香港中文大學
The Chinese University of Hong Kong

HONG KONG COMPUTER VISION WORKSHOP

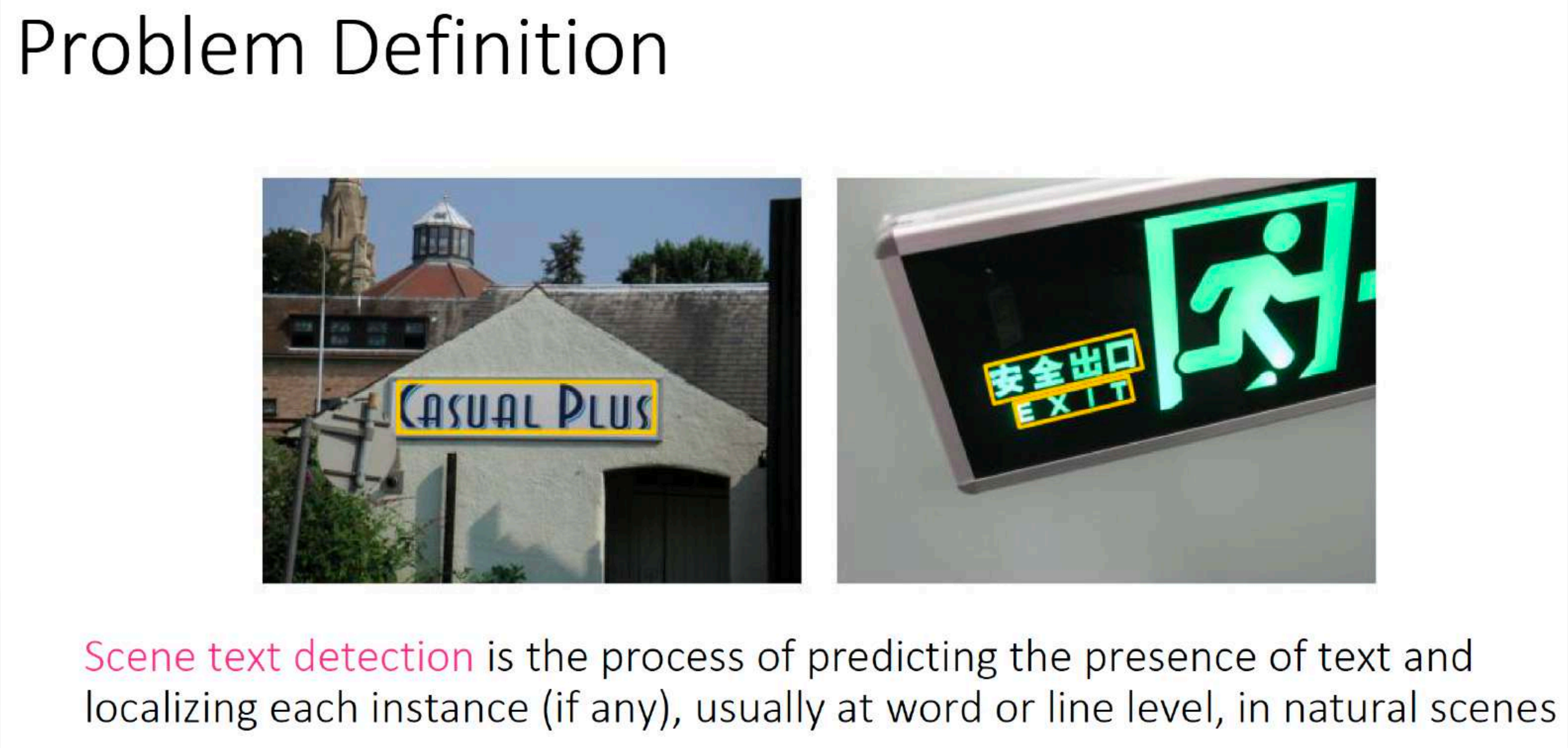
OCT 13，2019

Yasumoto International Academic Park（YIA）LT5（2/f）
The Chinese University of Hong Kong

Efficient and Accurate Arbitrary-Shaped Text Detection with Pixel Aggregation Network

The University of Hong Kong
Enze Xie Superviser: Dr. Luo Ping

TEXT DETECTION

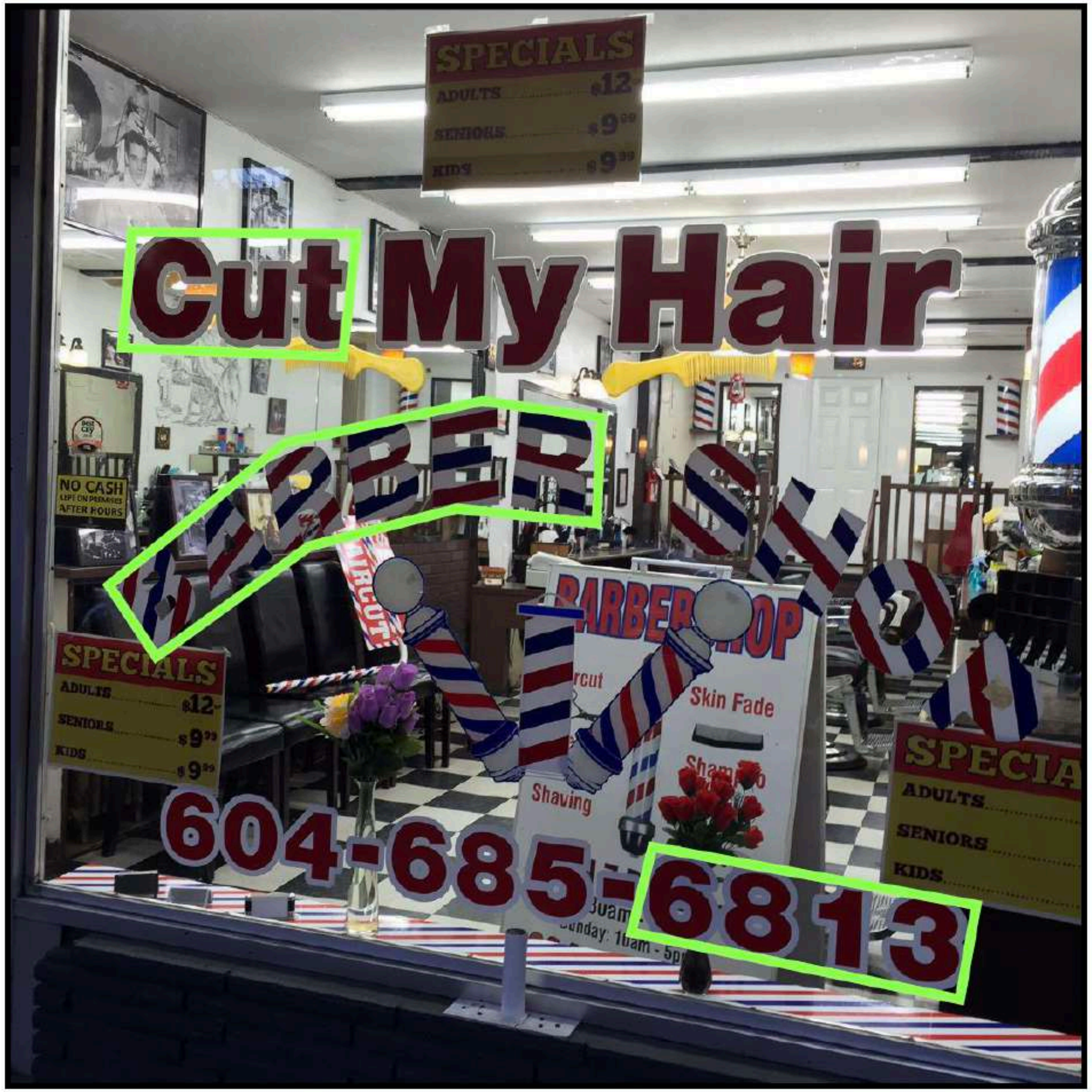


Problem Definition

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

TEXT DETECTION

Challenge:
Text with horizontal, multi-oriented, and curve shapes.



Speed vs Accuracy

Most STOA arbitrary-shape text detector
are too slow to be applied in real-world application.

We propose PAN to balance the speed and accuracy.

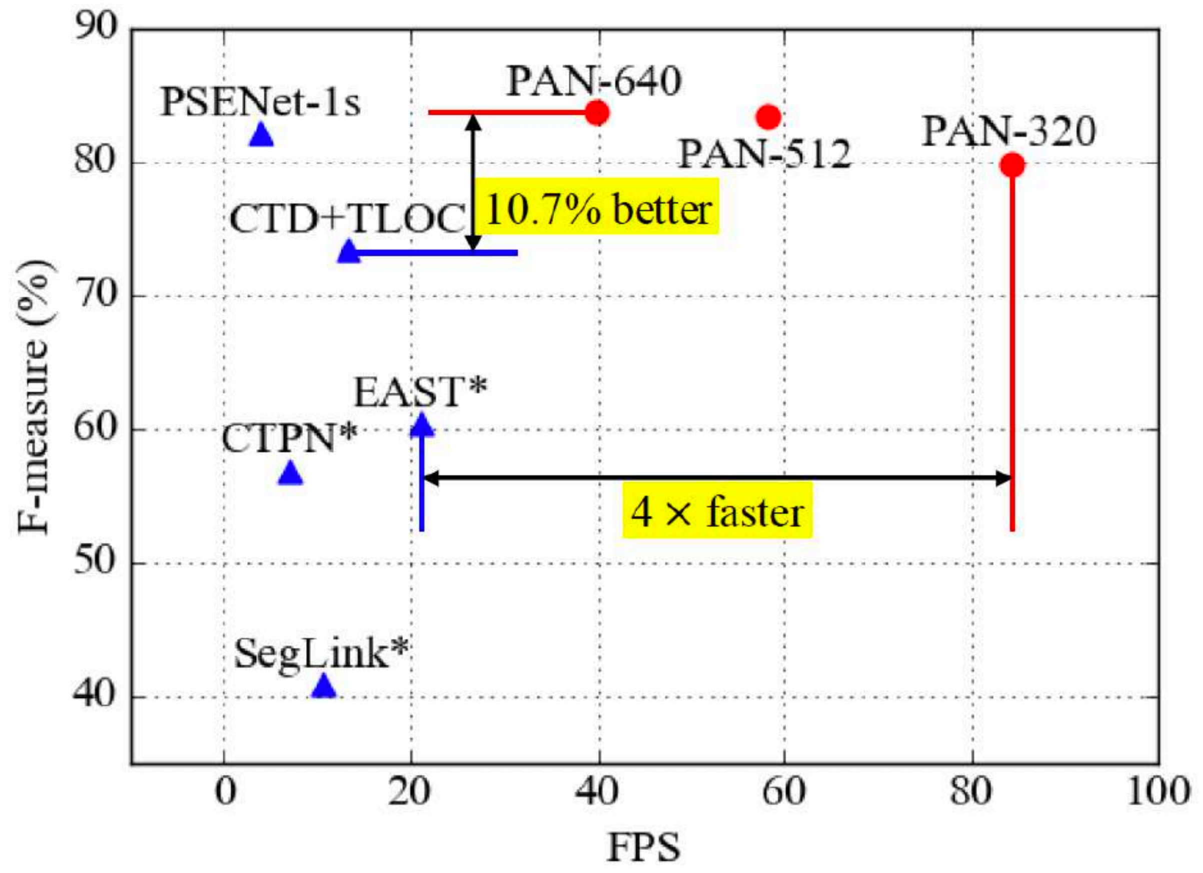


Figure 1. The performance and speed on curved text dataset CTW1500. PAN-640 is 10.7% better than CTD+TLOC, and PAN-320 is 4 times faster than EAST. * indicates the results from [31].

Pipeline

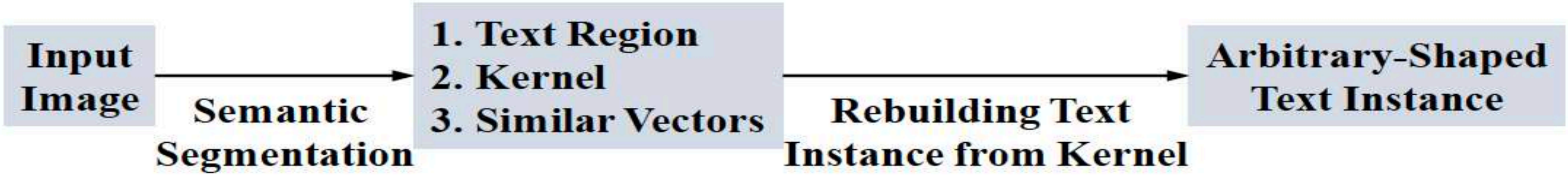


Figure 2. The overall pipeline of PAN.

Pipeline

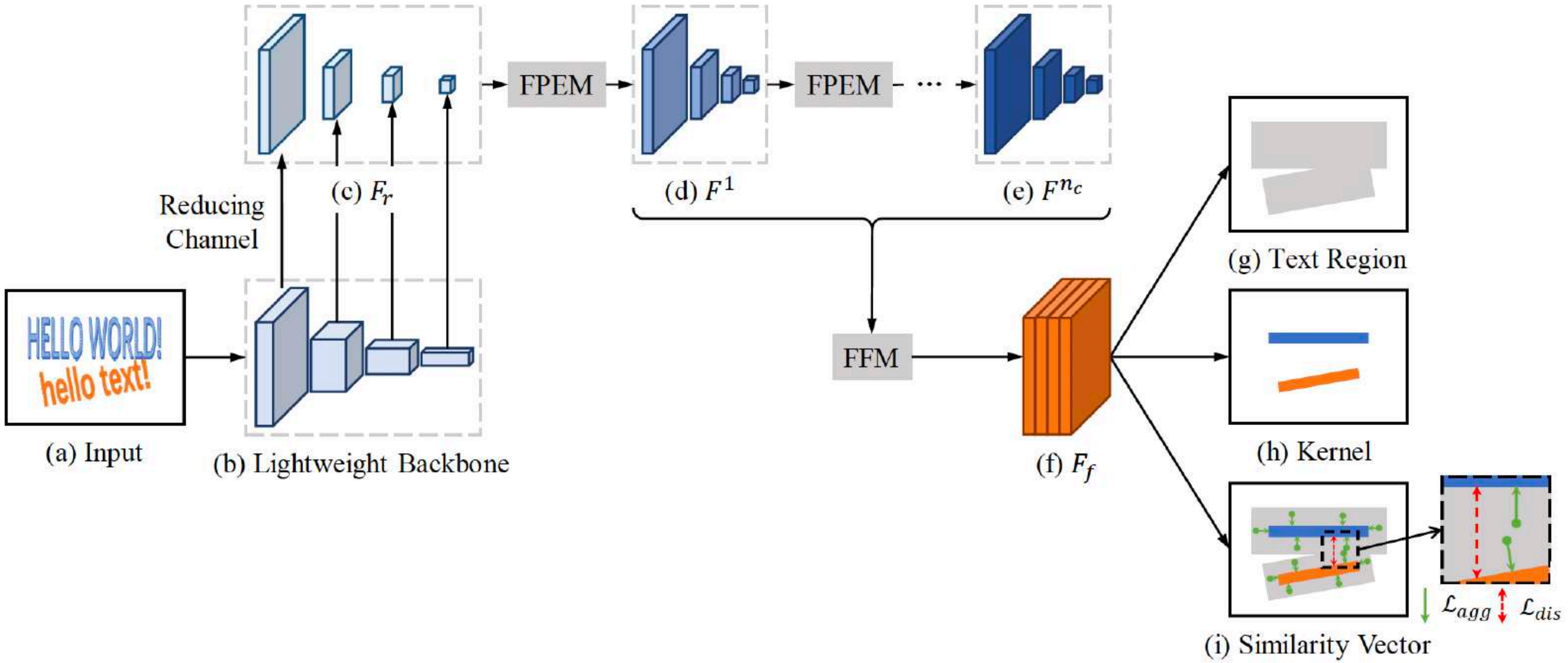


Figure 3. The overall architecture of PAN. The features from lightweight backbone network are enhanced by a low computational-cost segmentation head which is composed of Feature Pyramid Enhancement Module (FPEM) and Feature Fusion Module (FFM). The network predicts text regions, kernels and similarity vectors to describe the text instances.

FPEM

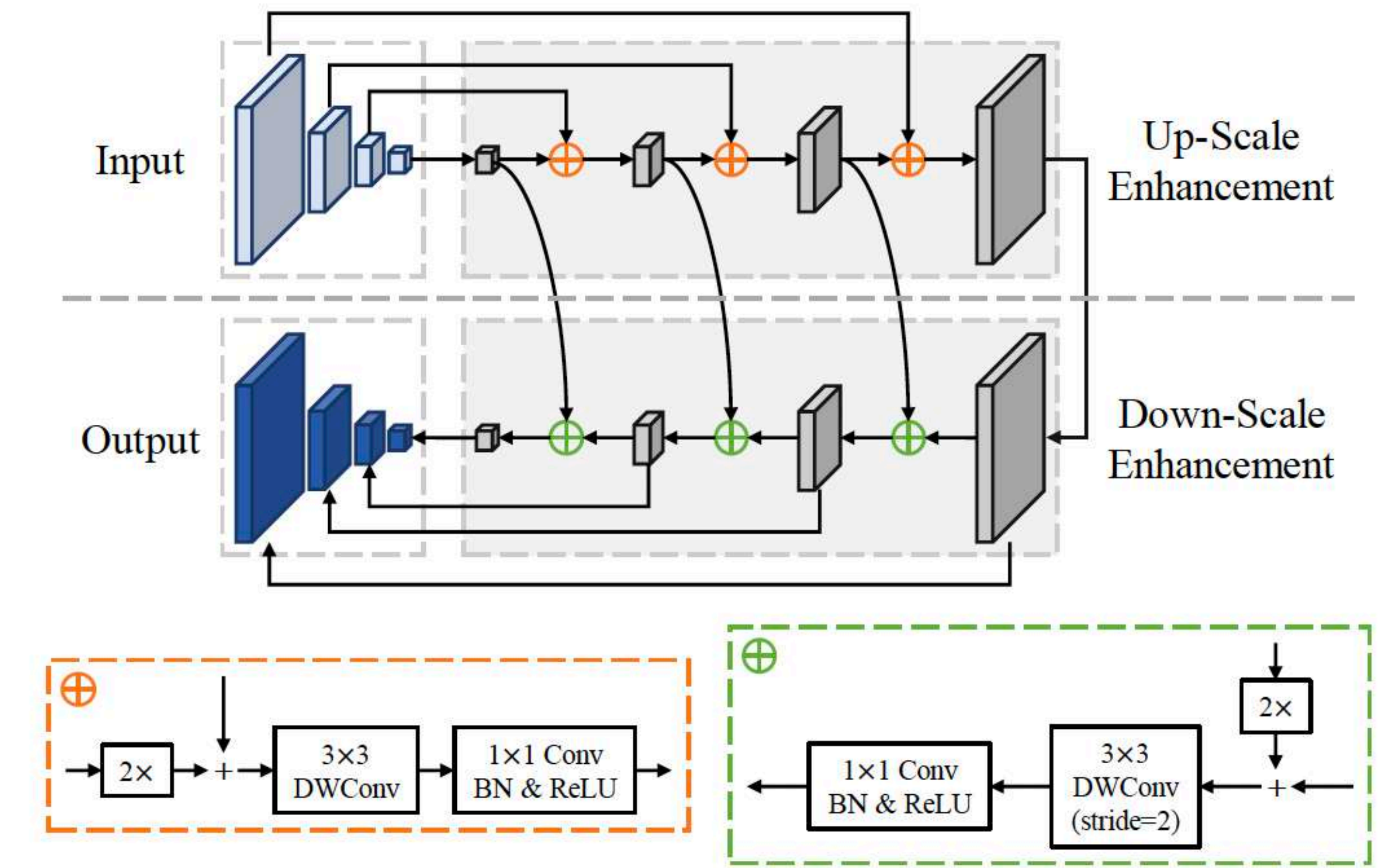


Figure 4. The details of FPEM. “+”, “2×”, “DWConv”, “Conv” and “BN” represent element-wise addition, 2× linear upsampling, depthwise convolution [18], regular convolution [23] and Batch Normalization [21] respectively.

FFM

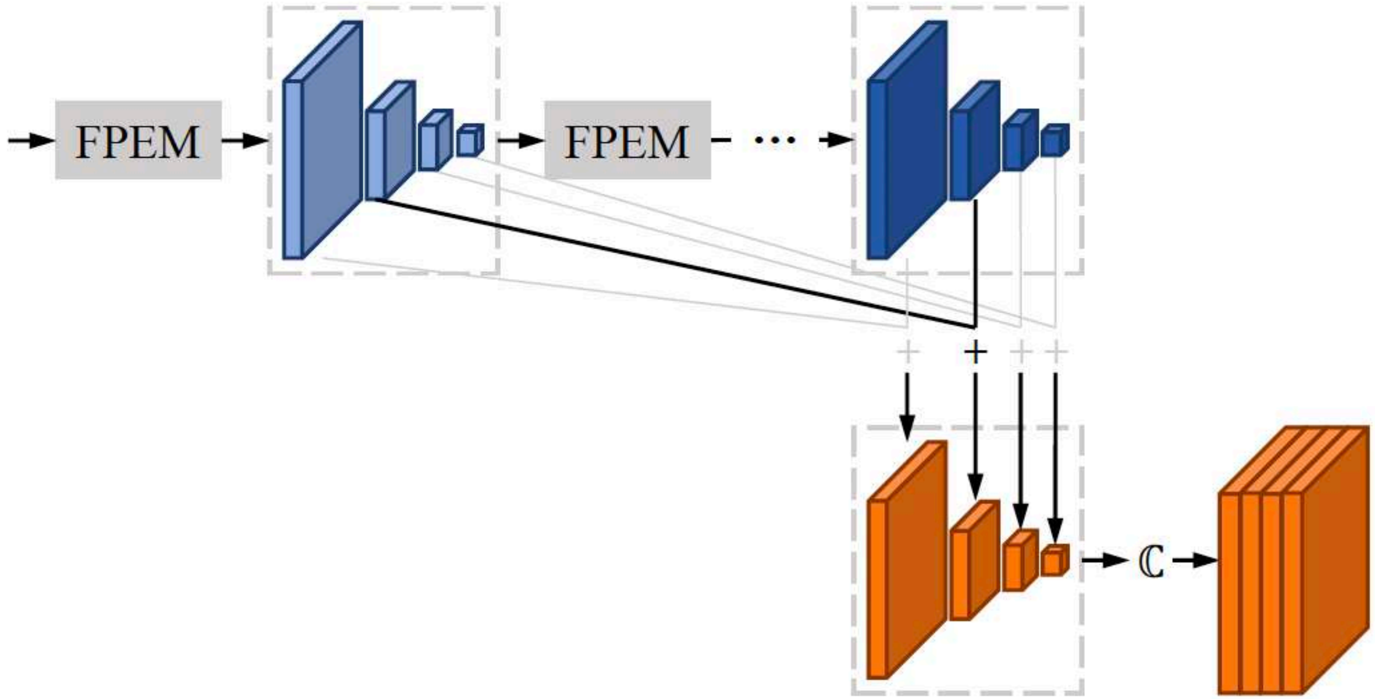


Figure 5. The detail of FFM. "+" is element-wise addition. "$\mathcal{C}$" is the operation of upsampling and concatenating.

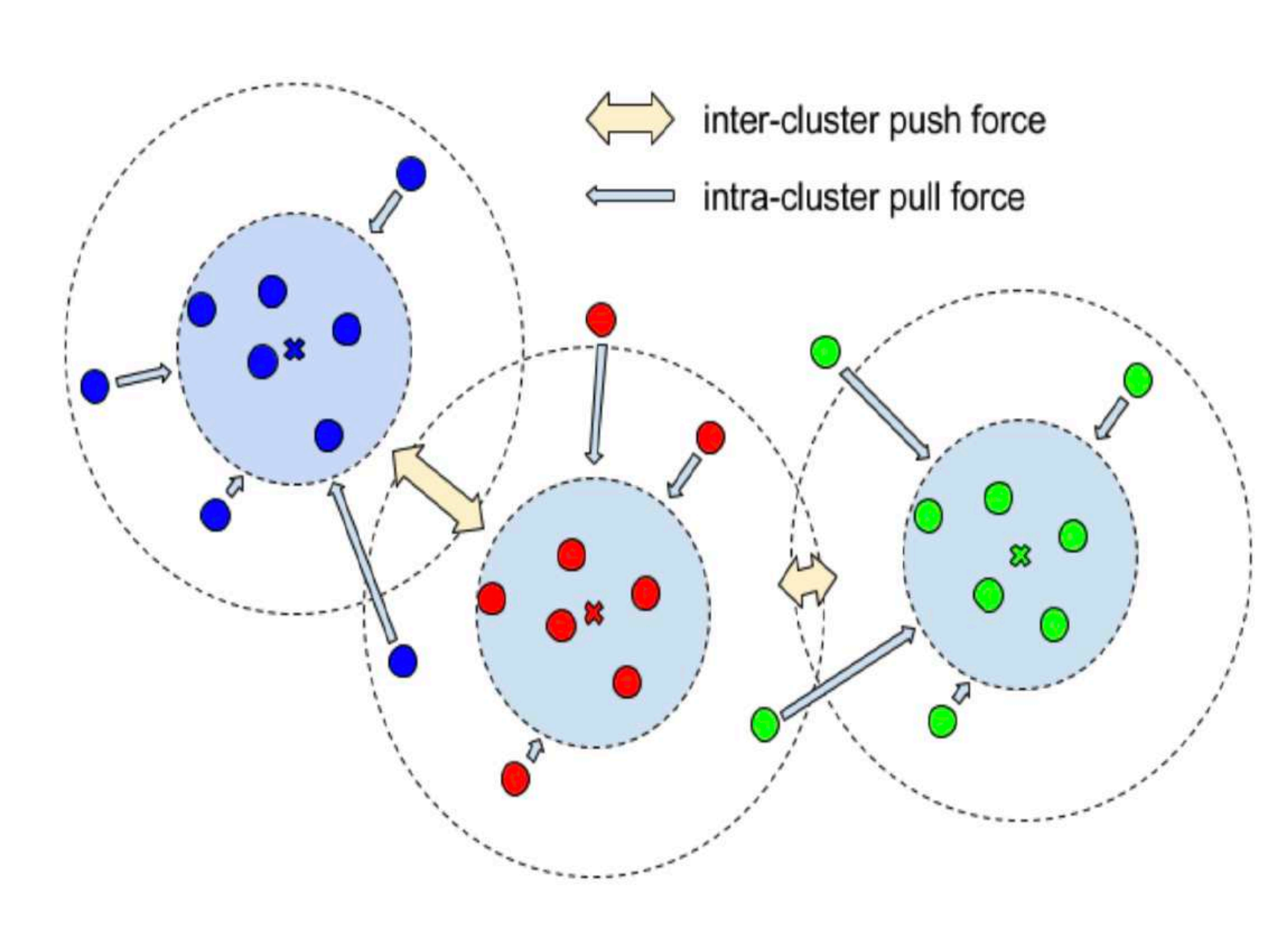
Pixel Aggregation（PA）

$$\mathcal{L}_{agg} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|T_i|}\sum_{p\in T_i} ln(\mathcal{D}(p, K_i) + 1),$$

$$\mathcal{D}(p, K_i) = max(\|\mathcal{F}(p) - \mathcal{G}(K_i)\| - \delta_{agg}, 0)^2,$$

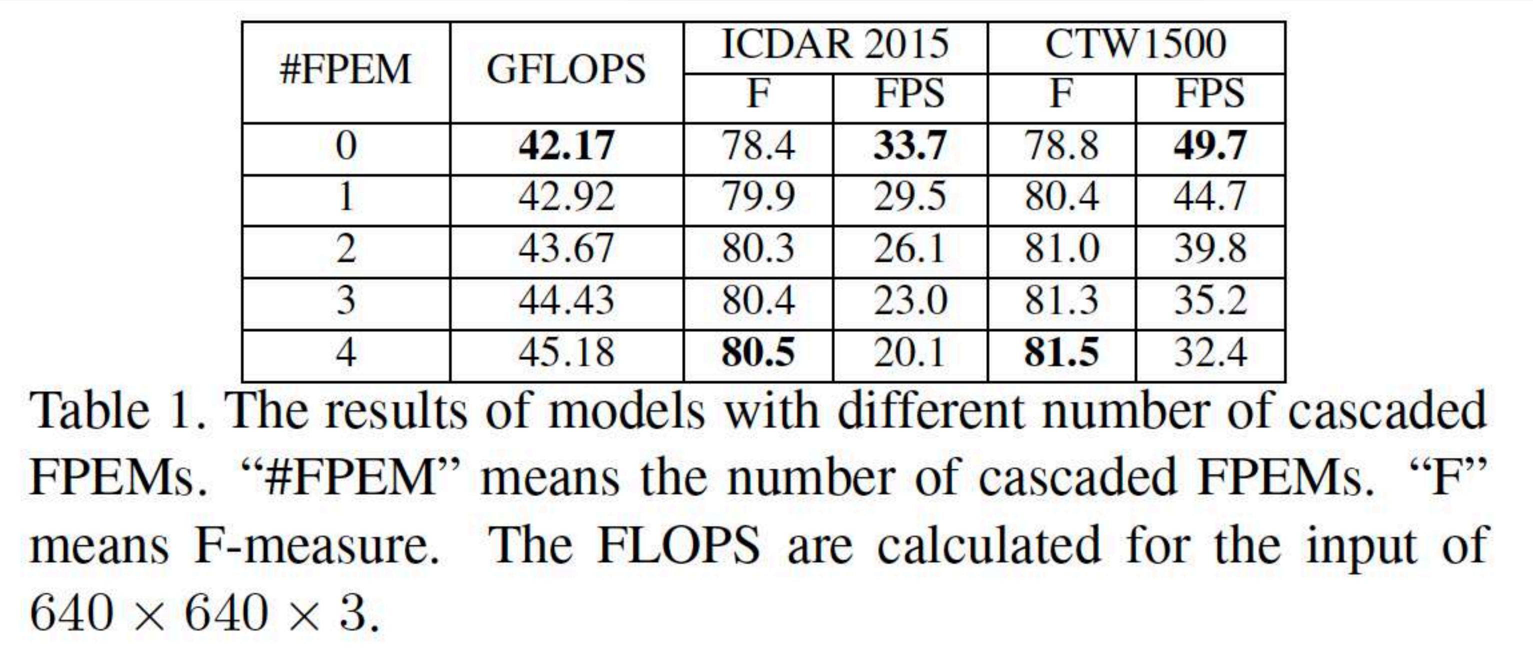
$$\mathcal{L}_{dis} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} ln(\mathcal{D}(K_i, K_j) + 1),$$

$$\mathcal{D}(K_i, K_j) = max(\delta_{dis} - \|\mathcal{G}(K_i) - \mathcal{G}(K_j)\|, 0)^2.$$



Semantic Instance Segmentation with a Discriminative Loss Function

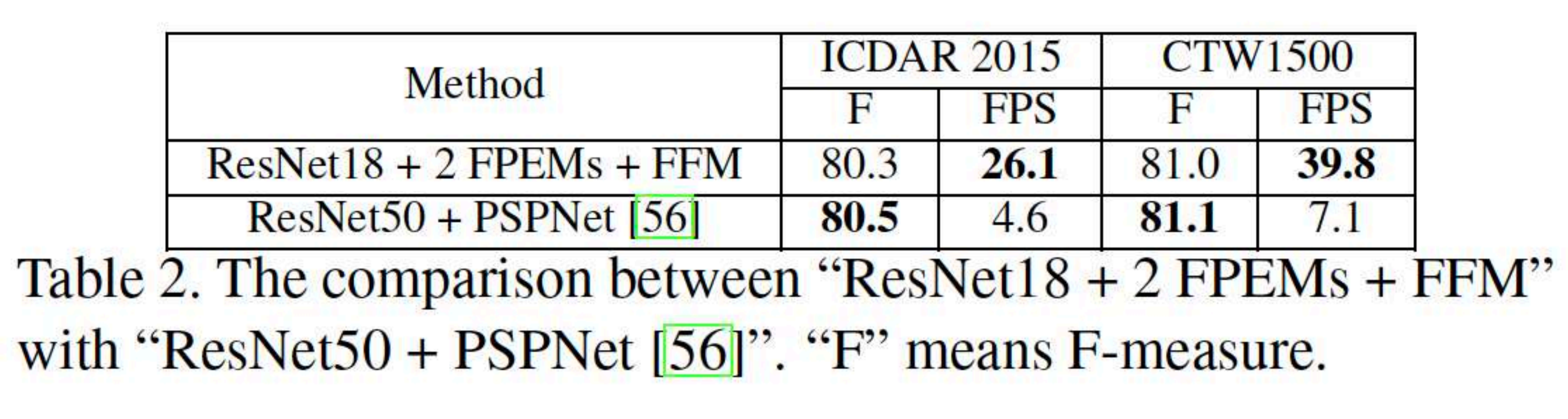
Ablation study on numbers of FPEM



#FPEM	GFLOPS	ICDAR 2015		CTW1500	
		F	FPS	F	FPS
0	**42.17**	78.4	**33.7**	78.8	**49.7**
1	42.92	79.9	29.5	80.4	44.7
2	43.67	80.3	26.1	81.0	39.8
3	44.43	80.4	23.0	81.3	35.2
4	45.18	**80.5**	20.1	**81.5**	32.4

Table 1. The results of models with different number of cascaded FPEMs. “#FPEM” means the number of cascaded FPEMs. “F” means F-measure. The FLOPS are calculated for the input of $640 \times 640 \times 3$.

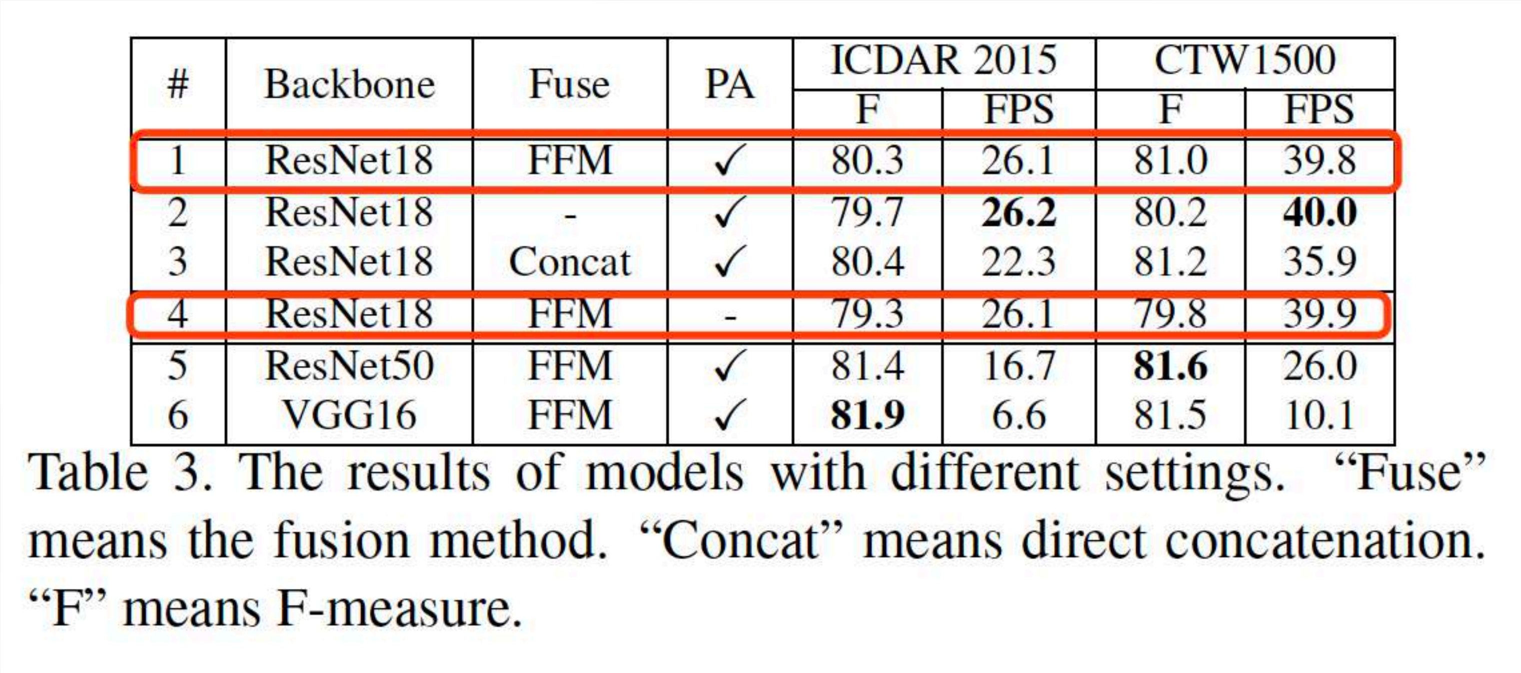
Ablation study on numbers of FPEM



Method	ICDAR 2015		CTW1500	
	F	FPS	F	FPS
ResNet18 + 2 FPEMs + FFM	80.3	**26.1**	81.0	**39.8**
ResNet50 + PSPNet [56]	**80.5**	4.6	**81.1**	7.1

Table 2. The comparison between “ResNet18 + 2 FPEMs + FFM” with “ResNet50 + PSPNet [56]”. “F” means F-measure.

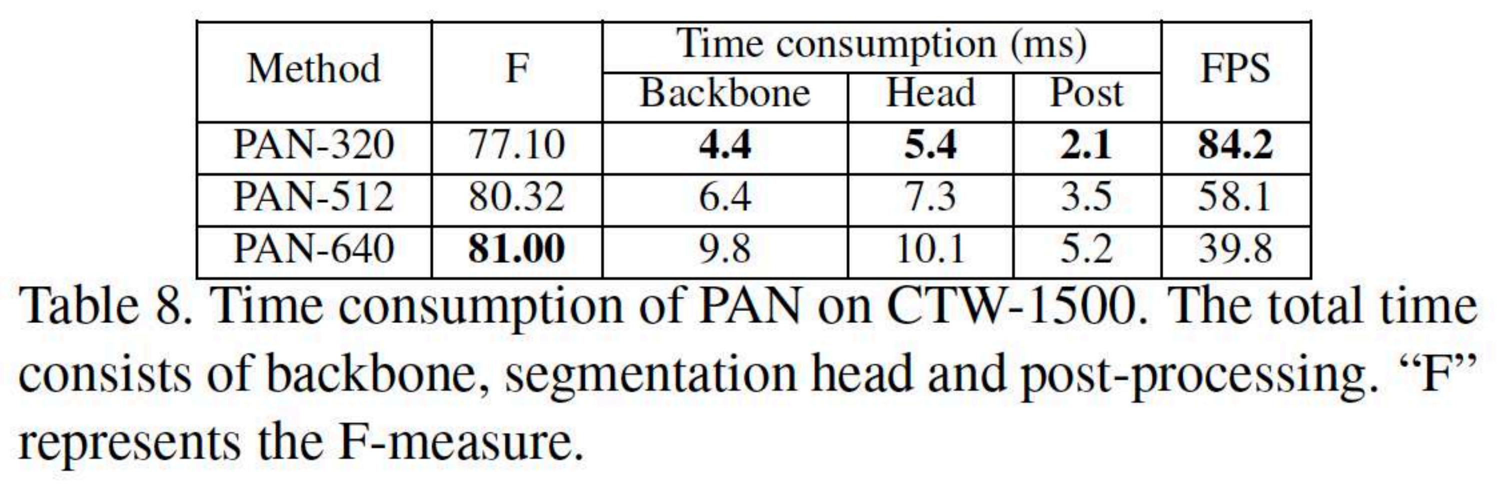
The effectiveness of PA.



#	Backbone	Fuse	PA	ICDAR 2015		CTW1500	
				F	FPS	F	FPS
1	ResNet18	FFM	✓	80.3	26.1	81.0	39.8
2	ResNet18	-	✓	79.7	**26.2**	80.2	**40.0**
3	ResNet18	Concat	✓	80.4	22.3	81.2	35.9
4	ResNet18	FFM	-	79.3	26.1	79.8	39.9
5	ResNet50	FFM	✓	81.4	16.7	**81.6**	26.0
6	VGG16	FFM	✓	**81.9**	6.6	81.5	10.1

Table 3. The results of models with different settings. “Fuse” means the fusion method. “Concat” means direct concatenation. “F” means F-measure.

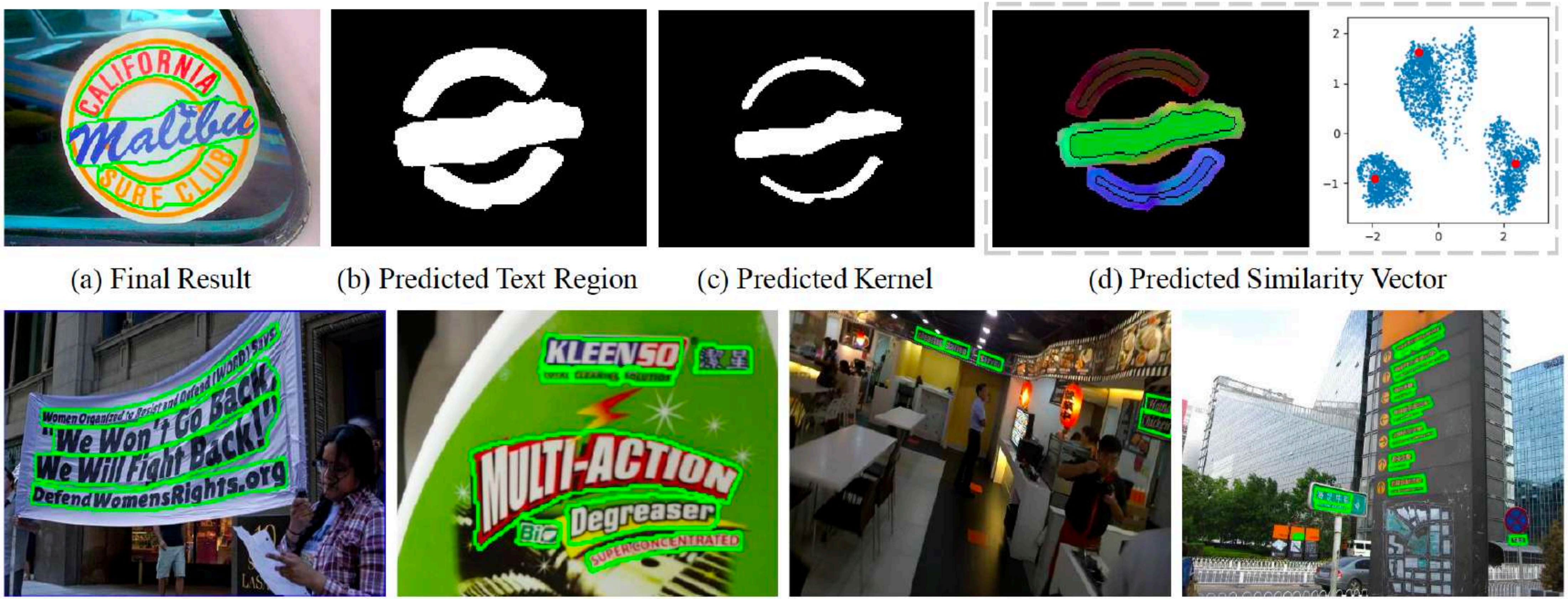
Speed analyze



Method	F	Time consumption (ms)			FPS
		Backbone	Head	Post	
PAN-320	77.10	**4.4**	**5.4**	**2.1**	**84.2**
PAN-512	80.32	6.4	7.3	3.5	58.1
PAN-640	**81.00**	9.8	10.1	5.2	39.8

Table 8. Time consumption of PAN on CTW-1500. The total time consists of backbone, segmentation head and post-processing. “F” represents the F-measure.

visualize



(a) Final Result (b) Predicted Text Region (c) Predicted Kernel (d) Predicted Similarity Vector

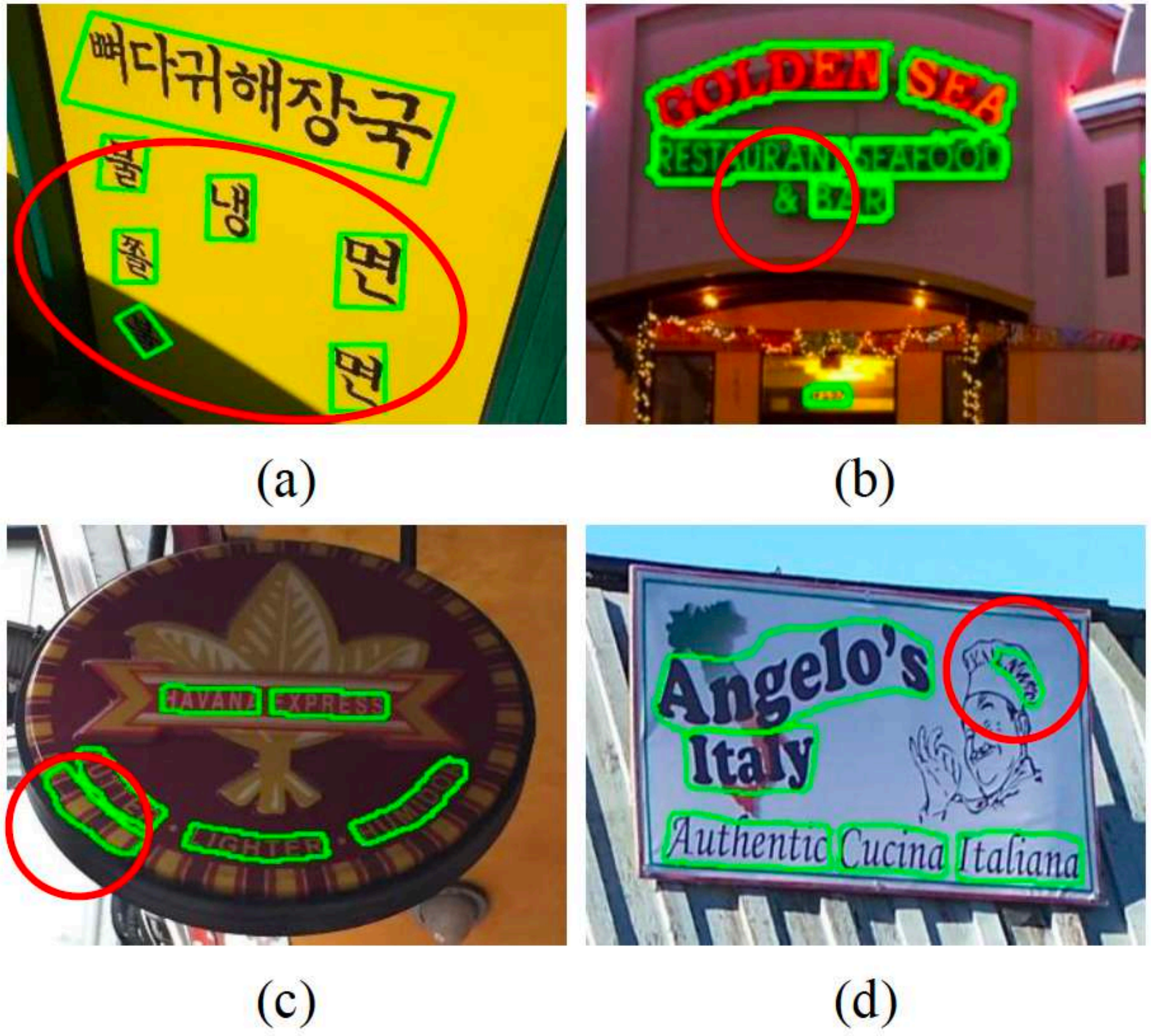
(e) CTW1500 (f) Total-Text (g) ICDAR 2015 (h) MSRA-TD500

Figure 6. Qualitative results of PAN. (a) is the final result of PAN. (b) is the predicted text regions. (c) is the predicted kernels. (d) is the visualization of similarity vectors, which is the best viewed in color and scatter diagram. (e)-(h) are results on four standard benchmarks.

Failure case

1 False Positive
2 Word Spilt

Need more NLP information to handle this problem



(a) (b)

(c) (d)

Figure 7. Failure Samples.

THANKS

THANK YOU FOR WATCHING