SHANGHAI JIAO TONG UNIVERSITY

Self-supervised Implicit Glyph Attention for Text Recognition

Tongkun Guan, Chaochen Gu*, Jingzheng Tu, Xue Yang, Qi Feng, Yudi Zhao, Wei Shen*
Shanghai Jiao Tong University

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction :

A. Motivation

(a) Context text

(b) Contextless text

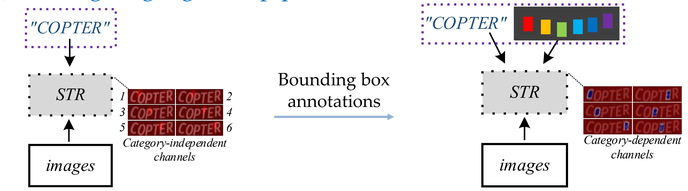
How well do recognition models generalize to arbitrary texts?

☒ language-aware methods ☑ language-free methods

B. Objective

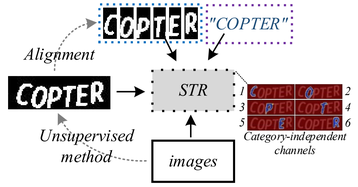
Extracting the distinctive visual features of characters on arbitrary texts.

(a) Existing language-free pipeline



(a) Implicit attention (b) Supervised attention

(b) Adding glyph attention without character-level annotations



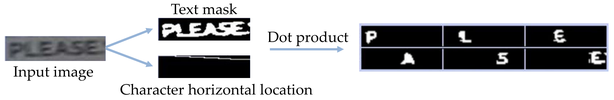
(c) Self-supervised implicit glyph attention (SIGA)

C. How to obtain glyph pseudo-labels?

☒

(a) Pixel-level annotations

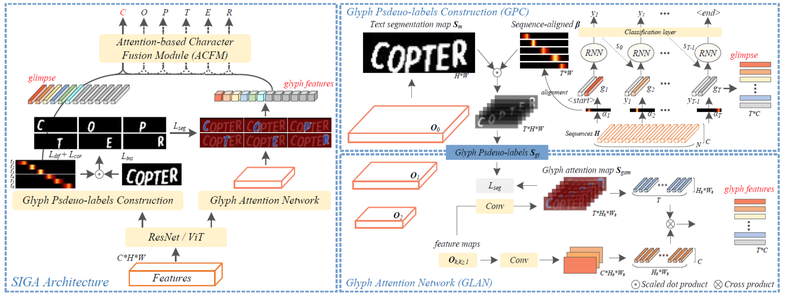
☑



(b) A simplified method to obtain the glyph structures

Methodology :

A. Our framework

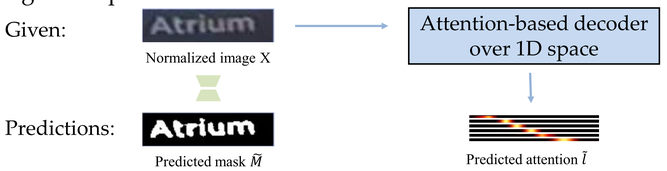


B. Details

We construct glyph pseudo-labels online by jointly:

(a) self-supervised text segmentation, producing text pseudo-labels;

(b) implicit attention alignment, producing sequence-aligned attentions serving as the position information.



I. Orthogonal constraint

$\mathcal{L}_{cor}$: $\sum_{1\le i<j\le T} \widetilde{l_i}\cdot\widetilde{l_j}\to 0$, $\mathcal{L}_{dif}$: $\sum_{i=1}^{T}(\psi(\widetilde{l_i})\cdot\widetilde{M})\to\widetilde{M}$, ψ: $\mathbb{R}^W\to\mathbb{R}^{W\times H}$

II. Theoretical basis

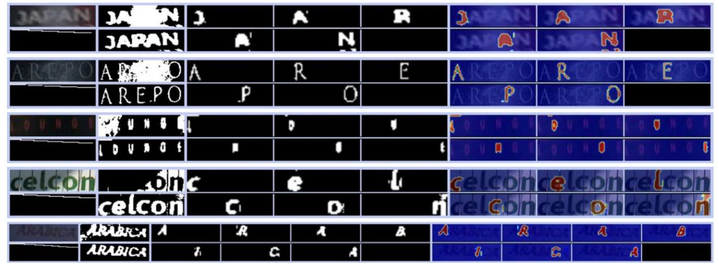
Suppose $\widetilde{M}=M$, the target $\widetilde{l_t}=l_t, \forall t\in\{1,\dots,T\}$ is a feasible solution as:

$\mathcal{L}_{cor}$: $\sum_{1\le i<j\le T} l_i\cdot l_j=0$, $\mathcal{L}_{dif}$: $\sum_{i=1}^{T}(\psi(l_i)\cdot M)=M$, ψ: $\mathbb{R}^W\to\mathbb{R}^{W\times H}$

Visualization :

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		

A. Visualization results on blurred text images



B. Visualization results on curved text images

Result :

type	Context										Contextless	
dataset	IIIT	SVT	IC03		IC13		IC15		SP	CT	MPSC	ArbitText
number	3000	647	860	867	857	1015	1811	2077	645	288	15003	1M
$SIGA_R$	95.9	92.7	96.5	95.9	97.0	95.6	85.1	81.7	87.1	91.7	65.6	66.0
$SIGA_S$	96.9	93.7	-	-	97.0	-	87.6	-	89.5	92.0	72.7	81.0
$SIGA_T$	96.6	95.1	96.9	97.0	97.8	96.8	86.6	83.0	90.5	93.1	72.0	71.7

Code :

https://github.com/TongkunGuan/SIGA