



华中科技大学  
Huazhong University of Science & Technology



# Deep Neural Networks for Scene Text Reading

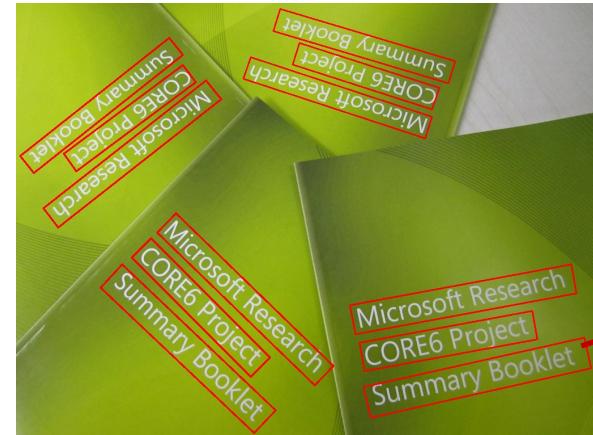
---

Xiang Bai

Huazhong University of Science and Technology

# Problem definitions

## □ Definitions



Summary Booklet

Scene text  
detection

Predicting the presence of text  
and localizing each instance  
(if any), usually at word or  
line level, in natural scenes

End-to-end  
recognition

Scene text  
recognition

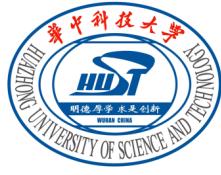
Converting text regions into  
computer readable and  
editable symbols



# Outline

---

- **Background**
- Scene Text Detection
- Scene Text Recognition
- Applications
- Future Trends



# Background

1822 OPTICS LETTERS / Vol. 20, No. 17 / September 1, 1995

the intensity. We use this result to evaluate the total intensity  $N(t) = \int \int \int dx dy dz |A(x, y, t; z)|^2$  to obtain  $N(t) = 2P(t)/n_0^{(j)} \epsilon_0 c$ , where  $P(t)$  is the instantaneous power. Note that  $N$  and  $P$  are functions of  $t$  but not of  $z$  because temporal dispersion and loss are assumed negligible. The coefficient  $n_2^{(j)}$ , defined by the relation  $\Delta n_j = n_2^{(j)} I_j$ , is related to the coefficient  $n_2^{(r)}$  by  $n_2^{(j)} = n_2^{(r)} \epsilon_0 c n_0^{(j)}$ , from which it follows that  $n_2^{(j)} N = 2n_2^{(r)} P$ . We use this, along with the definition of the critical power  $P_c = 2\pi/k_b n_0^{(j)} n_2^{(j)}$ , and the definition of the normalization field amplitude,  $u(x, y, z) = A(x, y, t; z)/\sqrt{N(t)}$ , to rewrite the nonlinear term in Eq. (2) as  $[2\pi P/\beta_{c,j}^2 P_c] |u|^2$ . (Note that  $P_c$  is defined can be negative.) We substitute this result into Eq. (2) along with a new variable,  $\xi = z/k_b$ , to obtain

$$in_0^{(j)} \frac{\partial}{\partial \xi} u - \frac{1}{2} \frac{\partial^2}{\partial x^2} u - \frac{1}{2} \frac{\partial^2}{\partial y^2} u - 2\pi \frac{P}{P_c} |u|^2 u.$$

(3)

Now let us consider the hypothetical situation in which two beams of light with identical normalized amplitudes  $u(x, y)$  enter two different samples, which we denote by the superscripts  $j = r$  (reference sample) and  $j = t$  (test sample). We let the samples have linear indices of refraction  $n_r^{(r)}$  and  $n_t^{(t)}$  and thicknesses  $L_r$  and  $L_t$ . If the power is small enough that the last term in Eq. (3) can be neglected, and if the sample lengths are chosen so that  $L_t/n_t^{(t)} = L_r/n_r^{(r)}$ , it follows from Eq. (3) that the normalized amplitudes are identical at the exit faces of the two samples. Furthermore, the normalized amplitudes will be nearly identical at the exit faces of the two samples if  $|L_t/n_t^{(t)} - L_r/n_r^{(r)}| \ll z_{d0}$ , where  $z_{d0}$  is the Rayleigh range<sup>11</sup> in free space. If the input power is increased to a very large value,  $P$ , and  $P_c$  is small, if the nonlinear indices of refraction of the samples are  $n_r^{(r)}$  and  $n_t^{(t)}$ , we see from Eq. (3) that to obtain the same  $u(x, y)$  at the exit faces of the two samples, we should adjust the powers so that  $[L_r/n_r^{(r)}]^2 P_r/P_c = [L_t/n_t^{(t)}]^2 P_t/P_c$ . For two samples of the same thickness  $L_t = L_r = L$ , this condition is equivalent to  $P_r/n_r^{(r)} = P_t/n_t^{(t)}$ . With the sample thicknesses properly selected and the powers properly adjusted,  $u(x, y)$  will be the same for both samples at any given distance from the exit faces, and therefore the measured normalized peak-to-valley transmittances  $\Delta T_{p/v} = [P_{p/v}^{(det)} - P_{p/v}^{(det)}]/P_{p/v}^{(det)}$  will also be the same. Here  $P_{p/v}^{(det)}$  and  $P_{p/v}^{(det)}$  are the maximum (peak) and minimum (valley) powers that are registered for the  $j$ th sample of the detector (det) after it passes through the aperture. The average or baseline power is  $P_{p/v}^{(det)} = [P_{p/v}^{(det)} + P_{p/v}^{(det)}]/2$ .

Following this analysis, we see that a simple procedure for making a Z-scan measurement is as follows: (1) Obtain reference and test samples of equal thicknesses  $L$  for which  $|L/n_0^{(r)} - L/n_0^{(t)}| \ll z_{d0}$ . (2) Make a Z-scan measurement of one of the samples. The exact size and shape of the aperture do not matter. For example, an obscuration disk (as in an eclipsing Z scan<sup>12</sup>) can be used. (3) Insert the second

sample and adjust the input power until the normalized peak-to-valley transmittance  $\Delta T_{p/v}$  matches that obtained for the first sample. (4) Calculate the nonlinear index of refraction using the following formula:

$$n_2^{(r)} = n_2^{(r)} P_r/P_t. \quad (4)$$

For a thin sample, it is not necessary to match the lengths as indicated in step (1) above, since the beam does not evolve appreciably (in either size or shape) in traversing the sample. For the special case in which the nonlinear phase shift is much less than unity, step (3) may also be simplified. To see how, we first note that  $I(x, y, t; z) = P(t)|u(x, y, z)|^2$ . The nonlinear phase shift for a thin sample can then be written as  $\Delta \phi_{\text{ref}}(\gamma_r, \gamma_t) = -\omega_0 n_2^{(r)} L_r P_r(t) |u(x, y, z)|^2 c$ . If  $\Delta \phi_{\text{ref}} \ll \pi$ , we let the electric-field amplitude at the entrance face of the sample is

$A(x, y, z = 0) = A_0(x, y, z = 0)$  at the  $\sqrt{N(t)}$  position.

For a thin sample,  $\gamma_r = \gamma_t$  to the exit face of the sample located at some arbitrary position is

as  $\Delta T_{p/v} = [P_{p/v}^{(det)} - P_{p/v}^{(det)}]/P_{p/v}^{(det)}$ . We can evaluate this quantity by using  $\Delta T_{p/v} = \int \int \int dx dy dz |B_2|^2 / \int \int \int dx dy dz |u_0|^2$ . In this expression, the integration over  $x$  and  $y$  is performed over the extent of the aperture, the integration over  $t$  represents the action of the detector. Since the quantity  $B_2$  is the same for the test and reference samples, it follows that  $\Delta T_{p/v} = \Delta T_{p/v}(\gamma_r)$ , and from this that  $\Delta T_{p/v}(\gamma_r) = \Delta T_{p/v}(\gamma_t)$ . Substituting into this equation the expressions for  $\gamma_r$  and  $\gamma_t$ , we get

$$n_2^{(r)} = n_2^{(r)} \frac{\Delta T_{p/v} L_r P_r}{\Delta T_{p/v} L_t P_t}. \quad (5)$$

When applicable, this formula permits a simplification of the measurement procedure since the power can be set to any convenient value. In other words,

Table 1. Ratio of  $n_2$  Values for Two Pairs of Liquids as Measured at  $\lambda_0 = 1064$  nm with Five Cuvette Thicknesses

Cuvette Thickness (mm)	$n_2(\text{toluene})/n_2(\text{glycerine})$	$n_2(\text{methanol})/n_2(\text{water})$
1	14.1	1.05
2	14.6	1.07
5	14.4	1.06
10	14.2	1.07
20	14.0	1.07
Average	14.3	1.06

Document image

VS

Scene text image



- Scattered and sparse
- Multi-oriented
- Multi-lingual

# Background

## Scene text detection methods before 2016

Proposals



Filtering



Regression



- Generate candidates using **hand-craft features**
- Text / non-text classification using **CNN/Random forest**
- Refine locations using **CNN**

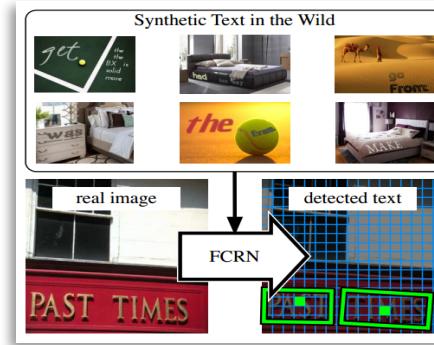
- [1] Jaderberg et al. Deep features for text spotting. ECCV, 2014.
- [2] Jaderberg et al. Reading text in the wild with convolutional neural networks. IJCV, 2016.
- [3] Huang et al. Robust scene text detection with convolution neural network induced mser trees. ECCV, 2014.
- [4] Zhang et al. Symmetry-based text line detection in natural scenes. CVPWR, 2015.
- [5] LGómez, D Karatzas. Textproposals: a text-specific selective search algorithm for word spotting in the wild. Pattern Recognition 70, 60-74

# Background

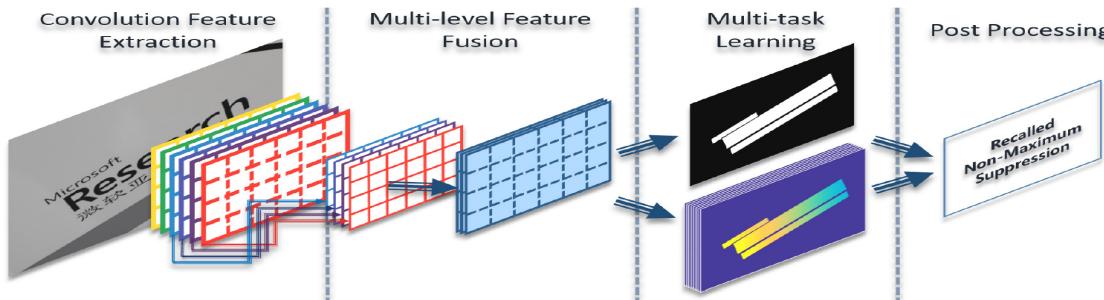
## Scene text detection methods after 2016



Segmentation-based method[1]



Proposal-based method[2]

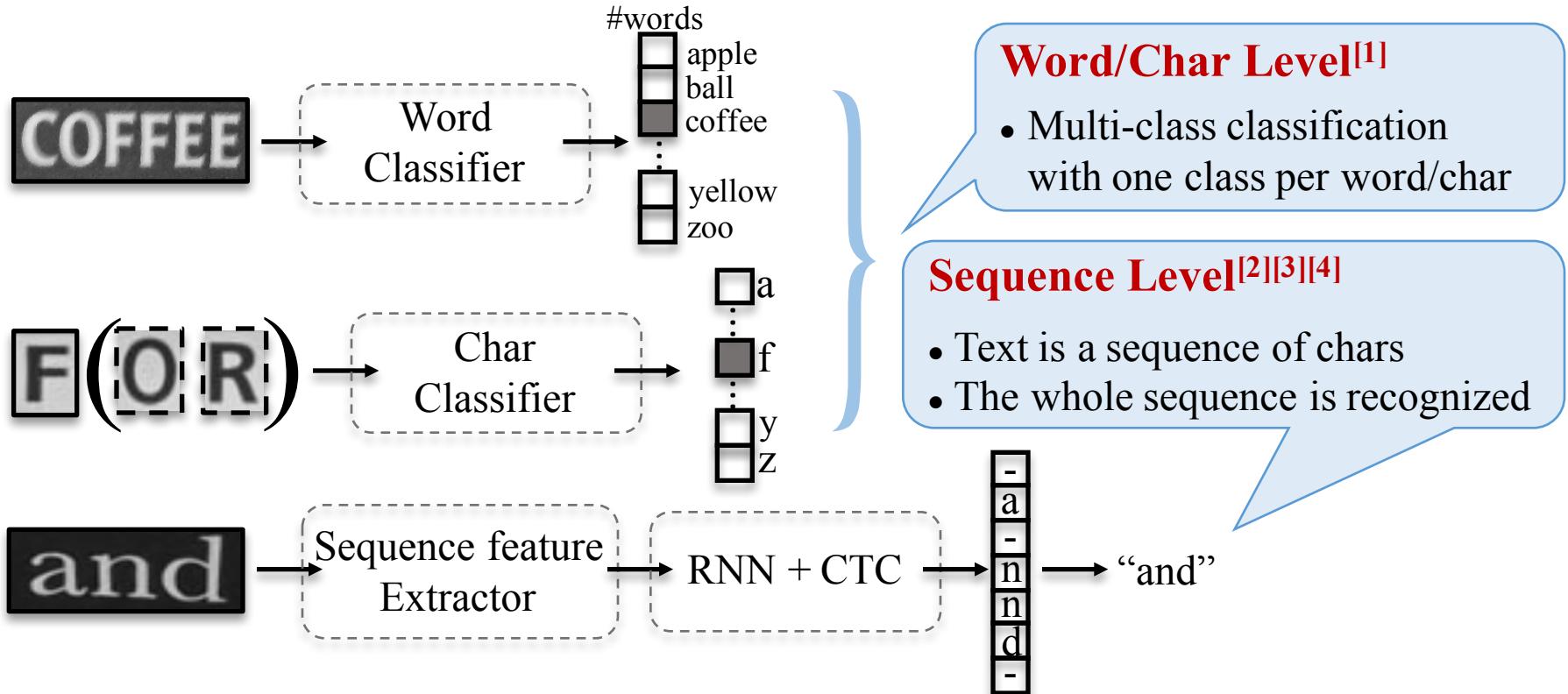


Hybrid method[3]

- [1] Zhang Z, et al. Multi-oriented text detection with fully convolutional networks. CVPR, 2016.
- [2] Gupta A, et al. Synthetic data for text localisation in natural images. CVPR, 2016.
- [3] He W, et al. Deep Direct Regression for Multi-Oriented Scene Text Detection. ICCV, 2017
- [4] Liao et al. TextBoxes: A fast text detector with a single deep neural network. AAAI, 2017.

# Background

## Scene text recognition methods



[1] M. Jaderberg et al. Reading text in the wild with convolutional neural networks. IJCV, 2016.

[2] B. Su et al. Accurate scene text recognition based on recurrent neural network. ACCV, 2014.

[3] He et al. Reading Scene Text in Deep Convolutional Sequences. AAAI, 2016.

[4] Shi B et al. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. TPAMI, 2017.



# Background

## Recent Trend

Statistics of related papers published in 2017 top conferences

Conference	Detection	Recognition	End-to-end recognition
AAAI-17	0	0	2
IJCAI-17	0	1	0
NIPS-17	0	1	0
ICCV-17	5	1	2
CVPR-17	3	0	0
ICDAR-17	8	2	1
<b>TOTAL</b>	<b>16</b>	<b>5</b>	<b>5</b>

- ❑ Over 80% text detection papers focus on multi-oriented text detection .
- ❑ **Scene text recognition** and **end-to-end recognition** are paid less attention to.
- ❑ Most papers focus on **English** text.

# Background

## Latin text vs. Non-Latin text



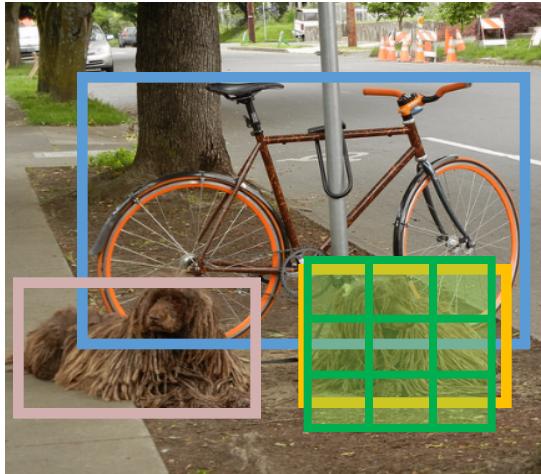
English: there is always a blank space  
between neighbor words

Chinese: no vision cues for partition,  
while semantic information is needed.

**Line-based detection and sequence labeling are appropriate for both Latin and Non-Latin text**

# Background

## Challenges in Non-Latin text detection



General objects



Words



Text Lines

- ❑ Unlike general objects and English words, text lines have **larger aspect ratios**
- ❑ Given the fixed size of convolutional filters, text lines cannot be totally covered.



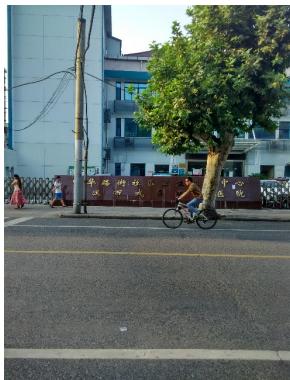
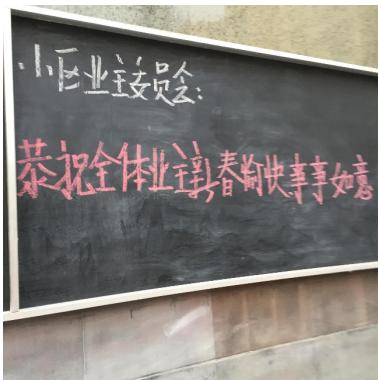
# Background

## Performance comparison on English / Chinese datasets

Dataset	Language	Num. Train/Test	Best F-measure
ICDAR 2013	English	229/233	0.90
ICDAR 2015	English	1000/500	0.81
RCTW 2017	Mainly Chinese	8034/4229	0.66

The performance of Chinese dataset is much lower.

### ICDAR 2017 Competition on Reading Chinese Text in the Wild

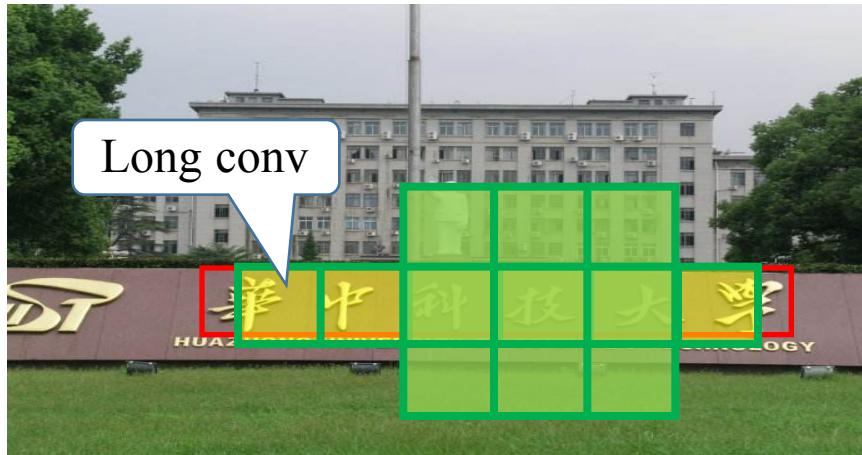


Link: <http://mclab.eic.hust.edu.cn/icdar2017chinese/>

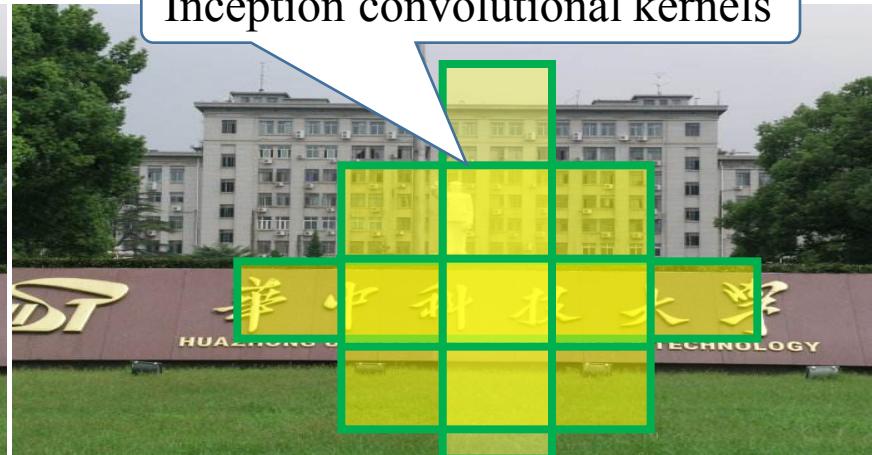
# Background

## Possible solutions for Non-Latin text detection

- Long convolutional kernel.
- Inception convolutional kernels.
- Part detection and grouping.



Inception convolutional kernels

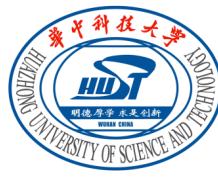




# Outline

---

- Background
- **Scene Text Detection**
- Scene Text Recognition
- Applications
- Future Trends



# Scene Text Detection

## ➤ Proposal-based method:

- Detecting text with a single deep neural network (**TextBoxes**)[1]

## ➤ Part-based method:

- Detecting text with Segments and Links (**SegLink**)[2]

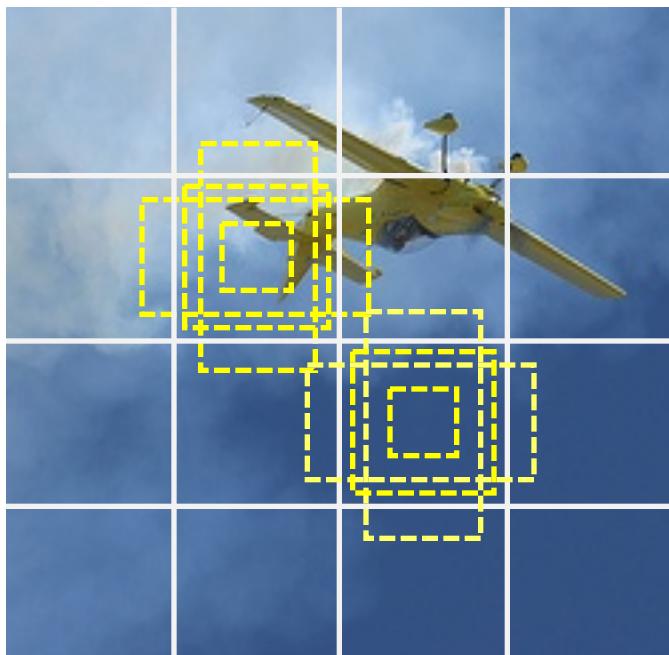
[1] M. Liao et al. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. AAAI, 2017.

[2] B. Shi et al. Detecting Oriented Text in Natural Images by Linking Segments. IEEE CVPR, 2017.

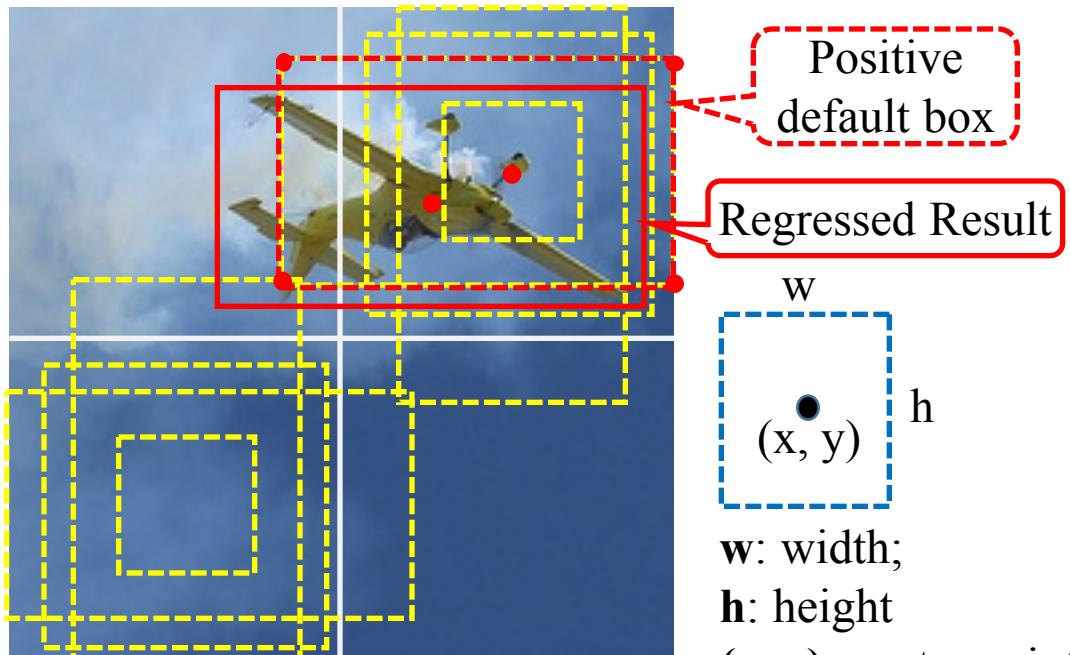
# TextBoxes: Horizontal text detection

## SSD: Single Shot MultiBox Detector

- Default boxes of different ratios and sizes
- Classify the default boxes
- Regress the matched default boxes



4 feature map

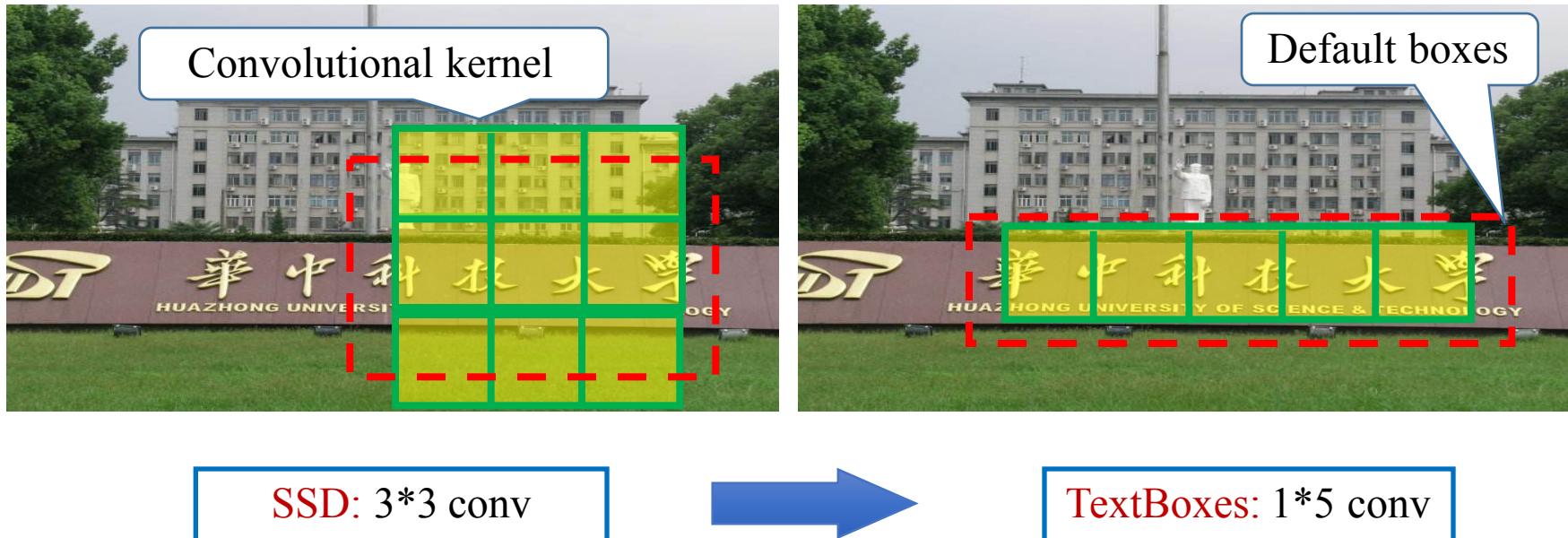


[1] W. Liu et al. SSD: Single Shot MultiBox Detector. ECCV, 2016.



# TextBoxes: Horizontal text detection

## Long convolutional kernels and default boxes



- ❑ Use SSD as the backbone.
- ❑ Long default boxes.
- ❑ Long convolutional kernels.

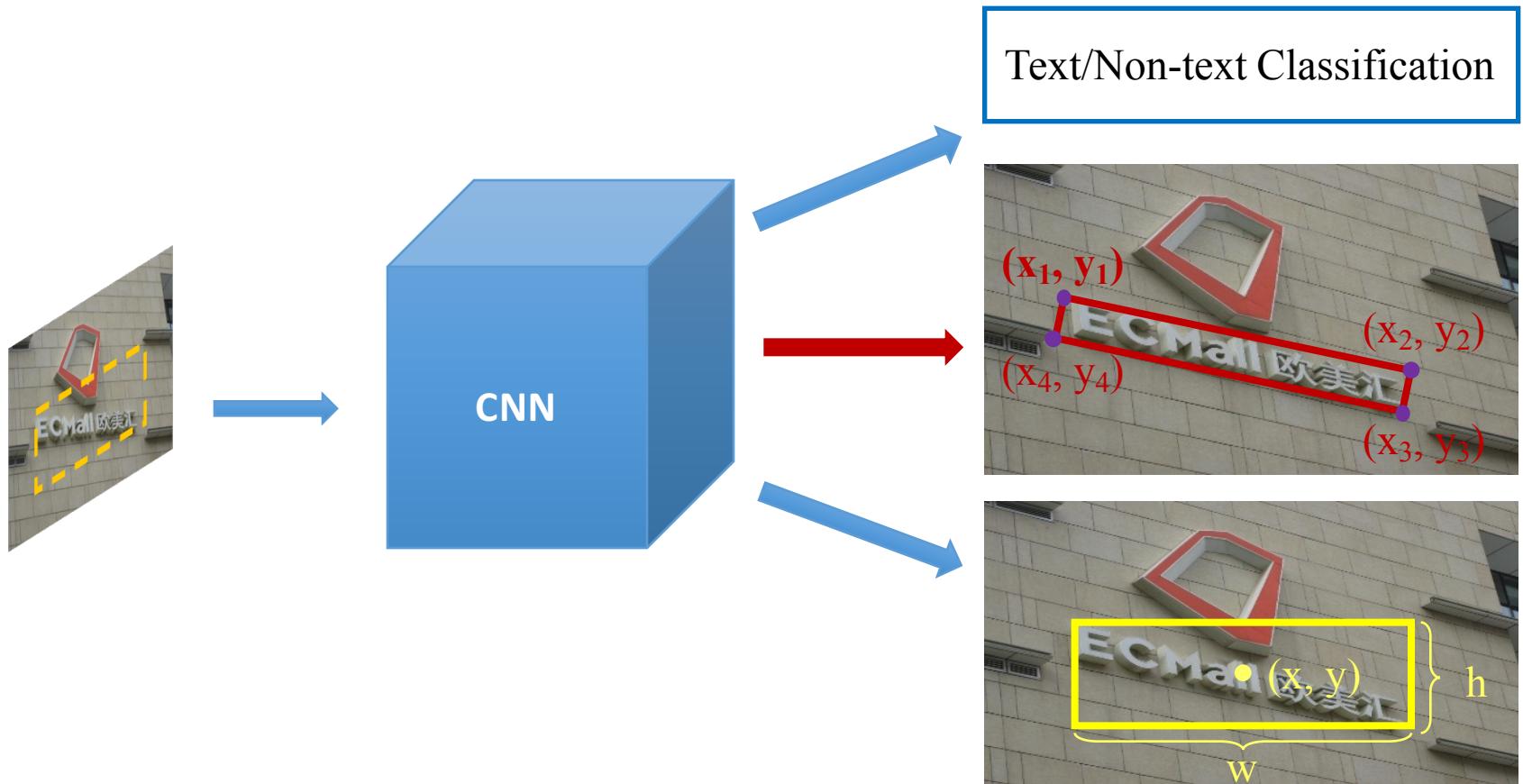
# TextBoxes: Horizontal text detection

## Experimental Results on ICDAR 2013

Methods	Precision	Recall	F-measure
Jaderberg IJCV16	0.89	0.68	0.77
FCRN CVPR16	0.92	0.76	0.83
Zhang CVPR16	0.88	0.8	0.84
SSD	0.80	0.60	0.69
<b>TextBoxes</b>	<b>0.89</b>	<b>0.83</b>	<b>0.86</b>



# TextBoxes++: Multi-oriented text detection



$(x_i, y_i)$  ( $i = 1, 2, 3, 4$ ) denote coordinates of the bounding box

# TextBoxes++: Multi-oriented text detection

## Text detection results on ICDAR 2015 Incidental Text

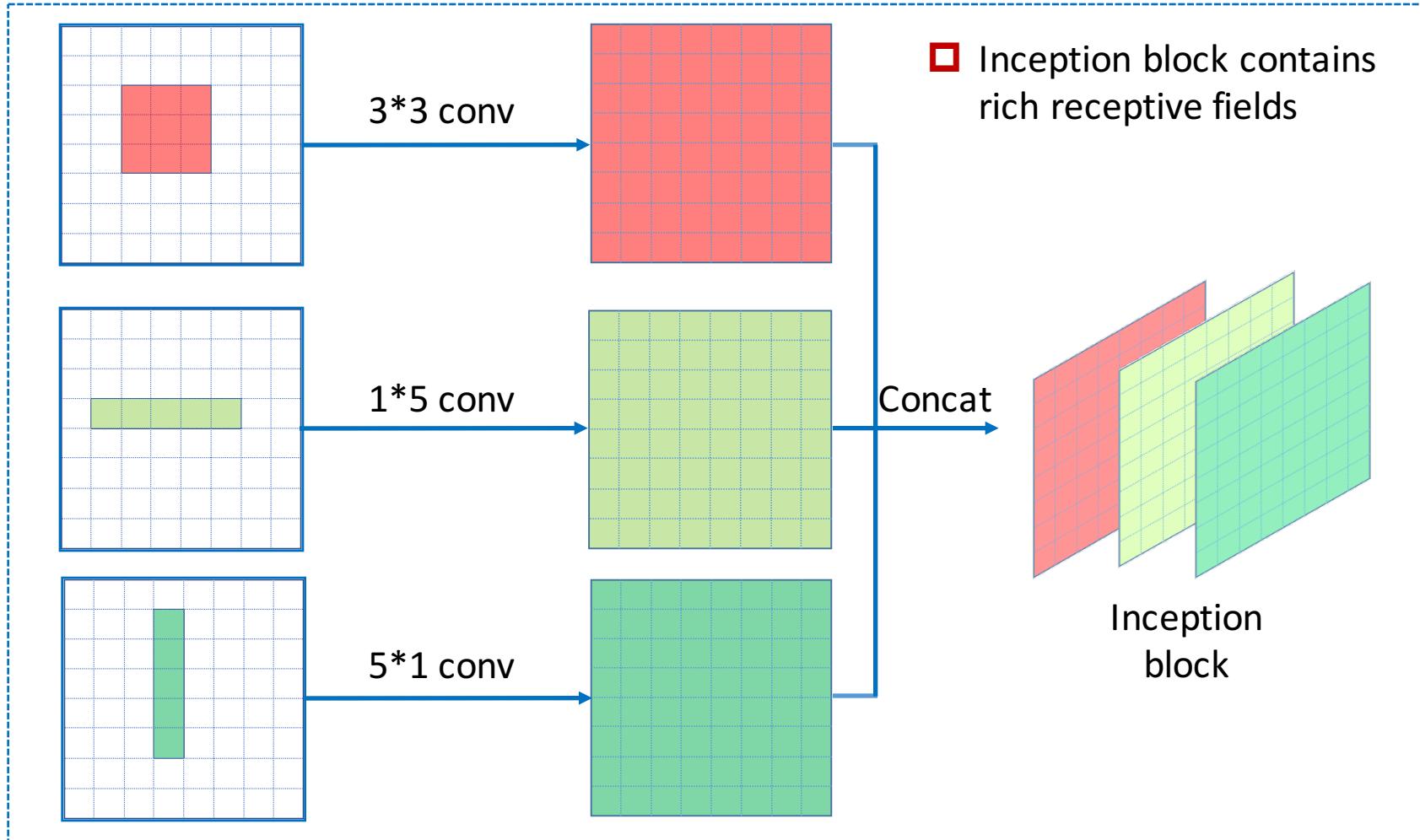
Methods	Recall	Precision	F-measure	FPS
SegLink CVPR17	0.768	0.731	0.75	8.9
EAST CVPR17	0.735	0.836	0.782	13.2
EAST multi-scale CVPR17	0.783	0.833	0.807	--
<b>TextBoxes++</b>	0.767	0.872	0.817	11.6
<b>TextBoxes++_multi-scale*</b>	<b>0.785</b>	<b>0.878</b>	<b>0.829</b>	--



\* multi-scale: Testing image with multi-scale inputs

# TextBoxes++: Long text line detection

## Inception block for long text lines



# TextBoxes++: Long text line detection



Without inception block



With inception block

Performances on RCTW (long)

A subset of RCTW which mainly consists of images with long text lines.

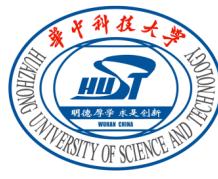
Method	F-measure
Baseline	0.6902
Baseline + inception block	0.7532

# TextBoxes++: Long text line detection

## Comparison with competition winners

Team Name	Max F-measure	FM-Rank
Foo & Bar	0.661054	1
NLPR_PAL	0.657598	2
gmh	0.636024	3
<b>TextBoxes++ with inception block</b>	<b>0.665295</b>	--





# Scene Text Detection

## ➤ Proposal-based method:

- Detecting text with a single deep neural network (TextBoxes)[1]

## ➤ Part-based method:

- Detecting text with Segments and Links (**SegLink**)[2]

[1] M. Liao et al. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. AAAI, 2017.

[2] B. Shi et al. Detecting Oriented Text in Natural Images by Linking Segments. IEEE CVPR, 2017.

# SegLink: Detect long text with segments and links

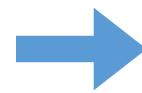


Segments  
(yellow boxes)



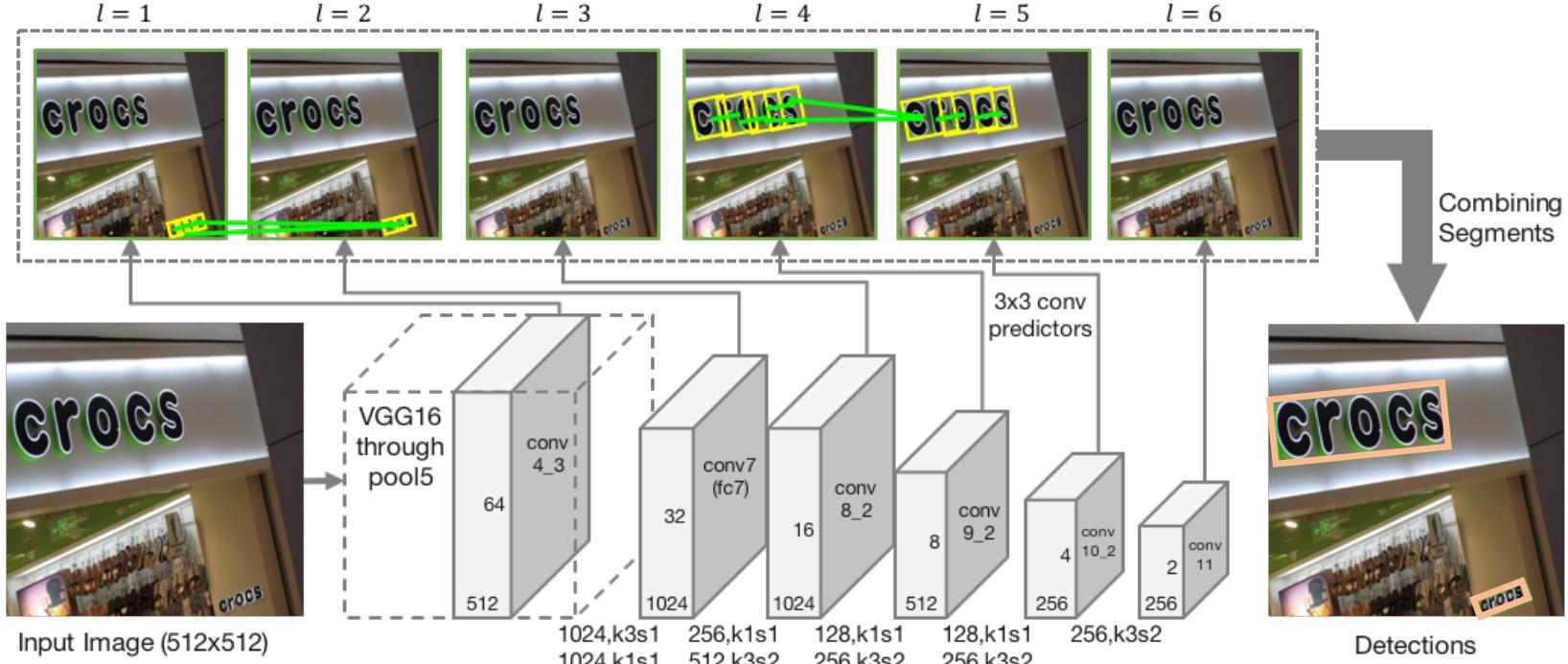
Links  
(green edges)

Large aspect ratio text lines can be detected using limited respective field with **Segments** and **Links**



Combined detection boxes

# SegLink: Detect long text with segments and links

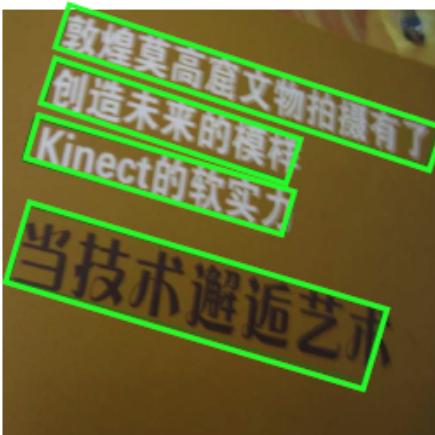


- Fully connected networks based on SSD and VGG16.
- Multiscale **Segments and Links** prediction
- Alternative solution to the limited respective field problem of long text lines

# SegLink: Detect long text with segments and links

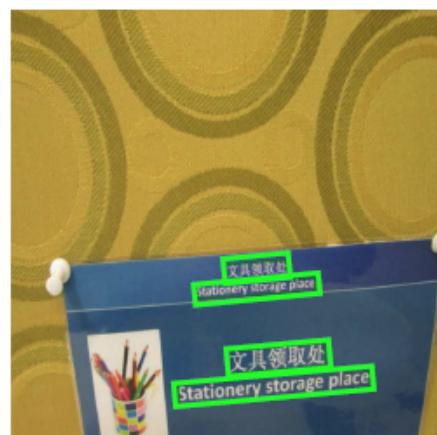
## Results on MSRA-TD500

Methods	Precision	Recall	F-measure
Kang et al. (CVPR 2014)	71	62	66
Yin et al. (TPAMI 2015)	81	63	74
Zhang et al. (CVPR 2016)	83	67	74
<b>SegLink</b>	<b>86</b>	<b>70</b>	<b>77</b>



## Results on ICDAR2015

Methods	Precision	Recall	F-measure
StradVision-2	<b>77.5</b>	36.7	49.8
CTPN	51.6	74.2	60.9
Megvii-Image++	72.4	57.0	63.8
<b>SegLink</b>	73.1	<b>76.8</b>	<b>75.0</b>



# SegLink: Detect long text with segments and links



**Seglink can detect text of curved shape**



# Outline

---

- Background
- Scene Text Detection
- **Scene Text Recognition**
- Applications
- Future Trends



# Scene Text Recognition

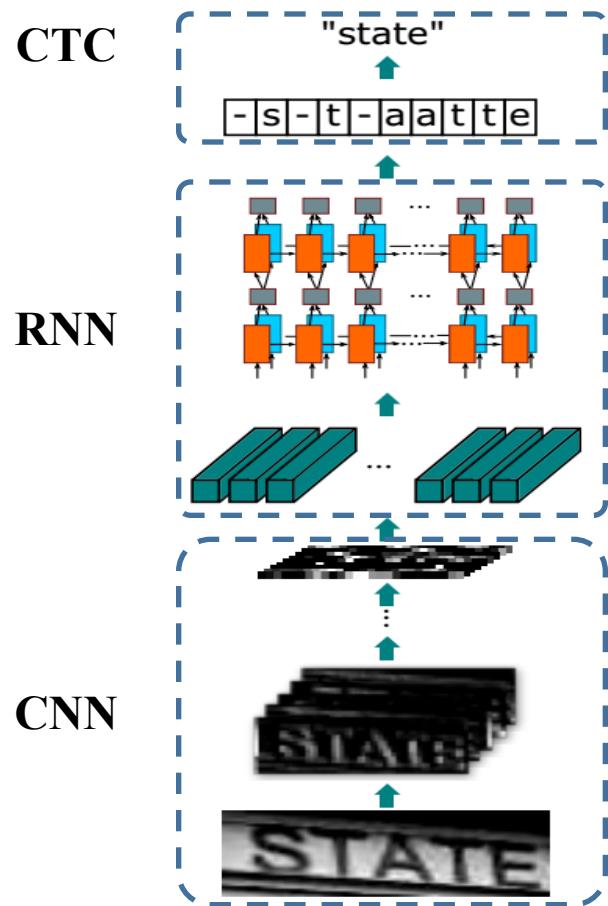
- CRNN model for Regular Text Recognition
- RARE model for Irregular Text Recognition

[1] CRNN: Shi B et al. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. TPAMI, 2017.

[2] RARE: Shi B et al. Robust scene text recognition with automatic rectification. CVPR, 2016.

# CRNN for Regular Text Recognition

## The Network Architecture

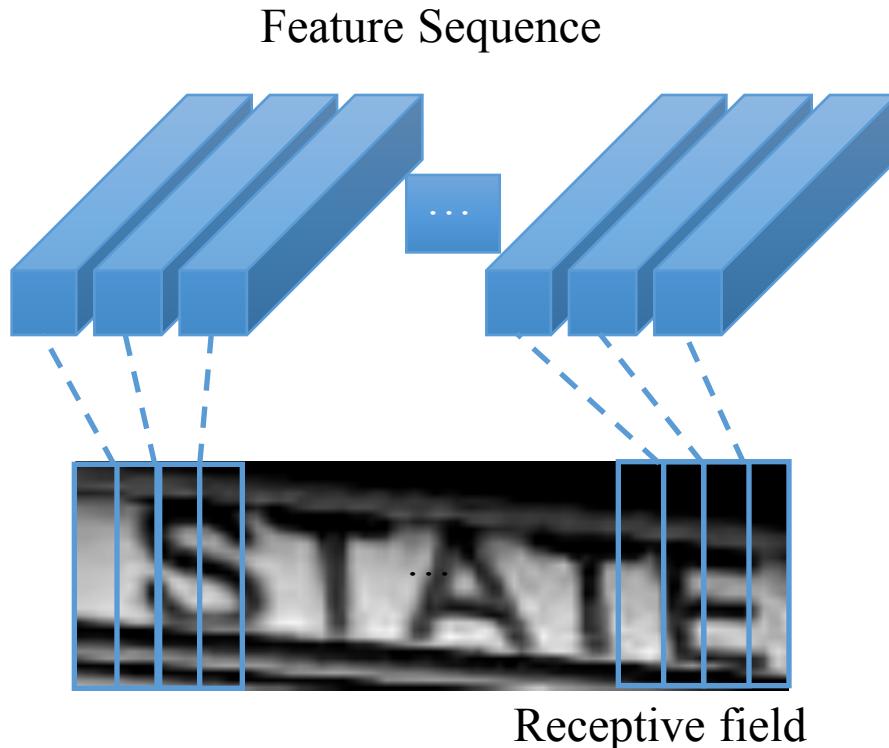


### Network Structure

- ❑ Convolutional layers extract feature maps
- ❑ Convert feature maps into feature sequence
- ❑ Sequence labeling with LSTM
- ❑ Translate labels to text

# CRNN for Regular Text Recognition

## Sequence Modeling





# CRNN for Regular Text Recognition

## Comparisons

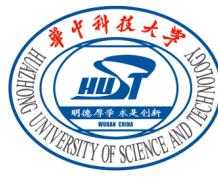
### Advantages

- End-to-end trainable
- Free of char-level annotations
- Unconstrained to specific lexicon
- 40~50 times less parameters than mainstream models
- Better or comparable performance with state-of-the-arts

### Results(lexicon-free)

Method	IIIT5K	SVT	IC03	IC13
Bissacco et al. (ICCV13)	-	78.0	-	87.6
Jaderberg et al. (IJCV15)*	-	80.7	93.1	90.8
Jaderberg et al. (ICLR15)	-	71.7	89.6	81.8
<b>Proposed</b>	81.2	<b>82.7</b>	<b>91.9</b>	<b>89.6</b>

\*is not lexicon-free, as its outputs are constrained to a 90k dictionary



# Scene Text Recognition

- CRNN model for Regular Text Recognition
- RARE model for Irregular Text Recognition

[1] CRNN: Shi B et al. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. TPAMI, 2017.

[2] RARE: Shi B et al. Robust scene text recognition with automatic rectification. CVPR, 2016.

# RARE for Irregular Text Recognition

## Motivation

Perspective and curved texts are hard to recognize!



SVT-Perspective

(a) Perspective texts

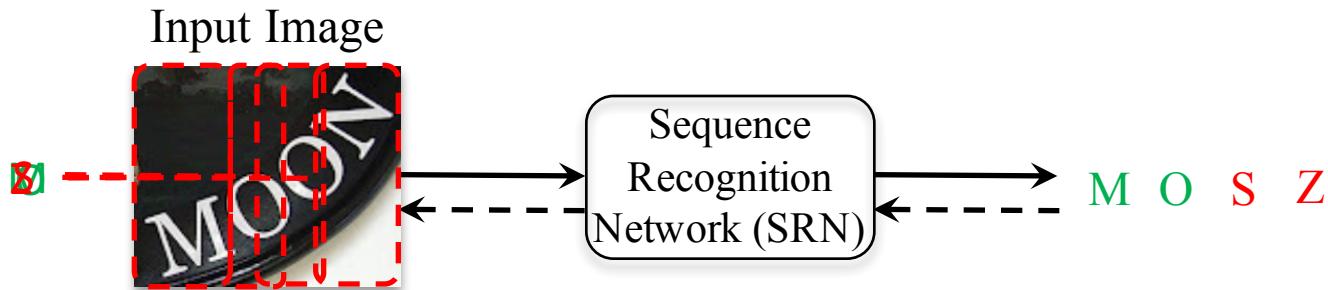


CUTE80

(b) Curved texts

# RARE for Irregular Text Recognition

## Attention-based Sequence Recognition



- **SRN**: an **attention-based** encoder-decoder framework
  - Encoder: ConvNet + Bi-LSTM
  - Decoder: Attention-based character generator

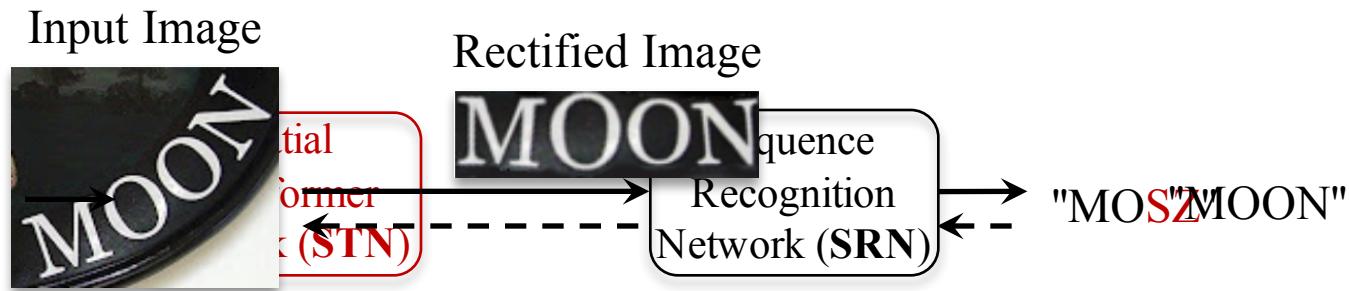
### Results

Method	IIIT5K	SVT	IC03	IC13	SVT-Per	CUTE80
SRN	83.6	84.9	93.6	91.8	68.2	62.5



# RARE for Irregular Text Recognition

## STN (Spatial Transform Network)<sup>[1]</sup> for Text Rectification

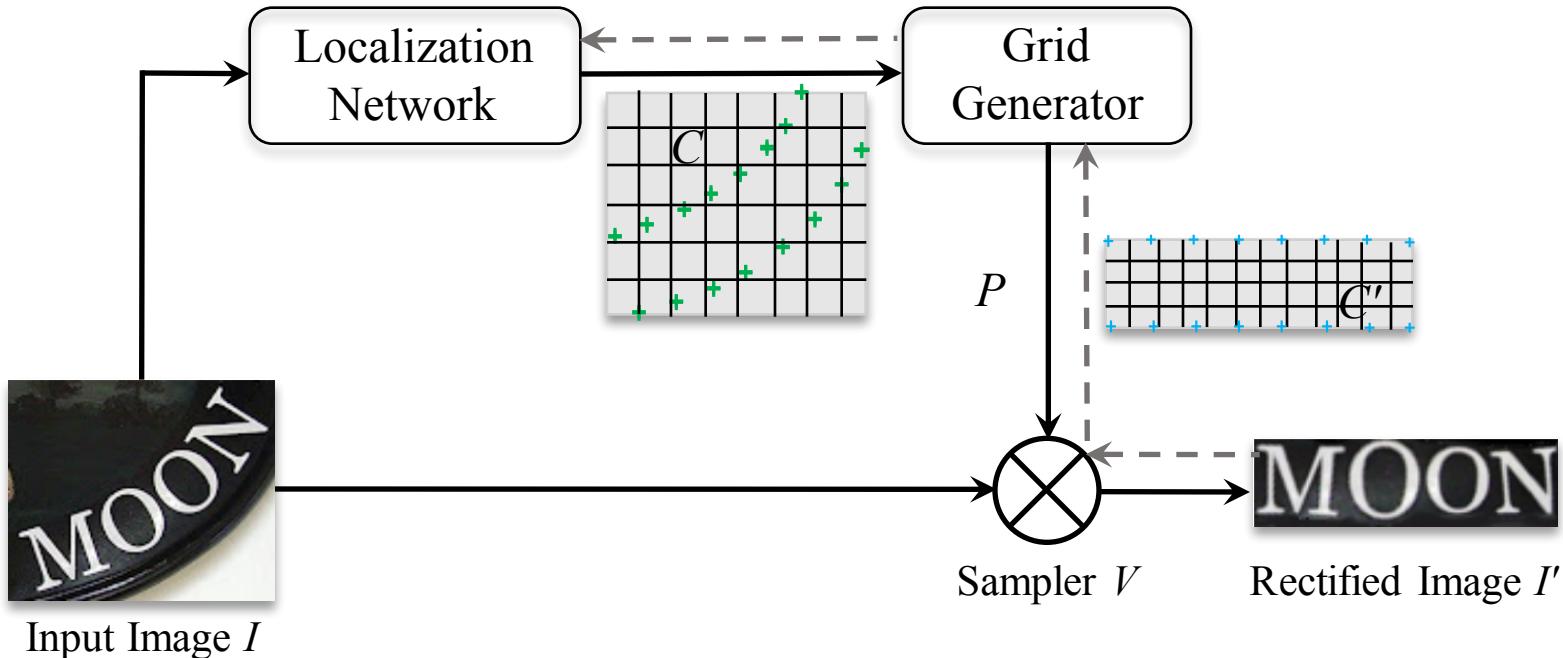


- An end-to-end trainable network
  - **STN:** rectifies images with spatial transformation
  - **SRN:** an attention-based encoder-decoder framework

[1] Jaderberg M et al. Spatial transformer networks. NIPS, 2015.

# RARE for Irregular Text Recognition

## Spatial Transformer Network (STN)



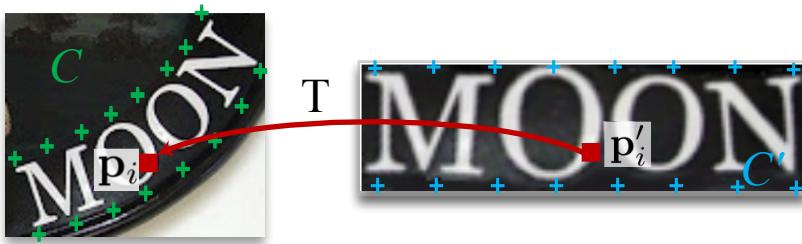
- **Localization Network:** A CNN that predicts the fiducial points.

[1] Jaderberg M et al. Spatial transformer networks. NIPS, 2015.



# RARE for Irregular Text Recognition

## Spatial Transformer Network (STN)



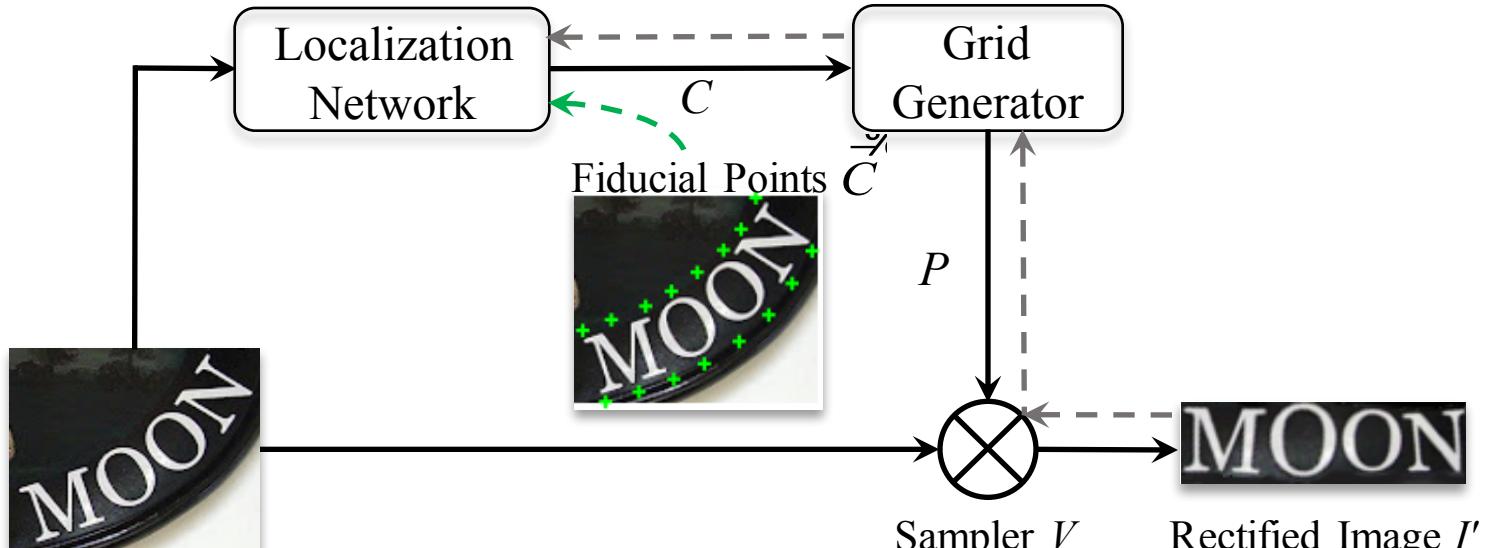
- Grid Generator: Computes a Thin-Plate-Spline (TPS) transform,  $\mathbf{T}$ , from the fiducial points  $C$ .
- Sampler: TPS-Transform input image  $I$  into rectified  $I'$ .

Method	Standard datasets			Deformable text datasets		
	IIIT5K	SVT	IC03	IC13	SVT-Per	CUTE80
SRN	83.6	84.9	<b>93.6</b>	91.8	68.2	62.5
<b>STN+SRN</b>	<b>88.2</b>	<b>86.7</b>	93.4	<b>92.7</b>	<b>76.8</b>	<b>76.7</b>

[1] Jaderberg M et al. Spatial transformer networks. NIPS, 2015.

# RARE for Irregular Text Recognition

## Supervised STN



- Synthetic dataset with fiducial points  $\tilde{C}$  to supervise the predicted  $C$ .

Method	IIIT5K	SVT	IC03	IC13	SVT-Per	CUTE80
SRN	83.6	84.9	93.6	91.8	68.2	62.5
STN+SRN	88.2	86.7	93.4	92.7	76.8	76.7
<b>STN(Supervised)+SRN</b>	<b>88.8</b>	<b>87.9</b>	<b>94.1</b>	<b>94.0</b>	<b>77.7</b>	<b>78.8</b>



# RARE for Irregular Text Recognition

## Rectification Visualization

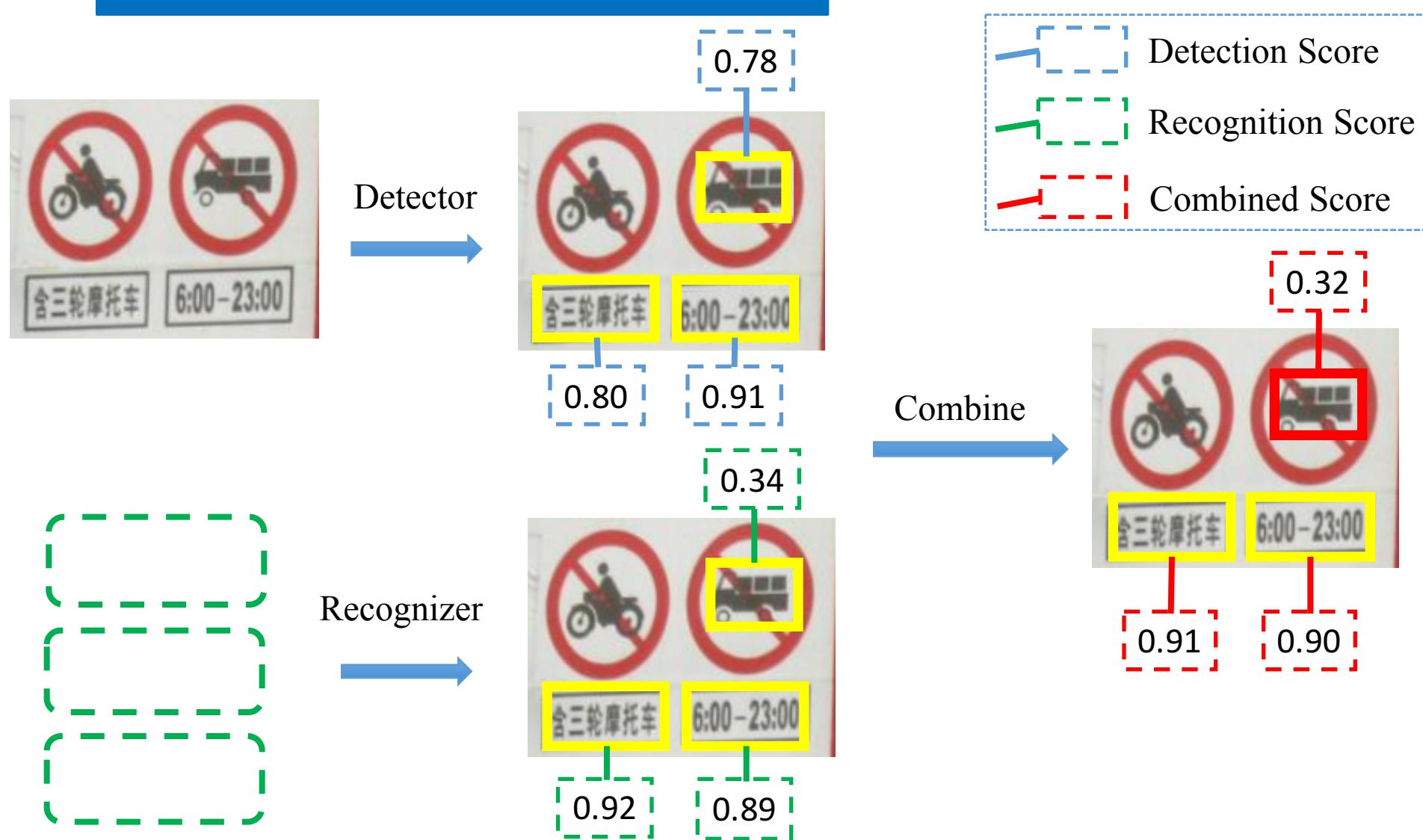
### SVT-Perspective

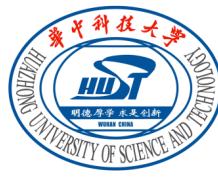
Input	Rectified	Prediction Groundtruth
		restaurant restaurant
		quiznos quiznos
		sheraton sheraton
		mobil mobil
		jewelry jewelry
		public public

### CUTE80

Input	Rectified	Prediction Groundtruth
		mercato marcato
		football football
		naval naval
		grove grove
		loka loka

# Recognition is helpful to detection





# Combination of TextBoxes++ and CRNN

- Detection and recognition are combined by

$$S = \frac{2 * \exp(S_d + S_r)}{\exp(S_d) + \exp(S_r)},$$

Detection score

Recognition score

Text detection results on ICDAR 2015 Incidental Scene Text dataset

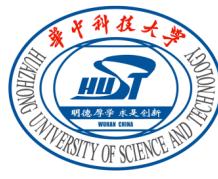
Methods	Recall	Precision	F-score
<b>Detection</b>	0.785	0.878	0.829
<b>Detection + Recognition</b>	<b>0.792</b>	<b>0.912</b>	<b>0.848</b>



# Outline

---

- Background
- Scene Text Detection
- Scene Text Recognition
- **Applications**
- Future Trends

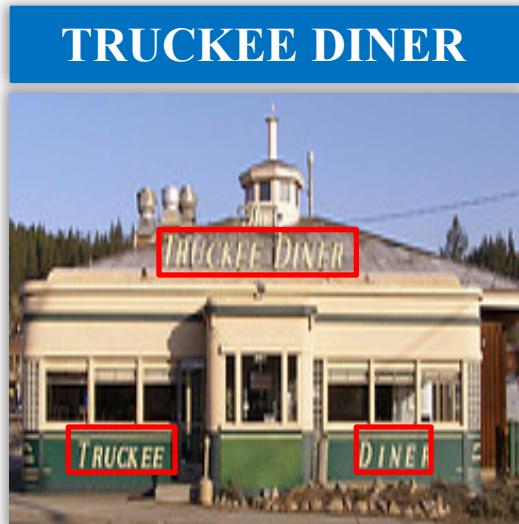


# Applications

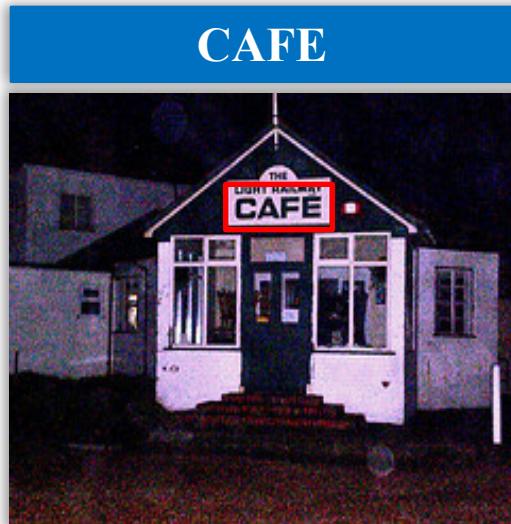
- Fine-Grained Image Classification with Textual Cue
- Number-based Person Re-Identification
- From Text Recognition to Person Re-Identification

# Fine-Grained Image Classification with Textual Cue

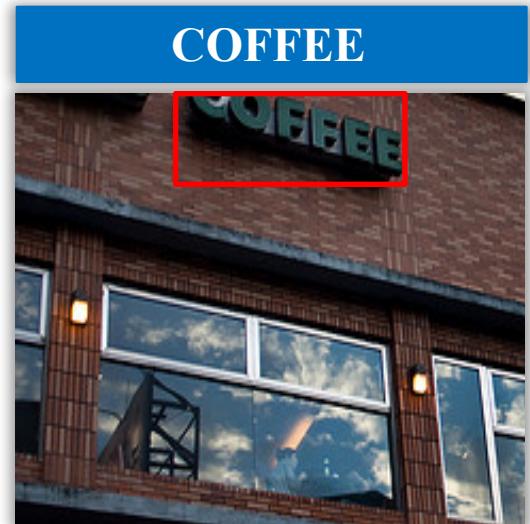
## Motivations



(a)



(b)



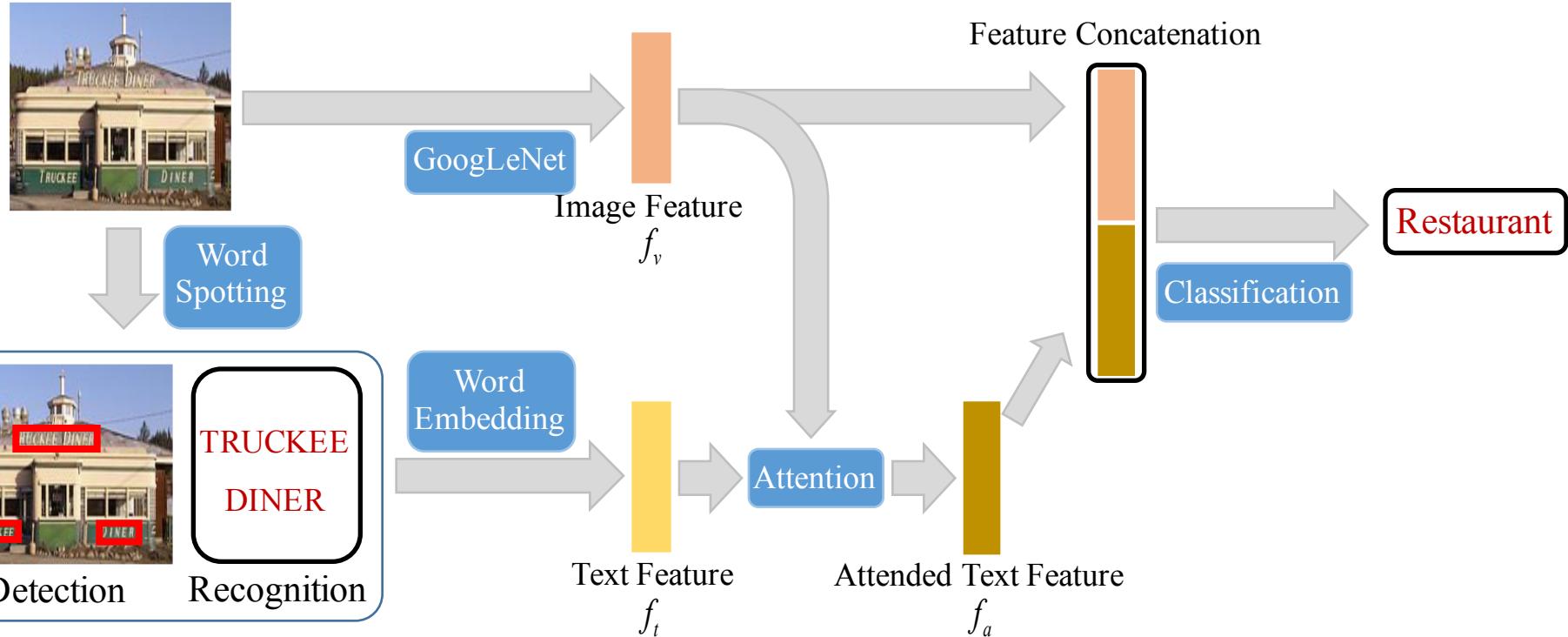
(c)

- ❑ Visual cues would group (a)-(b) whereas scene would group (b)-(c).
- ❑ Texts in images can improve the performance of fine-grained image classification.

[1] Bai X. et al. Integrating Scene Text and Visual Appearance for Fine-Grained Image Classification with Convolutional Neural Networks[J]. arXiv:1704.04613, 2017.

# Fine-Grained Image Classification with Textual Cue

## Pipeline



[1] Bai X. et al. Integrating Scene Text and Visual Appearance for Fine-Grained Image Classification with Convolutional Neural Networks[J]. arXiv:1704.04613, 2017.

# Fine-Grained Image Classification with Textual Cue

## Attention Model to Select Relevant Words



Repair shop



Hotel

- Some **irrelevant words** to this Category



# Fine-Grained Image Classification with Textual Cue

Con-Text dataset<sup>[1]</sup>



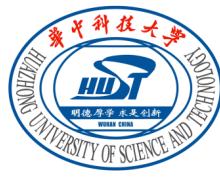
Drink Bottledataset<sup>[2]</sup>



- ❑ 28 categories of **Scenes**
- ❑ 24,255 images in total

- ❑ Selected from ImageNet
- ❑ 20 categories of **Drink Bottles**
- ❑ 18,488 images in total

[1] S. Karaoglu. et al. Con-text: text detection using background connectivity for fine-grained object classification. ACM2013  
[2] Bai X. et al. Integrating Scene Text and Visual Appearance for Fine-Grained Image Classification with Convolutional Neural Networks[J]. arXiv2017.



# Fine-Grained Image Classification with Textual Cue

## Results: mAP(%) improvement on different datasets

Method	Dataset	
	Con-Text	Drink Bottle
GoogLeNet <sup>[1]</sup>	61.3	63.1
<b>GoogLeNet + Textual Cue</b>	<b>79.6 (+18.3)</b>	<b>72.8 (+9.7)</b>

[1] C. Szegedy, et al. Going deeper with convolutions. CVPR2015

# Fine-Grained Image Classification with Textual Cue

## Visualization: learned weights of recognized words



### BAKERY

- CAKES: 0.57
- PASTRIES: 0.43
- OPEN: 5.5e-9
- ...

### CAFE

- STARBUCKS: 1
- SCOFF: 1.1e-8
- ...

### ROOTBEER

- ROOT: 0.89
- BEER: 0.11
- BREWED: 1.3e-6
- ...

### CHABLIS

- CHABLIS: 0.99
- FRANCE: 8.7e-12
- ...

□ **Filter** the incorrect recognized words

□ **Select** more related words to the category

# Fine-Grained Image Classification with Textual Cue

## Results of Image Search

Visual cue only



Root Beer      Cream Soda      Guinness      Slivovitz      Ginger ale

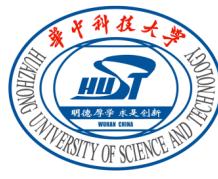
Visual and Textual Cues



Root Beer      Sarsaparilla

## Retrieval Results

Method	mAP(%)
GoogLeNet	48.0
GoogLeNet+Textual Cue	60.8 (+12.8)

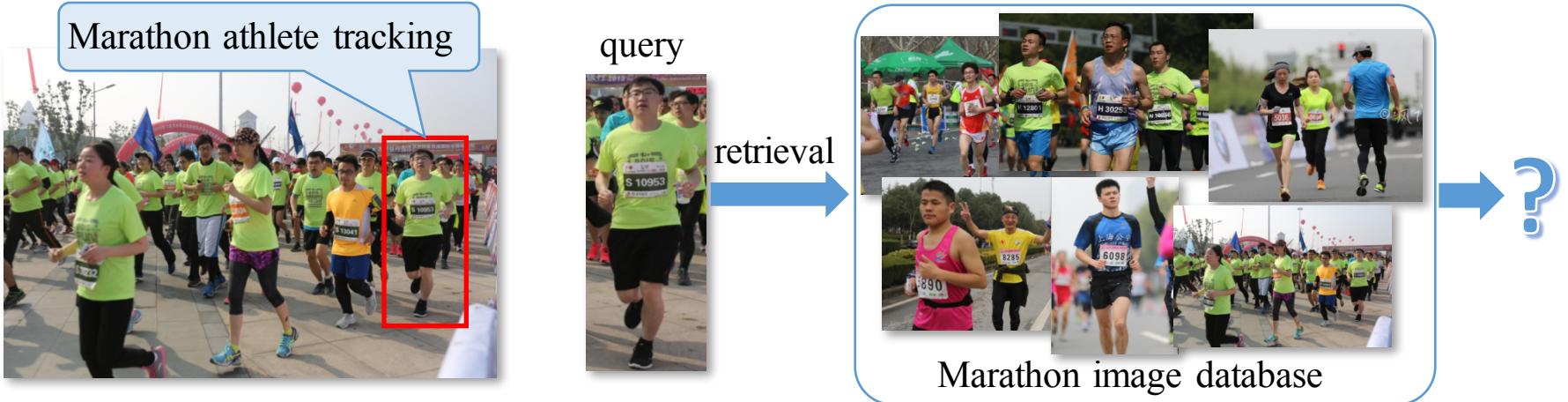


# Applications

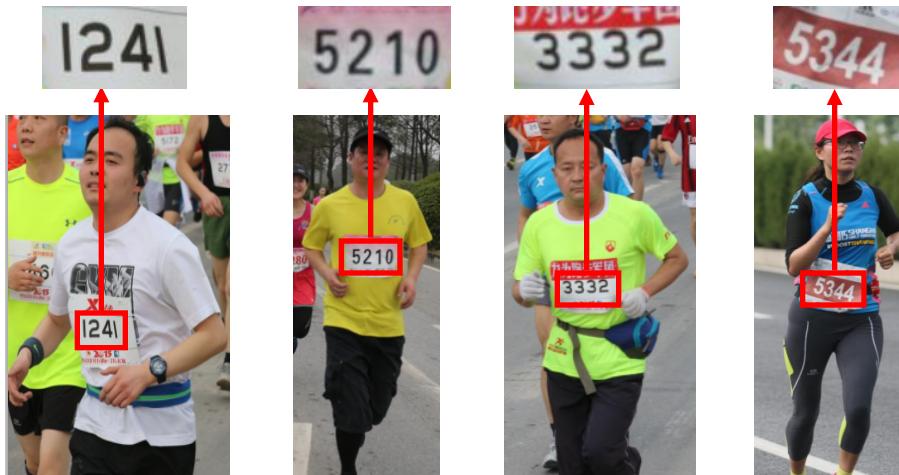
- Fine-Grained Image Classification with Textual Cue
- Number-based Person Re-Identification
- From Text Recognition to Person Re-Identification

# Number-based Person Re-Identification

- Problem: hard to track and retrieve an athlete in a marathon game

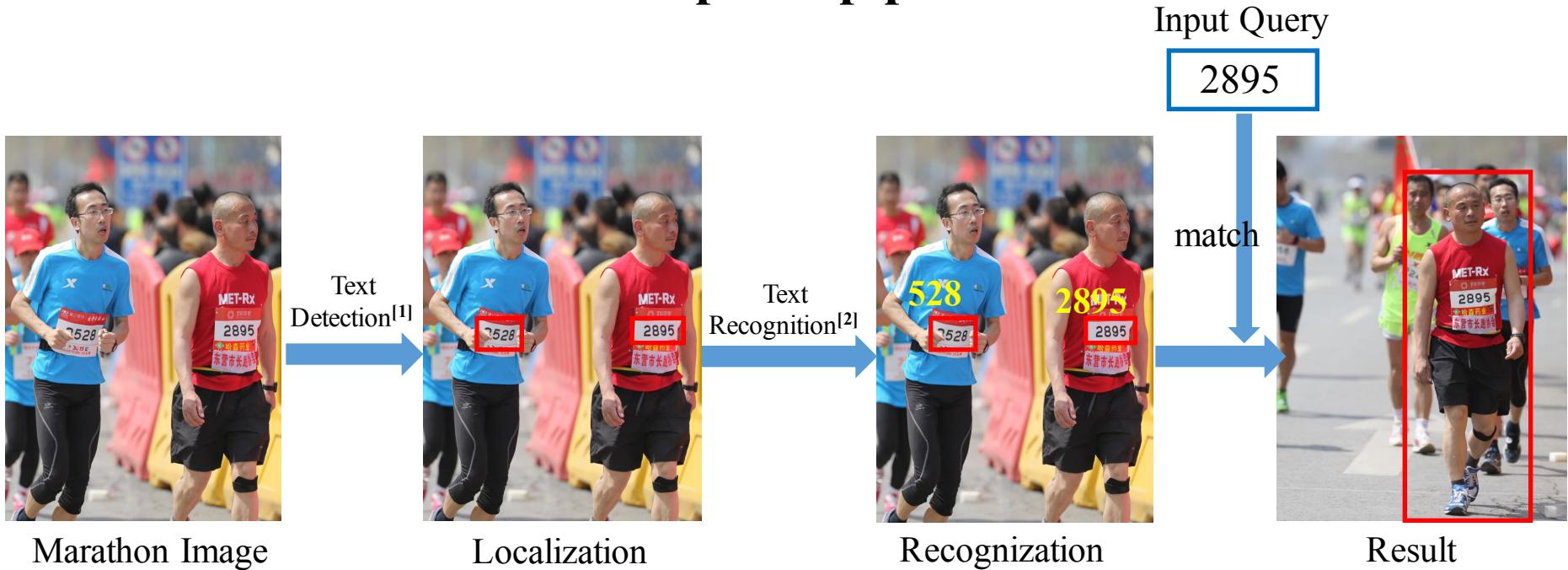


- Motivation: every athlete has a unique racing bib number



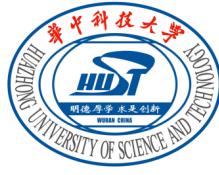
# Number-based Person Re-Identification

## Proposed pipeline



[1] M. Liao et al. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. AAAI, 2017.

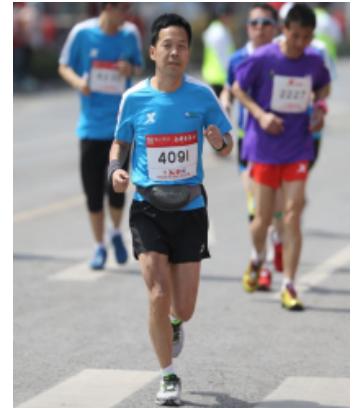
[2] Shi B, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. TPAMI, 2017.



# Number-based Person Re-Identification

## Marathon Dataset

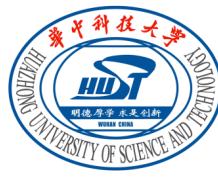
8706 training images, 1000 testing images



## Experimental Results

*Identification accuracy rate(Id\_acc): 85%*

$$Id\_acc = \frac{Num(\text{correctly recognized persons})}{Num(\text{total persons})}$$



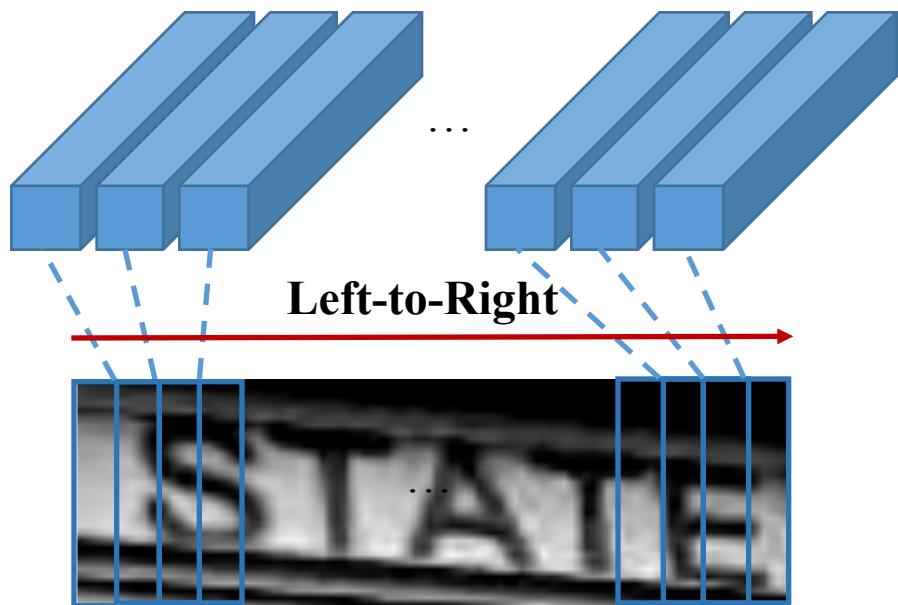
# Applications

- Fine-Grained Image Classification with Textual Cue
- Number-based Person Re-Identification
- From Text Recognition to Person Re-Identification

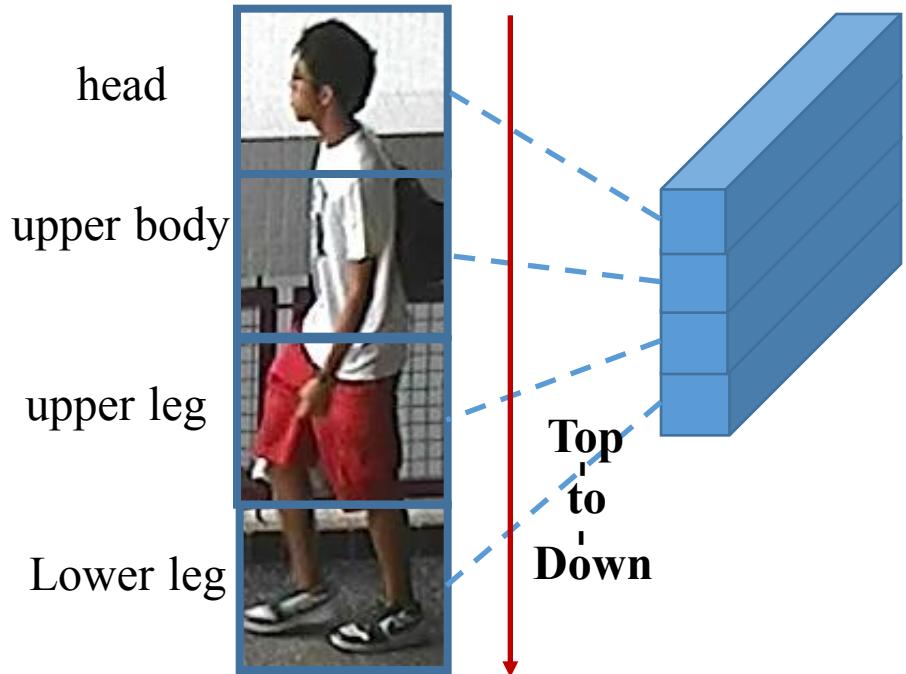
# From Text Recognition to Person Re-Identification

## Sequence Modeling

Text Recognition (CRNN)



Person Re-Identification

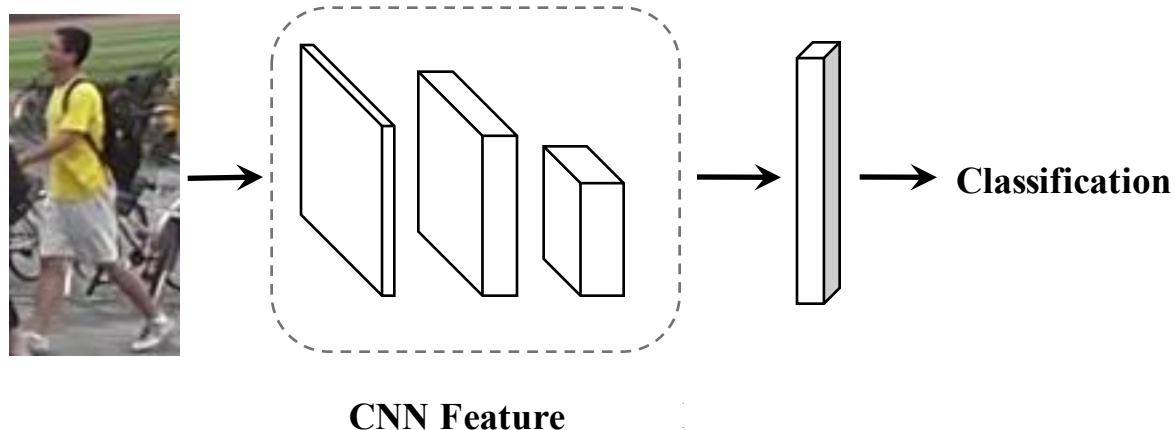


[1] CRNN: Shi B et al. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. TPAMI, 2017.

# From Text Recognition to Person Re-Identification

## Model Architecture

CNN + LSTM



Results on Market1501<sup>[1]</sup>

Method	mAP(%)	R1(%)
CNN	59.8	81.4
<b>CNN + LSTM</b>	<b>65.5</b>	<b>85.8</b>

R1: given a query, precision of the top-1 similar image from gallery discriminated by model.

[1] Zheng et al. Scalable Person Re-identification: A Benchmark. ICCV 2015

# From Text Recognition to Person Re-Identification

## Retrival Results

CNN



.... ....

query

CNN+LSTM



.... ....

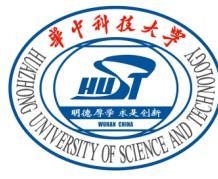
query



# Outline

---

- Background
- Scene Text Detection
- Scene Text Recognition
- Applications
- Future Trends



# Future Trends

- Irregular text detection (Curved & Perspective Text Lines )
- Multilingual End-to-end text recognition
- Semi-supervised or weakly supervised text detection and recognition
- Text image synthesis (GAN)
- Unified framework for OCR and NLP
- Integrating Scene text and Image/Videos for many applications.



# Resources (Papers & Datasets & Codes)

- ❑ B. Shi, C. Yao, M. Liao, M Yang, P Xu, L Cui, S Belongie, S Lu, X Bai.  
**ICDAR2017 Competition on Reading Chinese Text in the Wild ( RCTW-17).** ICDAR'17  
Dataset : <http://mclab.eic.hust.edu.cn/icdar2017chinese>
- ❑ B. Shi, X. Bai, S. Belongie.  
**Detecting Oriented Text in Natural Images by Linking Segments.** CVPR'17  
Code: <https://github.com/bgshih/semlink>
- ❑ M. Liao, B. Shi, X. Bai, X. Wang, W. Liu.  
**TextBoxes: A fast text detector with a single deep neural network.** AAAI'17  
Code: <https://github.com/MhLiao/TextBoxes>
- ❑ B. Shi, X. Bai, C. Yao.  
**An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition.** TPAMI'17  
Code: [http://mclab.eic.hust.edu.cn/~xbai/CRNN/crnn\\_code.zip](http://mclab.eic.hust.edu.cn/~xbai/CRNN/crnn_code.zip)
- ❑ B. Shi, , X. Wang, P. Lyu, C. Yao, X. Bai.  
**Robust scene text recognition with automatic rectification.** CVPR'16
- ❑ X. Bai, M. Yang, P. Lyu, et al.  
**Integrating Scene Text and Visual Appearance for Fine-Grained Image Classification**  
arXiv2017.



# Literature review (Papers & PPTs)

## ❑ [Survey Paper]

Scene text detection and recognition: Recent advances and future trends.

Y Zhu, C Yao, X Bai.

Frontiers of Computer Science 10 (1), 19–36, 2016.

[http://mclab.eic.hust.edu.cn/UpLoadFiles/Papers/FCS\\_TextSurvey\\_2015.pdf](http://mclab.eic.hust.edu.cn/UpLoadFiles/Papers/FCS_TextSurvey_2015.pdf)

## ❑ [Talk PPT in 2014]

Representation in Scene Text Detection and Recognition.

[http://mclab.eic.hust.edu.cn/~xbai/Talk\\_slice/Representation%20in%20Scene%20Text%20Detection%20and%20Recognition\\_20150207.pdf](http://mclab.eic.hust.edu.cn/~xbai/Talk_slice/Representation%20in%20Scene%20Text%20Detection%20and%20Recognition_20150207.pdf)

## ❑ [Talk PPT in 2017]

Oriented Scene Text Detection Revisited.

[http://mclab.eic.hust.edu.cn/~xbai/Talk\\_slice/Oriented-Scene-Text-Detection-Revisited\\_VALSE2017.pdf](http://mclab.eic.hust.edu.cn/~xbai/Talk_slice/Oriented-Scene-Text-Detection-Revisited_VALSE2017.pdf)



# Collaborators



Cong Yao.  
Cloud Team chief, Megvii Inc.



Baoguang Shi.  
PHD Candidate, HUST



Zheng Zhang.  
Associate Researcher, MSRA



Chengquan Zhang.  
Researcher, Baidu IDL



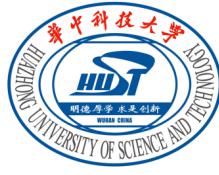
Minghui Liao.  
PHD student, HUST



Mingkun Yang.  
Master student, HUST



Serge Belongie.  
Professor, Cornell



# Refer to my homepage for more details

---



## Thank you !