SHANGHAI JIAO TONG UNIVERSITY

Self-supervised Character-to-Character Distillation for Text Recognition

Tongkun Guan, Wei Shen*, Xue Yang, Qi Feng, Zekun Jiang, Xiaokang Yang
Shanghai Jiao Tong University

ICCV23 PARIS

Introduction :

A. Motivation

When handling complicated text images, existing supervised text recognition methods are data-hungry.

(a) WordArt (b) Occluded text (c) More scene texts

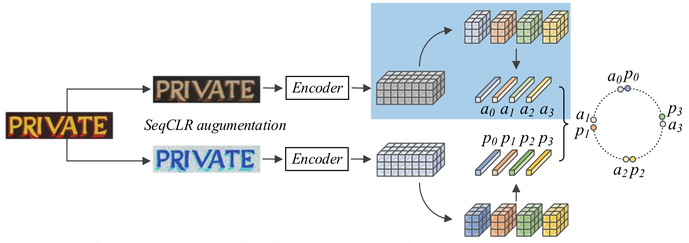
How well do recognition models generalize to arbitrary texts?

☒ supervised methods ☑ self-supervised methods

B. Objective

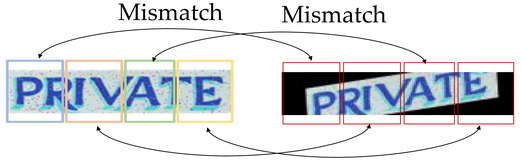
Extracting the robust visual features of characters on arbitrary texts.

Existing self-supervised pipeline



Sequence-level self-supervised learning

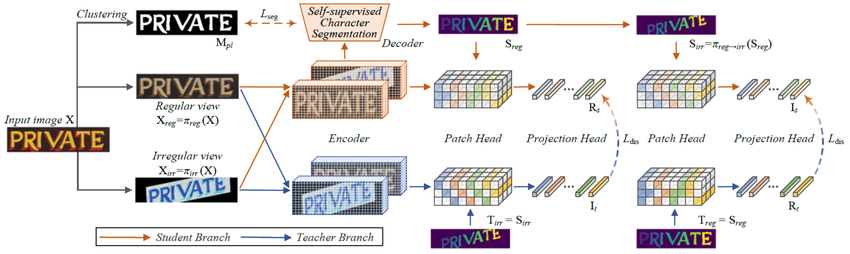
1) Inflexible data augmentation strategy, as large geometric transformations may cause inconsistency among the corresponding items



2) Neglecting character structures, which confuses networks to cause inter-character mixture

Methodology :

A. Our framework



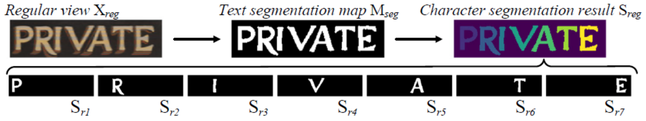
Character-level self-supervised learning

B. Details

We construct character pseudo-labels online by jointly:

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

(b) clustering-based character segmentation, producing a mask for each of its characters.



(c) Corresponding Character Regions Alignment

Problem:

$$given\ \mathbf{X}_{reg} = \pi_{reg}(\mathbf{X}), \mathbf{X}_{irr} = \pi_{irr}(\mathbf{X}),\ and\ \mathbf{S}_{reg},\ compute\ \mathbf{S}_{irr} = \pi_{reg\to irr}(\mathbf{S}_{reg}).$$

Solution:

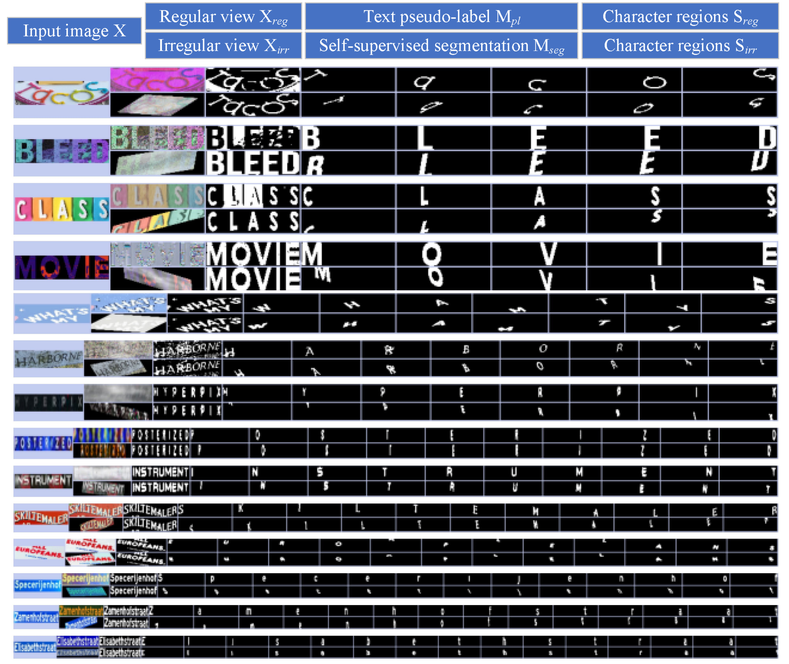
$$\pi_{reg\to irr} = \pi_{irr}(\pi_{reg}^{-1}) = \pi_{irr}$$
$$\mathbf{S}_{irr} = \pi_{irr}(\mathbf{S}_{reg})$$

(d) Character-to-character Distillation

I. Obtain character feature representations from different views and different branches (Rs, Is, Rt, and It).

II. Following DINO, establish distillation loss.

Visualization :



Result (For more results, please see our paper) :

Label Fraction	Method	IIIT	SVT	IC13	IC15	SVTP	CUTE	COCO	CTW	TT	HOST	WOST	Avg.
1%(27.8K)	DiG-ViT-Small	88.4	86.2	89.9	79.0	76.6	77.8	54.8	67.9	67.2	33.2	53.3	62.9
	CCD-ViT-Small	89.3	86.5	88.8	76.5	80.1	74.7	54.9	65.5	67.8	38.4	55.9	**63.7**
10%(278K)	DiG-ViT-Small	95.3	94.4	95.9	85.3	87.9	91.7	67.1	80.5	81.1	42.1	64.0	73.5
	CCD-ViT-Small	95.9	94.1	96.6	87.1	89.9	94.1	69.2	81.6	84.3	63.4	76.2	**78.2**
100%(2.78M)	DiG-ViT-Small	97.7	96.1	97.3	88.6	91.6	96.2	75.0	86.3	88.9	56.0	75.7	80.7
	CCD-ViT-Small	98.0	96.4	98.3	90.3	92.7	98.3	76.7	86.5	91.3	77.3	86.0	**84.9**

Code link

Blog Interpretation