

I3CL: Intra- and Inter-Instance Collaborative Learning for Arbitrary-shaped Scene Text Detection

Bo Du · Jian Ye · Jing Zhang · Juhua Liu* · Dacheng Tao

Received: date / Accepted: date

Abstract Existing methods for arbitrary-shaped text detection in natural scenes face two critical issues, *i.e.*, 1) fracture detections at the gaps in a text instance; and 2) inaccurate detections of arbitrary-shaped text instances with diverse background context. To address these issues, we propose a novel method named **Intra- and Inter-Instance Collaborative Learning** (I3CL). Specifically, to address the first issue, we design an effective convolutional module with multiple receptive fields, which is able to collaboratively learn better character and gap feature representations at local and long ranges inside a text instance. To address the second issue, we devise an instance-based transformer module to exploit the dependencies between different text instances and a global context module to exploit the semantic context from the shared background, which are able to collaboratively learn more discriminative text feature representation. In this way, I3CL can effectively exploit the intra- and inter-instance dependencies together in a unified end-to-end trainable frame-

B. Du and J. Ye are with National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, School of Computer Science and Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan, China. (e-mail: dubo@whu.edu.cn, leaf-yey@whu.edu.cn).

J. Zhang is with School of Computer Science, Faculty of Engineering, The University of Sydney, Sydney, Australia (e-mail: jing.zhang1@sydney.edu.au).

J. Liu is with Research Center for Graphic Communication, Printing and Packaging, and Institute of Artificial Intelligence, Wuhan University, Wuhan, China (e-mail: liuhuah@whu.edu.cn) (*Corresponding author*).

D. Tao is with JD Explore Academy, China and School of Computer Science, Faculty of Engineering, The University of Sydney, Sydney, Australia (e-mail: dacheng.tao@gmail.com). This work was done during Jian Ye's internship at JD Explore Academy.

work. Besides, to make full use of the unlabeled data, we design an effective semi-supervised learning method to leverage the pseudo labels via an ensemble strategy. Without bells and whistles, experimental results show that the proposed I3CL sets new state-of-the-art results on three challenging public benchmarks, *i.e.*, an F-measure of 77.5% on ArT, 86.9% on Total-Text, and 86.4% on CTW-1500. Notably, our I3CL with the ResNeSt-101 backbone ranked the 1st place on the ArT leaderboard. Code is available at github.com/ViTAE-Transformer/ViTAE-Transformer-Scene-Text-Detection.

Keywords Text Detection · Collaborative Learning · Semi-supervised Learning · Deep Learning · Transformer.

1 Introduction

As a key procedure for text reading, scene text detection has gradually become an active topic in the computer vision community due to its wide range of applications (Zhang and Tao, 2020), such as autonomous driving, scene parsing, and visual-impaired navigation. Many excellent methods have been proposed recently thanks to the success of deep learning (Chen et al., 2021; Dai et al., 2021; He et al., 2017; Liu et al., 2020a; Zhou et al., 2017; Zhu et al., 2021). However, many issues in this task remain open and challenging, such as fracture detections at the gaps in a text instance and inaccurate detections of text instances with diverse background context, due to various factors including irregular shapes, complex fonts, and variable scales.

Most of the previous methods (Liao et al., 2017; Shi et al., 2017; Zhou et al., 2017) are designed for horizontal or multi-oriented text detection and have encountered troubles in dealing with arbitrary-shaped text.

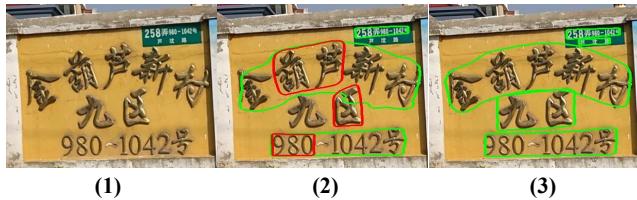


Fig. 1: From left to right, a text image, the result of Mask R-CNN, and the result of our I3CL. Existing instance segmentation-based methods suffer from fracture detections due to the gaps inside the text (the bottom digital nameplate) and inaccurate detection due to the arbitrary shapes of different instances. Our I3CL produces much better results thanks to the intra- and inter-instance collaborative learning.

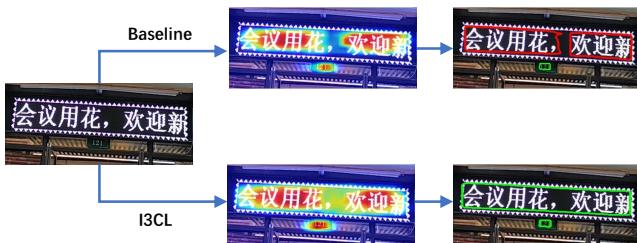


Fig. 2: Existing Mask R-CNN based methods fail to detect the gaps inside the text and produce fracture detection while our I3CL model can address this issue through collaborative learning.

Some methods propose to represent curved text with a set of characters, which is time-consuming and requires complex post-processing. In recent studies, inspired by Mask R-CNN (He et al., 2017), instance segmentation-based approach is proposed to address the problem of detecting text of various shapes. Nevertheless, simply applying Mask R-CNN to scene text detection also has some thorny problems.

As illustrated in Figure 2, one of the main problems is the fracture detection at the gaps in a text instance. When detecting text with extremely scattered and misaligned characters, the detection model may produce low text feature responses in the regions of gaps between characters because of its weak text feature representation capacity in these regions. As a result, the text detector will suffer from fracture detections. Therefore, how to learn a strong text feature representation for both characters and gaps in the text instance matters for improving the detection performance. Besides, another problem is the inaccurate detection of text instances due to diverse background, such as false positives, missed detections, as well as incomplete contours. Although existing methods learn to detect all text instances within an image through end-to-end modeling,

they treat them as individual instances during training. Consequently, existing methods have difficulties in distinguishing texts from the complex background and are prone to generate inaccurate detection results. In this paper, we argue that the text instances within an image probably have some kind of commonness. It refers to the common properties between different text instances due to similar font, color, size, and shared background context, which represent the semantic information of text instances and are completely different from the background semantic. Similar to the term of long-range dependencies between pixels within texture regions or objects in context, we term the relationship between text instances sharing the commonness as the long-range dependencies. How to exploit the dependencies between text instances and leverage the global context from the same background matters for learning a strong text feature representation.

To address these issues, we proposed a novel scene text detector based on Intra- and Inter-Instance Collaborative Learning (I3CL), which can effectively detect arbitrary-shaped scene texts. On the one hand, we first observe that the gaps in a text contain useful semantic information distinct from the background, since they are connected to the characters on both sides. We suspect that existing methods have limited performance because they are trapped by the limited receptive fields and thus have weak representation capacity for these gap regions. Based on the observation, we propose an intra-instance collaborative learning module, which treats a text as a combination of characters and gaps and learns discriminative features for them. Specifically, it consists of a cascade of three convolutional blocks, in each of which we use two convolutional layers with asymmetric horizontal and vertical convolutional kernels, and a convolutional layer with a regular convolutional kernel in parallel to them. In this way, it can model both character and gap regions in multi-oriented texts via an ensemble of paths with different receptive fields. On the other hand, to exploit the dependencies between different text instances, we propose an inter-instance collaborative learning module based on an instance-based transformer structure and a global context module, where the texture instance features are used as a token sequence to model the dependencies while the global context from the same background will be learned to supplement the above text features. By integrating these modules into a unified end-to-end trainable network, I3CL can learn a more discriminative feature representation for arbitrary-shaped scene text detection. In addition, to use unlabeled data to improve the performance, we design a simple yet effective pseudo label generation method based on an ensemble

strategy, which can mitigate the problems of missed and false detections when producing reliable pseudo labels.

The contribution of this work is four-fold. Firstly, we devise an intra-instance collaborative learning module to learn a unified feature representation for both character and gap regions in the text instance. Secondly, we devise an inter-instance collaborative learning module to exploit the dependencies between text instances within an image. Thirdly, we propose a pseudo label generation method based on an ensemble strategy to harvest the unlabeled data in a semi-supervised learning (SSL) framework. Finally, Our I3CL model outperforms existing methods and sets new state-of-the-art results on three challenging public benchmarks.

2 Related Work

2.1 Scene Text Detection

Regression-based methods follow the generic object detection framework and localize texts by directly regressing the bounding boxes of text instances. For example, EAST (Zhou et al., 2017) used efficient pixel-level regression for text objects without using the anchor mechanism and proposal generation. Based on SSD (Liu et al., 2016), TextBoxes (Liao et al., 2017) modified the aspect ratio of anchors and added a text-box layer using a horizontal convolutional kernel. Further, TextBoxes++ (Liao et al., 2018) applied quadrilaterals regression for multi-oriented text instances. SegLink (Shi et al., 2017) proposed to employ fully convolutional networks to detect text segments and model their link relationships. DGGR (Zhang et al., 2020b) first used a graph convolutional network to model relational reasoning of text components, and then grouped them into text results by linking merging. Although these methods have achieved good performance for quadrilateral text detection, most of them can not handle irregular shaped texts well due to the limited geometric representation ability.

Segmentation-based methods can accurately describe scene texts in various shapes using pixel-level segmentation masks. For example, TextSnake (Long et al., 2018) proposed a flexible and general text representation for arbitrary-shaped texts by predicting the text center line and text regions with geometry attributes. PSENet (Wang et al., 2019b) generated whole text boundary by performing progressive scale expansion of text regions using different scale kernels. Inspired by Mask R-CNN (He et al., 2017), SPCNet (Xie et al., 2019) proposed a supervised pyramid context network to detect arbitrary-shaped texts based on instance segmentation. CRAFT (Baek et al., 2019) detected the text

by clustering characters boxes according to exploring affinity between characters. For real-time detection, DB (Liao et al., 2020) designed a differentiable binarization module to perform the binarization process in a segmentation network. TextFuseNet (Ye et al., 2020) adopt a multi-path fusion architecture to fuse three levels of features for text detection. However, these methods still suffer from fracture detections and inaccurate detections. Moreover, these methods treat each text instance as an individual object for learning and training, and pays no attention to the adverse influence caused by gaps in a text, which make them suffer from fracture detections and inaccurate detections. In contrast to them, we propose a novel idea of intra- and inter-instance collaborative learning to learn better feature representation by exploiting the intra-instance characteristics and inter-instance dependencies.

2.2 Collaborative Learning

Collaborative learning (CL) has been widely used in different visual tasks. For example, Wang et al. (Wang et al., 2018) proposed a collaborative learning framework of object detection by enforcing partial feature sharing and prediction consistency to train a weakly supervised learner and a strongly supervised learner jointly. CDCL (Wang et al., 2021) presented a Cross-Dataset Collaborative Learning method to improve the generalization and discrimination of feature representations for semantic segmentation. Song et al. (Song and Chai, 2018) introduced a collaborative learning network where multiple classifier heads of the same network are simultaneously trained to improve generalization and robustness to label noise. Zhang et al. (Zhang et al., 2020c) proposed to improve text detection via collaborative training of weakly supervised text classification network and supervised text detection network. In the context of scene text detection, existing methods pay little attention to the gaps in the text and handle text instances separately, resulting in a weak text feature representation ability. By contrast, we propose a novel collaborative learning model to learn a unified feature representation for both characters and gaps in the text and exploit the dependencies between different instances, which is an instance-level collaborative learning different from the task-level collaborative learning in (Zhang et al., 2020c).

2.3 Self-training with Pseudo Labels

Self-training using pseudo labels is a learning paradigm associated with constructing models in semi-superv-

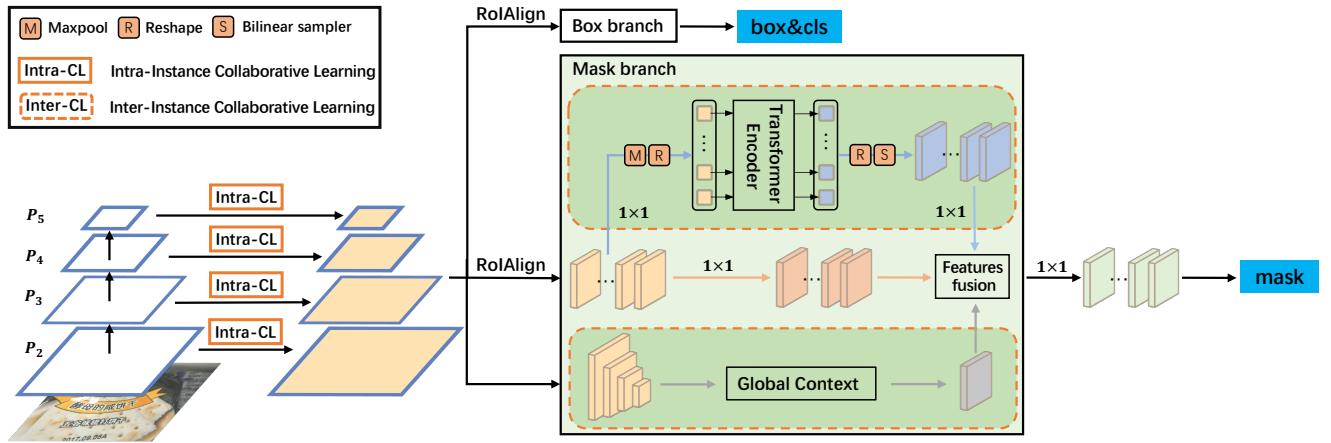


Fig. 3: The overall pipeline of the proposed I3CL. Based on the Mask R-CNN, it refines the feature map at each scale of the feature pyramid via the Intra-Instance Collaborative Learning module, and further embeds the features of text instances and global context using the Inter-Instance Collaborative Learning module in the mask branch.

ised learning, which leverages the model’s own confident predictions to produce the pseudo labels for unlabeled data (Xie et al., 2020; Zhang et al., 2021a, 2019b). Xie et al. (Xie et al., 2020) proposed a Noise Student method inspired by knowledge distillation with equal-or-larger student models. Zhang et al. (Zhang et al., 2019b) proposed to use the category centers of the source domain features as guiding anchors, which can be used to determine the active features of the target domain and generate pseudo labels for semantic segmentation. Zou et al. (Zou et al., 2019) proposed the confidence regularized self-training to avoid putting overconfident pseudo labels on wrong classes, which may lead to deviated solutions with propagated errors. Zhang et al. (Zhang et al., 2021b) proposed to use the feature distances from prototypes to estimate the likelihood of pseudo labels to facilitate online correction in the course of training. Yang et al. (Yang et al., 2021) designed multiple detection heads that predict pseudo labels for each other to provide complementary information. Unlike the above methods that may suffer from the erroneous pseudo labels from a single model, we proposed a new pseudo labeling method based on an ensemble strategy to produce reliable pseudo labels for text detection.

3 Methodology

3.1 Overview

The overall framework of the proposed I3CL model is illustrated in Figure 3. As shown, the basis of I3CL pipeline is built upon the Mask R-CNN framework. Firstly, the input text image is fed into the backbone network with an FPN (Lin et al., 2017) architecture to

generate a multi-scale feature pyramid, denoted as $\{P_2, P_3, P_4, P_5\}$, which have the same size as the input image with the down-sampling factors of $\{4, 8, 16, 32\}$, respectively. Secondly, the Intra-Instance Collaborative Learning (Intra-CL) module is used to further refine the text fracture of both characters and gaps implicitly on the feature maps at each scale of the feature pyramid. The detailed network structure of Intra-CL will be presented in Section 3.2. Next, we use a region proposal network (RPN) to produce text proposals for subsequent procedures. After that, box regression and mask prediction are carried out in two parallel branches. The box branch further refines and classifies text proposals. In the mask branch, we devise an Inter-Instance Collaborative Learning (Inter-CL) module to perform collaborative learning among all positive text instances. The text features from the Inter-CL module and the original ROIAlign module are fused into more discriminative features and used in instance segmentation to generate precise text contours. The detailed network structure of Inter-CL will be described in Section 3.3. Besides, a pseudo label generation method based on ensemble strategy for semi-supervised learning will be introduced in Section 3.4 in detail.

3.2 Intra-Instance Collaborative Learning

Existing methods focus on learning discriminative features for the character regions in the text instance while paying little attention to the gap regions between the characters, which may result in fracture detections (*i.e.*, false positives detection) at the gaps due to a weak feature representation ability. In this paper, we treat the text instance as a spaced combination of both

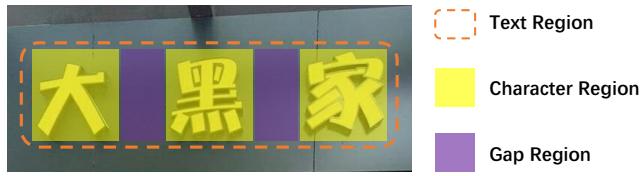


Fig. 4: The text region consists of spaced character regions and gap regions. Compared with the background, the gap regions are surrounded by characters in both sides, and contain rich text-related information in a long range. Therefore, it requires exploiting long-range dependencies between characters, between gaps, as well as between characters and gaps to learn a complete representation for the whole text instance.

characters and gaps, as illustrated in Figure 4. In other words, the characters are spaced by the gaps while the gaps are also surrounded by characters on both sides, indicating that there are long-range dependencies between characters, between gaps, as well as between characters and gaps. To exploit the dependencies and learn a unified discriminative feature representation for both characters and gaps, we propose the Intra-CL module consists of a cascade of three convolutional blocks with multiple receptive fields.

As shown in Figure 5, the Intra-CL module is composed of a cascade of three convolutional blocks. Each block contains three parallel convolutional layers with asymmetric horizontal and vertical convolutional kernels, *i.e.*, $k \times 1$ and $1 \times k$, as well as a regular $k \times k$ convolutional kernel. In our work, k is set to 7, 5, and 3, respectively. The features from the three layers are summed and fed into the subsequent block. In this way, the Intra-CL module indeed contains an ensemble of paths with multiple receptive fields. We also add a residual connection between the input feature and the fused feature of the last block since it has been proved that learning with residual connections is much easier and converges faster. It is noteworthy that we use asymmetric kernels to enable the Intra-CL module adapt to multi-oriented texts. Besides, we employ a large kernel at the first block and smaller ones in the subsequent blocks because the Intra-CL module is expected to learn long-range dependencies between characters and gaps at first and then gradually focus on the central region of either the character or gap to learn a discriminative feature representation. An ablation study of the design of the Intra-CL module is conducted in Section 4.3.

Unlike the inception-like modules in IncepText (Yang et al., 2018) enlarging received field for horizontal text detection by stacking separable convolutions ($1 \times k$ and $k \times 1$) sequentially and convolution layers with different

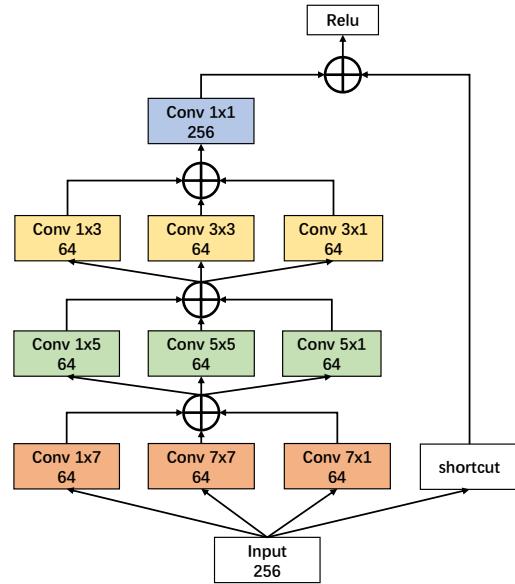


Fig. 5: The architecture of the Intra-Instance Collaborative Learning module. It consists of a cascade of three convolutional blocks, each of which contains two convolutional layers with asymmetric horizontal and vertical convolutional kernels, *i.e.*, $k \times 1$ and $1 \times k$, and a convolutional layer with a regular $k \times k$ convolutional kernel in parallel to them. In this way, it can model both character and gap regions in multi-oriented texts via an ensemble of paths with different receptive fields.

kernels in parallel, I3CL aims to learn text representations in longer ranges for arbitrary-shaped text detection. To this end, first, we utilize separable convolutions in parallel and stack convolution layers with different kernels sequentially. Second, we use large kernels at the shallow layers and small kernels at the deep layers to make the network gradually focus on the central region of either the character or gap to learn a complete text representation. Third, the parallel and serial structure in Intra-CL implies an ensemble of $3 \times 3 \times 3 = 27$ paths, each of which corresponds to a unique combination of convolution kernels and has a specific receptive field.

By deploying the proposed Intra-CL module at each scale of the feature pyramid, our detection model can exploit the long-range dependencies between characters and gaps in text region through the information flows among different paths and implicitly learn a unified feature representation for both characters and gaps, therefore effectively mitigating the fracture detection issue due to the gaps in a text instance.

3.3 Inter-Instance Collaborative Learning

Following Mask R-CNN (He et al., 2017), we use the RoIAlign to extract the RoI features of size $H \times W \times C$ from the multi-scale feature pyramid for M positives proposals, which will be fed into the box branch and mask branch, respectively. To model the dependencies between text instances, we applied the transformer structure in Inter-CL module as shown in Figure 3. Firstly, the M RoI features are fed into a 1×1 convolution layer to reduce their channel dimension from C to C_0 . Then, their spatial resolution is also reduced from $H \times W$ to $h \times w$ by using Adaptive Max-Pooling. Next, we flatten each feature into a vector of size $1 \times (h \times w \times C_0)$. In this way, we obtain a sequence of M token features (denoting as q), whose feature dimension has been significantly reduced. The sequence q is fed into a transformer encoder, which has three regular encoder layers with four heads of self-attention layers. And the output feature dimension of the feed-forward network in the transformer is $h \times w \times C_0$. Via the multi-head attention module, the long-range dependencies between different text instances in an image can be captured by adaptively attending to specific text instances that have similar background context or font appearance for any text instance. In this collaborative learning way, the representation ability of learned features can be improved. Afterward, the sequence q will be reshaped to a set of enhanced 2D visual features of size $h \times w \times C_0$, which will be upsampled using bilinear interpolation and transformed using a 1×1 convolution layer to recover the feature dimension as $H \times W \times C$. In this paper, the typical setting of the aforementioned parameters are $C=256$, $C_0=32$, $H=W=14$, $h=w=3$, and M is the number of positive instance proposals. The whole process can be described as follows:

$$q = \text{Reshape}(\text{AdaptiveMaxpool}(\text{Conv}_{1 \times 1}(f))), \quad (1)$$

$$q^{TE} = \text{TransformerEncoder}(q), \quad (2)$$

$$q^* = \text{Conv}_{1 \times 1}(\text{BilinearInterp}(\text{Reshape}(q^{TE}))), \quad (3)$$

where f denotes the RoI features of M text instances, q^{TE} denotes the learned features by the transformer encoder, and q^* is the recovered 2D visual features.

Existing methods typically detect texts according to the RoI feature that hardly pay attention to the global context from shared background for text instances within an image, which may tend to produce inaccurate detection results. We introduce two different structure designs in Inter-CL module to extract global context, which can be seen in Figure 7. 1) The first structure is built upon a pixel-based transformer. After obtaining the unified representation from all the levels of the

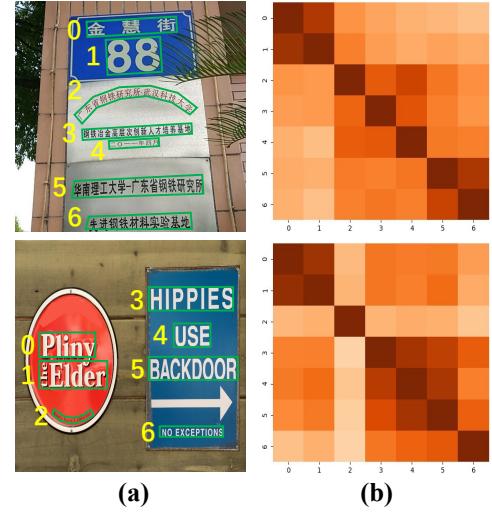


Fig. 6: (a) Different text instances on an image. (b) Attention map of the dependencies between text instances. The darker the color, the closer the dependency between two instances. As shown, the dependencies between different text instances are influenced by background context, font, and scale, etc.

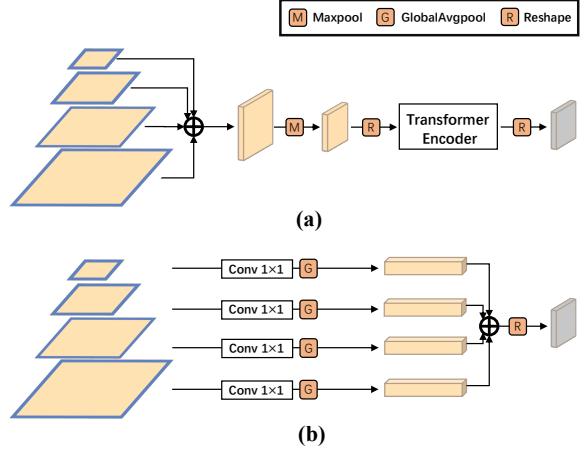


Fig. 7: (a) Global context extraction structure based on a Transformer encoder. (b) Global context extraction structure based on global average pooling.

feature pyramid like in TextFuseNet (Ye et al., 2020), we flatten the feature maps into a sequence of tokens, where each token is a feature vector at a specific pixel position on the feature maps as shown in Figure 7(a). In this way, we extract the global context by modeling the long-range dependencies between different pixels on the feature maps. 2) For the second structure shown in Figure 7(b), each level of feature pyramid is aggregated into a global context vector by a 1×1 convolution layer and global average pooling, and then we fuse global context from different scales through element-

wise summation. The global context will be fused with the original RoI features and the enhanced RoI features from the transformer encoder via element-wise summation to generate the discriminative text representation for text instance segmentation as shown in Figure 3.

As we have discussed, the text instances within an image probably have strong dependencies since they may share a same background or have a same font style, color, and scale, as illustrated in Figure 6(a). Based on the Inter-CL module, we can effectively model dependencies between text instances via the self-attention mechanism as demonstrated in Figure 6(b), which is beneficial for learning discriminative feature representation. It is noteworthy that we have not utilized the Inter-CL module in the box branch for the following three reasons. First, there are both positive and negative text samples in the box branch, while we only need to model the dependencies between different positive text instances rather than those negative ones, which are primarily used for training the classifier. Second, there are a lot of negative samples which will result in a long sequence and a bulk of computations if we directly apply the Inter-CL module based on them. Third, since we derive the detection results from the predicted masks rather than the quadrilateral bounding boxes for the arbitrary-shaped scene texts, therefore we only deploy the Inter-CL module in the mask branch.

3.4 Semi-supervised Learning

SSL has been widely applied in various deep learning tasks, which can effectively use unlabeled data to improve performance. Among them, self-training based on pseudo labels is one of the most common methods in SSL. However, existing methods obtain pseudo labels from the detection results only via a customized confidence threshold, ignoring the errors of missed and false detections in object regression. To mitigate the side effect of the problem, we propose a more reliable pseudo label generation method as described below.

Specifically, we first train three teacher models A , B , and C with different data augmentations on labeled data, by which these models will focus on respective corresponding scenes and learn different text representations. Second, the three models perform multi-scale testing on unlabeled data to avoid missed detections as much as possible. Third, reliable pseudo labels for unlabeled data will be generated from the three sets of detection results through an ensemble strategy, which is described in Algorithm 1.

As shown, Det_A , Det_B , and Det_C denote the detection result sets of model A , B , and C respectively. We define a text instance as a triplet of (m, b, s) , which

Algorithm 1: Pseudo-label Generation

```

Input:  $Det_A = \{text_i = (m_i, b_i, s_i)\}_{i=0}^I$ ,
         $Det_B = \{text_j = (m_j, b_j, s_j)\}_{j=0}^J$ ,
         $Det_C = \{text_k = (m_k, b_k, s_k)\}_{k=0}^K$ 
Output:  $L_p = \{text_q = (m_q, b_q, w_q)\}_{q=0}^Q$ 
begin
     $L_p = \{\}$ ;
    for  $text_i \in Det_A$  do
        if  $\exists text_j \in Det_B$  and  $text_k \in Det_C$ ,
             $iou_{ij} > T$  and  $iou_{ik} > T$  then
                 $m_q = \text{Overlap-Mask}(m_i, m_j, m_k)$ 
                 $b_q = \text{Soft-Box}(b_i, b_j, b_k)$ 
                 $w_q = s_i * s_j * s_k$ 
                 $L_p = L_p \cup (m_q, b_q, w_q)$ 
            end
        else if  $\exists text_j \in Det_B$  and  $iou_{ij} > T$  then
             $m_q = \text{Overlap-Mask}(m_i, m_j)$ 
             $b_q = \text{Soft-Box}(b_i, b_j)$ 
             $w_q = s_i * s_j * \alpha$ 
             $L_p = L_p \cup (m_q, b_q, w_q)$ 
        end
        else if  $\exists text_j \in Det_C$  and  $iou_{ik} > T$  then
             $m_q = \text{Overlap-Mask}(m_i, m_k)$ 
             $b_q = \text{Soft-Box}(b_i, b_k)$ 
             $w_q = s_i * s_k * \alpha$ 
             $L_p = L_p \cup (m_q, b_q, w_q)$ 
        end
        else
            | continue
        end
    end

```

are the mask, bounding box, and score of the text. L_p represents the final pseudo labels set including triplets of (m, b, w) , in which w is the corresponding loss weight of proposals matched with this pseudo label during the training. For each text instance in Det_A , we retrieve the presence of text instances with high similarity in Det_B and Det_C . If similar text instances appear in both Det_B and Det_C , we consider these instances to be highly reliable, and then fuse them into a pseudo label with the multiplication of scores as w . If similar text instances exist only in Det_B or Det_C , we consider these text instances to be weakly reliable and the w of fused text instance will be decayed. In contrast, the text instance will be ignored when there is no similar text instance in Det_B or Det_C . We calculate the Intersection over Union (IoU) score to evaluate the similarity between text instances. T and α represent the IoU threshold and decay weight, and are set to 0.8 and 0.5 respectively. Soft-Box means the process of boxes merging by calculating the average of coordinates of boxes as the new coordinates of the pseudo labeled box. Meanwhile, we obtain the overlap of the text masks as the pseudo labeled mask in Overlap-Mask process. Finally, we train a student

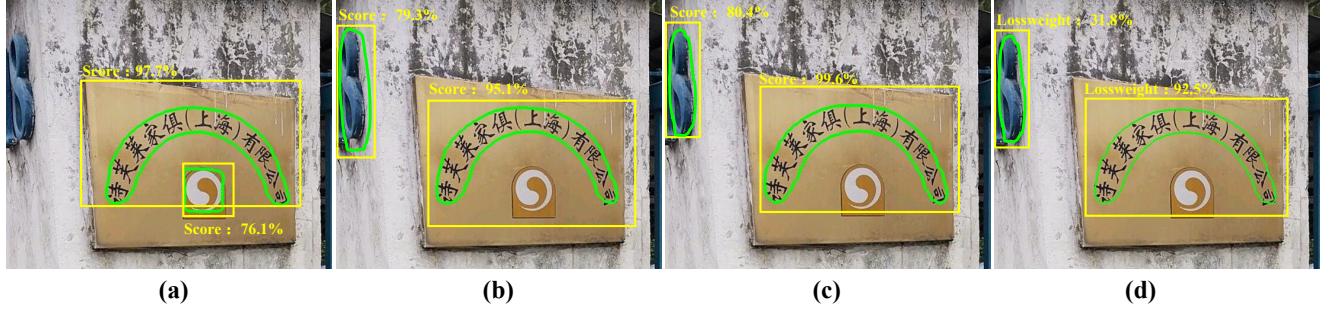


Fig. 8: (a)-(c) are the detection results from different teacher models, and (d) is the generated reliable pseudo label based on above results.

model on the combination of labeled data and pseudo labeled data.

Producing pseudo labels through multiple models mitigates the issue of missed detections caused by a single model with insufficient detection capability. Moreover, a pseudo label is jointly determined by multiple models, minimizing the problem of false positives and generating more accurate pseudo labels. The visualization example of the proposed pseudo labels generation method is shown in Figure 8.

3.5 Loss Function

Following Mask R-CNN, our model is trained in a multi-task manner, where a classification task, a box regression task, and an instance segmentation task are involved. Specifically, the final loss function is defined as follows:

$$L = L_{rpn} + L_{box} + L_{mask} \quad (4)$$

where L_{rpn} and L_{box} denote the loss functions in the RPN and box branch, both of which consist of a Cross-Entropy (*Binary* or *Softmax*) loss L_{cls} and a Smooth L1 loss L_{reg} for classification and box regression, respectively. L_{mask} denotes the loss function in the mask branch, which is a Binary Cross-Entropy loss.

4 Experiments

In this section, we evaluate the performance of I3CL model on three public benchmarks, *i.e.*, ArT (Chng et al., 2019), Total-Text (Ch'ng and Chan, 2017), and CTW-1500 (Yuliang et al., 2017), in which horizontal, quadrilateral, and curved texts exist simultaneously in most of the images. We first conduct comprehensive ablation studies to verify the effectiveness of proposed modules, and then compare I3CL with state-of-the-art methods.

4.1 Datasets and Evaluation Metrics

SynthText (Gupta et al., 2016) is a dataset consisting of 800k synthetic images and 8 million text instances. We use it to pre-train our I3CL model.

ArT (Chng et al., 2019) is newly released dataset in the ICDAR2019 Robust Reading Challenge on Arbitrary-Shaped Text. It is the most challenging arbitrary-shaped text dataset containing Chinese texts, English texts, and other mixed symbols. It has a total of 10,166 images, including 5,603 training images and 4,563 testing images. Text instances in ArT are labeled by polygons with adaptive number of key points.

Total-Text (Ch'ng and Chan, 2017) is a dataset that includes English texts of various shapes. It contains 1,255 images for training and 300 images for testing. All text instances are labeled by word-level polygons.

CTW-1500 (Yuliang et al., 2017) is an English dataset focusing on curved texts, which consists of 1,000 training images and 500 testing images. Different from Total-Text, text instances in CTW-1500 are labeled by text-line-level polygons.

We follow the same standard evaluation protocols by using Recall, Precision, and F-measure as the evaluation metrics, which are provided by the dataset creators or competition organizers. F-measure is the major evaluation metric.

4.2 Implementation Details

We implement our proposed I3CL model based on the Detectron2 codebase with PyTorch. All experiments are performed using Nvidia Tesla V100 (16G) GPUs. The model is trained on 4 GPUs and tested on 1 GPU. As in the most of previous methods, we choose the ResNet-50 with the FPN as the backbone encoder.

Training. The whole training can be roughly divided into three main stages. Firstly, we pre-train a

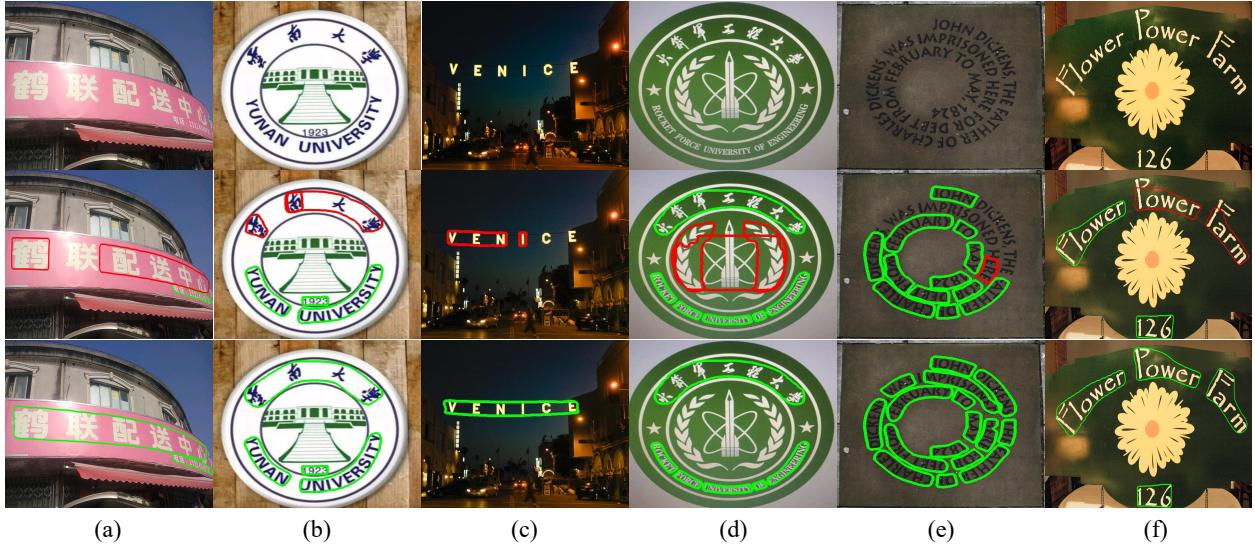


Fig. 9: Detection results of Mask R-CNN (second row) and our I3CL model (third row). Mask R-CNN produces fracture detections(a-c), and inaccurate detections such as false positives(d), missed detections(e), as well as incomplete text contours(f), while our I3CL model can effectively mitigate these issues and generate more accurate detection results.

base model on SynthText dataset for about 300k iterations. Secondly, for each benchmark dataset, we fine-tune the base model using the corresponding real-world images for 30 epochs. In particular, considering that SynthText only contains English texts, we further pre-train the base model on the LSVT (Sun et al., 2019) and ICDAR2019-MLT (Nayef et al., 2019) datasets before fine-tuning on ArT to enhance the ability of the model regarding Chinese, following the common practice in (Baek et al., 2020). Finally, based on the proposed pseudo-label generation method, we choose the test data of LSVT as the unlabeled data for ArT, and make Total-Text and CTW-1500 act as unlabeled data for each other to further fine-tune the model in a semi-supervised learning way.

The batch size during pre-training is set to 8, and reduced to 4 during fine-tuning. We adopt the Stochastic gradient descent (SGD) optimizer to optimize our network with a weight decay of 0.0001 and a momentum of 0.9. During pre-training, the initial learning rate is set to 0.001 in the first 100k iterations and divided by 10 for the remaining iterations. For all experiments during fine-tuning, the initial learning rate is set to 0.0005 in the first 10 epochs and divided by 10 at 20 and 30 epochs. The shorter sides of images are randomly resized to different scales (*i.e.*, 800, 1,000, 1,200), and the upper limit of longer side is 1,800. Data augmentation strategies such as random noise, brightness adjusting, and color change are applied to increase the diversity of training data.

Inference. During inference, we only perform a single-scale test to evaluate our model. The shorter side of the test images was scaled to 1,000 while keeping the aspect ratio unchanged, and the maximum size of the longer side is limited to 2,100, 1,875, and 1,800 on three datasets. The function of *findContours* in *OpenCV* is used to generate polygon contours of text instances from predicted masks as the final detection results.

4.3 Ablation Study

Effectiveness of each module. We conduct an ablation study on ArT, Total-Text, and CTW-1500 to verify the effectiveness of each proposed module in this paper. For each dataset, we trained four models by adding the proposed modules gradually. “**Baseline**” denotes the original Mask R-CNN baseline model. “**Intra-CL**” denotes the model using the Intra-Instance Collaborative Learning module. “**Inter-CL(GCT)**” and “**Inter-CL(GCG)**” denote the model using the Inter-Instance Collaborative Learning module with the global context structure based on transformer or global average pooling, respectively. “**SSL**” refers to that I3CL model is trained in the semi-supervised learning way. The results are summarized in Table 1.

As can be seen, **Intra-CL** improves the performance of the baseline model consistently on all three datasets, *e.g.*, 1.0%, 1.4%, and 1.3% gains in terms of the F-measure on ArT, Total-Text, and CTW-1500, respectively. In addition, integrating it with **Inter-CL(GCT)**

Table 1: Ablation study of the key components in our I3CL model on different datasets. “Intra-CL” represents the Intra-Instance Collaborative Learning module. “Inter-CL(GCT)” and “Inter-CL(GCG)” refer to Inter-Instance Collaborative Learning module with the global context structure based on transformer or global average pooling, respectively. “R”, “P”, and “F” represent Recall, Precision, and F-measure, respectively.

Method	ArT			Total-Text			CTW-1500			Parameter	GFLOPs
	R	P	F	R	P	F	R	P	F		
Baseline	68.9	80.0	74.0	80.1	86.8	83.3	81.3	84.4	82.8	44.3M	204.8
Intra-CL	69.4	81.7	75.0	82.2	87.5	84.7	82.8	85.5	84.1	46.7M	245.6
Intra-CL + Inter-CL(GCT)	70.9	82.8	76.4	83.4	88.8	86.0	84.4	87.3	85.8	58.5M	248.8
Intra-CL + Inter-CL(GCG)	71.3	82.7	76.6	83.7	89.2	86.3	84.5	87.4	85.9	52.2M	247.3
Intra-CL + Inter-CL(GCG&GCT)	71.4	82.3	76.5	83.4	89.4	86.3	84.4	87.6	86.0	59.0M	250.1
Intra-CL + Inter-CL(GCG) + SSL	72.2	83.6	77.5	84.2	89.8	86.9	84.6	88.4	86.5	52.2M	247.3

Table 2: Comparison results between the Mask R-CNN baseline and I3CL on three difficult subsets.

Method	ArT subset			Total-Text subset			CTW-1500 subset		
	R	P	F	R	P	F	R	P	F
Baseline	61.6	52.3	56.6	70.6	84.8	77.1	74.0	75.0	74.5
I3CL	68.8(↑7.2)	61.2(↑8.9)	64.8(↑8.2)	80.0(↑9.4)	89.8(↑5.0)	84.6(↑7.5)	82.1(↑8.1)	82.6(↑7.6)	82.4(↑7.9)

further brings absolute performance gains in terms of the F-measure increase by 1.4%, 1.3%, and 1.7%, respectively. By contrast, the combination of **Intra-CL** and **Inter-CL(GCG)** achieves a better gain of 1.6%, 1.6%, and 1.8% on F-measure but contains fewer parameters. In terms of F-measure, there is no significant gap between the two modules. The reason why GCG module works slightly better may be that each pixel of text RoI features integrates the complete global context. In terms of parameter, GCG module is much smaller than GCT module. Moreover, the combination of GCG module and GCT module in Inter-CL has no obvious advantage on F-measure and parameter. To this end, we choose the **Inter-CL(GCG)** as the default setting of Inter-CL module in I3CL. Finally, the I3CL model trained in a semi-supervised learning way achieves a gain of 3.5%, 3.6%, and 3.7% in terms of the F-measure over the baseline on the three datasets, respectively. Moreover, there is a similar trend of improvement in the Precision and Recall.

To prove that the improvements are obtained by addressing the two previous limitations, we select 100 images in each dataset, on which Mask R-CNN baseline is prone to produce a large number of fracture detections and inaccurate detections, as the difficult subset, and compare the performance gap between the Mask R-CNN baseline and the proposed I3CL on the three subsets. As shown in Table 2, I3CL achieves a remarkable gain of 8.2%, 7.5%, and 7.9% in terms of F-measure on the three subsets respectively. Some visual results of the Mask R-CNN baseline and our I3CL model are shown in Figure 9. As can be seen, Mask R-CNN produces fracture detections, missed detections, as well as

incomplete text contours indicated by the red boxes, while our I3CL model can effectively mitigate these issues and produce complete and accurate text masks. These quantitative and qualitative results demonstrate that our I3CL model benefits from the intra- and inter-instance collaborative learning and learns a better and more discriminative feature representation than Mask R-CNN baseline model, which help the text detector do well in detecting difficult texts.

Different settings of the Intra-CL module. Ablation studies are also conducted on the ArT dataset to investigate the impact of different settings of the Intra-CL module, *e.g.*, the number of convolutional branches in each block, feature fusion method, and the order of convolutional kernels in the cascade. The results are listed in Table 2. “**1-path**” denotes using a single convolutional branch with a regular $k \times k$ convolutional kernel. “**2-path**” denotes using two convolutional branches with asymmetric horizontal and vertical convolutional kernels. “**3-path**” denotes the default setting that contains all three branches as shown in Figure 5. “**cat**” and “**sum**” denote the feature fusion method, *i.e.*, concatenation and element-wise sum, respectively. “**kernel:753**” denotes using the 7×7 convolutional kernel in the first block and 5×5 and 3×3 kernels subsequently.

As shown, although both “**1-path**” and “**2-path**” can improve the performance marginally, “**3-path**” brings more improvement over the baseline model in terms of the F-measure, *i.e.*, a gain of 1.0%, confirming the value of using different kernels for modeling the multi-oriented texts. Besides, there is no significant difference between the concatenation and element-wise sum for feature fusion. We choose the latter one as the de-

Table 3: Ablation study of different settings of the Intra-CL module on the ArT dataset.

Method	R	P	F
Baseline	68.9	80.0	74.0
Intra-CL (1-path)	69.2	80.6	74.4
Intra-CL (2-path)	69.1	80.9	74.5
Intra-CL (3-path, kernel:753, cat)	69.1	81.8	74.9
Intra-CL (3-path, kernel:753, sum)	69.4	81.7	75.0
Intra-CL (3-path, kernel:357, sum)	69.0	80.5	74.3

Table 4: Ablation study on the setting of kernel sizes in the Intra-CL module on the ArT dataset.

Method	R	P	F	Parameter
kernel:753	69.4	81.7	75.0	46.7M
kernel:975	69.3	81.6	75.0	48.20M
kernel:777	69.6	80.9	74.8	48.10M
kernel:555	69.5	81.0	74.8	46.4M
kernel:333	69.5	80.7	74.7	45.1M
kernel:775	69.6	80.6	74.7	47.8M
kernel:773	69.7	81.5	75.1	47.5M
kernel:755	69.1	81.5	74.7	47.0M
kernel:733	69.6	81.0	74.9	46.3M

Table 5: Ablation study of different settings of pseudo label generation on the ArT dataset.

Method	R	P	F
Baseline	71.3	82.7	76.6
Threshold Filter	70.4	81.5	75.5
Ensemble	71.8	83.0	77.0
Ensemble + Overlap-Mask	71.5	84.1	77.3
Ensemble + Overlap-Mask + Soft-Box	72.2	83.6	77.5

Table 6: Evaluation results on the ArT dataset. "†", "‡", and "§" indicate that the results are collected from (Chng et al., 2019), official website of ArT, and our experiments using official released code, respectively.

Method	Venue	Backbone	R	P	F
PSENet† (Wang et al., 2019b)	CVPR'19	Res50	52.2	75.9	61.9
TextMountain † (Zhu and Du, 2021)	PR'21	Res50	53.5	86.2	66.0
TextRay (Wang et al., 2020a)	MM'20	Res50	58.6	76.0	66.2
ContourNet§ (Wang et al., 2020c)	CVPR'20	Res50	62.1	73.2	67.2
PAN§ (Wang et al., 2019c)	CVPR'19	Res18	61.1	79.4	69.1
CRAFT† (Baek et al., 2019)	CVPR'19	VGG16	68.9	77.2	72.9
PCR (Dai et al., 2021)	CVPR'21	DLA34	66.1	84.0	74.0
TextFuseNet† (Ye et al., 2020)	IJCAI'20	Res50	69.4	82.6	75.4
I3CL	-	Res50	71.3	82.7	76.6
I3CL + SSL	-	Res50	72.2	83.6	77.5

fault setting. When the order of convolutional kernels in the cascade is reversed, *i.e.*, from "**kernel:753**" to "**kernel:357**", we observe a performance drop of 0.7% in terms of the F-measure. We suspect the reason is that using a large kernel at the first block can effectively model long-range dependencies between characters, between gaps, and between characters and gaps, and using smaller kernels subsequently can gradually guide the Intra-CL module focus on the central region of the receptive field to learn a discriminative fea-

ture representation for either character or gap regions. However, when reversing the order, the Intra-CL module may struggle in extracting long-range context and be prone to noisy features in later blocks due to the large receptive fields. Besides, we conducted an ablation study on the setting of kernel sizes in the Intra-CL module. As shown in Table 4, compared with other settings, "**kernel:753**" achieves a better trade-off between performance and parameters. Although "**kernel:773**" delivers slightly better results in terms of recall and F-measure, its precision score is worse than that of "**kernel:753**" while increasing the number of parameters by 0.8M. As a result, we use "**kernel:753**" as the default setting in the Intra-CL module.

Different settings of pseudo label generation.

Moreover, we also conduct an ablation study to compare different settings of pseudo label generation on the ArT dataset. "**Threshold Filter**" denotes the common practice that selects the detection results from a single model as pseudo labels via a fixed threshold of confidence score. "**Ensemble**" represents the ensemble strategy based on multiple models without Inter-Mask and Soft-Box to refine the masks and boxes respectively. We use "**Ensemble + Overlap-Mask + Soft-Box**" as the default setting in our pseudo label generation. As shown in Table 5, due to the missed and false detections, "**Threshold Filter**" results in a severe side effect on the performance of the baseline model, *i.e.* 0.9%, 1.2%, and 1.1% drop on the three metrics respectively. In contrast, "**Ensemble + Overlap-Mask**" achieves the highest Precision of 84.1% among all models with ResNet-50 backbone on ArT. Furthermore, the default "**Ensemble + Overlap-Mask + Soft-Box**" can bring considerable performance gains of 0.9%, 0.9%, and 0.9% on the three metrics respectively, which verifies the effectiveness of our method.

4.4 Comparison with State-of-the-art Methods

Evaluation on ArT. We evaluate the effectiveness of the proposed I3CL model in detecting arbitrary-shaped mixed-language text on ArT dataset. The evaluation results of I3CL and previous methods are presented in Table 6. I3CL has achieved the best performance in terms of Recall and F-measure without using semi-supervised learning, which surpasses the current best method, *i.e.*, TextFuseNet† (Ye et al., 2020), by a large margin of 1.2% in terms of the F-measure. When applying the semi-supervised learning, a more compelling result can be achieved, *i.e.*, 72.2% on Recall and 77.5% on F-measure. To the best of our knowledge, I3CL is the first method achieving an F-measure over 77.0% on ArT using the ResNet-50 backbone.

Evaluation on Total-Text. We evaluate the effectiveness of the proposed I3CL model in detecting word-level arbitrary-shaped English text on Total-Text dataset. As shown in Table 7, similarly, I3CL sets a new state-of-the-art result of 86.3% F-measure on Total-Text. Furthermore, our detector with full implementations outperforms the latest state-of-the-art method, *i.e.*, FCENet (Zhu et al., 2021), by 1.1% F-measure. Moreover, I3CL achieves the highest Precision of 89.8%, which has a gain of 0.5% over the previous best method. In addition, our I3CL is the only one exceeding 86.0% in terms of the F-measure in all contenders.

Evaluation on CTW-1500. We evaluate the effectiveness of the proposed I3CL model in detecting text-line-level arbitrary-shaped English text on CTW-1500 dataset. The results are summarized in Table 10. As can be seen, our model achieves the best results with Precision of 88.4% and F-measure of 86.5% while keeping highly competitive results on Recall. Compared with the previous best method FCENet (Zhu et al., 2021), I3CL outperforms it by 1.2% on Recall, 0.8% on Precision, and 1.0% on F-measure. Note that since the text instances in CTW-1500 are labelled by text-line-level polygons rather than word-level annotations in Total-text, our method can learn better local and long-range features to handle such challenging cases and produce more accurate detections.

4.5 Competition on ArT Leaderboard

To explore the upper limit of the detection performance of I3CL, we join the melee on the ArT leaderboard. More complex experiments are conducted to improve the performance of I3CL on ArT dataset, including using stronger backbones and other common tricks during the training and testing.

Backbone. We adopt ResNet (He et al., 2016) and its variants, *i.e.*, ResNeXt (Xie et al., 2017) and ResNeSt (Zhang et al., 2020a), with different depths of {50,101,152} as the backbone. Besides, we also adopt ResNet-50 pre-trained with RegionCL (Xu et al., 2021a) pretext task and transformer-based ViTAEv2-S (Zhang et al., 2022) backbone for further experiments. The comparisons of different backbones can be seen in Table 9. As shown, RegionCL contrastive learning on ResNet-50 backbone apparently assists the downstream scene text detection task. Due to the superior feature representation by incorporating transformers with intrinsic inductive bias, ViTAEv2-S surpasses the base ResNet-50 with a large margin using similar parameters, *i.e.*, 2.3% gain on F-measure. What’s more, deeper backbones bring effective gains on all three evaluation metrics compared to the base ResNet-50. Among all the backbones we have

tried, ResNeSt-101 stands out with the highest values on Recall, Precision, and F-measure at single-scale testing, *i.e.*, 75.1%, 86.3%, and 80.3%, which are even better than the deeper ResNeXt-152. After comprehensive consideration about the size, training speed, and memory consumption of the model, we chose the ResNeSt-101 as the final backbone.

Table 7: Evaluation results on the Total-Text dataset.

Method	Venue	Backbone	R	P	F
Mask-TTD (Liu et al., 2019a)	TIP’19	Res50	74.5	79.1	76.7
TextSnake (Long et al., 2018)	ECCV’18	VGG16	74.5	82.7	78.4
ATTR (Wang et al., 2019d)	CVPR’19	VGG16	76.2	80.9	78.5
MSR (Xue et al., 2019)	IJCAI’19	Res50	74.8	83.8	79.0
CSE (Liu et al., 2019c)	CVPR’19	Res34	79.1	81.4	80.2
SAST (Wang et al., 2019a)	MM’19	Res50	76.9	83.8	80.2
TextDragon (Feng et al., 2019)	ICCV’19	VGG16	75.7	85.6	80.3
TextRay (Wang et al., 2020a)	MM’20	Res50	77.9	83.5	80.6
TextField (Xu et al., 2019)	TIP’19	VGG16	79.9	81.2	80.6
PSENet (Wang et al., 2019b)	CVPR’19	Res50	78.0	84.0	80.9
SegLink++ (Tang et al., 2019)	PR’19	VGG16	80.9	82.1	81.5
MS-CAFA (Dai et al., 2019)	TMM’19	Res50	78.6	84.6	81.5
SPCNet (Xie et al., 2019)	AAAI’19	Res50	82.8	83.0	82.9
LOMO (Zhang et al., 2019a)	CVPR’19	Res50	79.3	87.6	83.3
CRAFT (Baek et al., 2019)	CVPR’19	VGG16	79.9	87.6	83.6
CRNet (Zhou et al., 2020)	MM’20	Res50	82.5	85.8	84.1
Boundary (Wang et al., 2020b)	AAAI’20	Res50	83.5	85.2	84.3
ABCNet (Liu et al., 2020b)	CVPR’20	Res50	81.3	87.9	84.5
DB (Liao et al., 2020)	AAAI’20	Res50-DCN	82.5	87.1	84.7
PAN (Wang et al., 2019c)	ICCV’19	Res18	81.0	89.3	85.0
TextPerception (Qiao et al., 2020)	AAAI’20	Res50	81.8	88.8	85.2
PCR (Dai et al., 2021)	CVPR’21	DLA34	82.0	88.5	85.2
TextFuseNet (Ye et al., 2020)	IJCAI’20	Res50	83.2	87.5	85.3
ContourNet (Wang et al., 2020c)	CVPR’20	Res50	83.9	86.9	85.4
DGGR (Zhang et al., 2020b)	CVPR’20	VGG16	84.9	86.5	85.7
FCENet (Zhu et al., 2021)	CVPR’21	Res50-DCN	82.5	89.3	85.8
I3CL	-	Res50	83.7	89.2	86.3
I3CL + SSL	-	Res50	84.2	89.8	86.9

Table 8: Evaluation results on the CTW-1500 dataset.

Method	Venue	Backbone	R	P	F
CTD (Liu et al., 2019b)	PR’19	Res50	69.8	77.4	73.4
TextSnake (Long et al., 2018)	ECCV’18	VGG16	85.3	67.9	75.6
CSE (Liu et al., 2019c)	CVPR’19	Res34	76.0	81.1	78.4
Mask-TTD (Liu et al., 2019a)	TIP’19	Res50	79.0	79.7	79.4
ATTR (Wang et al., 2019d)	CVPR’19	VGG16	80.2	80.1	80.1
SAE (Tian et al., 2019)	CVPR’19	Res50	77.8	82.7	80.1
LOMO (Zhang et al., 2019a)	CVPR’19	Res50	76.5	85.7	80.8
SAST (Wang et al., 2019a)	MM’19	Res50	77.1	85.3	81.0
SegLink++ (Tang et al., 2019)	PR’19	VGG16	79.8	82.8	81.3
TextField (Xu et al., 2019)	TIP’19	VGG16	79.8	83.0	81.4
ABCNet (Liu et al., 2020b)	CVPR’20	Res50	78.5	84.4	81.4
MSR (Xue et al., 2019)	IJCAI’19	Res50	78.3	85.0	81.5
TextRay (Wang et al., 2020a)	MM’20	Res50	80.4	82.8	81.6
PSENet (Wang et al., 2019b)	CVPR’19	Res50	79.7	84.8	82.2
DB (Liao et al., 2020)	AAAI’20	Res50-DCN	80.2	86.9	83.4
CRAFT (Baek et al., 2019)	CVPR’19	VGG16	81.1	86.0	83.5
TextDragon (Feng et al., 2019)	ICCV’19	VGG16	82.8	84.5	83.6
PAN (Wang et al., 2019c)	ICCV’19	Res18	81.2	86.4	83.7
CRNet (Zhou et al., 2020)	MM’20	Res50	80.9	87.0	83.8
ContourNet (Wang et al., 2020c)	CVPR’20	Res50	84.1	83.7	83.9
DGGR (Zhang et al., 2020b)	CVPR’20	VGG16	83.0	85.9	84.5
TextPerception (Qiao et al., 2020)	AAAI’20	Res50	81.9	87.5	84.6
PCR (Dai et al., 2021)	CVPR’21	DLA34	82.3	87.2	84.7
TextFuseNet (Ye et al., 2020)	IJCAI’20	Res50	85.0	85.8	85.4
FCENet (Zhu et al., 2021)	CVPR’21	Res50-DCN	83.4	87.6	85.5
I3CL	-	Res50	84.5	87.4	85.9
I3CL + SSL	-	Res50	84.6	88.4	86.5

Multi-scale Testing. We resized the maximum size of the longer side to {1,500, 1,800, 2,100, 2,400, 2,700}, and the shorter side is scaled to {1,000, 1,300, 1,500}.



ICDAR2019-ArT

Total-Text

CTW-1500

Fig. 10: Some visual results of our I3CL model on the ArT, Total-Text, and CTW-1500 datasets, respectively.

Results from different scales are aggregated and NMS is used to suppress the redundant text instances to get final detection results. As can be seen in Table 10, compared to single-scale testing, multi-scale testing achieves a margin of 0.9% on F-measure over the baseline.

Soft-NMS (Bodla et al., 2017). NMS is replaced by Soft-NMS during the testing. We tried two decay strategies of confidence score, including linear decay and Gaussian decay. A finding in our experiment is that Soft-NMS with linear decay is better for ArT, and the F-measure was increased from 81.2% to 82.0%.

Mixup (Zhang et al., 2017). We implemented Mixup by pasting two text images and their labels in a fixed transparency ratio, *i.e.*, 1:1. Besides, we paste text images and non-text images together in a random transparency ratio. Those non-text images include landscape, architectural and animal images, increasing the diversity of background. In this way, I3CL achieves 80.8% in terms of F-measure at single-scale testing without Soft-NMS and 82.3% at multi-scale testing with Soft-NMS.

Model Ensemble. We ensemble the detected text boxes of different models to obtain better final results, such as models training with different backbones, different data augmentation strategies, and different iterations. Similarly in multi-scale testing, detection results from models using different training strategies are aggregated and then we use Soft-NMS to remove redundant text instances. As shown in Table 10, our I3CL

ultimately achieves an extremely impressive F-measure of 84.0% on ArT dataset.

In summary, the proposed I3CL sets new state-of-the-art on the ArT, Total-Text, and CTW-1500 for arbitrary-shaped scene text detection. Specifically, I3CL with ResNeSt-101 backbone achieves an impressive detection performance and ranks the 1st place on the ArT leaderboard. Some visual results are shown in Figure 10. As can be seen, I3CL can well handle different challenging cases including various shapes, extremely small scales, large gaps between characters, diverse font styles, and backgrounds, showing great potential for real-world applications.

Table 9: Results of I3CL without SSL using different backbones on ArT dataset. † and ‡ represent the RegionCL (Xu et al., 2021a) with finetuning and without finetuning on the ImageNet training data, respectively. * indicates that the whole detection model is implemented in MMDetection (Chen et al., 2019).

Backbone	R	P	F
ResNet-50	71.3	82.7	76.6
ResNet-50 w/ RegionCL†	72.6	81.9	77.0
ResNet-50 w/ RegionCL‡	73.5	81.6	77.3
ViTAEv2-S*	75.4	82.8	78.9
ResNeXt-101	74.1	85.5	79.4
ResNeSt-101	75.1	86.3	80.3
ResNeXt-152	74.9	86.0	80.1

Table 10: Results of I3CL using different tricks with ResNeSt-101 backbone on ArT dataset.

MS Testing	Soft-NMS	Mixup	Model Ensemble	F
✓				80.3
✓	✓			81.2
✓	✓	✓		82.0
✓	✓	✓	✓	82.3
✓	✓	✓		84.0

5 Discussion about Model Complexity

In this section, we discuss the model complexity of our I3CL, including parameter, computation, and inference speed.

Parameter. As shown in Table 1, I3CL brings considerable performance improvements of over 3% F-measure on different datasets with 17.8% parameters increase. The main parameter increase comes from transformer in Inter-CL module. ContourNet (Wang et al., 2020c) using Deformable ROI pooling with 256.3M parameters achieved 67.2%, 85.4%, and 83.9% in terms of F-measure on ArT, Total-Text, and CTW-1500, respectively. In contrast, I3CL sets new state-of-the-art results with 77.5%, 86.9%, and 86.5% on the three datasets and maintains a better trade-off between the model size and the performance with only 52.2M parameters, which proves that effectively model design for specific problems is important.

Computation. As known, transformer often brings heavy computation due to the self-attention mechanism. Unlike the image classification task in (Xu et al., 2021b), (Zhang et al., 2022) and (Liu et al., 2021), the computation of transformer encoder for modeling the dependencies between text instances in Inter-CL module depends on the number of text instances and the dimension of sequence features. After statistics, the average number of text instances on the images of the three datasets is 9. Benefit by the dimension reduction of the input sequence features and reasonable depth of transformer structure, transformer in Inter-CL module only increases about 0.05 GFLOPs computation on average. Compare with the 204.8 GFLOPs computation of the Mask R-CNN baseline, we consider that the computation increases of transformer structure when modeling the dependencies between text instances in Inter-CL module is acceptable. The total computation of I3CL can be seen in Table 1. Overall, I3CL brings considerable performance improvements of over 3% F-measure on different datasets with 20.7% computation increase.

Inference Speed. I3CL achieves an inference speed of 7.6 fps on CTW-1500 dataset, which is slightly slower

than 9.1 fps of the Mask R-CNN baseline. The comparison results of speed between I3CL and some previous methods can be seen in Table 11. When testing on CTW-1500, I3CL surpasses PSENet (Wang et al., 2019b) and ContourNet (Wang et al., 2020c) both on F-measure and speed (*i.e.*, 86.5% *vs* 82.2% and 83.9% and 7.6 fps *vs* 3.9 fps and 4.5 fps). However, compared with DB (Liao et al., 2020), though I3CL outperforms it by a large margin on F-measure (*i.e.*, 86.5% *vs* 83.5%), our method lags behind on speed (*i.e.*, 7.6 fps *vs* 22.0 fps). Overall, Although the inability to achieve real-time detection is a congenital limitation of the two-stage detector, I3CL still has obvious advantages over some previous methods on inference speed.

Table 11: Comparision results of speed between I3CL and some previous methods on CTW-1500 dataset.

Method	Venue	Backbone	FPS
CSE (Liu et al., 2019c)	CVPR'19	Res34	2.6
PSENet (Wang et al., 2019b)	CVPR'19	Res50	3.9
MSR (Xue et al., 2019)	IJCAI'19	Res50	4.3
LOMO (Zhang et al., 2019a)	CVPR'19	Res50	4.4
ContourNet (Wang et al., 2020c)	CVPR'20	Res50	4.5
TextField (Xu et al., 2019)	TIP'19	VGG16	6.0
TextFuseNet (Ye et al., 2020)	IJCAI'20	Res50	7.3
PCR (Dai et al., 2021)	CVPR'21	DLA34	11.8
DB (Liao et al., 2020)	AAAI'20	Res50	22.0
I3CL	-	Res50	7.6

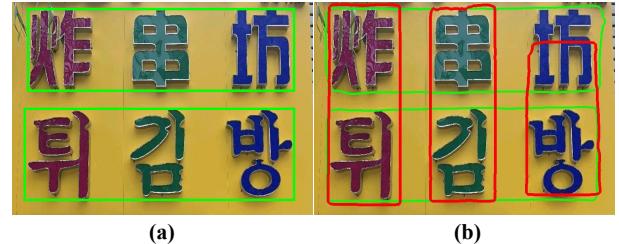


Fig. 11: (a) A test image with two horizontal text regions. (b) Failure detection. In (b), green polygons represent true positives, while red polygons represent false positives.

6 Limitation

In this section, we discuss the limitations of the proposed I3CL model. Although achieving state-of-the-art performance on three challenging benchmarks, our method is not outstanding enough in terms of speed, which can not meet the requirement for real-time applications. In the future, we plan to investigate efficient

instance segmentation pipelines and fast implementation as well as other effective and lightweight modules for collaborative learning. In addition, our model may generate linguistically ambiguous text proposals when detecting text arranged in multiple rows and columns. A failure detection example is shown in Figure 11. In the future, domain knowledge of linguistics can be utilized to design more effective modules as well as grouping strategies for proposal generation and filtering to mitigate the issue.

7 Conclusion

In this paper, we first identify two issues in arbitrary-shaped text detection, *i.e.*, fracture detection and inaccurate detection, and then argue that collaborative learning of both character and gap regions in text and long-dependencies between text instances within an image matters for mitigating the two issues. To validate the idea, we make the first attempt to propose a novel intra- and inter-instance collaborative learning model named I3CL, where an Intra-CL module based on a cascade of convolutional blocks with multiple receptive fields and an Inter-CL module based on a text instance transformer are devised. Besides, a new method of pseudo label generation based on ensemble strategy is proposed for semi-supervised learning of scene text detection. Comprehensive empirical studies on three public benchmarks demonstrate the effectiveness of the proposed I3CL model and its superiority over existing methods. We hope this study can open a new perspective for text detection and encourage more follow-up work in modeling long-range dependencies within and between text instances.

Acknowledgements This work was supported in part by National Natural Science Foundation of China: Grant No. 62076186, 62141112 and 41871243, in part by Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies): Grant No. 2019AEA170, and in part by ARC FL-170100117. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

- Baek Y, Lee B, Han D, Yun S, Lee H (2019) Character region awareness for text detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 9365–9374
- Baek Y, Shin S, Baek J, Park S, Lee J, Nam D, Lee H (2020) Character region attention for text spot-
- ting. In: Proceedings of the European Conference on Computer Vision, Springer, pp 504–521
- Bodla N, Singh B, Chellappa R, Davis LS (2017) Soft-nms—improving object detection with one line of code. In: Proceedings of the IEEE international conference on computer vision, pp 5561–5569
- Chen K, Wang J, Pang J, Cao Y, Xiong Y, Li X, Sun S, Feng W, Liu Z, Xu J, Zhang Z, Cheng D, Zhu C, Cheng T, Zhao Q, Li B, Lu X, Zhu R, Wu Y, Dai J, Wang J, Shi J, Ouyang W, Loy CC, Lin D (2019) MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:190607155
- Chen Z, Zhang J, Tao D (2021) Recursive context routing for object detection. International Journal of Computer Vision 129(1):142–160
- Chng CK, Liu Y, Sun Y, Ng CC, Luo C, Ni Z, Fang C, Zhang S, Han J, Ding E, et al. (2019) Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In: Proceedings of International Conference on Document Analysis and Recognition (ICDAR), IEEE, pp 1571–1576
- Ch'ng CK, Chan CS (2017) Total-text: A comprehensive dataset for scene text detection and recognition. In: Proceedings of International Conference on Document Analysis and Recognition, pp 935–942
- Dai P, Zhang H, Cao X (2019) Deep multi-scale context aware feature aggregation for curved scene text detection. IEEE Transactions on Multimedia 22(8):1969–1984
- Dai P, Zhang S, Zhang H, Cao X (2021) Progressive contour regression for arbitrary-shape scene text detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7393–7402
- Feng W, He W, Yin F, Zhang XY, Liu CL (2019) Textdragon: An end-to-end framework for arbitrary shaped text spotting. In: Proceedings of the IEEE International Conference on Computer Vision, pp 9076–9085
- Gupta A, Vedaldi A, Zisserman A (2016) Synthetic data for text localisation in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2315–2324
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778
- He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2961–2969
- Liao M, Shi B, Bai X, Wang X, Liu W (2017) Textboxes: A fast text detector with a single deep neural network. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 3070–3076

- ence on Artificial Intelligence, AAAI Press, pp 4161–4167
- Liao M, Shi B, Bai X (2018) Textboxes++: A single-shot oriented scene text detector. *IEEE Transactions on Image Processing* 27(8):3676–3690
- Liao M, Wan Z, Yao C, Chen K, Bai X (2020) Real-time scene text detection with differentiable binarization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 34, pp 11474–11481
- Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2117–2125
- Liu J, Chen Z, Du B, Tao D (2020a) Asts: A unified framework for arbitrary shape text spotting. *IEEE Transactions on Image Processing* 29:5924–5936
- Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: Single shot multibox detector. In: Proceedings of European conference on computer vision, Springer, pp 21–37
- Liu Y, Jin L, Fang C (2019a) Arbitrarily shaped scene text detection with a mask tightness text detector. *IEEE Transactions on Image Processing* 29:2918–2930
- Liu Y, Jin L, Zhang S, Luo C, Zhang S (2019b) Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recognition* 90:337–345
- Liu Y, Chen H, Shen C, He T, Jin L, Wang L (2020b) Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9809–9818
- Liu Z, Lin G, Yang S, Liu F, Lin W, Goh WL (2019c) Towards robust curve text detection with conditional spatial expansion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7269–7278
- Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:210314030
- Long S, Ruan J, Zhang W, He X, Wu W, Yao C (2018) Textsnake: A flexible representation for detecting text of arbitrary shapes. In: Proceedings of European Conference on Computer Vision, pp 20–36
- Nayef N, Patel Y, Busta M, Chowdhury PN, Karatzas D, Khelif W, Matas J, Pal U, Burie JC, Liu Cl, et al. (2019) Icdar2019 robust reading challenge on multilingual scene text detection and recognition—rrcmlt-2019. In: Proceedings of the International Conference on Document Analysis and Recognition, IEEE, pp 1582–1587
- Qiao L, Tang S, Cheng Z, Xu Y, Niu Y, Pu S, Wu F (2020) Text perceptron: Towards end-to-end arbitrary-shaped text spotting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 34, pp 11899–11907
- Shi B, Bai X, Belongie S (2017) Detecting oriented text in natural images by linking segments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2550–2558
- Song G, Chai W (2018) Collaborative learning for deep neural networks. arXiv preprint arXiv:180511761
- Sun Y, Ni Z, Chng CK, Liu Y, Luo C, Ng CC, Han J, Ding E, Liu J, Karatzas D, et al. (2019) Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In: Proceedings of the International Conference on Document Analysis and Recognition, IEEE, pp 1557–1562
- Tang J, Yang Z, Wang Y, Zheng Q, Xu Y, Bai X (2019) Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping. *Pattern recognition* 96:106954
- Tian Z, Shu M, Lyu P, Li R, Zhou C, Shen X, Jia J (2019) Learning shape-aware embedding for scene text detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4234–4243
- Wang F, Chen Y, Wu F, Li X (2020a) Textray: Contour-based geometric modeling for arbitrary-shaped scene text detection. In: Proceedings of the 28th ACM International Conference on Multimedia, pp 111–119
- Wang H, Lu P, Zhang H, Yang M, Bai X, Xu Y, He M, Wang Y, Liu W (2020b) All you need is boundary: Toward arbitrary-shaped text spotting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 34, pp 12160–12167
- Wang J, Yao J, Zhang Y, Zhang R (2018) Collaborative learning for weakly supervised object detection. arXiv preprint arXiv:180203531
- Wang L, Li D, Zhu Y, Tian L, Shan Y (2021) Cross-dataset collaborative learning for semantic segmentation. arXiv preprint arXiv:210311351
- Wang P, Zhang C, Qi F, Huang Z, En M, Han J, Liu J, Ding E, Shi G (2019a) A single-shot arbitrarily-shaped text detector based on context attended multi-task learning. In: Proceedings of the 27th ACM International Conference on Multimedia, pp 1277–1285
- Wang W, Xie E, Li X, Hou W, Lu T, Yu G, Shao S (2019b) Shape robust text detection with progressive scale expansion network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern

- Recognition, pp 9336–9345
- Wang W, Xie E, Song X, Zang Y, Wang W, Lu T, Yu G, Shen C (2019c) Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 8440–8449
- Wang X, Jiang Y, Luo Z, Liu CL, Choi H, Kim S (2019d) Arbitrary shape scene text detection with adaptive text region representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6449–6458
- Wang Y, Xie H, Zha ZJ, Xing M, Fu Z, Zhang Y (2020c) Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 11753–11762
- Xie E, Zang Y, Shao S, Yu G, Yao C, Li G (2019) Scene text detection with supervised pyramid context network. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, pp 9038–9045
- Xie Q, Luong MT, Hovy E, Le QV (2020) Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10687–10698
- Xie S, Girshick R, Dollár P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1492–1500
- Xu Y, Wang Y, Zhou W, Wang Y, Yang Z, Bai X (2019) Textfield: Learning a deep direction field for irregular scene text detection. *IEEE Transactions on Image Processing* 28(11):5566–5579
- Xu Y, Zhang Q, Zhang J, Tao D (2021a) Regioncl: Can simple region swapping contribute to contrastive learning? arXiv preprint arXiv:211112309
- Xu Y, Zhang Q, Zhang J, Tao D (2021b) Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems* 34
- Xue C, Lu S, Zhang W (2019) MSR: multi-scale shape regression for scene text detection. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, ijcai.org, pp 989–995
- Yang Q, Cheng M, Zhou W, Chen Y, Qiu M, Lin W (2018) Inceptext: A new inception-text module with deformable PSROI pooling for multi-oriented scene text detection. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, pp 1071–1077
- Yang Q, Wei X, Wang B, Hua XS, Zhang L (2021) Interactive self-training with mean teachers for semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5941–5950
- Ye J, Chen Z, Liu J, Du B (2020) Textfusenet: Scene text detection with richer fused features. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, pp 516–522
- Yuliang L, Lianwen J, Shuaitao Z, Sheng Z (2017) Detecting curve text in the wild: New dataset and new solution. arXiv preprint arXiv:171202170
- Zhang C, Liang B, Huang Z, En M, Han J, Ding E, Ding X (2019a) Look more than once: An accurate detector for text of arbitrary shapes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 10552–10561
- Zhang H, Cisse M, Dauphin YN, Lopez-Paz D (2017) mixup: Beyond empirical risk minimization. arXiv preprint arXiv:171009412
- Zhang H, Wu C, Zhang Z, Zhu Y, Lin H, Zhang Z, Sun Y, He T, Mueller J, Manmatha R, et al. (2020a) Resnest: Split-attention networks. arXiv preprint arXiv:200408955
- Zhang J, Tao D (2020) Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things. *IEEE Internet of Things Journal* 8(10):7789–7817
- Zhang J, Chen Z, Tao D (2021a) Towards high performance human keypoint detection. *International Journal of Computer Vision* 129(9):2639–2662
- Zhang P, Zhang B, Zhang T, Chen D, Wang Y, Wen F (2021b) Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 12414–12424
- Zhang Q, Zhang J, Liu W, Tao D (2019b) Category anchor-guided unsupervised domain adaptation for semantic segmentation. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada, pp 433–443
- Zhang Q, Xu Y, Zhang J, Tao D (2022) Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. arXiv preprint arXiv:220210108
- Zhang SX, Zhu X, Hou JB, Liu C, Yang C, Wang H, Yin XC (2020b) Deep relational reasoning graph network for arbitrary shape text detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 9699–9708
- Zhang X, Yue Y, Yang Y, Zhang X, Wang W, Zou Q (2020c) Collaborative learning network for scene text

- detection. In: 2020 Chinese Automation Congress (CAC), IEEE, pp 6788–6793
- Zhou X, Yao C, Wen H, Wang Y, Zhou S, He W, Liang J (2017) East: an efficient and accurate scene text detector. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp 5551–5560
- Zhou Y, Xie H, Fang S, Li Y, Zhang Y (2020) Crnet: A center-aware representation for detecting text of arbitrary shapes. In: Proceedings of the 28th ACM International Conference on Multimedia, pp 2571–2580
- Zhu Y, Du J (2021) Textmountain: Accurate scene text detection via instance segmentation. *Pattern Recognition* 110:107336
- Zhu Y, Chen J, Liang L, Kuang Z, Jin L, Zhang W (2021) Fourier contour embedding for arbitrary-shaped text detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3123–3131
- Zou Y, Yu Z, Liu X, Kumar BVKV, Wang J (2019) Confidence regularized self-training. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE, pp 5981–5990