Language Matters: A Weakly Supervised Vision-Language Pre-training Approach for Scene Text Detection and Spotting

Chuhui Xue[1,2], Wenqing Zhang[2], Yu Hao[2],
Shijian Lu[1], Philip Torr[3], and Song Bai[2]

[1] Nanyang Technological University
[2] ByteDance Inc.
[3] University of Oxford

Abstract. Recently, Vision-Language Pre-training (VLP) techniques have greatly benefited various vision-language tasks by jointly learning visual and textual representations, which intuitively helps in Optical Character Recognition (OCR) tasks due to the rich visual and textual information in scene text images. However, these methods cannot well cope with OCR tasks because of the difficulty in both instance-level text encoding and image-text pair acquisition (i.e. images and captured texts in them). This paper presents a weakly supervised pre-training method, oCLIP, which can acquire effective scene text representations by jointly learning and aligning visual and textual information. Our network consists of an image encoder and a character-aware text encoder that extract visual and textual features, respectively, as well as a visual-textual decoder that models the interaction among textual and visual features for learning effective scene text representations. With the learning of textual features, the pre-trained model can attend texts in images well with character awareness. Besides, these designs enable the learning from weakly annotated texts (i.e. partial texts in images without text bounding boxes) which mitigates the data annotation constraint greatly. Experiments over the weakly annotated images in ICDAR2019-LSVT show that our pre-trained model improves F-score by +2.5% and +4.8% while transferring its weights to other text detection and spotting networks, respectively. In addition, the proposed method outperforms existing pre-training techniques consistently across multiple public datasets (e.g., +3.2% and +1.3% for Total-Text and CTW1500).

Keywords: Vision-Language Pre-training; Scene Text Detection; Scene Text Spotting

1 Introduction

Optical Character Recognition (OCR) (including scene text detection, recognition, and spotting) has attracted increasing interests in recent years in both computer vision and deep learning research communities due to its wide range

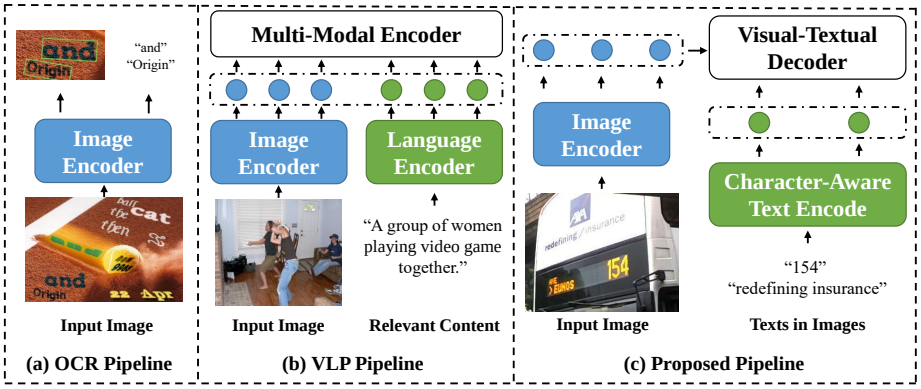

Fig. 1. Illustration of general Optical Character Recognition (OCR), Vision-Language Pre-training (VLP) pipeline, and the proposed pipeline (oCLIP): General OCR pipelines focus only on visual features from images. In addition, general VLP models extract image and language features from input images and corresponding sentence-level text, and model the interaction among all visual and textual features through a multi-modal encoder. Differently, oCLIP extracts instance-level textual features from texts instances in images. It models the interactions between each text instance and its extracted image features which can be trained with weak supervision only (i.e. partial texts in images without text bounding boxes). Our pre-trained model weights can be directly transferred to various scene text detectors and spotters with significant performance improvement.

of applications in multilingual translation, autonomous driving, etc. Most of the existing OCR techniques follow general computer vision pipelines that first extract visual features from the input image and and then perform feature regression or classification for text detection or recognition, as shown in Fig. 1 (a). However, we human usually read texts by utilizing not only the visual features of each text but also our linguistic knowledge in our memory. For example, we usually locate and read texts faster and more easily with the knowledge of the corresponding text language. Therefore, both visual and textual information are useful to robust reading of texts from natural scene images.

Recently, joint learning visual and textual representations has been studied in many Vision-Language Pre-training (VLP) techniques [40, 68, 5], and it greatly promotes various Vision-Language (VL) tasks such as Visual Question Answering (VQA), Image-Text Retrieval, etc. As a language-related task, OCR can intuitively benefit from these VLP techniques. However, most existing VLP methods usually suffer from two typical constraints while being applied to OCR tasks. **(1)** Each image in VL tasks is usually associated with one sentence or paragraph where words or phrases (i.e. tokens) are arranged in reading orders. Instead, an image in OCR tasks often contains many text instances each of which consists of one or multiple tokens. The tokens within one text instance are often closely related to each other (e.g. 'redefining' and 'insurance' in Fig. 1(c)) while

those from different text instances are completely irrelevant (e.g. 'insurance' and '154' in Fig. 1(c)). This makes it difficult to encode the textual information in a general sequential way. **(2)** Most VLP models learn from image-text pairs in which images and texts are correlated with each other at content-level (e.g. images and captions) as illustrated in Fig. 1(b). These content-relevant image-text pairs can be easily obtained from web, social media, etc., which has been proven to be effective for various VL tasks [40]. In contrast, OCR tasks aim to detect and recognize text instances that appear in images as shown in Fig. 1(c). The image-text pairs (i.e. images and texts in them) are more difficult to obtain as compared to VL tasks, requiring expensive and inefficient annotations.

We present an **O**CR **C**ontrastive **L**anguage-**I**mage **P**re-training (oCLIP) technique that exploits textual information for learning effective visual text representations for better scene text detection and spotting. Different from the text encoder in the existing VLP methods [40], we design a character-aware text encoder as illustrated in Fig. 1(c). It extracts language features by encoding textual information from the sequence of characters in each text instance without considering the relations among irrelevant text instances. In addition, we introduce a visual-textual decoder that models the relations between the input image and each labelled text instance only instead of all captured texts in the input image. With the two designs, oCLIP can learn effective visual text representations from weakly-annotated data (i.e. partial text instances in images without text bounding boxes) which greatly mitigates the data acquisition challenge and enables exploitation of large amounts of weakly-annotated images.

The contributions of this paper are three-fold. First, it introduces an end-to-end trainable pre-training network that allows to exploit language supervision to learn effective visual text representations. Second, we design a character-aware text encoder and a visual-textual decoder that can extract effective instance-level textual information and learn from partial text transcriptions without requiring text bounding boxes. Third, extensive experiments over multiple public datasets show that the proposed weakly supervised pre-trained network achieves superior performance on various scene text detection and spotting datasets.

2 Related Work

2.1 Scene Text Detection and Spotting

Most of recent scene text detectors are trained on fully-annotated data which can be broadly classified into two categories. The first category takes a bottom-up approach which first detects low-level text elements like characters [2], text segments [41, 47] and text keypoints [65] and then groups them into words or text lines. The second category treats words as one specific type of objects, and many scene text detectors like EAST [76], TextBoxes++ [25], RRD [28] and PSENet [54] are designed to detect text bounding boxes directly with generic object detection or segmentation techniques. Besides, many researchers study the text-specific features for robust text detection through text border or counter [66, 59, 77, 8], deformation convolution [52, 61], local refinement [73, 15] and so on.

Besides, many methods are designed to address the data bias. Some works [11, 71, 26] aim to synthesize scene text images that can be used for training scene text detection, recognition and spotting models. In addition, WeText [49] and OPM [44] design different weakly supervised mechanisms to use different types of data for training. GA-DAN [72] and TST [60] study the domain adaptation that adapt the synthetic scene text images to real. More recently, STKM [51] is proposed to pre-train a general model backbone for different scene text detectors.

Besides, many end-to-end trainable scene text spotters have been designed in which text detector and recognizer are complementary to each other. Li et al. [21] first integrates the scene text detector and RNN-based recognizer in to a unified network. Liu et al. [29] and He et al. [16] leverage more advanced scene text detectors or recognizers for better text spotting performances. More recently, Mask TextSpotters [36, 23, 24] adopt Mask R-CNN [13] as text detector and character segmentation or attention module for recognition. ABCNet [30, 31] proposes to detect texts with Bezier curves. TextDragon [10] detects center lines of texts along which characters are recognized in sequence. Baek et al. [3] proposes to detect characters by training with weakly supervised mechanism. Xing et al. [62] propose to detects and recognizes characters simultaneously. MANGO [39] is designed for text spotting with mask attention guidance. TextTranSpotter [19] is proposed to leverage the weakly-annotated images for training. Additionally, text recognition with less annotation have been studied in [48, 1].

2.2 Vision-Language Pre-training

As inspired by the advanced Transformer-based pre-training techniques [9] in Natural Language Processing (NLP) community, many vision-language pre-training methods have been studied in recent years, which greatly promotes the many multi-modal tasks in computer vision community. ViLBERT [35] and LXMERT [46] present a two-stream framework with a vision-language co-attention module for cross-modal feature fusion. On the other hand, VisualBERT [22], Unicoder-VL [20], VL-BERT [43], and UNITER [4] follow a single-stream framework (i.e. vanilla BERT structure), focusing on generic VL tasks including VCR and VQA. Besides, many VLP methods have been proposed for VL tasks such as RVL-BERT [6] for visual relationship detection, PERVALENT [12] and VLN-BERT [37] for visual navigation, VisualID [38] and VD-BERT [58] for visual dialog, etc.

3 Methodology

We present oCLIP that learns better scene text visual representations by feature alignment with textual information. As shown in Fig. 2, the proposed network first extracts image embeddings from input images by using an image encoder (including a network backbone ResNet-50 [14] followed by a multi-head attention layer). A character-aware text encoder is designed to extract the textual information from the transcriptions of text instances in input images by encoding the sequence of characters in each text instance. The extracted textual and

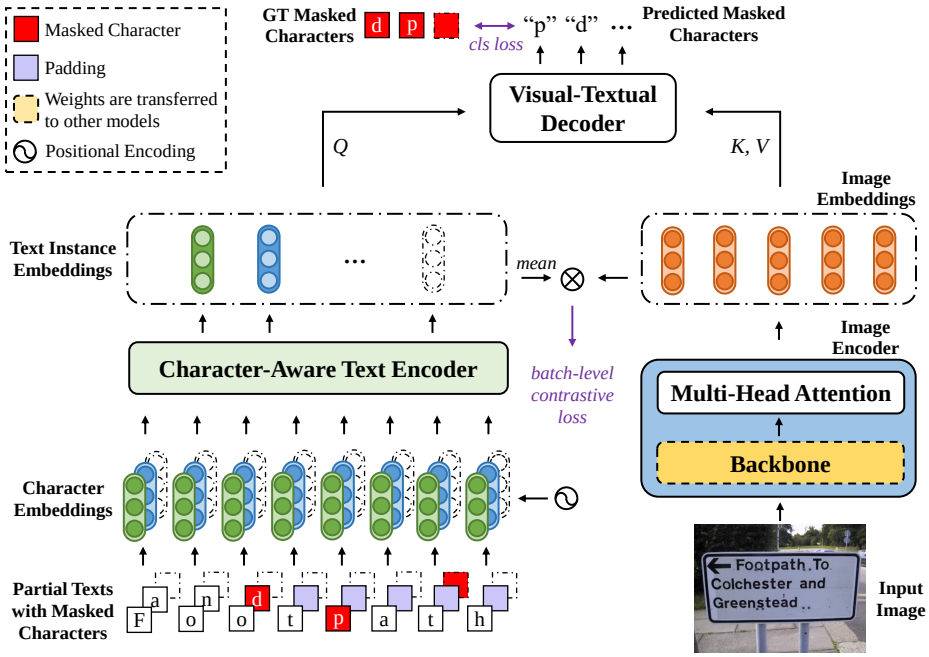

Fig. 2. The framework of oCLIP: Given an input image, an image encoder (including a backbone followed by a multi-head attention layer) first extracts the visual features. Meanwhile, the characters in each text instance are transformed to character embeddings, and a character-aware text encoder further extracts text instance embeddings from the character embeddings. A visual-textual decoder models the interactions between the text instance embeddings and the corresponding image embeddings. During training, a random character in each text instance will be masked (as highlighted by red boxes) and the overall network is optimized by predicting the masked characters.

visual features are passed into a visual-textual decoder which models the interactions among the visual features of input image and the textual features of each individual text instance. During training, we randomly mask a character in each text instance and the network is optimized by predicting the masked characters leveraging the extracted visual and textual features.

3.1 Character-Aware Text Encoder

In general VL tasks, texts (e.g. titles, captions, etc.) are usually sentences that consist of sequences of text tokens. As such, the text encoders for VL tasks are often designed to encode texts in a sequential way. However, the natural scene images in OCR tasks usually contain one or multiple text instances. The text tokens within each text instance are sequentially related to each other while those from different text instances are often completely irrelevant. This makes

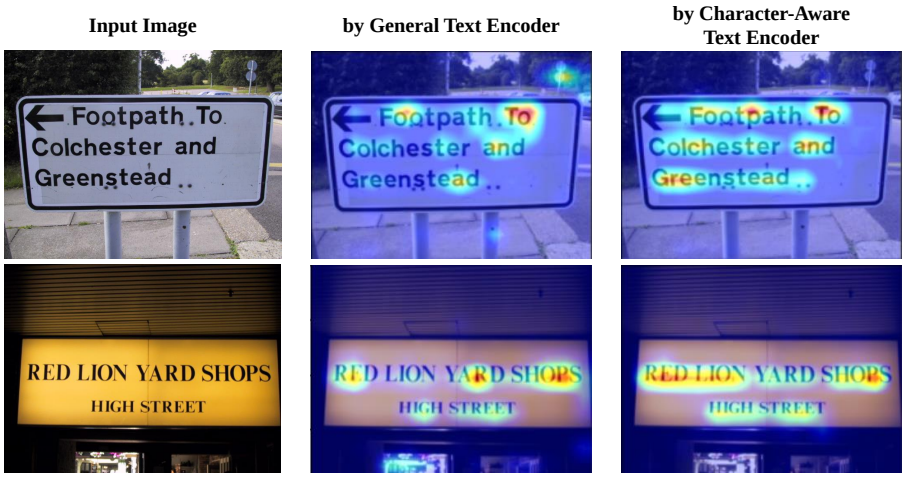

Fig. 3. Illustration of the proposed character-aware text encoder: Given sample images in the first column, columns 2-3 show the attention maps (from the attention layer in the image encoder) that are obtained from models with the general sentence-level text encoder and the proposed character-aware text encoder, respectively. The proposed character-aware text encoder attends better to text regions as compared with the general text encoder, leading to better learning of the scene text visual representations of the network backbone.

it difficult to encode these text instances by using a general text encoder. To address this issue, we design a character-aware text encoder.

The proposed character-aware text encoder extracts instance-level text embeddings with the input text instances as sequences of characters. Given n annotated text instances $T = \{t_0, t_1, ...t_{n-1}\}$ in an image, each text instance t_i consists of a sequence of characters $t_i = [c_0^i, c_1,^i, ..., c_{k-1}^i]$. We embed the characters into fixed-sized vectors and add a set of learnt positional encoding [50] $PE = [PE_0, PE_1, ..., PE_k]$ to capture the sequential information of characters in each text instance only, which can be formulated by:

$$ce_j^i = W_c \cdot c_j^i + PE_j, \quad i \in [0, n-1], \quad j \in [0, k-1], \tag{1}$$

where W_c is the character embedding matrix. The encoded character embeddings of $i-th$ text instance $ce^i = [ce_0^i, ce_1^i, ..., ce_{k-1}^i]$ are hence passed into a Transformer [50] encoder which models the interaction among all characters in the text instance and extracts the text instance embeddings te_i from its character embeddings ce_i. As a result, the character-aware text encoder extracts the text instance embeddings $te = \{te_0, te_1, ..., te_{n-1}\}$ from the annotated text instances $t = \{t_0, t_1, ...t_{n-1}\}$. Note a randomly selected character in each text instance is masked during training by setting it to the mask category.

The proposed character-aware text encoder effectively encodes the instance-level textual information and neglects the relations between each pair of text

instances. In addition, it can help to learn better visual text representations. Fig. 3 shows two sample images accompanied with the attention maps from the attention layer in the image encoder (details in Fig. 2). As Fig. 3 shows, by extracting textual information from the general text encoder, the overall model only focuses on partial text instances (e.g. 'Foo' and 'th' of 'Footpath'). This is because the tokens in general text encoder usually contain multiple characters (e.g. the token 'Footpath' contains 8 characters) and the model thus tends to focus on the most important parts only in the token according to the linguistic knowledge. Instead, the proposed text encoder can attend better to all text regions in images with the awareness of each character, demonstrating the superiority of the proposed encoder on learning visual text representations for scene text detection and spotting tasks.

3.2 Visual-Textual Decoder

The existing scene text pre-training techniques require fully-annotated data for training where the bounding boxes or transcriptions of all text instances are provided. However, such annotations are often extremely expensive and difficult to obtain. To address the data annotation bias, we present a visual-textual decoder that models the interaction between the input image and each individual annotated text while ignoring the unlabelled texts. The model thus can be trained by using the annotations of partial text instances in the images.

Given an input image I as shown in Fig. 2, we first extract the image embeddings ie and the the textual information te by using an image encoder (including a network backbone followed by a multi-head attention layer) and a character-aware text encoder, respectively. The visual-textual decoder hence learns the relationships among ie and each item in te (i.e. embeddings of each text instance) to enhance the learning of visual representations. Specifically, the visual-textual decoder consists of 6 stacked decoder layers each of which contains a multi-head attention layer and a feed-forward network. The text instance embeddings te are passed into the visual-textual decoder as queries and the image embeddings ie are passed into the decoder as keys and values. This allows every text instance alone to attend over all positions in the image embeddings. Note that we don't adopt the self-attention layer in the visual-textual decoder in order to neglect the relationships between each pair of text instances and eliminates the effects of unlabelled text instances. The model thus can effectively learn from partial annotated text instances. Finally, the visual-textual decoder predicts the masked characters in each text instance for optimization.

The masked characters can be predicted by learning the language knowledge from textual information only. We illustrate the attention maps of the decoder in Fig. 4 to demonstrate the effectiveness of the proposed visual-textual decoder. For each sample image in Fig. 4, we pass three text instances (with masked characters [M]) into our network, and we obtain three attention maps and three predicted masked characters each of which corresponds to an input text instance. As Fig. 4 shows, the visual-textual decoder not only predicts the masked characters (e.g. 'I' for 'ST[M]RLING') but also attends the regions of corresponding

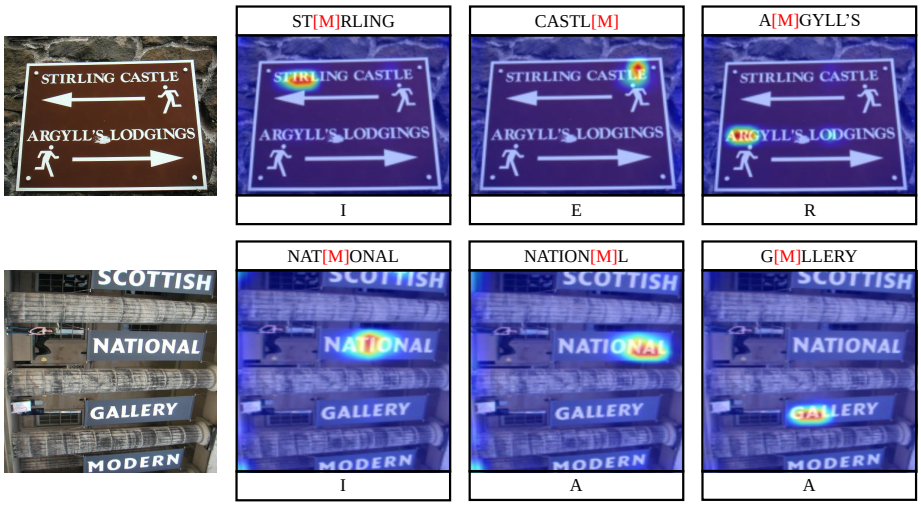

Fig. 4. Illustration of the proposed visual-textual decoder: Given two sample images in the first column, the input text instances (masked characters are highlighted by [M]), corresponding attention maps in the decoder and the predicted masked characters are shown from top to bottom in each box in columns 2-4, respectively. The proposed visual-textual decoder aligns the visual and textual features well, which effectively attends and predicts the masked characters in images.

masked characters well in the images. It can be seen that the proposed decoder aligns the visual and textual features to predict the masked characters (instead of using textual information alone), demonstrating the effectiveness of the proposed visual-textual decoder.

3.3 Network Optimization

During training, the proposed model takes text instances T (with masked characters $\boldsymbol{y}^{msk}$) and images I as inputs, and predicts the masked characters $\boldsymbol{p}^{msk}(I,T)$ for optimization. We consider the masked character prediction as a classification problem and adopt cross-entropy H for optimization:

$$\mathcal{L}_{cls} = \mathbb{E}_{(I,T)\sim D}\mathrm{H}(\boldsymbol{y}^{msk}, \boldsymbol{p}^{msk}(I,T)). \tag{2}$$

Besides, as inspired by CLIP [40], we adopt a batch-level contrastive loss for faster convergence. Given N images and N texts in a training batch, we form N^2 (text, image) pairs from all texts and images, where N pairs of texts and images are correlated with each other and $N^2 - N$ pairs are unrelated. For each image and text, we calculate the softmax-normalized image-to-text and text-to-image similarity as:

$$p_b^{i2t}(I) = \frac{\exp(I, T_b)}{\sum_{b=1}^{B}\exp(I, T_b)}, \quad p_b^{t2i}(T) = \frac{\exp(T, I_b)}{\sum_{b=1}^{B}\exp(T, I_b)}. \tag{3}$$

Let $\boldsymbol{y}^{i2t}(I)$ and $\boldsymbol{y}^{t2i}(T)$ denote the ground-truth one-hot similarity, where negative pairs have a probability of 0 and the positive pair has a probability of 1. The batch-level contrastive loss is thus defined by:

$$\mathcal{L}_{bc} = \mathbb{E}_{(I,T)\sim D}[\mathrm{H}(\boldsymbol{y}^{i2t}(I), \boldsymbol{p}^{i2t}(I)) + \mathrm{H}(\boldsymbol{y}^{t2i}(T), \boldsymbol{p}^{t2i}(T))]. \tag{4}$$

The full pre-training objective is defined by:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{bc}. \tag{5}$$

4 Experiments

4.1 Datasets

We use a number of public datasets in our experiments including SynthText[11], ICDAR2019-LSVT[45], CTW1500[70], Total-Text[7], and ICDAR2015[17]. More details are available in the supplementary material.

4.2 Implementation Details

Pre-training: We use ResNet-50 [14] as the backbone in the image encoder of the proposed network. The input images are resized to 512×512 during training. We adopt the Adam optimizer [18] with decoupled weight decay regularization [34] applied to all weights that are not gains or biases. The initial learning rate is $1e^{-4}$ which decays using a cosine schedule [33]. The model is trained end-to-end for 100 epochs on 8 Tesla V100 GPUs with batch size of 640. The length of each text instance is set as 25 following [69, 64].

Fine-tuning: We fine-tune several scene text detectors and spotters for evaluation of oCLIP including: 1) PSENet [54], 2) DB [27], 3) FCENet [77], 4) TextBPN [75], and 5) Mask TextSpotter-v3 [24]. More details are available in the supplementary material.

4.3 Experimental Results

We evaluate the proposed oCLIP from three aspects. First, we evaluate the performances of the proposed method by training with weakly annotated data (i.e. with partial annotated text instances available in each image). Second, we compare the proposed method with the existing pre-training techniques in scene text community. Third, we compare the proposed method with the state-of-the-art scene text detectors and spotters.

Table 1. Scene text detection performances of different models on **ICDAR2019-LSVT** dataset. '+oCLIP': Our pre-trained model with 400,000 weakly annotated images in ICDAR2019-LSVT dataset is adopted for fine-tuning.

Model	Precision	Recall	F-score
MSR [67]	86.4	63.4	73.1
Keypoint [65]	78.5	70.1	74.1
DB [27]	76.5	70.5	73.4
DB+oCLIP	**81.5**	**70.9**	**75.8**
PSENet [54]	90.4	63.5	74.6
PSENet+oCLIP	**90.7**	**67.0**	**77.1**

Table 2. Scene text spotting performances of different models on **ICDAR2019-LSVT** dataset. '+oCLIP': Our pre-trained model with 400,000 weakly annotated images in ICDAR2019-LSVT dataset is adopted for fine-tuning. 'P', 'R', 'F', '1-NED', and 'E2E' refer to Precision, Recall, F-score, Normalized metric in terms of Normalized Edit Distance, and end-to-end, respectively.

Method	Detection			E2E Spotting			
	P	R	F	1-NED	P	R	F
Mask TextSpotter-V3	80.5	61.0	69.4	35.7	32.1	24.4	27.7
Mask TextSpotter-V3+oCLIP	**80.6**	**61.9**	**70.1**	**39.0**	**37.4**	**28.7**	**32.5**

Weakly Supervised Pre-training: We evaluate the performances of oCLIP on learning visual text representations from weakly annotated data. We first conduct the experiments by pre-training our model on 400,000 weakly annotated images (i.e. only the transcription of the text-of-interest in each image is provided), and fine-tuning different scene text detectors and spotters on 30,000 fully annotated images from ICDAR2019-LSVT dataset. As Table 1 and 2 show, oCLIP improves the performances of different scene text detectors and spotters, demonstrating that the proposed method effectively learns the visual representations from weakly annotated data. Note that most previous approaches are designed to train on fully annotated images and they can't utilize the weakly annotated images from ICDAR2019-LSVT dataset well.

In addition, we conduct an experiment on SynthText dataset to show the effects of the amount of annotated texts on model performances. We first prepare four sets of text annotations from SynthText dataset by randomly selecting different proportions of text instances (i.e. 25%, 50%, 75%, and 100%) in each image (e.g. 1 out of 4 text instances in each image are used for training '25%' model). Next, we pre-train four models on all images in SynthText dataset by using the four sets of text annotations, and then transfer the backbone weights to fine-tune PSENet on Total-Text dataset. For comparison, we report the performances of two additionally models including: 1) 'No Pre-train' model in which no pre-training is adopted, and 2) 'Baseline' model that first trains PSENet on

Table 3. The effectiveness of the proposed weakly supervised pre-training technique: We pre-train four models by using different proportions of text instances in SynthText dataset (e.g. 1 out of 4 text instances in each image are used for training for '25%' model), and transfer the models weights to fine-tune PSENet on Total-Text dataset. 'Baseline': Train PSENet on SynthText and then fine-tune on Total-Text.

% annotated texts	Precision	Recall	F-score
No Pre-train	81.8	75.1	78.3
Baseline	87.8	79.0	82.6
25%	90.2	80.1	84.8
50%	91.1	80.0	85.2
75%	90.6	80.8	85.4
100%	**90.7**	**80.8**	**85.5**

Table 4. Comparison with existing scene text pre-training techniques: by pre-training on the same set of data (i.e. SynthText dataset), the proposed pre-training method outperforms the existing pre-training techniques consistently across different datasets. '+SynthText': Train PSENet with SynthText and then fine-tune with Total-Text.

Model	Total-Text			CTW1500		
	P	R	F	P	R	F
PSENet [54]	81.8	75.1	78.3	80.6	75.6	78.0
PSENet+SynthText	87.8	79.0	82.6	81.8	77.8	79.7
PSENet+STKM[51]	86.3	78.4	82.2	85.1	78.2	81.5
PSENet+oCLIP[SynthText]	90.7	80.8	85.5	86.3	79.6	82.8
PSENet+oCLIP[Web Images]	**92.2**	**82.4**	**87.0**	**87.5**	**79.9**	**83.5**

SynthText and then fine-tunes on Total-Text, respectively. As Table 3 shows, all four pre-train models help to improve the performances of PSENet, which outperforms the 'No Pre-train' and 'Baseline' models significantly. Besides, the four models achieve comparable performances on scene text detection task by pre-training on different amount of annotated texts, demonstrating the effectiveness of the proposed weakly supervised learning.

Comparing with Existing Scene Text Pre-training Strageties: We compare the oCLIP with two scene text pre-training strategies including: (1) training PSENet on SynthText dataset and then fine-tuning on real dataset, and (2) pre-training on SynthText by using STKM [51] and transferring the pre-trained weights to fine-tune PSENet on real dataset. For a fair comparison, we pre-train our model on SynthText with full annotations and transfer the backbone weights for fine-tuning PSENet on real datasets. As Table 4 shows, by pre-training on the same set of data, oCLIP outperforms the existing pre-training techniques by +3.3% and +1.3% in F-score on Total-Text and CTW1500 datasets, respectively.

Table 5. Comparison with state-of-the-art scene text detection techniques on **CTW1500** dataset. '+oCLIP' refers to that our pre-trained model on SynthText dataset is adopted for fine-tuning. 'RN50', 'PD', 'Syn', and 'MLT' refer to ResNet-50, pre-training data, SynthText dataset, and ICDAR2027-MLT dataset, respectively

Model	PD	Precision	Recall	F-score
TextSnake [32]	Syn	67.9	**85.3**	75.6
ATRR [57]	-	80.1	80.2	80.1
TextField [63]	Syn	83.0	79.8	81.4
Keypoint [65]	Syn	**88.3**	77.7	82.7
PAN [56]	Syn	88.0	79.4	83.5
CRAFT [2]	Syn	86.4	81.4	83.7
ContourNet [59]	-	83.7	84.1	83.9
SD [61]	MLT	85.8	82.3	84.0
DRRG [74]	MLT	85.9	83.0	84.5
TextBPN [75]	Syn	87.8	81.5	84.5
DB-RN50 [27]	-	81.1	80.6	80.8
DB-RN50+oCLIP	Syn	82.5	81.5	**82.0** **(+1.2)**
FCENet-RN50 [77]	-	85.7	80.7	83.1
FCENet-RN50+oCLIP	Syn	87.2	83.9	**85.6** **(+2.5)**

Automatic Data Acquisition and Training from Web Images: The proposed oCLIP can be simply applied to an automatic data acquisition and training pipeline due to the success of learning from weakly-annotated images. We extracted texts from 40 million web images and filtered out less-confident ones by using the existing scene text detector and recognizer form model pre-training. As Table 4 shows, by learning from the automatically extracted data from web images, oCLIP significantly improves the performances of PSENet on Total-Text and CTW1500 datasets. More details are available in supplementary material.

Comparing with State-of-the-Art Scene Text Detectors and Spotters: We further conduct experiments to compare oCLIP with state-of-the-art scene text detection and spotting techniques. For a fair comparison, we pre-train a model by our method on SynthText with full annotations and transfer the backbone weights to fine-tune DB, FCENet, TextBPN, and Mask TextSpotter-V3 on real datasets. As Table 5-8 show, the proposed pre-trained model effectively promote the existing scene text detectors to state-of-the-art performances on different dataset. In addition, by transferring the pre-trained weights from our model, the performances of different scene text detectors and spotters are consistently improved by large margins.

4.4 Ablation Studies

We study the contributions of different modules in our method including a character-aware encoder (CAE), a visual-textual decoder (VTD), and a batch-

Table 6. Comparison with state-of-the-art scene text detection techniques on **Total-Text** dataset. '+oCLIP' refers to that our pre-trained model on SynthText dataset is adopted for fine-tuning. 'RN50', 'PD', 'Syn', and 'MLT' refer to ResNet-50, pre-training data, SynthText dataset, and ICDAR2027-MLT dataset, respectively

Model	PD	Precision	Recall	F-score
TextSnake [32]	Syn	82.7	74.5	78.4
ATRR [57]	-	80.9	76.2	78.5
MSR [67]	Syn	83.8	74.8	79.0
TextField [63]	Syn	81.2	79.9	80.6
PAN [56]	Syn	88.0	79.4	83.5
CRAFT [2]	MLT	87.6	79.9	83.6
Keypoint [65]	Syn	86.1	82.6	84.4
ContourNet [59]	-	86.5	84.9	85.4
DRRG [74]	MLT	86.5	84.9	85.7
SD [61]	MLT	**89.2**	84.7	86.9
DB-RN50 [27]	-	81.7	75.0	78.2
DB-RN50+oCLIP	Syn	86.1	82.1	**84.1** (+5.9)
TextBPN [75]	-	88.0	82.9	85.4
TextBPN+oCLIP	Syn	89.0	**85.3**	**87.1** (+1.7)

Table 7. Comparison with state-of-the-art scene text detection techniques on **ICDAR2015** dataset. '+oCLIP' refers to that our pre-trained model on SynthText dataset is adopted for fine-tuning. 'RN50', 'PD', 'Syn', and 'MLT' refer to ResNet-50, pre-training data, SynthText dataset, and ICDAR2027-MLT dataset, respectively.

Model	PD	Precision	Recall	F-score
SegLink [42]	Syn	76.1	76.8	75.0
TextField [63]	Syn	84.3	80.1	82.4
TextSnake [32]	Syn	84.9	80.4	82.6
PAN [56]	Syn	84.0	81.9	82.9
ATRR [57]	-	90.4	83.3	86.8
CRAFT [2]	MLT	89.8	84.3	86.9
ContourNet [59]	-	87.6	86.1	86.9
SD [61]	MLT	88.7	**88.4**	88.6
DB-RN50 [27]	-	89.3	74.0	80.9
DB-RN50+oCLIP	Syn	89.1	82.0	**85.4** (+4.5)
FCENet-RN50 [77]	-	88.0	81.9	84.9
FCENet-RN50+oCLIP	Syn	**91.2**	82.7	**86.7** (+1.8)

level contrastive loss (BCL). We train four models with different modules included on fully annotated SynthText dataset and fine-tune PSENet on Total-Text dataset. As Table 9 shows, with the inclusion of different modules in our network, the performances of PSENet can be improved consistently, demonstrating the effectiveness of different modules in of network.

Table 8. Comparison with state-of-the-art scene text spotting techniques on **ICDAR2015** and **Total-Text** dataset. '+oCLIP' refers to that the model are fine-tuned from the our pre-trained model on SynthText dataset. 'S', 'W', and 'G' refer to end-to-end recognition with strong, weak, generic lexicon for ICDAR2015. 'Full' refers to full lexicon for Total-Text.

Model	**ICDAR2015**			**Total-Text**
	S	**W**	**G**	**Full**
CharNet [62]	80.1	74.5	62.2	-
FOTS [29]	83.6	74.5	62.2	-
TextDragon [10]	82.5	78.3	65.2	74.8
Boundary TextSpotter [53]	79.7	75.2	64.1	-
PAN++ [55]	82.7	78.2	69.2	78.6
ABCNet-V2 [31]	82.7	78.5	73.0	78.1
Mask TextSpotter-V3 [24]	83.3	78.1	74.2	78.4
Mask TextSpotter-V3+oCLIP	**84.1**	**78.6**	**74.3**	**79.6**

Table 9. Ablation study of the proposed method for scene text detection over Total-Text dataset. We fine-tune PSENet by using the pre-trained models with different modules. 'CAE', 'VTD', and 'BCL' refer to character-aware encoder, visual-textual decoder, and batch-level contrastive loss, respectively.

	CAE	**VTD**	**BCL**	**Precision**	**Recall**	**F-score**
No Pretrain				81.8	75.1	78.3
1			✓	88.1	77.7	82.6
2	✓		✓	89.6	78.9	83.9
3	✓	✓		89.3	77.4	82.9
4	✓	✓	✓	**90.7**	**80.8**	**85.5**

5 Conclusion

This paper presents a weakly supervised pre-training technique for scene text detection and spotting tasks. It focuses on the joint learning of visual and textual information from images and text transcriptions to enhance the learning of visual representations. It designs a character-aware text encoder and a visual-textual decoder that improves the feasibility of oCLIP on learning from partial text transcriptions only without text bounding boxes. Experimental results show that the proposed method can effectively learn from weakly-annotated scene text datasets which greatly mitigates the data acquisition challenge and significantly promotes different scene text detectors and spotters.

References

1. Baek, J., Matsui, Y., Aizawa, K.: What if we only use real datasets for scene text recognition? toward scene text recognition with fewer labels. In: Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3113–3122 (2021)
2. Baek, Y., Lee, B., Han, D., Yun, S., Lee, H.: Character region awareness for text detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9365–9374 (2019)
3. Baek, Y., Shin, S., Baek, J., Park, S., Lee, J., Nam, D., Lee, H.: Character region attention for text spotting. In: European Conference on Computer Vision. pp. 504–521. Springer (2020)
4. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Learning universal image-text representations (2019)
5. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European conference on computer vision. pp. 104–120. Springer (2020)
6. Chiou, M.J., Zimmermann, R., Feng, J.: Visual relationship detection with visual-linguistic knowledge from multimodal representations. IEEE Access **9**, 50441–50451 (2021)
7. Ch'ng, C.K., Chan, C.S.: Total-text: A comprehensive dataset for scene text detection and recognition. In: Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on. vol. 1, pp. 935–942. IEEE (2017)
8. Dai, P., Zhang, S., Zhang, H., Cao, X.: 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 (2021)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Feng, W., He, W., Yin, F., Zhang, X.Y., Liu, C.L.: Textdragon: An end-to-end framework for arbitrary shaped text spotting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9076–9085 (2019)
11. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2315–2324 (2016)
12. Hao, W., Li, C., Li, X., Carin, L., Gao, J.: Towards learning a generic agent for vision-and-language navigation via pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13137–13146 (2020)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. He, M., Liao, M., Yang, Z., Zhong, H., Tang, J., Cheng, W., Yao, C., Wang, Y., Bai, X.: Most: A multi-oriented scene text detector with localization refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8813–8822 (2021)
16. He, T., Tian, Z., Huang, W., Shen, C., Qiao, Y., Sun, C.: An end-to-end textspotter with explicit alignment and attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5020–5029 (2018)
17. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., et al.: Icdar 2015 competition on robust reading. In: 2015 13th international conference on document analysis and recognition (ICDAR). pp. 1156–1160. IEEE (2015)

18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Kittenplon, Y., Lavi, I., Fogel, S., Bar, Y., Manmatha, R., Perona, P.: Towards weakly-supervised text spotting using a multi-task transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4604–4613 (2022)
20. Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D.: Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020)
21. Li, H., Wang, P., Shen, C.: Towards end-to-end text spotting with convolutional recurrent neural networks. In: Proceedings of the IEEE international conference on computer vision. pp. 5238–5246 (2017)
22. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
23. Liao, M., Lyu, P., He, M., Yao, C., Wu, W., Bai, X.: Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(2), 532–548 (2021). https://doi.org/10.1109/TPAMI.2019.2937086
24. Liao, M., Pang, G., Huang, J., Hassner, T., Bai, X.: Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In: European Conference on Computer Vision. pp. 706–722. Springer (2020)
25. Liao, M., Shi, B., Bai, X.: Textboxes++: A single-shot oriented scene text detector. IEEE transactions on image processing **27**(8), 3676–3690 (2018)
26. Liao, M., Song, B., Long, S., He, M., Yao, C., Bai, X.: Synthtext3d: synthesizing scene text images from 3d virtual worlds. Science China Information Sciences **63**(2), 1–14 (2020)
27. Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X.: Real-time scene text detection with differentiable binarization. In: Proc. AAAI (2020)
28. Liao, M., Zhu, Z., Shi, B., Xia, G.s., Bai, X.: Rotation-sensitive regression for oriented scene text detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5909–5918 (2018)
29. Liu, X., Liang, D., Yan, S., Chen, D., Qiao, Y., Yan, J.: Fots: Fast oriented text spotting with a unified network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5676–5685 (2018)
30. Liu, Y., Chen, H., Shen, C., He, T., Jin, L., Wang, L.: 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 (2020)
31. Liu, Y., Shen, C., Jin, L., He, T., Chen, P., Liu, C., Chen, H.: Abcnet v2: Adaptive bezier-curve network for real-time end-to-end text spotting. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2021). https://doi.org/10.1109/TPAMI.2021.3107437
32. Long, S., Ruan, J., Zhang, W., He, X., Wu, W., Yao, C.: Textsnake: A flexible representation for detecting text of arbitrary shapes. In: Proceedings of the European conference on computer vision (ECCV). pp. 20–36 (2018)
33. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
34. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

35. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019)
36. Lyu, P., Liao, M., Yao, C., Wu, W., Bai, X.: Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 67–83 (2018)
37. Majumdar, A., Shrivastava, A., Lee, S., Anderson, P., Parikh, D., Batra, D.: Improving vision-and-language navigation with image-text pairs from the web. In: European Conference on Computer Vision. pp. 259–274. Springer (2020)
38. Murahari, V., Batra, D., Parikh, D., Das, A.: Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In: European Conference on Computer Vision. pp. 336–352. Springer (2020)
39. Qiao, L., Chen, Y., Cheng, Z., Xu, Y., Niu, Y., Pu, S., Wu, F.: Mango: A mask attention guided one-stage scene text spotter. Proceedings of the AAAI Conference on Artificial Intelligence pp. 2467–2476
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
41. Shi, B., Bai, X., Belongie, S.: Detecting oriented text in natural images by linking segments. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
42. Shi, B., Bai, X., Belongie, S.: Detecting oriented text in natural images by linking segments. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2550–2558 (2017)
43. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530 (2019)
44. Sun, Y., Liu, J., Liu, W., Han, J., Ding, E., Liu, J.: Chinese street view text: Large-scale chinese text reading with partially supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9086–9095 (2019)
45. Sun, Y., Ni, Z., Chng, C.K., Liu, Y., Luo, C., Ng, C.C., Han, J., Ding, E., Liu, J., Karatzas, D., et al.: Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1557–1562. IEEE (2019)
46. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019)
47. Tang, J., Yang, Z., Wang, Y., Zheng, Q., Xu, Y., Bai, X.: Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping. Pattern recognition **96**, 106954 (2019)
48. Tensmeyer, C., Wigington, C.: Training full-page handwritten text recognition models without annotated line breaks. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1–8. IEEE (2019)
49. Tian, S., Lu, S., Li, C.: Wetext: Scene text detection under weak supervision. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1492–1500 (2017)
50. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
51. Wan, Q., Ji, H., Shen, L.: Self-attention based text knowledge mining for text detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5983–5992 (2021)

52. Wang, F., Zhao, L., Li, X., Wang, X., Tao, D.: Geometry-aware scene text detection with instance transformation network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1381–1389 (2018)
53. Wang, H., Lu, P., Zhang, H., Yang, M., Bai, X., Xu, Y., He, M., Wang, Y., Liu, W.: All you need is boundary: Toward arbitrary-shaped text spotting. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12160–12167 (2020)
54. Wang, W., Xie, E., Li, X., Hou, W., Lu, T., Yu, G., Shao, S.: Shape robust text detection with progressive scale expansion network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9336–9345 (2019)
55. Wang, W., Xie, E., Li, X., Liu, X., Liang, D., Zhibo, Y., Lu, T., Shen, C.: Pan++: Towards efficient and accurate end-to-end spotting of arbitrarily-shaped text. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
56. Wang, W., Xie, E., Song, X., Zang, Y., Wang, W., Lu, T., Yu, G., Shen, C.: 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 (2019)
57. Wang, X., Jiang, Y., Luo, Z., Liu, C.L., Choi, H., Kim, S.: Arbitrary shape scene text detection with adaptive text region representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6449–6458 (2019)
58. Wang, Y., Joty, S., Lyu, M.R., King, I., Xiong, C., Hoi, S.C.: Vd-bert: A unified vision and dialog transformer with bert. arXiv preprint arXiv:2004.13278 (2020)
59. Wang, Y., Xie, H., Zha, Z.J., Xing, M., Fu, Z., Zhang, Y.: Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11753–11762 (2020)
60. Wu, W., Lu, N., Xie, E., Wang, Y., Yu, W., Yang, C., Zhou, H.: Synthetic-to-real unsupervised domain adaptation for scene text detection in the wild. In: Proceedings of the Asian Conference on Computer Vision (2020)
61. Xiao, S., Peng, L., Yan, R., An, K., Yao, G., Min, J.: Sequential deformation for accurate scene text detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16. pp. 108–124. Springer (2020)
62. Xing, L., Tian, Z., Huang, W., Scott, M.R.: Convolutional character networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9126–9136 (2019)
63. Xu, Y., Wang, Y., Zhou, W., Wang, Y., Yang, Z., Bai, X.: Textfield: Learning a deep direction field for irregular scene text detection. IEEE Transactions on Image Processing **28**(11), 5566–5579 (2019)
64. Xue, C., Lu, S., Bai, S., Zhang, W., Wang, C.: I2c2w: image-to-character-to-word transformers for accurate scene text recognition. arXiv preprint arXiv:2105.08383 (2021)
65. Xue, C., Lu, S., Hoi, S.: Detection and rectification of arbitrary shaped scene texts by using text keypoints and links. Pattern Recognition **124**, 108494 (2022)
66. Xue, C., Lu, S., Zhan, F.: Accurate scene text detection through border semantics awareness and bootstrapping. In: Proceedings of the European conference on computer vision (ECCV). pp. 355–372 (2018)
67. Xue, C., Lu, S., Zhang, W.: Msr: Multi-scale shape regression for scene text detection. arXiv preprint arXiv:1901.02596 (2019)

68. Xue, H., Huang, Y., Liu, B., Peng, H., Fu, J., Li, H., Luo, J.: Probing intermodality: Visual parsing with self-attention for vision-and-language pre-training. Advances in Neural Information Processing Systems **34** (2021)
69. Yu, D., Li, X., Zhang, C., Liu, T., Han, J., Liu, J., Ding, E.: Towards accurate scene text recognition with semantic reasoning networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12113–12122 (2020)
70. Yuliang, L., Lianwen, J., Shuaitao, Z., Sheng, Z.: Detecting curve text in the wild: New dataset and new solution. arXiv preprint arXiv:1712.02170 (2017)
71. Zhan, F., Lu, S., Xue, C.: Verisimilar image synthesis for accurate detection and recognition of texts in scenes. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 249–266 (2018)
72. Zhan, F., Xue, C., Lu, S.: Ga-dan: Geometry-aware domain adaptation network for scene text detection and recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9105–9115 (2019)
73. Zhang, C., Liang, B., Huang, Z., En, M., Han, J., Ding, E., Ding, X.: Look more than once: An accurate detector for text of arbitrary shapes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10552–10561 (2019)
74. Zhang, S.X., Zhu, X., Hou, J.B., Liu, C., Yang, C., Wang, H., Yin, X.C.: Deep relational reasoning graph network for arbitrary shape text detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9699–9708 (2020)
75. Zhang, S.X., Zhu, X., Yang, C., Wang, H., Yin, X.C.: Adaptive boundary proposal network for arbitrary shape text detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1305–1314 (2021)
76. Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., Liang, J.: East: An efficient and accurate scene text detector. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
77. Zhu, Y., Chen, J., Liang, L., Kuang, Z., Jin, L., Zhang, W.: Fourier contour embedding for arbitrary-shaped text detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3123–3131 (2021)