Text Detection as Specialized Object Detection

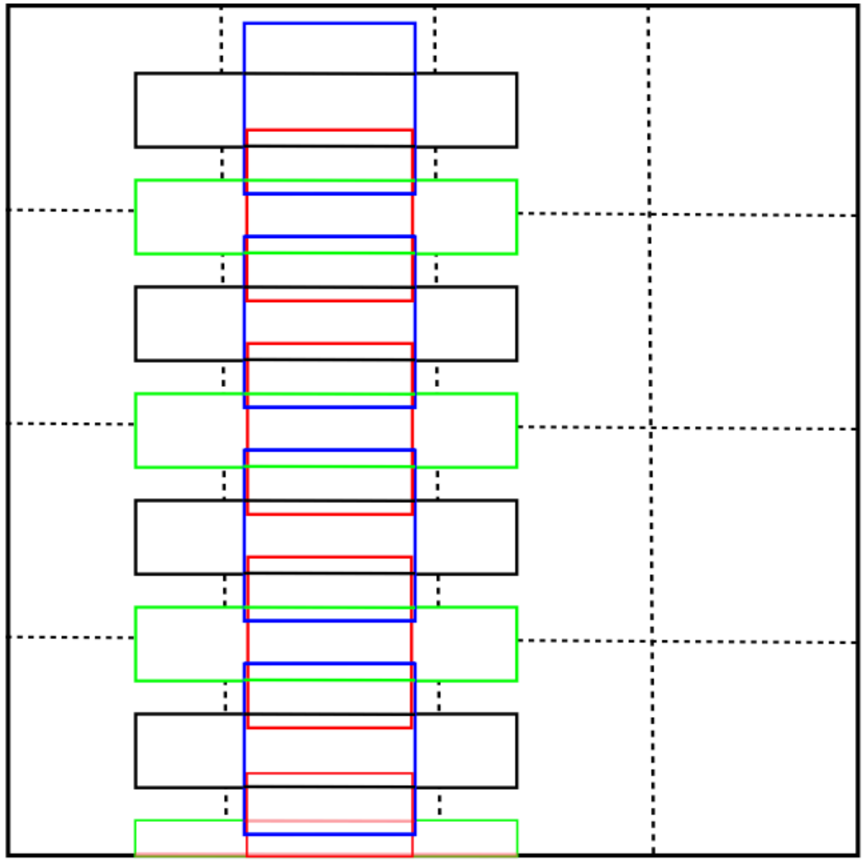
Methods	Paper
Faster RCNN	DeepText (Zhong et al., 2016) Rotation Proposal (Ma et al., 2017)
SSD	TextBoxes (Liao et al., 2017) SegLink (Shi et al., 2017) Deep Matching Prior Network (Liu and Jin, 2017)
Hourglass Structure	Deep Direct Regression (He et al., 2017) EAST (Zhou et al., 2017) Multi-Channel Prediction (Yao et al., 2016)

What’s Unique in Text Detection

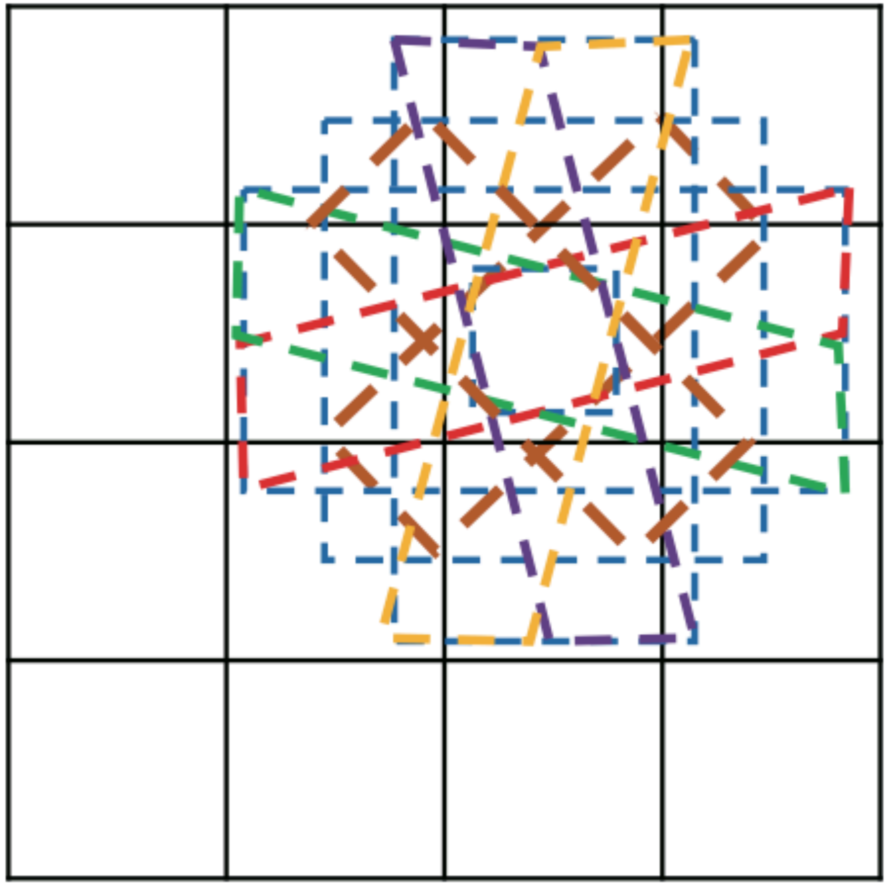


(Zhou et al., 2017)

Extended Anchor Boxes

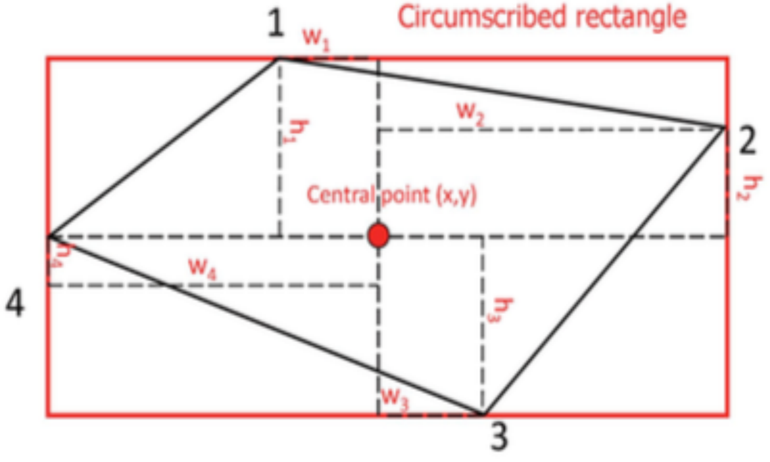


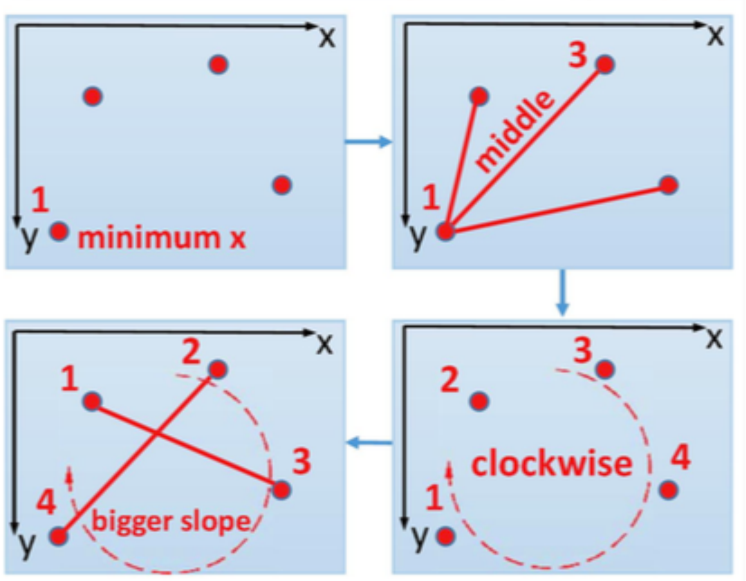
(Liao et al., 2017)

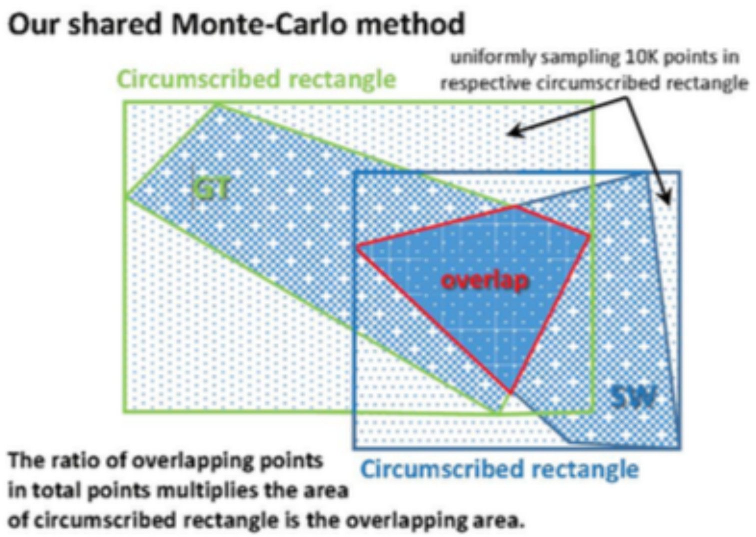


(Liu and Jin, 2017)

Quadrilateral Bounding Box





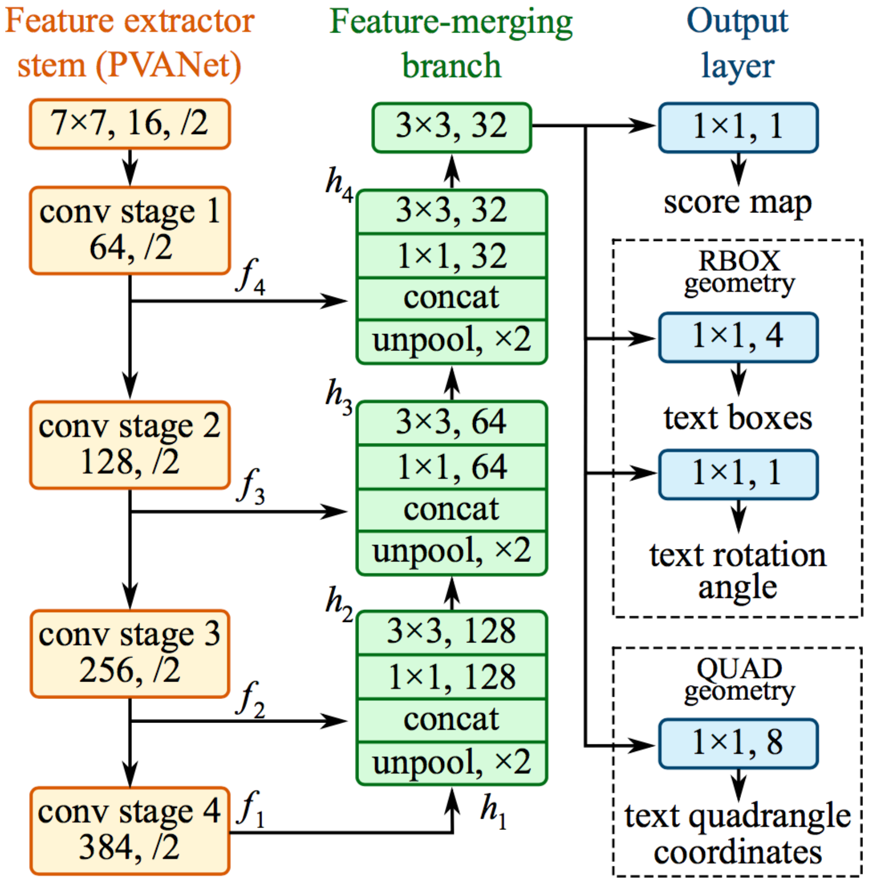

Our shared Monte-Carlo method

uniformly sampling 10K points in respective circumscribed rectangle

The ratio of overlapping points in total points multiplies the area of circumscribed rectangle is the overlapping area.

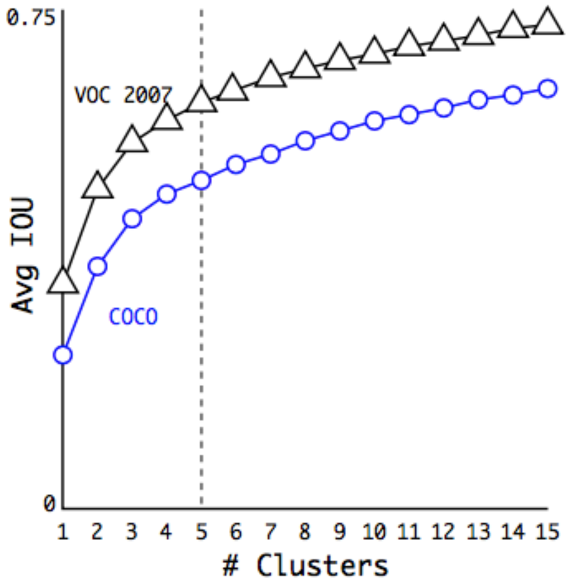
(Liu and Jin, 2017)

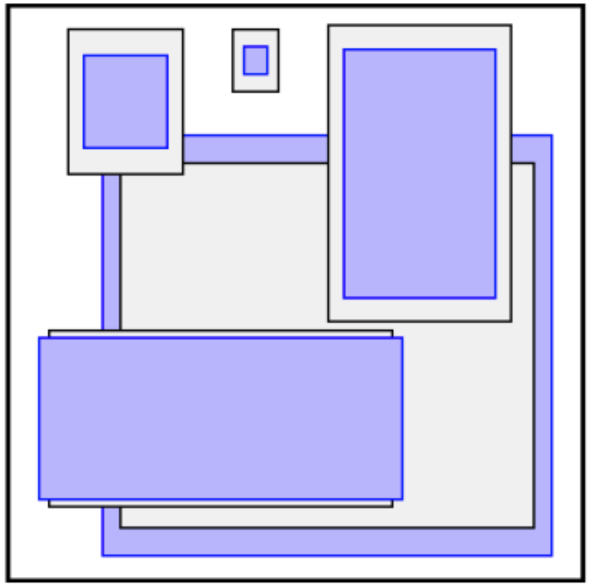
EAST (Zhou et al., 2017)



Anchor Clustering With K-means

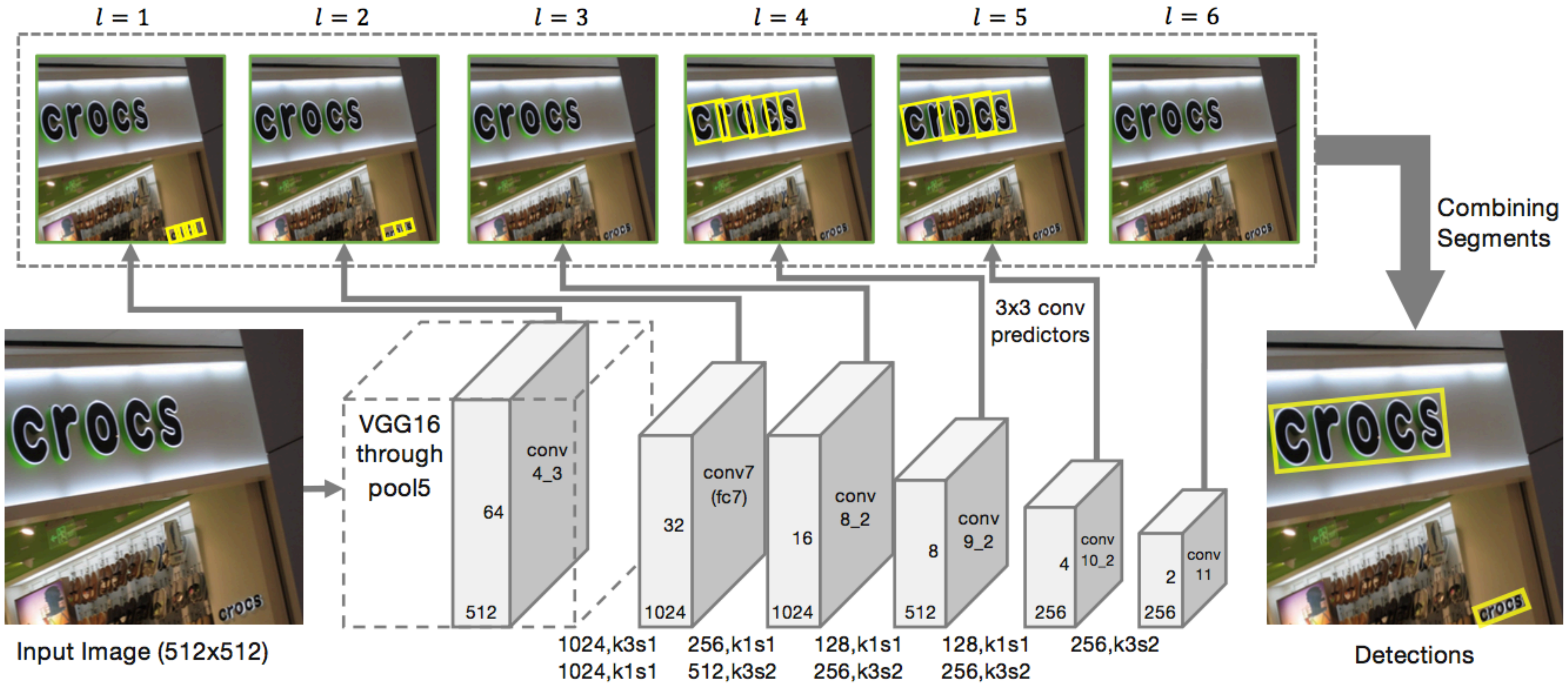
$$d(box, centroid) = 1 - IOU(box, centroid)$$





YOLO9000 (Redmon and Farhadi, 2016)

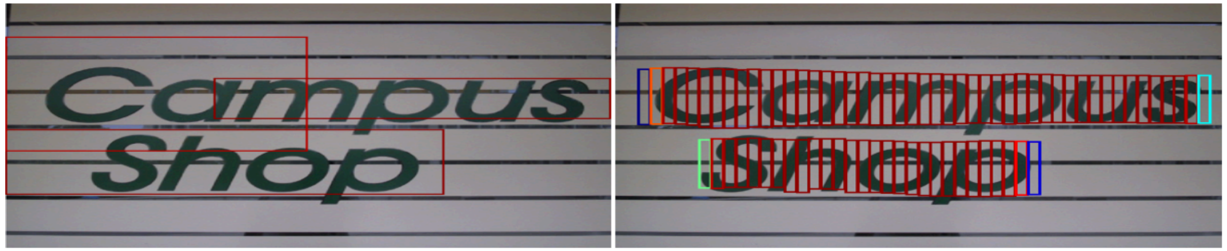
Long Text



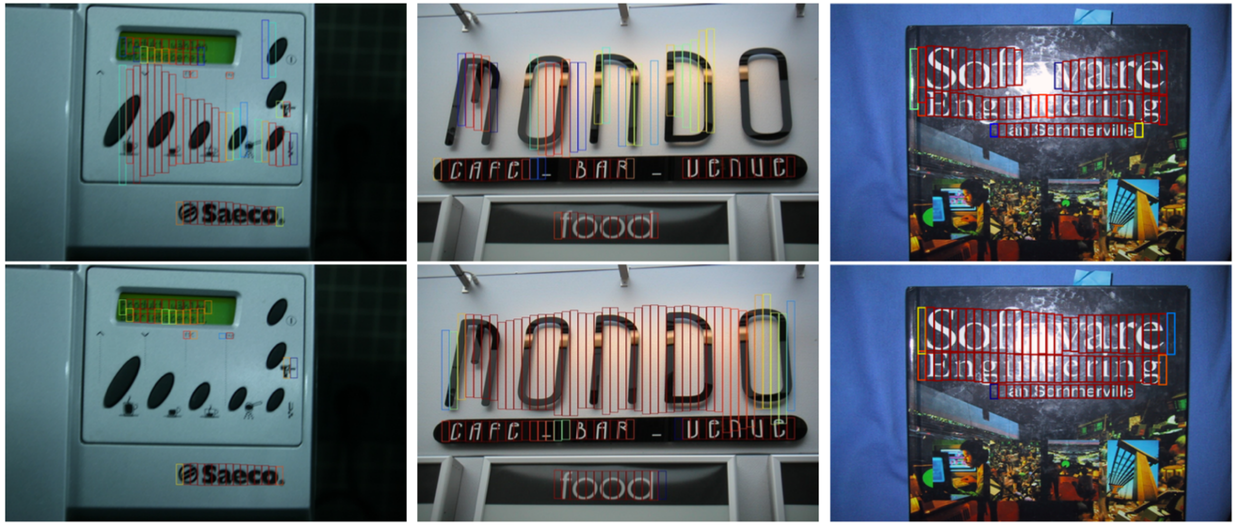
SegLink (Shi et al., 2017)

Long Text (cont.)

Fine-scale proposals



Recurrent connectionist text proposals



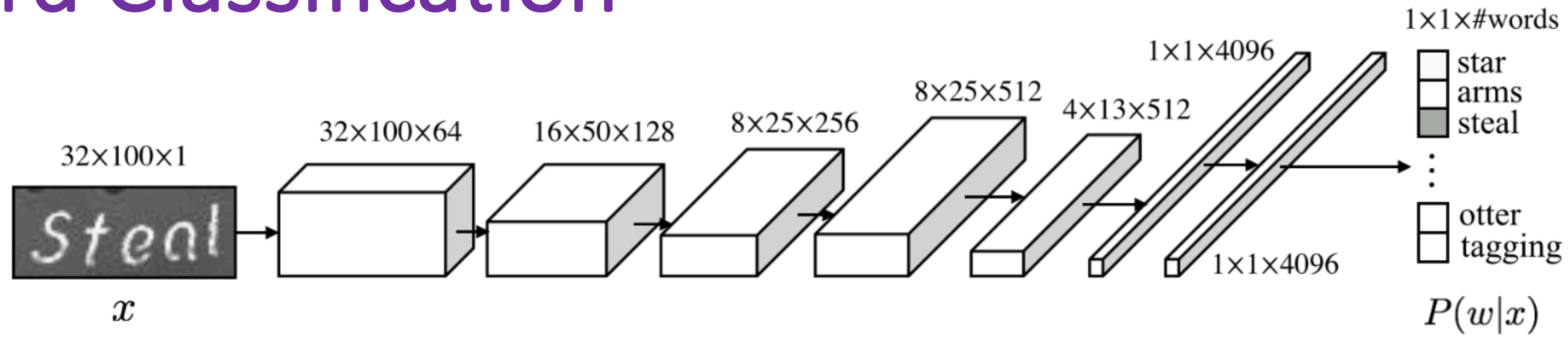
CTPN (Tian et al., 2016)

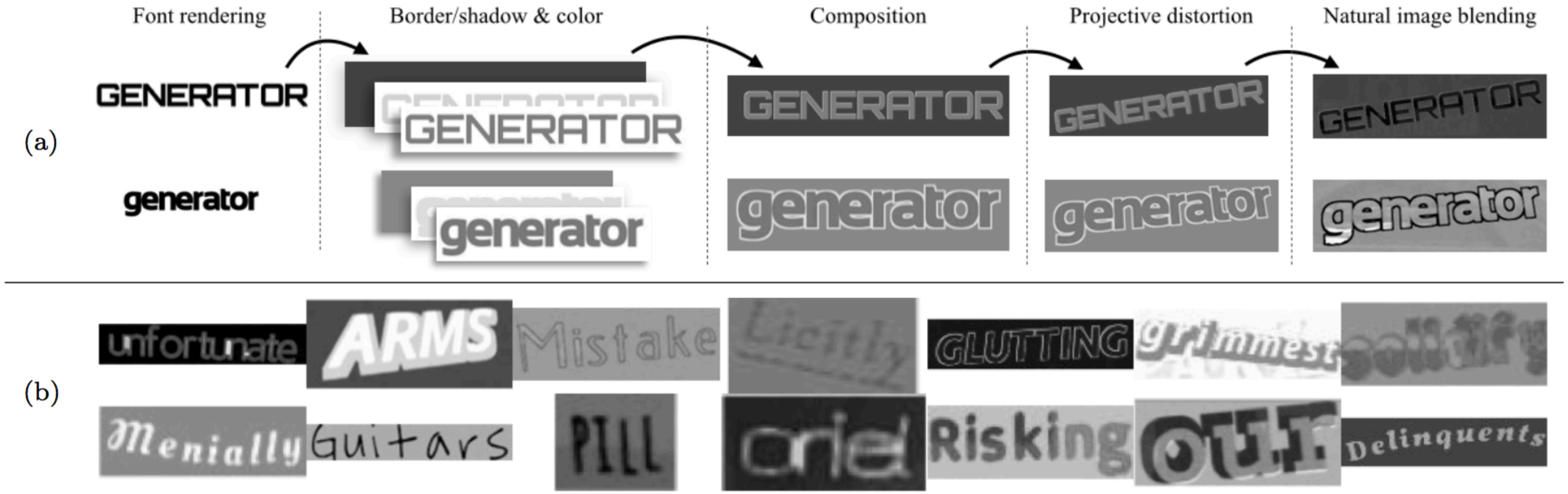
Weapons to Understand Images

- Object Detection
- Text Detection
- Text Recognition

Segmentation-free

Word Classification





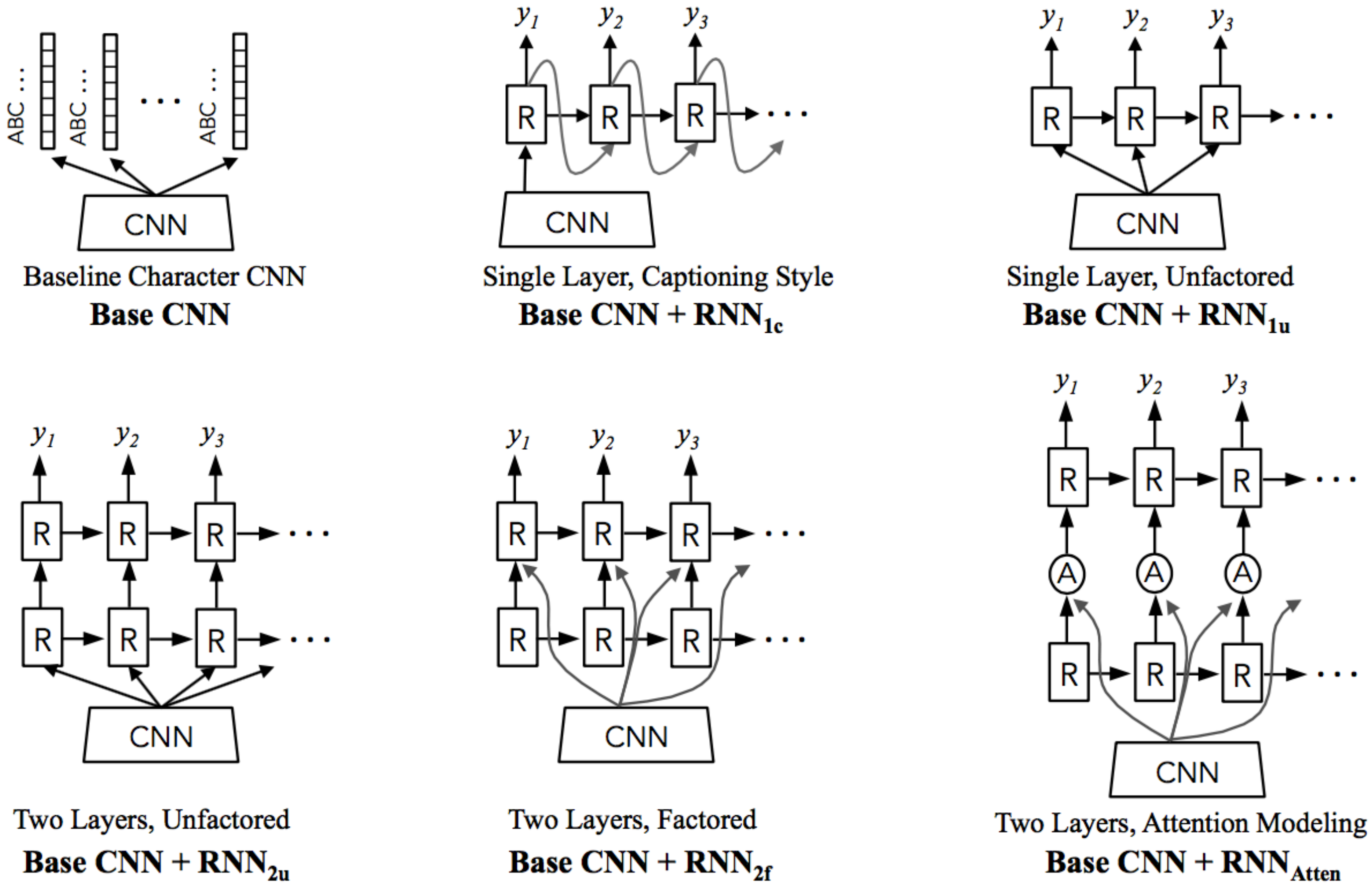
(Jaderberg et al., 2016)

RNN-based Method

- Seq2seq, encoder-decoder structure
  - (Shi et al., 2016b)
  - (Lee and Osindero, 2016)

- CTC loss
  - (He et al., 2016)
  - (Shi et al., 2016a)

Seq2seq



Baseline Character CNN
Base CNN

Single Layer, Captioning Style
Base CNN + RNN$_{1c}$

Single Layer, Unfactored
Base CNN + RNN$_{1u}$

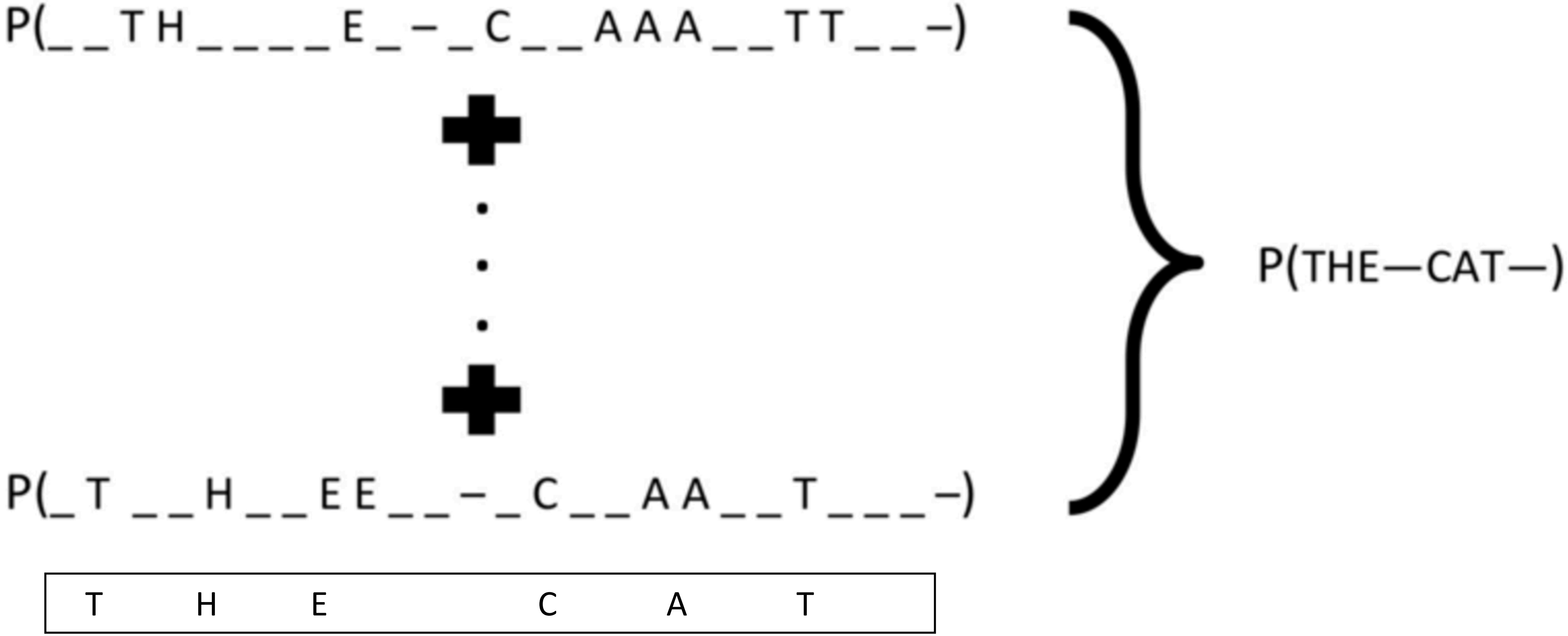
Two Layers, Unfactored
Base CNN + RNN$_{2u}$

Two Layers, Factored
Base CNN + RNN$_{2f}$

Two Layers, Attention Modeling
Base CNN + RNN$_{Atten}$

(Lee and Osindero, 2016)

Connectionist Temporal Classification



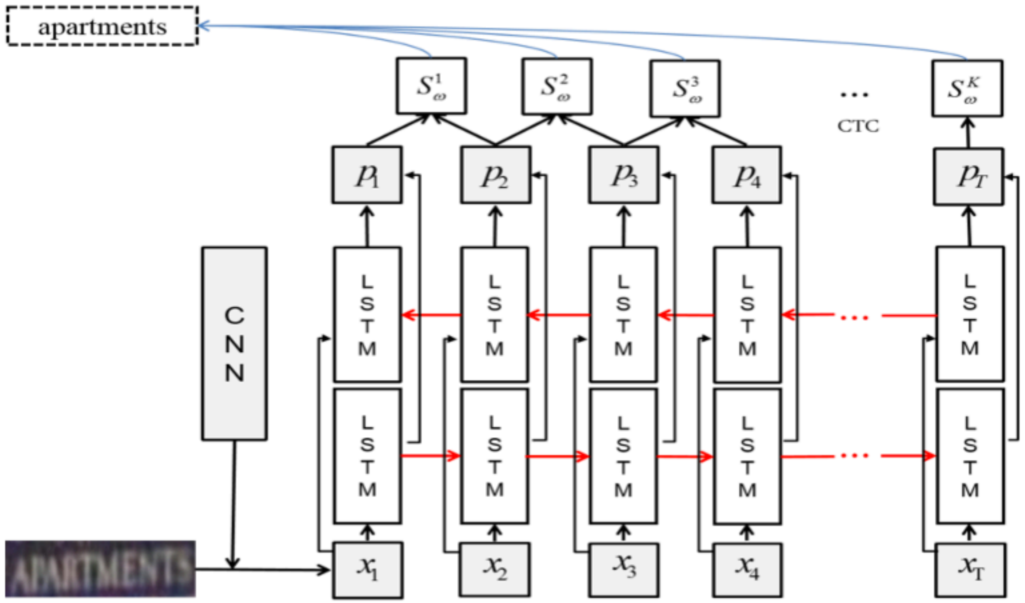
P(_ _ T H _ _ _ _ _ E _ – _ C _ _ A A A _ _ T T _ _ –)

P(_ T _ _ H _ _ E E _ _ – _ C _ _ A A _ _ T _ _ _ –)

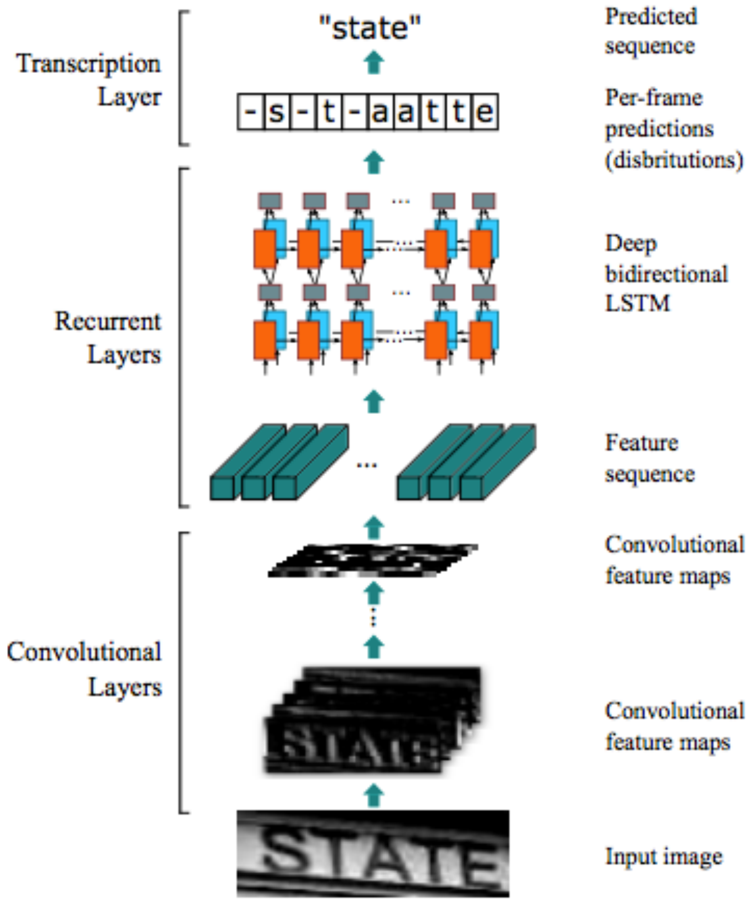
P(THE—CAT—)

T H E C A T

Connectionist Temporal Classification (cont.)



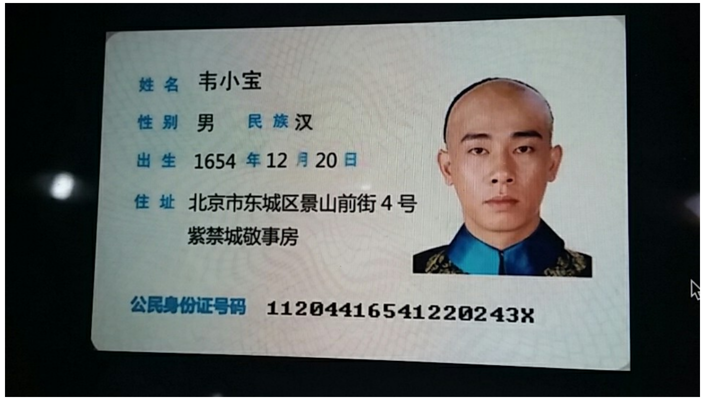
(He et al., 2016)



(Shi et al., 2016a)

饿了么实践

OCR



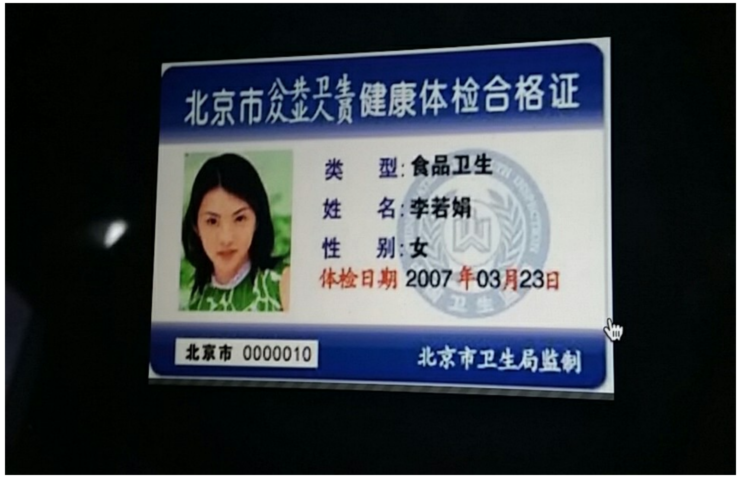
姓名 韦小宝

性别 男 民族 汉

出生 1654 年 12 月 20 日

住址 北京市东城区景山前街 4 号
紫禁城敬事房

公民身份证号码 11204416541220243X



北京市公共卫生从业人员健康体检合格证

类 型：食品卫生

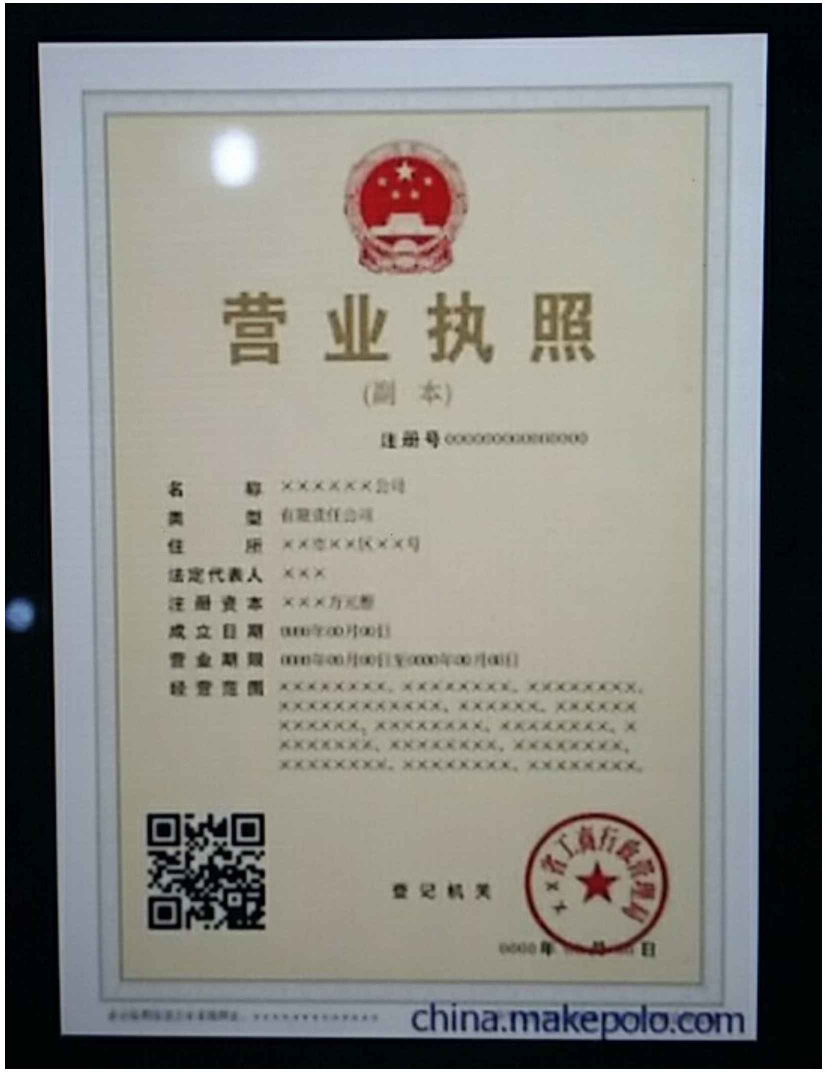
姓 名：李若娟

性 别：女

体检日期 2007 年03月23日

北京市 0000010

北京市卫生局监制



营 业 执 照

（副 本）

注册号 000000000000000

名 称 ×××××××公司

类 型 有限责任公司

住 所 ××市××区××号

法定代表人 ×××

注册资本 ×××万元整

成立日期 0000年00月00日

营业期限 0000年00月00日至0000年00月00日

经营范围 ×××××××××、×××××××××、×××××××××、
×××××××××××××、××××××、××××××
××××××、×××××××××、×××××××××、×
××××××××、×××××××××、×××××××××、
×××××××××、×××××××××、×××××××××、

登记机关

××省工商行政管理局

0000年00月00日

china.makepolo.com

违规图片审核

美团网
meituan.com

正规官网黄金会员
稳定1个月
每个ID限拍一个
不掉线 不改密

秒杀
限时抢购

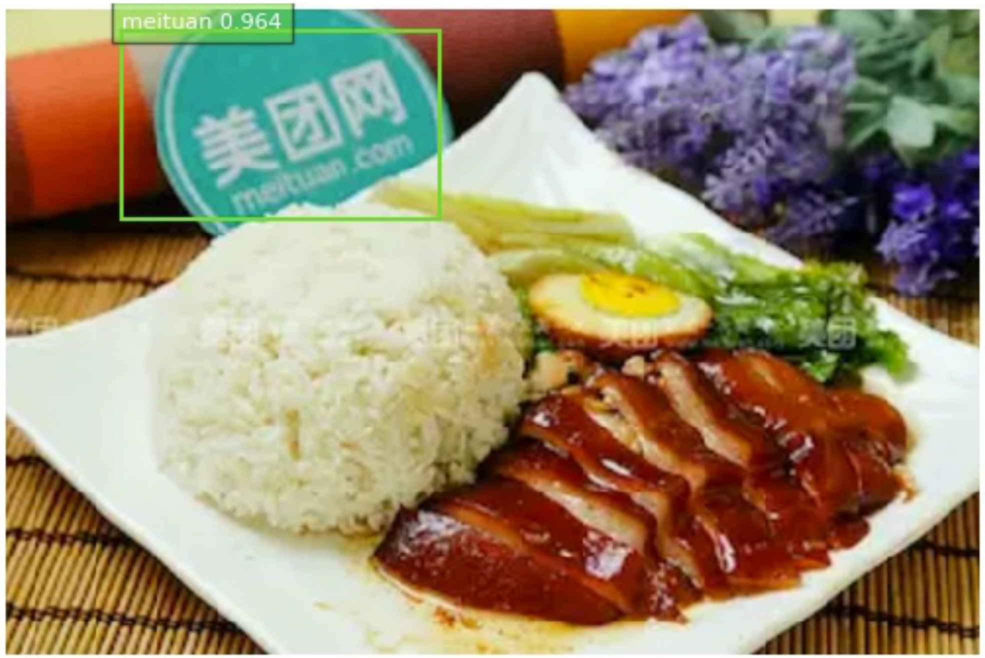
百度糯米 nuomi.com
省钱更省心

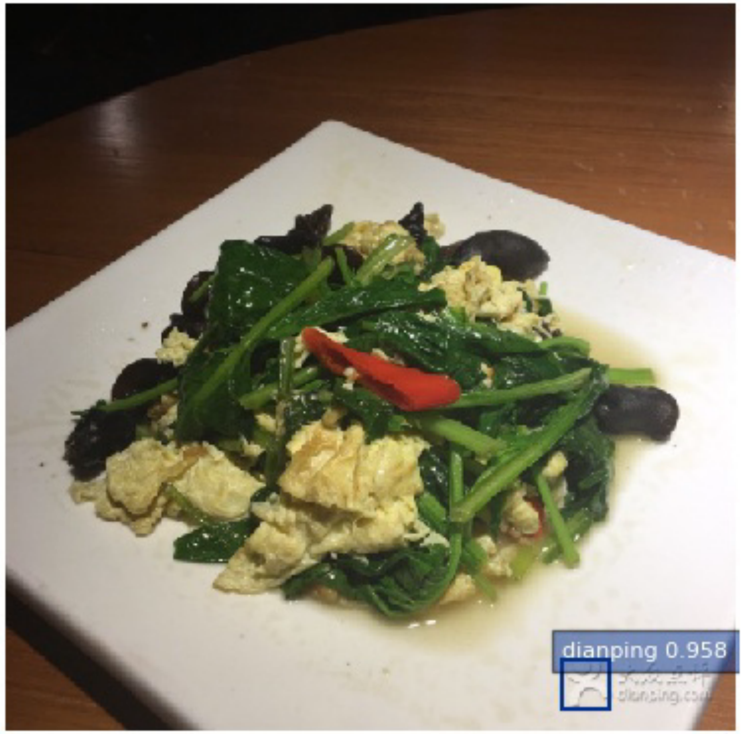
美团外卖
美团外卖 送啥都快
扫描下载客户端

手机充值
手机充值
手机充值 SHOUJI SHONGZHI

Logo检测

- Reduced-VGG-SSD+ResNet





证件OCR

- Detection： TextBoxes
- Recognition： CNN+LSTM+CTC

- Difficulties
  - Lack of labelled data
  - Chinese characters
  - CTC loss hard to converge

Reference

- Dai, J., Li, Y., He, K., Sun, J., 2016. R-fcn: Object detection via region-based fully convolutional networks, in: Advances in Neural Information Processing Systems. pp. 379–387.
- Fu, C.-Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.C., 2017. DSSD: Deconvolutional Single Shot Detector. arXiv preprint arXiv:1701.06659.
- Girshick, R., 2015. Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision. pp. 1440–1448.
- Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 580–587.
- Goodfellow, I.J., Bulatov, Y., Ibarz, J., Arnoud, S., Shet, V., 2013. Multi-digit number recognition from street view imagery using deep convolutional neural networks. arXiv preprint arXiv:1312.6082.
- Gupta, A., Vedaldi, A., Zisserman, A., 2016. Synthetic data for text localisation in natural images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2315–2324.
- He, P., Huang, W., Qiao, Y., Loy, C.C., Tang, X., 2016. Reading Scene Text in Deep Convolutional Sequences, in: AAAI. pp. 3501–3508.
- He, W., Zhang, X.-Y., Yin, F., Liu, C.-L., 2017. Deep Direct Regression for Multi-Oriented Scene Text Detection. arXiv preprint arXiv:1703.08289.
- Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., others, 2016. Speed/accuracy trade-offs for modern convolutional object detectors. arXiv preprint arXiv:1611.10012.
- Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A., 2016. Reading text in the wild with convolutional neural networks. International Journal of Computer Vision 116, 1–20.
- Lee, C.-Y., Osindero, S., 2016. Recursive recurrent nets with attention modeling for OCR in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2231–2239.
- Liao, M., Shi, B., Bai, X., Wang, X., Liu, W., 2017. TextBoxes: A Fast Text Detector with a Single Deep Neural Network., in: AAAI. pp. 4161–4167.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal Loss for Dense Object Detection. arXiv preprint arXiv:1708.02002.

Reference

- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C., 2016. SSD: Single shot multibox detector, in: European Conference on Computer Vision. Springer, pp. 21–37.
- Liu, Y., Jin, L., 2017. Deep matching prior network: Toward tighter multi-oriented text detection. arXiv preprint arXiv:1703.01425.
- Ma, J., Shao, W., Ye, H., Wang, L., Wang, H., Zheng, Y., Xue, X., 2017. Arbitrary-Oriented Scene Text Detection via Rotation Proposals. arXiv preprint arXiv:1703.01086.
- Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788.
- Redmon, J., Farhadi, A., 2016. YOLO9000: better, faster, stronger. arXiv preprint arXiv:1612.08242.
- Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems. pp. 91–99.
- Shi, B., Bai, X., Belongie, S., 2017. Detecting Oriented Text in Natural Images by Linking Segments. arXiv preprint arXiv:1703.06520.
- Shi, B., Bai, X., Yao, C., 2016a. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE transactions on pattern analysis and machine intelligence.
- Shi, B., Wang, X., Lyu, P., Yao, C., Bai, X., 2016b. Robust scene text recognition with automatic rectification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4168–4176.
- Shrivastava, A., Gupta, A., Girshick, R., 2016. Training region-based object detectors with online hard example mining, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 761–769.
- Tian, Z., Huang, W., He, T., He, P., Qiao, Y., 2016. Detecting text in natural image with connectionist text proposal network, in: European Conference on Computer Vision. Springer, pp. 56–72.
- Yao, C., Bai, X., Sang, N., Zhou, X., Zhou, S., Cao, Z., 2016. Scene text detection via holistic, multi-channel prediction. arXiv preprint arXiv:1606.09002.
- Zhong, Z., Jin, L., Zhang, S., Feng, Z., 2016. Deeptext: A unified framework for text proposal generation and text detection in natural images. arXiv preprint arXiv:1605.07314.
- Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., Liang, J., 2017. EAST: An Efficient and Accurate Scene Text Detector. arXiv preprint arXiv:1704.03155.

Q&A