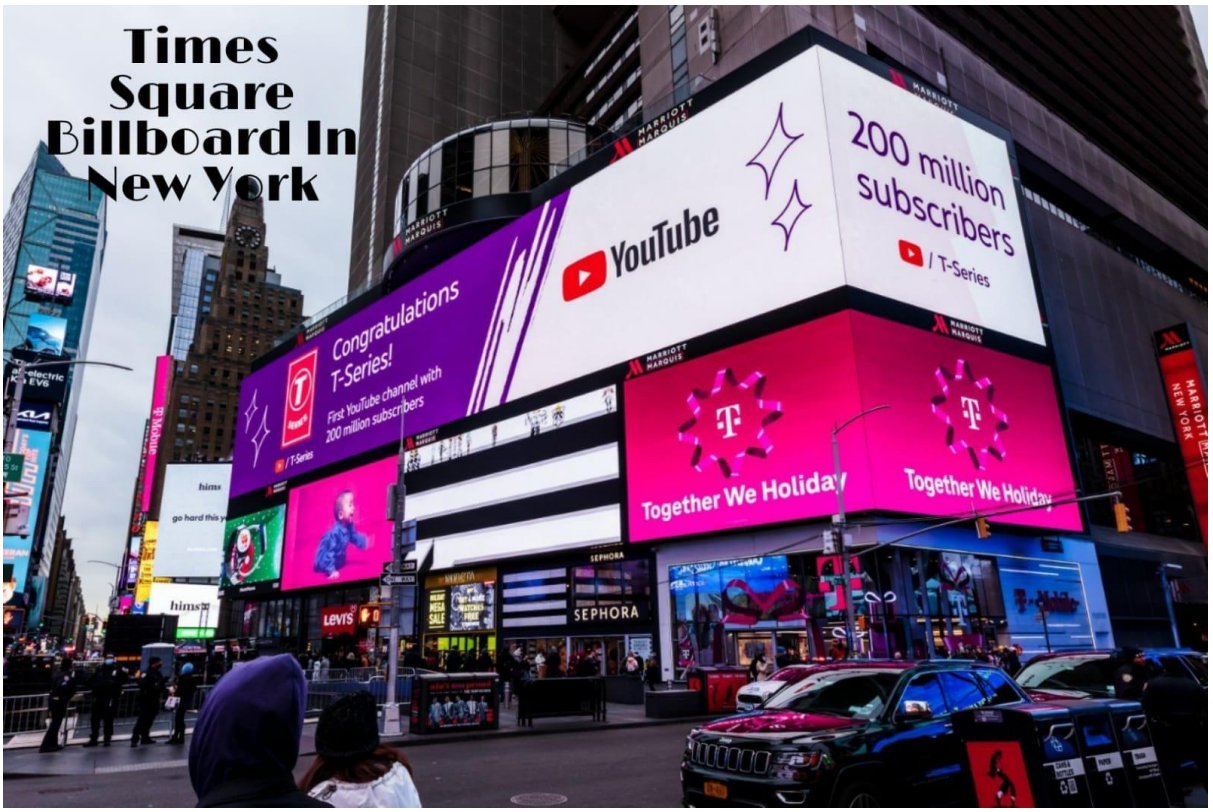
Paper	Year	Total Text	MSRA TD500	CTW 1500	IC-17	IC-15	IC-13
Boundary Transformer	2022	**90.13**	**90.10**	**86.49**			
Unified Text Detection and Layout Analysis	2022	87.94	87.70	85.97	**77.24**		
CRAFT	2019		82.9		73.9	86.9	95.2
PCR	2021	85.2	87.0	84.7			
GLASS	2022	86.2				78.8	
YAMTS	2021	81.5				**87.0**	95.2
Mask TextSpotter v3	2020	78.4	83.5			75.1	
MASTER	2021					79.4	**95.3**
Centripetal Text	2021	86.3	86.1	83.9			



1. Boundary Transformer (2022, paper, code)
   Total-Text: 90.13
   CTW-1500: 86.49
   MSRA-TD500: 90.10

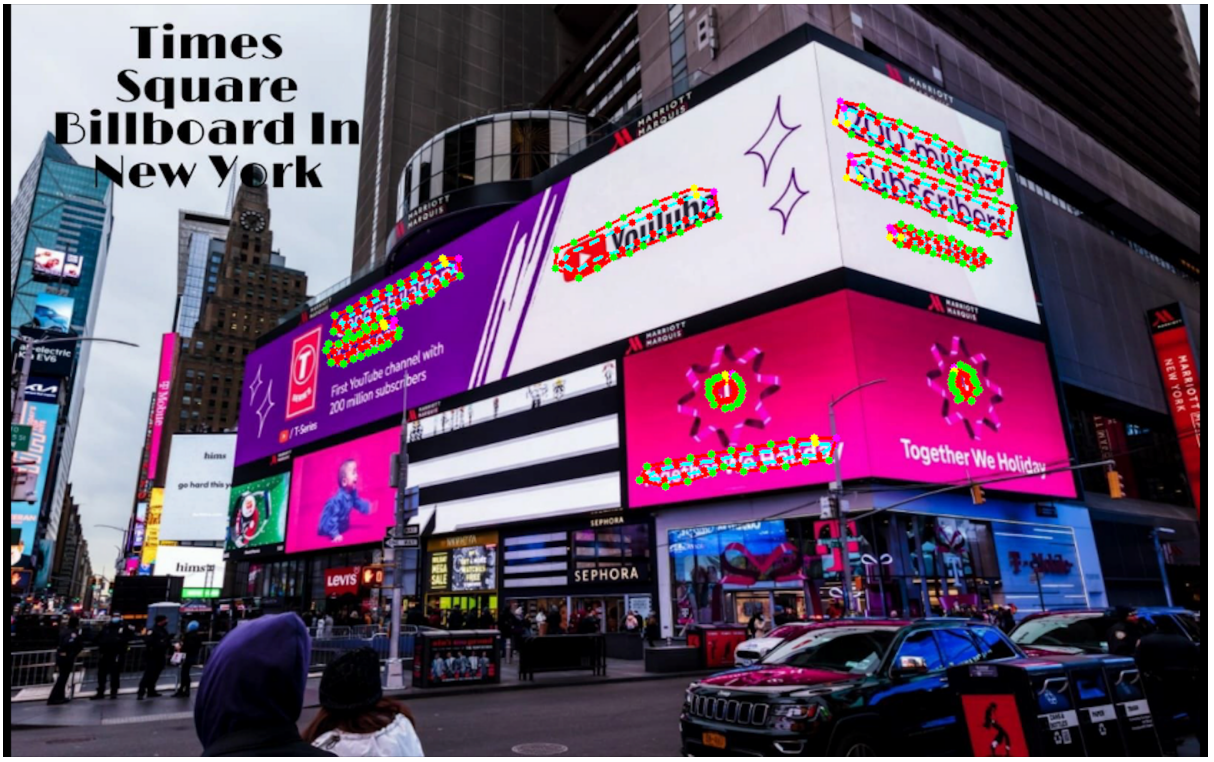
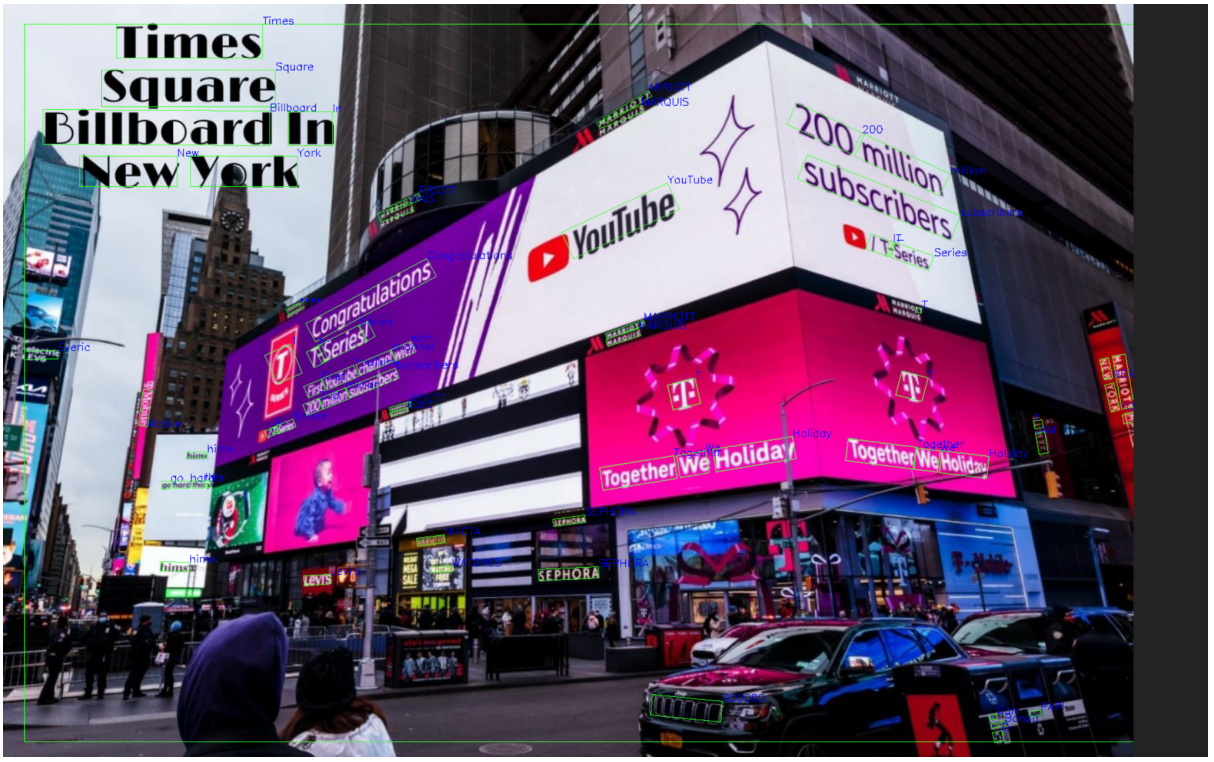


TABLE VII
EXPERIMENTAL RESULTS ON TOTAL-TEXT, CTW-1500 AND MSRA-TD500. "EXT" MEANS USING THE EXTERNAL DATASET TO PRETRAIN THE MODEL.† DENOTES THE END-TO-END SCENE TEXT SPOTTING. * DENOTES THE METHOD USING RESNET50 WITH DCN [40] AS A BACKBONE. THE BEST SCORE IS HIGHLIGHTED IN **BOLD**.

Methods	Published	Ext	Total-Text				CTW-1500				MSRA-TD500			
			R	P	F	FPS	R	P	F	FPS	R	P	F	FPS
SegLink [41]	CVPR'17	Syn	-	-	-	-	-	-	-	-	70.0	86.0	77.0	8.9
MCN [42]	CVPR'18	Syn	-	-	-	-	-	-	-	-	79	88	83	-
LSAE[31]	CVPR'19	Syn	-	-	-	-	77.8	82.7	80.1	-	81.7	84.2	82.9	-
ATTR[5]	CVPR'19	-	76.2	80.9	78.5	10.0	-	-	-	-	82.1	85.2	83.6	-
MSR[43]	IJCAI'19	Syn	73.0	85.2	78.6	-	79.0	84.1	81.5	-	76.7	87.4	81.7	-
CSE[44]	CVPR'19	MLT	79.7	81.4	80.2	0.4	76.1	78.7	77.4	0.38	-	-	-	-
TextDragon†[25]	ICCV'19	MLT+	75.7	85.6	80.3	-	82.8	84.5	83.6	-	-	-	-	-
TextField[28]	TIP'19	Syn	79.9	81.2	80.6	-	79.8	83.0	81.4	-	75.9	87.4	81.3	5.2
PSENet-1s [27]	CVPR'19	MLT	77.96	84.02	80.87	3.9	79.7	84.8	82.2	3.9	-	-	-	-
SegLink++ [23]	PR'19	Syn	80.9	82.1	81.5	-	79.8	82.8	81.3	-	-	-	-	-
LOMO[11]	CVPR'19	MLT+	79.3	87.6	83.3	-	76.5	85.7	80.8	-	-	-	-	-
CRAFT [24]	CVPR'19	MLT	79.9	87.6	83.6	-	81.1	86.0	83.5	-	78.2	88.2	82.9	8.6
DB*[4]	AAAI'20	Syn	82.5	87.1	84.7	32.0	80.2	86.9	83.4	22.0	79.2	91.5	84.9	32.0
PAN[3]	ICCV'19	Syn	81.0	89.3	85.0	**39.6**	81.2	86.4	83.7	**39.8**	83.8	84.4	84.1	30.2
TextPerception† [45]	AAAI'20	Syn	81.8	88.8	85.2	-	81.9	87.5	84.6	-	-	-	-	-
ContourNet [6]	CVPR'20	-	83.9	86.9	85.4	3.8	84.1	83.7	83.9	4.5	-	-	-	-
ABCNet† [7]	CVPR'20	MLT+	81.3	87.9	84.5	9.5	83.4	84.4	81.4	9.5	-	-	-	-
DRRG [2]	CVPR'20	MLT	84.93	86.54	85.73	-	83.02	85.93	84.45	-	82.30	88.05	85.08	-
Boundary†[46]	AAAI'20	Syn	85.0	88.9	87.0	-	-	-	-	-	-	-	-	-
TextRay [9]	MM'20	ArT−	77.9	83.5	80.6	-	80.4	82.8	81.6	-	-	-	-	-
TextFuseNet [47]	IJCAI'20	Syn	83.2	87.5	85.3	7.1	**85.0**	85.8	85.4	7.3	-	-	-	-
TextMountain[48]	PR'21	MLT	-	-	-	-	82.9	83.4	83.2	-	-	-	-	-
MOST[19]	CVPR'21	Syn	-	-	-	-	-	-	-	-	82.7	90.4	86.4	-
PCR(Res50)[10]	CVPR'21	-	80.2	86.1	83.1	-	79.8	85.3	82.4	-	77.8	87.6	82.4	-
PCR(DLA34)[10]	CVPR'21	MLT	82.0	88.5	85.2	-	82.3	87.2	84.7	11.8	83.5	90.8	87.0	-
FCENet [8]	CVPR'21	-	79.8	87.4	83.4	-	80.7	85.7	83.1	-	-	-	-	-
FCENet*[8]	CVPR'21	-	82.5	89.3	85.8	-	83.4	87.6	85.5	-	-	-	-	-
TextBPN[12]	ICCV'21	Syn	84.65	90.27	87.37	10.3	81.45	87.81	84.51	12.2	80.68	85.40	82.97	12.7
TextBPN[12]	ICCV'21	MLT	85.19	90.67	87.85	10.7	83.60	86.45	85.00	12.2	84.54	86.62	85.57	12.3
Ours(Res18-4s-1024)	-	MLT	81.90	89.88	85.70	32.5	81.62	87.55	84.48	35.3	**87.46**	92.38	89.85	**38.5**
Ours(Res50-1s-1024)	-	-	85.29	89.86	87.52	12.0	81.12	88.08	84.46	14.7	81.27	88.25	84.62	15.7
Ours(Res50-1s-1024)	-	MLT	85.34	91.81	88.46	13.3	83.77	87.30	85.50	14.1	85.40	89.23	87.27	15.2
Ours(Res50-1s-1024*)	-	MLT	**87.93**	**92.44**	**90.13**	13.2	84.71	**88.34**	**86.49**	16.5	86.77	**93.69**	**90.10**	15.3

2. GCP Vision API (link)

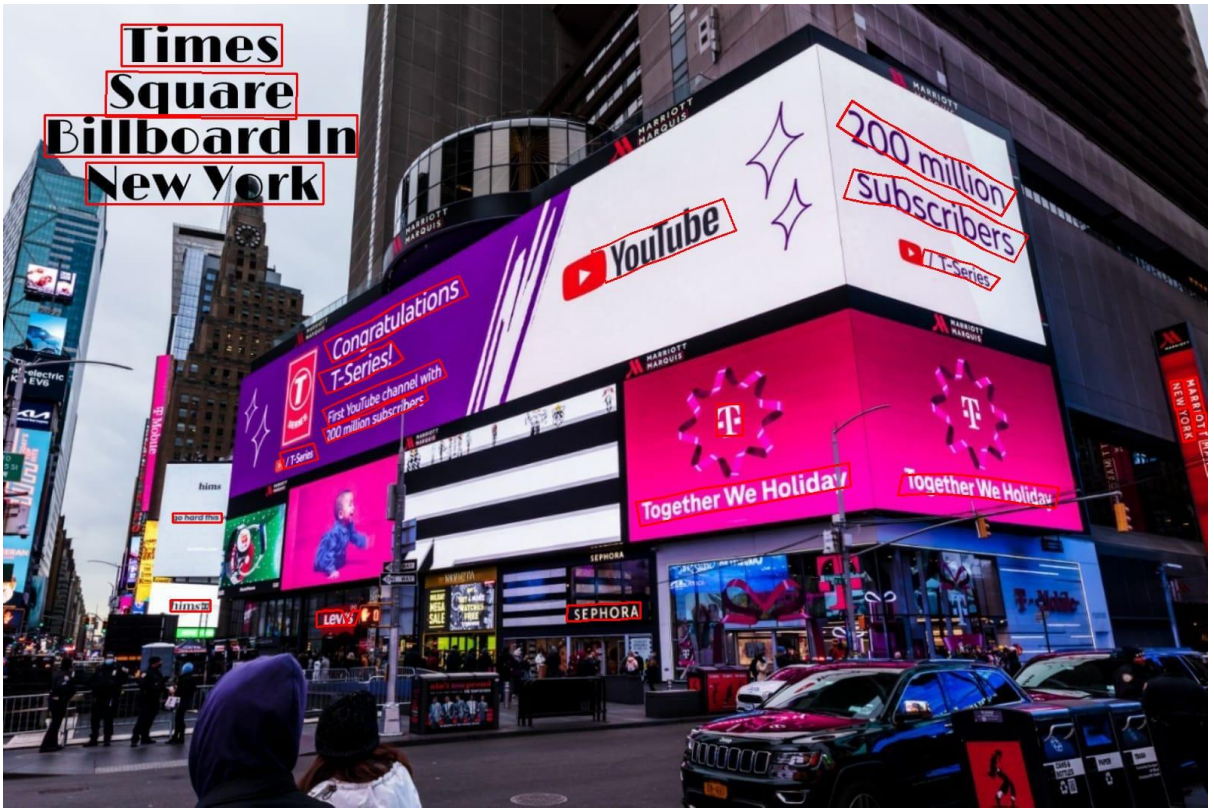


3. Unified Text Detection and Layout Analysis (2022, paper, code)
   ICDAR-17: 77.24
   Total-Text: 87.94
   CTW-1500: 85.97
   MSRA-TD500: 87.70

Method	Venue	Training Data		Word Detection						Line Detection					
		Pub	HierText	ICDAR 17 MLT			Total-Text			CTW1500			MSRA-TD500		
				P	R	F	P	R	F	P	R	F	P	R	F
CRAFT [3]	CVPR19	✓		80.6	68.2	73.9	87.6	79.9	83.6	86.0	81.1	83.5	88.2	78.2	82.9
PSENet [55]	CVPR19	✓		75.3	69.2	72.2	84.0	78.0	80.9	84.8	79.7	82.2			
FCE [64]	CVPR21	✓		-	-	-	89.3	82.5	85.8	87.6	83.4	85.5	-	-	-
MOST [18]	CVPR21	✓		**82.0**	72.0	76.7	-	-	-	-	-	-	90.4	82.7	86.4
ABPNet [61]	ICCV21	✓		-	-	-	**90.67**	85.19	87.85	87.66	80.57	83.97	86.62	84.54	85.57
CentripetalText [44]	NeurIPS21	✓		-	-	-	90.6	82.5	86.3	**88.3**	79.9	83.9	90.0	82.5	86.1
PCR [14]	CVPR21	✓					88.5	82.0	85.2	87.2	82.3	84.7	**90.8**	83.5	87.0
Ours (word)	-	✓		77.71	75.88	76.78	85.49	90.53	**87.94**	-	-	-	-	-	-
		✓	✓	78.05	**76.44**	**77.24**	84.96	**91.06**	87.90	-	-	-	-	-	-
Ours (line)	-	✓		-	-	-	-	-	-	83.92	85.87	84.88	86.59	86.81	86.69
		✓	✓	-	-	-	-	-	-	84.56	**87.44**	**85.97**	88.04	**87.44**	**87.70**

Table 5. Results of word and text line detection on public scene text datasets. Both our word and line detectors are outperforming the latest methods, even though our models are not fine-tuned for any target datasets. The proposed new dataset also proves to be a helpful complement to existing scene text datasets.

4. CRAFT (2019, paper, code)
   ICDAR-13: 95.2
   ICDAR-15: 86.9
   ICDAR-17: 73.9
   MSRA-TD500: 82.9



Method	IC13(DetEval)			IC15			IC17			MSRA-TD500			FPS
	R	P	H	R	P	H	R	P	H	R	P	H	
Zhang et al. [39]	78	88	83	43	71	54	-	-	-	67	83	74	0.48
Yao et al. [37]	80.2	88.8	84.3	58.7	72.3	64.8	-	-	-	75.3	76.5	75.9	1.61
SegLink [31]	83.0	87.7	85.3	76.8	73.1	75.0	-	-	-	70	86	77	20.6
SSTD [7]	86	89	88	73	80	77	-	-	-	-	-	-	7.7
Wordsup [10]	87.5	93.3	90.3	77.0	79.3	78.2	-	-	-	-	-	-	1.9
EAST* [40]	-	-	-	78.3	83.3	80.7	-	-	-	67.4	87.3	76.1	13.2
He et al. [9]	81	92	86	80	82	81	-	-	-	70	77	74	1.1
R2CNN [11]	82.6	93.6	87.7	79.7	85.6	82.5	-	-	-	-	-	-	0.4
TextSnake [23]	-	-	-	80.4	84.9	82.6	-	-	-	73.9	83.2	78.3	1.1
TextBoxes++* [16]	86	92	89	78.5	87.8	82.9	-	-	-	-	-	-	2.3
EAA [8]	*87*	*88*	*88*	*83*	*84*	*83*	-	-	-	-	-	-	-
Mask TextSpotter [24]	*88.1*	*94.1*	*91.0*	*81.2*	*85.8*	*83.4*	-	-	-	-	-	-	4.8
PixelLink* [3]	87.5	88.6	88.1	82.0	85.5	83.7	-	-	-	73.2	83.0	77.8	3.0
RRD* [18]	86	92	89	80.0	88.0	83.8	-	-	-	73	87	79	10
Lyu et al.* [25]	84.4	92.0	88.0	79.7	89.5	84.3	**70.6**	74.3	72.4	76.2	87.6	81.5	5.7
FOTS [20]	-	-	*87.3*	*82.0*	*88.8*	*85.3*	*57.5*	*79.5*	*66.7*	-	-	-	23.9
CRAFT(ours)	**93.1**	**97.4**	**95.2**	**84.3**	**89.8**	**86.9**	68.2	**80.6**	**73.9**	**78.2**	**88.2**	**82.9**	8.6

Table 1. Results on quadrilateral-type datasets, such as ICDAR and MSRA-TD500. * denote the results based on multi-scale tests. Methods in *italic* are results solely from the detection of end-to-end models for a fair comparison. R, P, and H refer to recall, precision and H-mean, respectively. The best score is highlighted in **bold**. FPS is for reference only because the experimental environments are different. We report the best FPSs, each of which was reported in the original paper.

5. PCR (2021, paper, code) - To test later
   CTW-1500: 84.7
   Total-Text: 85.2
   ArT: 74.0
   MSRA-TD500: 87.0

Table 4: Comparisons with related works on *CTW1500*. 'Ext' means using the external dataset to pretrain the model.'Hybrid' denotes integrating the regression and segmentation in a framework.

Type	method	Venue	Backbone	Ext	R (%)	P (%)	F (%)
Segmentation-based	PAN [44]	ICCV'19	Res18	×	77.7	84.6	81.0
	TextSnake [27]	ECCV'18	VGG16	√	**85.3**	67.9	75.6
	MSR [53]	IJCAI'19	Res50	√	78.3	85.0	81.5
	PSENet [43]	CVPR'19	Res50	√	79.7	84.8	82.2
	CRAFT [1]	CVPR'19	VGG16	√	81.1	86.0	83.5
	LOMO [58]	CVPR'19	Res50	√	69.6	**89.2**	78.4
	SAE [38]	CVPR'19	Res50	√	77.8	82.7	80.1
	PAN [44]	ICCV'19	Res18	√	81.2	86.4	83.7
	SAST [42]	MM'19	Res50	√	77.1	85.3	81.0
	TextField [51]	TIP'19	VGG16	√	79.8	83.0	81.4
	DB [17]	AAAI'20	Res50-DCN	√	80.2	86.9	83.4
	DRRGN [59]	CVPR'20	VGG16	√	83.0	85.9	84.5
	CRNet [63]	MM'20	Res50	√	80.9	87.0	83.8
Hybrid	CSE [25]	CVPR'19	Res34	×	76.0	81.1	78.4
	ContourNet [48]	CVPR'20	Res50	×	84.1	83.7	83.9
	Mask-TTD [23]	TIP'20	Res50	×	79.0	79.7	79.4
	SD [49]	ECCV'20	Res50	√	82.3	85.8	84.0
Regression-based	SLPR [64]	ICPR'18	Res50	×	70.1	80.1	74.8
	CTD-CLOC [24]	PR'19	Res50	×	69.8	77.4	73.4
	ATRR [45]	CVPR'19	SE-VGG16	×	80.2	80.1	80.1
	TextRay [39]	MM'20	Res50	×	80.4	82.8	81.6
	ICG [36]	PR'19	VGG16	√	79.8	82.8	81.3
	Our PCR	—	Res50	×	79.8	85.3	82.4
	Our PCR	—	DLA34	×	81.1	87.1	84.0
	Our PCR	—	DLA34	√	82.3	87.2	**84.7**

Table 5: Comparisons with related works on *Total-Text*. 'Ext' means using the external dataset to pretrain the model. † denotes the end-to-end scene text spotting.

Type	method	Venue	Backbone	Ext	R (%)	P (%)	F (%)
Segmentation-based	PAN [44]	ICCV'19	Res18	×	79.4	88.0	83.5
	TextSnake [27]	ECCV'18	VGG16	√	74.5	82.7	78.4
	LOMO [58]	CVPR'19	Res50	√	75.7	88.6	81.6
	PSENet [43]	CVPR'19	Res50	√	78.0	84.0	80.9
	CRAFT [1]	CVPR'19	VGG16	√	79.9	87.6	83.6
	MSR [53]	IJCAI'19	Res50	√	74.8	83.8	79.0
	PAN [44]	ICCV'19	Res18	√	81.0	**89.3**	85.0
	TextDragon [7]†	ICCV'19	VGG16	√	75.7	85.6	80.3
	SAST [42]	MM'19	Res50	√	76.9	83.8	80.2
	TextField [51]	TIP'19	VGG16	√	79.9	81.2	80.6
	DB [17]	AAAI'20	Res50-DCN	√	82.5	87.1	84.7
	CRNet [63]	MM'20	Res50	√	82.5	85.8	84.1
Hybrid	CSE [25]	CVPR'19	Res34	×	79.1	81.4	80.2
	Mask-TTD [23]	TIP'20	Res50	×	74.5	79.1	76.7
	FTSN [5]	ICPR'18	Res101	√	78.0	84.7	81.3
	Mask-TextSpotter [28] †	ECCV'18	Res50	√	55.0	69.0	61.3
	SPCNet [50]	AAAI'19	Res50	√	82.8	83.0	82.9
	Mask-TextSpotter-v2 [13] †	TPAMI'19	Res50	√	75.4	81.8	78.5
	MS-CAFA [4]	TMM'20	Res50	√	78.6	84.6	81.5
Regression-based	ATRR [45]	CVPR'19	SE-VGG16	×	76.2	80.9	78.5
	CTC-CLOC [24]	PR'19	Res50	×	71.0	74.0	73.0
	TextRay [39]	MM'20	Res50	×	77.9	83.5	80.6
	ICG [36]	PR'19	VGG16	√	80.9	82.1	81.5
	Boundary [41]†	AAAI'20	Res50	√	**83.5**	85.2	84.3
	Poly-FRCNN [2]	IJDAR'20	Inc-Res-v2	√	68.0	78.0	73.0
	MS-CAFA [4]	TMM'20	Res50	√	78.6	84.6	81.5
	Our PCR	—	Res50	×	80.2	86.1	83.1
	Our PCR	—	DLA34	×	81.5	86.4	83.9
	Our PCR	—	DLA34	√	82.0	88.5	**85.2**

Table 6: Comparisons with related works on *ArT*.

Method	Venue	Ext	R (%)	P (%)	F (%)
TextRay [44]	MM'20	√	58.6	76.0	66.2
Ours (DLA-34)	—	×	65.0	83.6	73.1
Ours (DLA-34)	—	√	**66.1**	**84.0**	**74.0**

Table 7: Comparisons with related works on *TD500*.

Type	method	Venue	Backbone	Ext	R (%)	P (%)	F (%)
Segmentation-based	EAST [62]	CVPR'17	PVANet	×	67.4	87.3	76.1
	PixelLink [6]	AAAI'18	VGG16	×	73.2	83.0	77.8
	Border [52]	ECCV'18	DesNet121	×	77.4	83.0	80.1
	TextSnake [27]	ECCV'18	VGG16	√	73.9	83.2	78.3
	MSR [53]	IJCAI'19	Res50	√	76.7	87.4	81.7
	CRAFT [1]	CVPR'19	VGG16	√	78.2	88.2	82.9
	SAE [38]	CVPR'19	Res50	√	81.7	84.2	82.9
	PAN [44]	ICCV'19	Res18	√	**83.8**	84.4	84.1
	TextField [51]	TIP'19	VGG16	√	75.9	87.4	81.3
	DB [17]	AAAI'20	Res50-DCN	√	79.2	**91.5**	84.9
	CRNet [63]	MM'20	Res50	√	82.0	86.0	84.0
	DRRGN [59]	CVPR'20	VGG16	√	82.3	88.1	85.1
Hybrid	DSRN [47]	IJCAI'19	Res50	×	71.2	87.6	78.5
	FTSN [5]	ICPR'18	Res101	√	77.1	87.6	82.0
	Corner [29]	CVPR'18	VGG16	√	76.2	87.6	81.5
	Mask-TextSpotter-v2 [13] †	TPAMI'19	Res50	√	68.6	80.8	74.2
	Mask-TextSpotter-v3 [14] †	ECCV'20	Res50	√	77.5	90.7	83.5
Regression-based	RRPN [30]	TMM'18	VGG16	×	69.0	82.0	75.0
	ATRR [45]	CVPR'19	SE-VGG16	×	82.1	85.2	83.6
	SegLink [34]	CVPR'17	VGG16	√	70.0	86.0	77.0
	RRD [18]	CVPR'18	VGG16	√	73.0	87.0	79.0
	Our PCR	—	Res50	×	77.8	87.6	82.4
	Our PCR	—	DLA34	×	79.2	90.0	84.3
	Our PCR	—	DLA34	√	83.5	90.8	**87.0**

6. GLASS (2022, paper, code) - To test later, they haven't made an inference+visualization yet
   ICDAR-15: 78.8 (this is the result using a generic lexicon)
   Total-Text: 86.2
   TextOCR: 67.1

Table 1. Results for ICDAR 2015, Total-Text and TextOCR datasets. 'S', 'W' and 'G' refer to strong, weak and generic lexicons. "None" refers to recognition without any lexicon. "Full" lexicon contains all the words in the test set. (*) refers to using specific lexicons from [20]. (†) indicates IoU of 0.1 was used instead of 0.5 during evaluation. (‡) represents results obtained using method's official source code.

Method	ICDAR 2015						Total-Text				TextOCR
	Word Spotting			End-to-End			Word Spotting		End-to-End		End-to-End
	S	W	G	S	W	G	None	Full	None	Full	
TextDragon [8]	86.2	81.6	68.0	82.5	78.3	65.2	-	-	48.8	71.8	-
ABCNet v2 [29]	-	-	-	82.7	78.5	73.0	70.4	78.1	-	-	-
MTSv3* [20]	83.1	79.1	75.1	83.3	78.1	74.2	-	-	71.2	78.4	50.8
Text Perc. [37]	84.1	79.4	67.9	80.5	76.6	65.1	69.7	78.3	-	-	-
CRAFTS [3]	-	-	-	83.1	**82.1**	74.9	-	-	**78.7**	-	-
MANGO*† [36]	85.2	81.1	74.6	**85.4**	80.1	73.9	72.9	83.6	68.9‡	78.9‡	-
YAMTS* [16]	**86.8**	82.4	76.7	85.3	79.8	74.0	-	-	71.1	78.4	-
Ours*	**86.8**	**82.5**	**78.8**	84.7	80.1	**76.3**	**79.9**	**86.2**	76.6	**83.0**	**67.1**

7. YAMTS (2021, paper, code) - OpenVino didn't work
   ICDAR-13: 95.2
   ICDAR-15: 87.0
   Total-Text: 81.5
   Open Images V5: 63.5

Method	Word Spotting			End-to-end recognition		
	S	W	G	S	W	G
TextBoxes++	**96.0**	95.0	87.0	93.0	92.0	85.0
CRAFTS (L-1280)	-	-	-	**94.2**	**93.8**	**92.2** †
MANGO* (L-1440)	92.9	92.7	88.3	93.4	92.3	88.7
YAMTS* (1280x768)	95.2	**95.0**	**93.3**	93.6	93.3	91.0
YAMTS*❖ (1280x768)	94.1	93.1	89.8	92.1	90.8	87.0

† means that generic evaluation is performed without the generic vocabulary set.
* means that the method uses the specific lexicons from Liao et al. (2019).
❖ means that the model is trained without Open Images V5 Text Annotation.

Table 3: Results on ICDAR 2013. 'S', 'W' and 'G' mean recognition with strong, weak and generic lexicon, respectively.

Method	Word Spotting			End-to-end recognition		
	S	W	G	S	W	G
TextBoxes++	76.5	69.0	54.4	73.3	65.9	51.9
Mask TextSpotter v3* (S-1440)	83.1	79.1	75.1	83.3	78.1	74.2
CRAFTS (2560x1440)	-	-	-	83.1	**82.1**	74.9†
MANGO* (L-1800)	85.2	81.1	74.6	85.4	80.1	73.9
YAMTS* (1280x768)	85.3	81.9	76.6	83.8	79.2	74.1
YAMTS* (1600x960)	**87.0**	**83.6**	**78.9**	**85.5**	80.7	**76.1**
YAMTS*❖ (1280x768)	84.4	79.9	73.4	82.8	77.2	70.5
YAMTS*❖ (1600x960)	86.8	82.4	76.7	85.3	79.8	74.0

† means that generic evaluation is performed without the generic vocabulary set.
* means that the method uses the specific lexicons from Liao et al. (2019).
❖ means that the model is trained without Open Images V5 Text Annotation.

Table 4: Results on ICDAR 2015. 'S', 'W' and 'G' mean recognition with strong, weak and generic lexicon, respectively.

Method	End-to-end recognition	
	None	Full
ABCNet (Multi-Scale)	69.5	78.4
Mask TextSpotter v3* (S-1000, Liao et al. (2020))	71.2	78.4
Mask TextSpotter v3* (S-1000, Singh et al. (2021))	74.5	81.6
MANGO (L-1600)	72.9	**83.6**
CRAFTS (L-1920)	**78.7**	-
YAMTS* (1280x768)	73.7	80.1
YAMTS*‡ (1280x768)	74.5	81.5
YAMTS*❖ (1280x768)	71.1	78.4

* means that the method uses the specific lexicons from Liao et al. (2019).
‡ means that the text recognition head is fine-tuned and the rest layers are frozen.
❖ means that the model is trained without Open Images V5 Text Annotation.

Table 5: Results on Total-Text validation. "None" refers to recognition without any lexicon. "Full" lexicon contains all words in test set.

Method	Word Spotting	End-to-end recognition
YAMTS (1280x768)	58.8	51.6
YAMTS (1600x960)	63.5	56.3

Table 6: Results on Open Images V5 validation. No lexicon is used.

8. Mask TextSpotterV3 (2020, paper, code)
   Rotated ICDAR-13: 84.2 (45° rotation), 84.7 (60° rotation)
   MSRA-TD500: 83.5
   Total-Text: 78.4
   ICDAR-15: 75.1 (with generic lexicon)

Table 1. Quantitative results on the RoIC13 dataset. The evaluation protocol is the same as the one in the IC15 dataset. The end-to-end recognition task is evaluated without lexicon. *CharNet is tested with the officially released pre-trained model; Mask TextSpotter v2 (MTS v2) is trained with the same rotation augmentation as Mask TextSpotter v3 (MTS v3). "P", "R", and "F" indicate precision, recall and F-measure. "E2E" is short for end-to-end recognition. More results are in the supplementary

Method	RoIC13 dataset (Rotation Angle: 45°)						RoIC13 dataset (Rotation Angle: 60°)					
	Detection			E2E			Detection			E2E		
	P	R	F	P	R	F	P	R	F	P	R	F
CharNet* [43]	57.8	56.6	57.2	34.2	33.5	33.9	65.5	53.3	58.8	10.3	8.4	9.3
MTS v2* [21]	64.8	59.9	62.2	66.4	45.8	54.2	70.5	61.2	65.5	68.2	48.3	56.6
MTS v3	**91.6**	**77.9**	**84.2**	**88.5**	**66.8**	**76.1**	**90.7**	**79.4**	**84.7**	**88.5**	**67.6**	**76.6**

Table 2. Quantitative detection results on the MSRA-TD500 dataset

Method	P	R	F
He et al. [14]	71	61	69
DeepReg [16]	77	70	74
RRD [25]	87	73	79
PixelLink [6]	83.0	73.2	77.8
Xue et al. [44]	83.0	77.4	80.1
CRAFT [1]	88.2	78.2	82.9
Tian et al. [38]	84.2	**81.7**	82.9
MSR [38]	87.4	76.7	81.7
DB (without DCN) [24]	86.6	77.7	81.9
Mask TextSpotter v2 [21]	80.8	68.6	74.2
Mask TextSpotter v3	**90.7**	77.5	**83.5**

Table 3. Quantitative end-to-end recognition results on the Total-Text dataset. "None" means recognition without any lexicon. "Full" lexicon contains all words in the test set. The values in the table are the F-measure. The evaluation protocols are the same as those in Mask TextSpotter v2

Method	None	Full
Mask TextSpotter v1 [30]	52.9	71.8
CharNet [43] Hourglass-57	63.6	-
Qin et al. [34] Inc-Res	63.9	-
Boundary TextSpotter [40]	65.0	76.1
ABCNet [28]	64.2	75.7
Mask TextSpotter v2 [21]	65.3	77.4
Mask TextSpotter v3	**71.2**	**78.4**

Table 4. Quantitative results on the IC15 dataset in terms of F-measure. "S", "W" and "G" mean recognition with strong, weak, and generic lexicon respectively. The values in the bracket (such as $1,600$ and $1,400$) indicate the short side of the input images. Note that in most real-world applications there are no such strong/weak lexicons with only 100/1000+ words. Thus, performance with the generic lexicon of 90k words is more meaningful

Method	Word Spotting			E2E Recognition			FPS
	S	W	G	S	W	G	
TextBoxes++ [22]	76.5	69.0	54.4	73.3	65.9	51.9	-
He *et al.* [15]	85.0	80.0	65.0	82.0	77.0	63.0	-
Mask TextSpotter v1 [30] (1600)	79.3	74.5	64.2	79.3	73.0	62.4	2.6
TextDragon [7]	**86.2**	**81.6**	68.0	82.5	**78.3**	65.2	2.6
CharNet [43] R-50	-	-	-	80.1	74.5	62.2	-
Boundary TextSpotter [40]	-	-	-	79.7	75.2	64.1	-
Mask TextSpotter v2 [21] (1600)	82.4	78.1	73.6	83.0	77.7	73.5	2.0
Mask TextSpotter v3 (1440)	83.1	79.1	**75.1**	**83.3**	78.1	**74.2**	2.5

9. MASTER (2021, paper, code) - No pre-trained model provided
   IIIT5K: 95
   SVT: 90.6
   ICDAR-03: 96.4
   ICDAR-13: 95.3
   ICDAR-15: 79.4
   SVTP: 84.5
   CUTE: 87.5

Table 2: Performance of our model and other state-of-the-art methods on public datasets. All values are reported as a percentage (%). "None" means no lexicon. * indicates using both word-level and character-level annotations to train the model. ** denotes the performance of SAR trained only on the synthetic text datasets. In each column, the best performance result is shown in **bold** font, and the second-best result is shown with an underline. Our model achieves competitive performance on most of the public datasets, and the distance between us and the first place [50] is very small on IIIT5k and SVT datasets.

Method	IIIT5K	SVT	IC03	IC13	IC15	SVTP	CUTE
	None	None	None	None	None	None	None
Jaderberg *et al.* [34]	-	80.7	93.1	90.8	-	-	-
Shi *et al.* [33]	81.9	81.9	90.1	88.6	-	71.8	59.2
STAR-Net [51]	83.3	83.6	-	89.1	-	73.5	-
Wang and Hu [52]	80.8	81.5	-	-	-	-	-
CRNN [5]	81.2	82.7	91.9	89.6	-	-	-
Focusing Attention [31]*	87.4	85.9	94.2	93.3	70.6	-	-
SqueezedText [53]*	87.0	-	-	92.9	-	-	-
Char-Net [54]*	92.0	85.5	-	91.1	74.2	78.9	-
Edit Probability [6]*	88.3	87.5	94.6	94.4	73.9	-	-
ASTER [7]	93.4	89.5	94.5	91.8	76.1	78.5	79.5
NRTR [37]	86.5	88.3	<u>95.4</u>	<u>94.7</u>	-	-	-
SAR** [8]	91.5	84.5	-	91.0	69.2	76.4	83.3
ESIR [35]	93.3	90.2	-	91.3	76.9	79.6	83.3
MORAN [36]	91.2	88.3	95.0	92.4	68.8	76.1	77.4
Wang *et al.* [39]	93.3	88.1	-	91.3	74.0	80.2	85.1
Mask TextSpotter [50]*	**95.3**	**91.8**	95.2	**95.3**	<u>78.2</u>	<u>83.6</u>	**88.5**
MASTER (Ours)	<u>95.0</u>	<u>90.6</u>	**96.4**	**95.3**	**79.4**	**84.5**	<u>87.5</u>

10. CentripetalText (2021, paper, code)
    Total-Text: 86.3
    CTW-1500: 83.9
    MSRA-TD500: 86.1

Table 3: Quantitative detection results on Total-Text and CTW1500. "P", "R" and "F" represent the precision, recall, and F-measure, respectively. "Ext." denotes external training data. * indicates the multi-scale testing is performed.

Method	Ext.	Venue	Total-Text				CTW1500			
			P	R	F	FPS	P	R	F	FPS
CTPN [34]	-	ECCV'16	-	-	-	-	60.4	53.8	56.9	7.1
SegLink [31]	-	CVPR'17	30.3	23.8	26.7	-	42.3	40.0	40.8	10.7
EAST [51]	-	CVPR'17	50.0	36.2	42.0	-	78.7	49.1	60.4	21.2
PSENet [37]	-	CVPR'19	81.8	75.1	78.3	3.9	80.6	75.6	78.0	3.9
PAN [38]	-	ICCV'19	88.0	79.4	83.5	39.6	84.6	77.7	81.0	39.8
CT-320	-	-	87.6	72.7	79.4	**93.2**	**85.7**	73.2	79.0	**107.2**
CT-512	-	-	87.9	80.8	84.2	57.0	85.2	78.4	81.7	59.8
CT-640	-	-	**88.8**	**81.4**	**84.9**	40.0	85.5	**79.2**	**82.2**	40.8
TextSnake [23]	✓	ECCV'18	82.7	74.5	78.4	-	67.9	**85.3**	75.6	-
MSR [43]	✓	IJCAI'19	83.8	74.8	79.0	-	85.0	78.3	81.5	-
SegLink++ [33]	✓	PR'19	82.1	80.9	81.5	-	82.8	79.8	81.3	-
PSENet [37]	✓	CVPR'19	84.0	78.0	80.9	3.9	84.8	79.7	82.2	3.9
SPCNet [40]	✓	AAAI'19	83.0	82.8	82.9	-	-	-	-	-
LOMO* [48]	✓	CVPR'19	87.6	79.3	83.3	-	85.7	76.5	80.8	-
CRAFT [1]	✓	CVPR'19	87.6	79.9	83.6	-	86.0	81.1	83.5	-
Boundary [36]	✓	AAAI'20	85.2	83.5	84.3	-	-	-	-	-
DB [16]	✓	AAAI'20	87.1	82.5	84.7	32.0	86.9	80.2	83.4	22.0
PAN [38]	✓	ICCV'19	89.3	81.0	85.0	39.6	86.4	81.2	83.7	39.8
DRRG [49]	✓	CVPR'20	86.5	**84.9**	85.7	-	85.9	83.0	**84.5**	-
CT-320	✓	-	88.0	75.4	81.2	**93.2**	87.7	74.7	80.7	**107.2**
CT-512	✓	-	90.2	81.5	85.6	57.0	87.8	79.0	83.2	59.8
CT-640	✓	-	**90.5**	82.5	**86.3**	40.0	**88.3**	79.9	83.9	40.8

Table 4: Quantitative detection results on MSRA-TD500. "P", "R" and "F" represent the precision, recall, and F-measure, respectively. "Ext." denotes external training data.

Method	Ext.	P	R	F	FPS
RRPN [25]	-	82.0	68.0	74.0	-
EAST [51]	-	**87.3**	67.4	76.1	13.2
PAN [38]	-	80.7	77.3	78.9	30.2
CT-736	-	87.1	**79.3**	**83.0**	**34.8**
SegLink [31]	✓	86.0	70.0	77.0	8.9
PixelLink [3]	✓	83.0	73.2	77.8	3.0
TextSnake [23]	✓	83.2	73.9	78.3	1.1
RRD [17]	✓	87.0	73.0	79.0	10.0
TextField [42]	✓	87.4	75.9	81.3	-
CRAFT [1]	✓	88.2	78.2	82.9	8.6
MCN [21]	✓	88.0	79.0	83.0	-
PAN [38]	✓	84.4	**83.8**	84.1	30.2
DB [16]	✓	**91.5**	79.2	84.9	32.0
DRRG [49]	✓	88.1	82.3	85.1	-
CT-736	✓	90.0	82.5	**86.1**	**34.8**