Self-supervised Implicit Glyph Attention for Text Recognition (Supplementary Material)

Tongkun Guan¹, Chaochen Gu^{2*}, Jingzheng Tu², Xue Yang¹, Qi Feng², Yudi Zhao², Wei Shen^{1*} ¹ MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University ² Department of Automation, Shanghai Jiao Tong University

{gtk0615,jacygu,tujingzheng,yangxue-2019-sjtu,fengqi,yudizhao,wei.shen}@sjtu.edu.cn

Our code and two large-scale contextless datasets (MPSC and ArbitText) will be released in the future: https://github.com/TongkunGuan/SIGA.

1. Further Details for Text Datasets

In this section, we present more visualizations of text datasets. As shown in Figure 1, the existing scene text recognition datasets are taken from natural scenes, including traffic signs, shopping mall trademarks, billboards, *etc*. These images have relatively clear texts with variable styles and colours against a chaotic background.

In contrast, the MPSC dataset contains many contextless texts with low visual contrast, corroded surfaces, and uneven illumination as shown in Figure 2, which poses a new challenge to contextless text recognition. Specifically, these text images are marked with Latin characters and Arabic numerals to record the serial number, production date, and other product information. Recognizing these texts plays an increasingly important role in intelligent industrial manufacturing, which is conducive to improving the assembly speed of industrial production lines and the efficiency of logistics transmission in the industrial scene. Besides, as shown in Figure 3, we employ the synthetic tool [8] by selecting the appropriate background images and various fonts and colours to generate these text images. Each text of the ArbitText dataset contains a random combination of Latin characters and Arabic numerals. The whole dataset contains 1M images, which is used to evaluate the generalizability and efficiency of language-free models on contextless texts.

2. Effectiveness of IAA Module

We measure the effects of our implicit attention alignment (IAA) module on the finely annotated dataset, TextSeg.

2.1. Metric

Let $\boldsymbol{b} \in \{0,1\}^{H \times W}$ be the character mask generated by assigning 1 to the locations in the ground-truth character



(a) CUTE80 [6]; (b) ICDAR2003 [3]; (c) ICDAR2013 [2]; (d) ICDAR2015 [1]; (e) IIIT5k [4]; (f) SVT [7]; (g) SVTP [5].



Seo, fisas	a N Rmdi	Q ^{FGI} D	intaz B	4 ce 9
Liphcz5z K	FNTEPE	PY6XS	Ezudan	2dka gjxsm
mtqvwgpej OX	n7qwezp 🔓	1LPO2	CBORKOT	1914dog
TRIXA	Xb	ODA CE	00.09	zwhojytic

Figure 3. Some examples of ArbitText dataset.

Table 1. Ablation results of different loss components.							
Loss - \mathcal{L}_{cor} \mathcal{L}_{dif} $\mathcal{L}_{cor} + \mathcal{L}_{d}$							
$\Theta\%(ACC\%)$	53.2(69.1)	55.2(69.4)	60.5(70.0)	63.6(70.5)			

box and 0 otherwise, we calculate its horizontal projections $l \in \{0,1\}^W$ by a max operation of b along with x-axis. We then assume that $\tilde{l} \in \{0,1\}^W$ denotes the thresholded network predictions (> 0.05 = 1) for the attention of corresponding character, the metric Θ is defined as: $\Theta = l \cdot \tilde{l} / || l + \tilde{l} - l \cdot \tilde{l} ||_1$. And then, we also evaluate their average recognition accuracies on the ten standard context benchmarks. Specifically, the detailed ablation results are as illustrated in Table 1.

^{*}Corresponding author.

Input image X	Text p	seudo-label S	pl	Glyp	n neeudo-lahel S	Glyph	Glyph attention S	
Sequence-aligned attention β	Text segn	nentation mas	sk S _m	Стурі	i pseudo-iabei s	_{gt} Orypn	Gryph attention S _{gam}	
Figure 4. The arrangement order.								
access access a		¢	С		access a	access	access	
access	6	\$		5	access a	access	access	
CEDELE CEDELE C		E	Ð		CEDELE	EDELE	CEDELE	
CEDELE	£	Ĺ		ε	CEDELE	EDELE	CEDELE	
JOBSITE JOBSITE J	Ō	B	8	7	OBSITE JOBS	ITE JOBSI	TE JOBSITE	
JOBSITE 1		° E		\mathcal{I}	DBSITE JOBS	ITE JOBSI	TE	
Analysis Analysis A	.n	а	l	A	alysis Analy	sis Analys	is Analysis	
Analysis y	S	Î		s Ar	alysis Analy	sis Analys	its Analysi s	
BATTERY BATTERY B	A	Л	T	8/	ATTERYBATTE	ERYBATTE	RYBATTERY	
BATTERY E	R	. Y		BA	ATTERY BATTE	RYBATTE	RY	
BRETAGNE BRETAGNE B	R	E	T	BR	ETAGNE BRETAG	GNE BRETAG	NE BRETAGNE	
BRETAGNE A	<u>.</u> G	N		BR	ETAGNE BRETAC	SNE BRETAG	NE BRETAGNE	
	A		Ļ	AVAIL	ABLEAVAILABLEAV	ALLABLEAVAL		
	•		•	Corn	ADLENVAILADLEAV	AILADLEATAT	ADLE	
	t "	0 T	ę	Conn	ector Connector Co	onnector Conn	ector	

Figure 5. Visualization results of the SIGA method on horizontal text images.

2.2. Theoretical Basis

1

Given a normalized image, let $l_t, \tilde{l_t} \in \{0, 1\}^W$ be its ground-truth horizontal projections and the thresholded network predictions for the attention at the decoding time t, we target on $\tilde{l_t} = l_t, \forall t \in \{1, ..., T\}$ to mitigate the alignment drift issue. Specifically, we propose a constraint function in implicit attention alignment module, which can be summarized as follows:

$$\sum_{\leqslant i < j \leqslant T} \tilde{l}_i \cdot \tilde{l}_j \to 0, \sum_{i=1}^T (\psi(\tilde{l}_i) \cdot \tilde{M}) \to \tilde{M}, \quad (1)$$

where $\tilde{\boldsymbol{M}} \in (0, 1)^{W \times H}$ is our network predictions for text mask and $\psi : \mathbb{R}^W \to \mathbb{R}^{W \times H}$ with a dimension expansion.

Ideally, define M as the ground-truth text mask, suppose $\tilde{M} = M$, the target $\tilde{l}_t = l_t, \forall t \in \{1, ..., T\}$ is a good feasible solution as:

$$\sum_{1 \leqslant i < j \leqslant T} \boldsymbol{l}_i \cdot \boldsymbol{l}_j = 0, \sum_{i=1}^T (\psi(\boldsymbol{l}_i) \cdot \boldsymbol{M}) = \boldsymbol{M}$$
 (2)

Although the target is a necessary but not sufficient condition for our constraint function as some extreme cases exist, the generality where the attention mechanism works in most images, ensures that SIGA can toward the target, which is also demonstrated by the above-mentioned ablation results.

3. Visualizations of Glyph Attention

In SIGA, five important items assist the text recognition network to obtain glyph features for improving performance. They are text pseudo-label S_{pl} , sequence-aligned weights β , text segmentation mask S_m , glyph pseudolabels S_{gt} , and glyph attention maps S_{gam} , respectively.

Specifically, given an input image X, SIGA first employs the K-means algorithm to generate a text pseudo-label S_{pl} , and further utilizes the text pseudo-label to optimize our designed self-supervised text segmentation module to generate a text segmentation mask S_m . Then, we follow an implicit attention method as the baseline structure to obtain implicit attention weights α , which are transformed into sequence-aligned attention vectors β by an orthogonal constraint, and served as the position information of characters in the input image X. Next, we obtain the glyph pseudo-label S_{pl} via the dot product operation between the sequence-aligned attention vectors β and the learned text segmentation mask S_m . Finally, supervised by the glyph pseudo-label S_{pl} , our text recognition network produces glyph attention maps S_{qam} .

To further illustrate the generation pipeline of glyph structures in SIGA, as shown in Figure 5-8, we visualize more examples of these items on horizontal, oriented, curved, and blurred text images. Specifically, every example follows the arrangement order in Figure 4.

MONACO MONAC	:0 ^M		0	4	MONACO	MONACO	MONACO
MONAC	:0	Ą	¢	0	MONACO	MONACO	MONACO
sehat seha	t s		e	**	sehat	sehat	sehat
seha	f _		£		sehat	sehal	
CARMEN CARME	EN C		A	R.	CARMEN	CARMEN	CARMEN
CARMI	N	м	ŧ,	N	CARMEN	CARMEN	CARMEN
ESPRIT	5		•	P	SPRIT	ESPRIT	SPRIT
ESPRI	۲	•	•	<u> </u>	SPRIT	SPRIT	SPRIT.
ENDORSES ENDORSES		N	Þ	0	NDORSES END	ORSES ENDORS	SES ENDORSES
ENDORSES	R	2	E	5 ^E	NDORSES END	ORSES ENDOR:	SES ENDORSES
BRITISH		R	1	т _В	RITISH	TISH BRITIS	BRITISH
BRITIST	ſ	5.	1-1	В	RITIST BRIT	BRITIS	572
Springer Springer S		P	T	i \$	pringer Spri	ngerSpring	gerSpringer
Springer	ŋ	-8	,e	T S1	pri n ger Spri	ngerSpring	SerSpringer
VICTORIA'S VICTORIA'S	ч	c	r	Ø VICT	ORIA'S METORIA'S	VICTORIA'S VICTO	VICIORIA'S
VICTORIAS		'	* *	VICT	ORIAS VICTORIA'S	VICTORIAS	DRIAS
Continues Continues	\$	•	4	t Coni	tinues COntinues	COMBINES CON	inues Continues
Continues		41	R 5.	con	Incer Controles	continee. cont	8

Figure 6. Visualization results of the SIGA method on oriented text images.

TERAL	TERRY	ч		É	R	TERAL TERAL TERAL
	TERRY		R	7		TERAL TERAL
ORRES	TORRES	4		0	R	ORRES ORRES ORRES
	JORRES		R	E	\$	CORRES CORRES CORRES
PHDY:	r ^{d D} Y	÷		T ²	Ð	FHDY FHDY FHDY S
	FHDY.		*	j.		FHDF'S FHDF'S
ESSIEN	ESSIEN	÷.		S	\$	ESSIEN ESSIEN ESSIEN
	ESSIEN		1	E,	N	ESSIEN ESSIEN ESSIEN
GARDEN	GARDEN	Q,		A	R	GARDEN GARDEN GARDEN
	GARDEN		D	B	14	GARDEN GARDEN GARDEN
STUBHU8 5	UBHU8 4		5	v	B 5	NBHUB STUBHUB STUBHUB STUBHUB
્ર	UBHUB	н			5	SUBHUS SUBHUS SUBHUS
Washing W	ashing 💭		'a	5	h 🕅	ashing <mark>Washing Washing Washing</mark>
W.	ashing	i.	г	n ş	W	ashing Washing Washing
OONOVAN OC	NOVAN O		0	11.	0	ONOVAN ONOVAN OONOVAN OONOVAN
	NOVAN	v	4	1	0	ONOVAL OONOVAL OONOVAL

Figure 7. Visualization results of the SIGA method on curved text images.

PLEASE P PLEASE A	۹L 5	E E	PLEASE PLEASE	PREASE PLEASE	PLEASE PLEASE
grves e		~			
JARAN A'	A N	R	<u>ک</u>	AN	
AREPO AREPO P	R O	E	AREPO AREPO	ARE O ARE O	AREPO
	, ,	۲ t		0	
celcon c	¢ D:	_i, nĭ	<u>celcon</u> celcon	celcon celcon	celcon
ARABICA A R ARABICA I	A C A	B. A		an a station - C station	8
SKINCARE S IK	R RJ	NH1 31		ARE SERVICAR	
	υ β		(° S	9	a j

Figure 8. Visualization results of the SIGA method on blurred text images.

References

- Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *ICDAR*, pages 1156–1160. IEEE, 2015. 1
- [2] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *ICDAR*, pages 1484–1493. IEEE, 2013.
- [3] Simon M Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, Robert Young, Kazuki Ashida, Hiroki Nagai, Masayuki Okamoto, Hiroaki Yamamoto, et al. Icdar 2003 robust reading competitions: entries, results, and future directions. *IJDAR*, 7:105–122, 2005. 1
- [4] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *BMVC*, pages 1–11, 2012. 1
- [5] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *ICCV*, pages 569–576, 2013. 1
- [6] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection

system for natural scene images. *Expert Systems with Applications*, 41(18):8027–8048, 2014. 1

- [7] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *ICCV*, pages 1457–1464. IEEE, 2011.
- [8] Moonbin Yim, Yoonsik Kim, Han-Cheol Cho, and Sungrae Park. Synthetic text image generator towards better text recognition models. In *ICDAR*, pages 109–124. Springer, 2021. 1