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

Tongkun Guan[1], Chaochen Gu[2*], Jingzheng Tu[2], Xue Yang[1], Qi Feng[2], Yudi Zhao[2], 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 box and 0 otherwise, we calculate its horizontal projections $\boldsymbol{l} \in \{0,1\}^{W}$ by a max operation of $\boldsymbol{b}$ along with x-axis. We then assume that $\tilde{\boldsymbol{l}} \in \{0,1\}^{W}$ denotes the thresholded network predictions ($> 0.05 = 1$) for the attention of corresponding character, the metric Θ is defined as: $\Theta = \boldsymbol{l}\cdot\tilde{\boldsymbol{l}} / \|\boldsymbol{l}+\tilde{\boldsymbol{l}}-\boldsymbol{l}\cdot\tilde{\boldsymbol{l}}\|_1$. 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.

Figure 1. Some examples of natural scene text datasets. (a) CUTE80 [6]; (b) ICDAR2003 [3]; (c) ICDAR2013 [2]; (d) ICDAR2015 [1]; (e) IIIT5k [4]; (f) SVT [7]; (g) SVTP [5].

Figure 2. Some examples of MPSC dataset.

Figure 3. Some examples of ArbitText dataset.

Table 1. Ablation results of different loss components.

Loss	-	$\mathcal{L}_{\text{cor}}$	$\mathcal{L}_{\text{dif}}$	$\mathcal{L}_{\text{cor}} + \mathcal{L}_{\text{dif}}$
Θ%(ACC%)	53.2(69.1)	55.2(69.4)	60.5(70.0)	63.6(70.5)

Input image X	Text pseudo-label S_{pl}	Glyph pseudo-label S_{gt}	Glyph attention S_{gam}
Sequence-aligned attention β	Text segmentation mask S_m		

Figure 4. The arrangement order.

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

2.2. Theoretical Basis

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

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

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

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

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

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 $\boldsymbol{S}_{pl}$, sequence-aligned weights $\boldsymbol{\beta}$, text segmentation mask $\boldsymbol{S}_m$, glyph pseudo-labels $\boldsymbol{S}_{gt}$, and glyph attention maps $\boldsymbol{S}_{gam}$, respectively.

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

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.

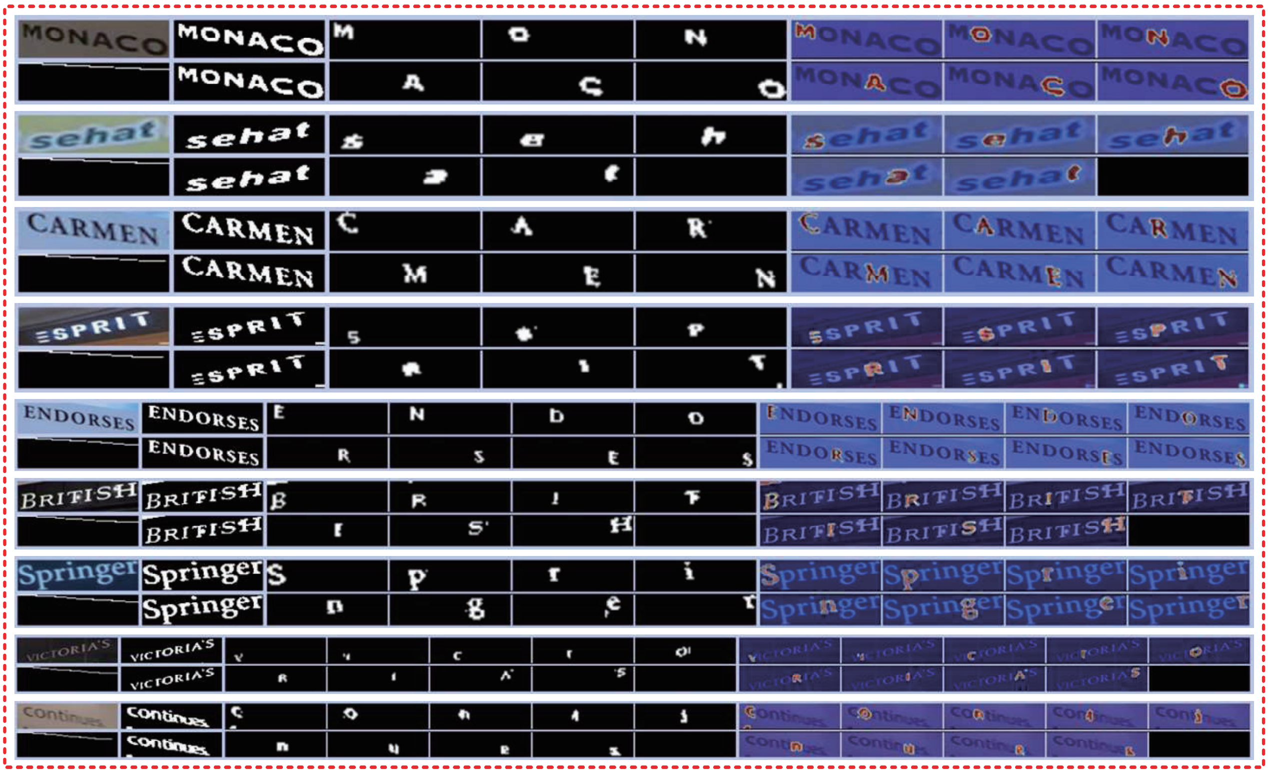

Figure 6. Visualization results of the SIGA method on oriented text images.

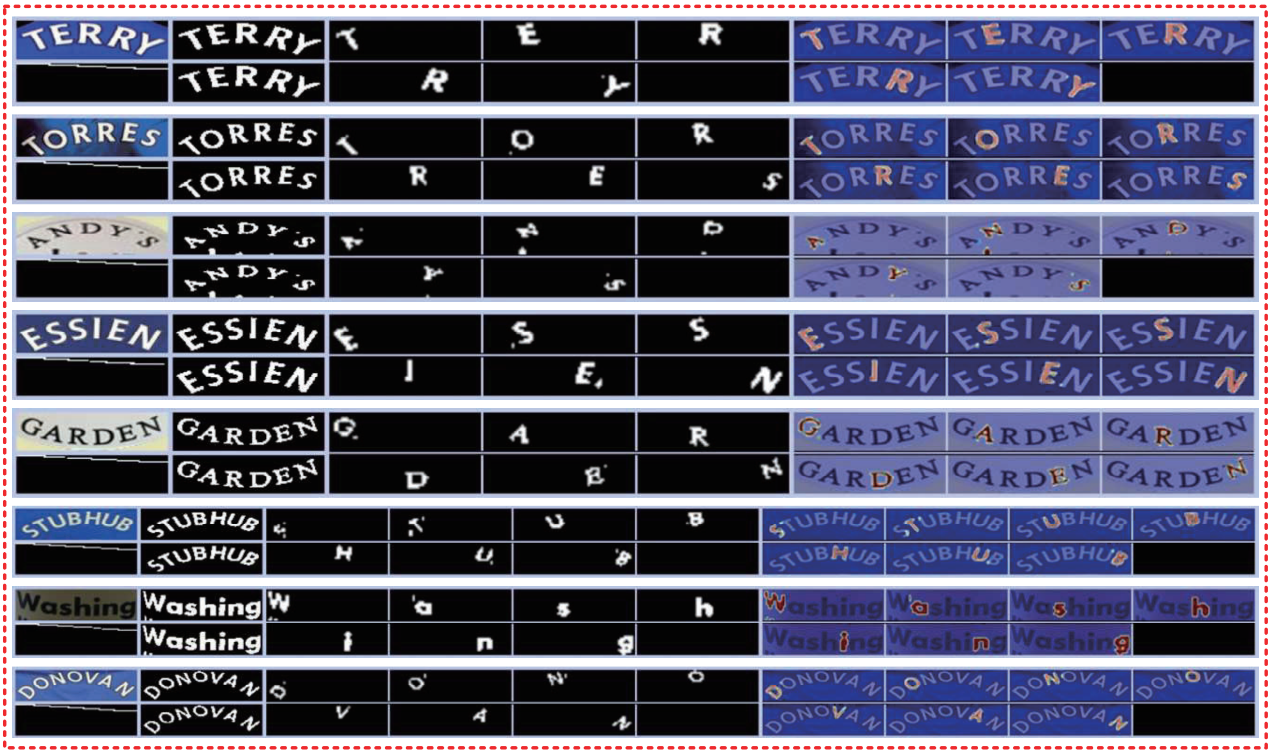

Figure 7. Visualization results of the SIGA method on curved text images.

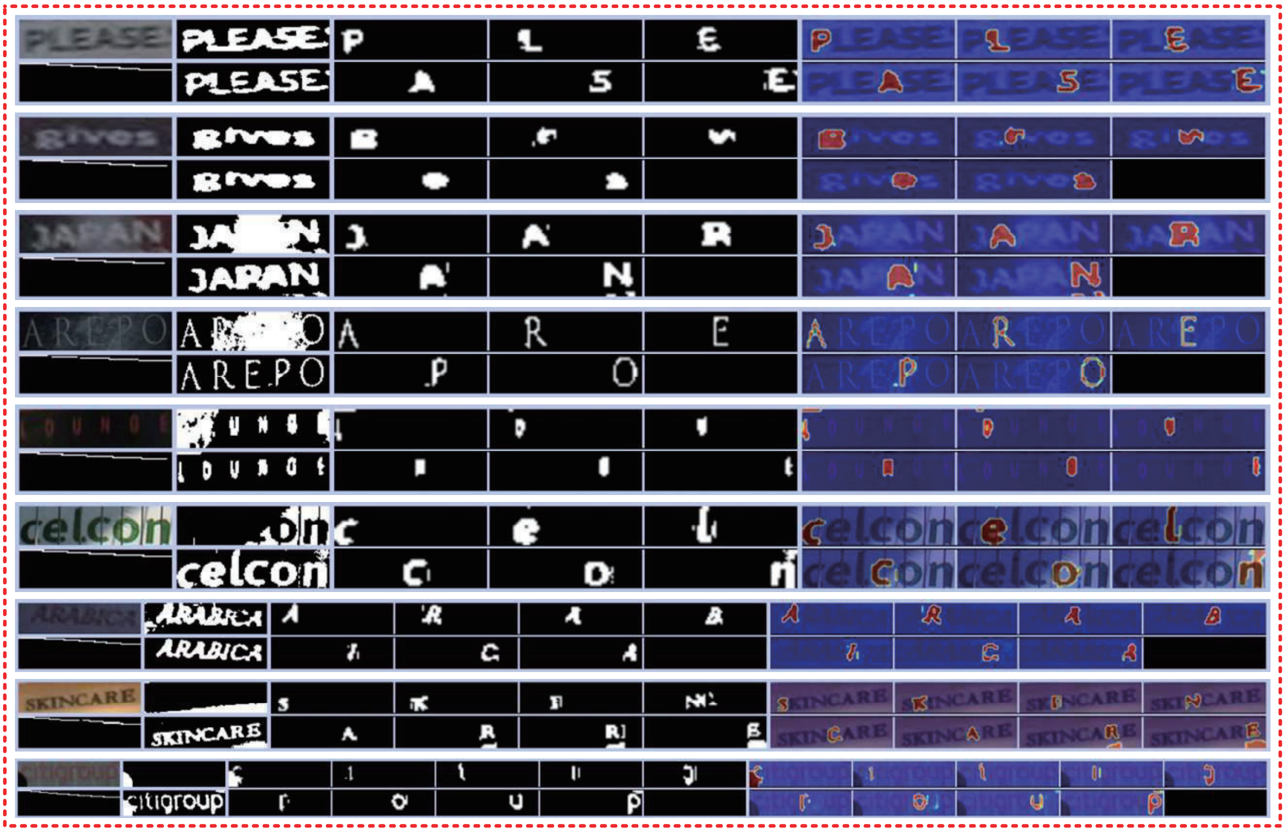

Figure 8. Visualization results of the SIGA method on blurred text images.

References

[1] 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. 1

[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. 1

[8] Moonbin Yim, Yoonsik Kim, Han-Cheol Cho, and Sungrae Park. Synthtiger: Synthetic text image generator towards better text recognition models. In *ICDAR*, pages 109–124. Springer, 2021. 1