WeCromCL: Weakly Supervised Cross-Modality Contrastive Learning for Transcription-only Supervised Text Spotting

Jingjing Wu[⋆1], Zhengyao Fang[⋆1], Pengyuan Lyu[⋆2], Chengquan Zhang[2], Fanglin Chen[1], Guangming Lu[1], and Wenjie Pei[†1]

[1] Harbin Institute of Technology , Shenzhen, China
[2] Department of Computer Vision Technology, Baidu Inc.
{jingjingwu_hit, zhengyaonineve, wenjiecoder}@outlook.com, lvpyuan@gmail.com, zhangchengquan@baidu.com, {chenfanglin, luguangm}@hit.edu.cn,

Abstract. Transcription-only Supervised Text Spotting aims to learn text spotters relying only on transcriptions but no text boundaries for supervision, thus eliminating expensive boundary annotation. The crux of this task lies in locating each transcription in scene text images without location annotations. In this work, we formulate this challenging problem as a **We**akly Supervised **Cro**ss-**m**odality **C**ontrastive **L**earning problem, and design a simple yet effective model dubbed ***WeCromCL*** that is able to detect each transcription in a scene image in a weakly supervised manner. Unlike typical methods for cross-modality contrastive learning that focus on modeling the holistic semantic correlation between an entire image and a text description, our *WeCromCL* conducts atomistic contrastive learning to model the character-wise appearance consistency between a text transcription and its correlated region in a scene image to detect an anchor point for the transcription in a weakly supervised manner. The detected anchor points by *WeCromCL* are further used as pseudo location labels to guide the learning of text spotting. Extensive experiments on four challenging benchmarks demonstrate the superior performance of our model over other methods. Code will be released.

Keywords: Transcription-only supervised text spotting · Weakly supervised cross-modality contrastive learning

1 Introduction

Scene text spotting aims to detect and recognize text instances in scene text images. Existing methods [7, 16, 17, 20, 22, 24, 33–36, 40, 42] for text spotting have achieved remarkable progress relying on fully supervised learning, whereas these

⋆ Authors contribute equally.
† Corresponding author.

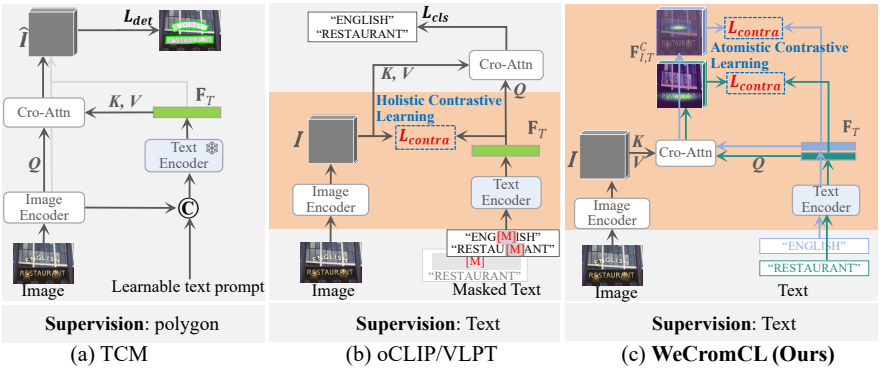

Fig. 1: Comparison between our *WeCromCL* and TCM [41], oCLIP [37] as well as VLPT [30]. (a) TCM distinguishes text regions from non-text regions in a scene image in a fully supervised manner using text polygon annotations. (b) Both oCLIP and VLPT perform holistic contrastive learning between the entire scene image and the text in a full supervised way w.r.t. the contrastive pairs to learn effective image encoder for downstream OCR tasks, while relying on the auxiliary task for optimization, namely predicting masked characters (oCLIP) or masked words (VLPT). (c) Our *WeCromCL* conducts atomistic contrastive learning to model the appearance consistency between a text transcription and its correlated region in the scene image for transcription-wise detection in a weakly supervised manner without text location annotations.

methods entail a large amount of annotations of text boundaries, which is extremely labor-consuming. In this work, we investigate transcription-only supervised text spotting, which only requires text transcriptions but no text boundaries for supervised learning, dramatically reducing the annotation overhead.

Transcription-only supervised text spotting is much more challenging than text spotting in full supervision, owing to the key difficulty of locating text transcriptions in scene text images without annotated text boundaries. A prominent method for transcription-only supervised text spotting is NPTS [26], which formulates the text spotting as a sequence prediction task. Specifically, it concatenates all text instances in a scene image into one sequence and seeks to predict all characters in an auto-regressive manner. While such modeling frees NPTS from text detection, a key limitation is that it suffers from arduous optimizing convergence. This is because there is no predefined order between different text instances when concatenating them together, the optimization of the model has to fit all potential permutations. Moreover, the model does not learn explicitly the mapping between text instances and correlated image regions without text detection, which further increases the difficulty of convergence. As pointed out as a primary limitation in the paper of NPTS, 'the training procedure requires a large number of computing resources'. Another state-of-the-art method for transcription-only supervised text spotting is TOSS [32], which draws inspiration from DETR [1] and locates text instances in scene images by pre-learning a set of text queries to probe transcriptions. However, the DETR-based method-

ology was initially designed for supervised object detection with location annotation. Although TOSS conducts modifications to adapt to weakly supervised text spotting, the absence of positional supervision still limits its effectiveness.

In this work we decompose the transcription-only supervised text spotting into two stages. First, our method detects an anchor point for each transcription in a scene image to locate the correlated image region in a weakly supervised manner. Second, the obtained anchor points are used as pseudo location labels to learn a single-point supervised text spotter which relies on only one single point instead of text boundary as detection supervision. The first step, namely detecting the anchor points to locate transcriptions without the groundtruth, is particularly challenging, and meanwhile its detection accuracy is crucial to the performance of text spotting in the second stage. To address this problem, we formulate it as a weakly supervised cross-modality contrastive learning problem, and design a simple yet effective model dubbed *WeCromCL* for it. Unlike typical methods for cross-modality contrastive learning that focus on modeling the holistic semantic correlation between a text description and an entire image, as oCLIP [37] and VLPT [30] behave in Figure 1, our *WeCromCL* conduct atomistic contrastive learning to learn the character-wise appearance consistency between a text transcription and its correlated region in a scene image in a weakly supervised manner. In particular, we design a soft modeling mechanism to learn an activation map by measuring the appearance correlation between a transcription and each pixel of a scene image. The activated region in the scene image is identified as the anchor point for this transcription and is associated with the transcription for contrastive learning to optimize *WeCromCL*.

Without the location annotations for text, our *WeCromCL* can still detect each transcription effectively. The rationale is that a transcription acts as a cluster center that associates all matched images containing it and the model is optimized to learn the similar appearance pattern for this transcription among all associated images, leading to precise location of transcriptions via learning the activation map. To conclude, our contributions are summarized as follows:

- We decompose the task of transcription-only supervised text spotting into two stages including weakly supervised text detection and single-point text spotting. Then we formulate the first and also challenging step as a cross-modality atomistic contrastive learning problem in an weakly supervised manner.
- To the best of our knowledge, we are the first to define and investigate the weakly supervised atomistic contrastive learning problem between image and text modalities. We particularly propose a simple yet effective method for it, called *WeCromCL*, which can learn an effective cross-modality character-wise consistency metric between a transcription and its visual appearance in a scene image, thereby detecting the correlated image regions for the transcription without annotated text boundaries.
- Leveraging the predicted anchor points by our *WeCromCL* as pseudo location labels, we learn an effective single-point supervised text spotter adapted from SRSTS v2 [35, 36], a state-of-the-art text spotter. Integrating the proposed *WeCromCL* and the learned single-point text spotter, we construct a pow-

erful system for transcription-only supervised text spotting, which compares favorably with existing methods on four challenging benchmarks.

2 Related Work

Vision-Language Contrastive Learning. Vision-language contrastive learning has attracted increasing attention in recent years. A variety of vision-language contrastive learning methods [3, 6, 8, 10, 14, 15, 29, 39] are proposed for representation learning of both visual information and language prompt. These methods typically focus on learning the semantic correlations between text and image modalities, whereas our method aims to model the character-wise appearance similarity between a text transcription and its correlated region around the anchor point in a scene image. Recently, contrastive learning has also been introduced to OCR. A prominent example is oCLIP [38], which conducts contrastive learning to optimize the image encoder for text spotting. It performs contrastive learning between an image and all the text instances appearing on the image in a holistic manner. Similarly, VLPT [30] also conducts holistic contrastive learning between an entire image and a transcription, and utilizes masked language modeling for auxiliary learning. Unlike oCLIP and VLPT, our *WeCromCL* seeks to learn the correlation between a transcription and the correlated image region for text location in a weakly supervised learning mode.

Fully Supervised Text Spotting. The mainstream text spotters need precise boundaries for supervision. The typical two-stage methods [5, 9, 16–20, 22, 24, 28, 33,34] conduct detection and recognition serially and bridge them by RoI pooling operation. Recently several one-stage methods have been proposed. MANGO [27] regards text spotting as a pure text recognition task by a designed position-aware attention module. SRSTS [35,36] decouples recognition from detection and proposes a sampling-based text recognition mechanism. Several works [7, 40, 42] modify Deformable DETR [43] to deal with text spotting. SPTS [26] represents text instance as a single point and tackles scene text spotting as a sequence prediction task.

Transcription-only Supervised Text Spotting. Currently, few works conduct text spotting under transcription-only supervision. Kittenplo *et al.* [13] refines Deformable DETR as an end-to-end text spotter named TTS. TTS is pre-trained on fully annotated synthetic data and fine-tuned on the transcription-only real-word data. It can be seen that TTS still uses a huge number of annotated synthetic data for training. Peng *et al.* [26] proposes no-point text spotting (NPTS) based on SPTS. NPTS takes transcription-only annotations as supervision and predicts randomly ordered transcriptions appearing in the scene text image. However, such design leads to arduous optimizing convergence and slow inference speed. TOSS [32] is transcription-only supervised and locates text instance by pre-learned queries, and its effectiveness is limited without detection supervision. Unlike the previous methods, we propose to conduct text spotting in two stages to ease transcription-only supervised text recognition problem: 1) detecting the anchor points for transcriptions; 2) conducting text spotting with

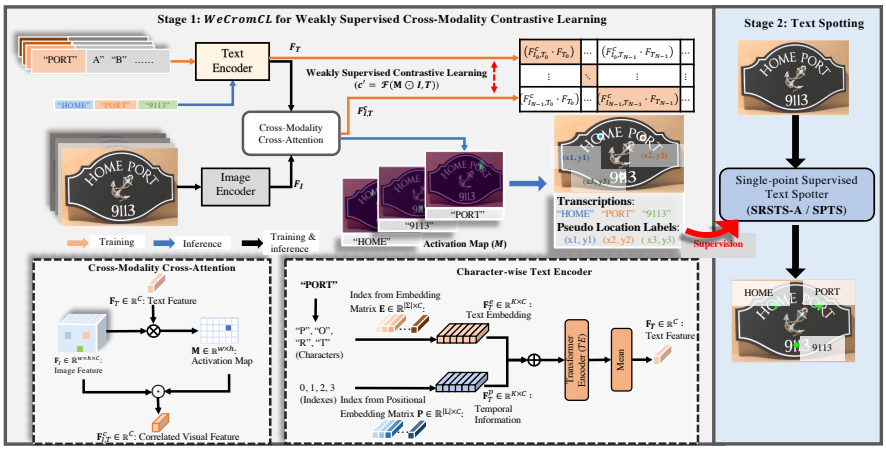

Fig. 2: Architecture of our proposed transcription-only supervised text spotter. Our method consists of two stages: 1) detecting the anchor point for each text instance as pseudo location label by *WeCromCL*; 2) conducting text spotting under the supervision of obtained pseudo location labels.

the obtained anchor points as pseudo labels. As a result, our method circumvents the limitations suffered by previous transcription-only supervised methods.

3 Method

Without annotations of text locations, it is difficult to apply the classical detect-and-recognize spotting paradigm [5,16,17,17,20,22,24,28,33,34] to transcription-only supervised text spotting. In light of this, we circumvent this difficulty by decomposing the task into two stages as shown in Figure 2: 1) detecting an anchor point in the scene image for each transcription to locate the correlated image region, and 2) leveraging the obtained anchor points as pseudo location labels to learn a single-point supervised text spotter which is learned based on only one single point as location annotation. The first step, namely detection of anchor points for transcriptions, is particularly challenging since the annotations of transcription locations are not available. Besides, The performance of text spotting in the second stage relies primarily on the predicting precision of the anchor points in the first stage. Thus, we focus on the first step and formulate it as a weakly supervised atomistic cross-modality contrastive learning problem, then we specifically design a simple yet effective framework dubbed *WeCromCL*.

3.1 Weakly Supervised Atomistic Cross-Modality Contrastive Learning

Typical cross-modality contrastive learning between text and image modalities, like CLIP [29] or oCLIP [37], aims to learn the holistic semantic compatibility

c between an entire image I and a text description T, which can be formulated as:

$$c = \mathcal{F}(I, T), \tag{1}$$

where $\mathcal{F}$ denotes the transformation function of a contrastive learning model. In the task of transcription-only supervised text spotting, only the transcriptions contained in a scene image are provided whilst the annotation of text locations for each transcription is not available. Thus, we aim to estimate the location for each transcription to serve as the pseudo location labels for supervised learning of a text spotter. Formally, given an image I containing a set of text transcriptions among which a text transcription T is only associated with its corresponding region in I, the correlation c' can be represented as:

$$c' = \mathcal{F}(\mathbf{M} \odot I, T), \tag{2}$$

where $\mathbf{M}$ is an activation map whose size is equal to that of I and $\odot$ denotes element-wise multiplication. All elements of $\mathbf{M}$ are binary values indicating whether the corresponding pixel is associated with the transcription T. Since the groundtruth of $\mathbf{M}$ is not provided, we have to optimize the contrastive learning model $\mathcal{F}$ in a weakly supervised manner. Thus, we refer to such contrastive learning setting as weakly supervised atomistic contrastive learning between image and text modalities.

Formulating the detection of transcription in an image as the weakly supervised atomistic contrastive learning across modalities defined in Equation 2 involves two crucial challenges:

- *Challenge 1*: effective modeling of $\mathcal{F}$ entails precise estimation of the activation map $\mathbf{M}$ in weakly supervised learning without the groundtruth.
- *Challenge 2*: unlike typical cross-modality contrastive learning such as CLIP that models the holistic semantic correlations between an entire image and a text, we aim to learn the atomistic correlation between a text transcription and its visual appearance in the correlated region in the scene image.

To address these challenges, we design a simple yet effective model, namely *WeCromCL*, for weakly supervised atomistic contrastive learning.

3.2 WeCromCL

We propose *WeCromCL* to detect an anchor point for each text transcription to locate its correlated region in the scene image, which serves as the pseudo location label for optimizing the text spotter in the second stage. *WeCromCL* follows weakly supervised atomistic cross-modality contrastive learning framework. As formulated in Equation 2, it takes a text transcription T and an image I as input, and predicts whether the image contains the transcription by measuring the correlation c' between them. Meanwhile, *WeCromCL* predicts the activation map $\mathbf{M}$ in which the highly activated region corresponds to the associated image region for the text transcription and is identified as the anchor point.

As shown in Figure 2, similar to CLIP, *WeCromCL* employs an image encoder and a text encoder to learn latent embeddings for the input image and text

transcription, respectively. In particular, we design a soft modeling mechanism to learn the activation map and thereby deal with *Challenge 1*. Besides, we devise a character-wise text encoder for tackling *Challenge 2*, which enables *WeCromCL* to learn the character-wise appearance similarity between the input transcription and its correlated region in the paired image. Finally, atomistic cross-modality contrastive learning is conducted to optimize the whole model of *WeCromCL*, using the constructed positive and negative training pairs based on the proposed negative-sampling mining scheme.

Image Encoder. The image encoder of our *WeCromCL* first employs BiFPN [31] to extract multi-scale convolutional features, then enhances the feature learning by applying the deformable transformer encoder [43]. The encoded image embeddings for the image I are denoted as $\mathbf{F}_I \in \mathbb{R}^{w \times h \times C}$.

Character-Wise Text Encoder. Typical cross-modality contrastive learning between image and text modalities, like CLIP, focuses on modeling the semantic correlation between two inputs. Thus, the text encoder of such models is designed to learn the holistic semantics of the input text. In contrast, our *WeCromCL* aims to learn the character-wise appearance consistency between the input text transcription and its visual appearance in the correlated image region. Thus, we devise the text encoder in the similar way as oCLIP [37] so that the encoded text embeddings 1) are distinguishable between different characters and 2) contain the temporal sequence information among characters in the text.

To learn text embeddings distinguishable between different characters, we learn individual vectorial embeddings with C dimensions for each character in the alphabet Σ, which is equivalent to learning an embedding matrix $\mathbf{E} \in \mathbb{R}^{|\Sigma| \times C}$. Then we can encode the text transcription T containing K characters by indexing the corresponding embeddings from $\mathbf{E}$ for each character of T sequentially and obtain the text embedding $\mathbf{F}_T^e \in \mathbb{R}^{K \times C}$.

To learn the temporal sequence information among characters in the text transcription, we learn extra positional embedding for each character position, resulting in an embedding matrix $\mathbf{P} \in \mathbb{R}^{L \times C}$ where L indicates the maximum number of characters in a transcription. As a result, we can encode the temporal information $\mathbf{F}_T^p \in \mathbb{R}^{K \times C}$ for the transcription T by indexing the positional embedding from $\mathbf{P}$ for all characters sequentially. We fuse the text embedding and the positional embedding by character-wise feature addition, and then adopt the Transformer encoder (TE) to perform feature propagation between characters in the transcription to model the correlation between them:

$$\mathbf{F}_T = \text{Mean}(\text{TE}(\mathbf{F}_T^e + \mathbf{F}_T^p)), \tag{3}$$

where $\mathbf{F}_T \in \mathbb{R}^C$ is the averaged text embedding over all characters by 'Mean'.

Soft Modeling of Activation Map by Cross-Modality Cross-Attention. The key to estimating the activation map ($\mathbf{M}$ in Equation 2) is how to measure the appearance correlation between the transcription and each pixel of the input image. To this end, we propose a soft modeling mechanism to learn such appearance correlation by measuring the cosine similarity between them in a projected feature space:

$$\begin{aligned} \mathbf{M}_{(i,j)} &= (\mathbf{W}_T^\top \mathbf{F}_T) \cdot (\mathbf{W}_I^\top \mathbf{F}_{I,(i,j)}), \\ \mathbf{M} &= \text{softmax}(\mathbf{M}). \end{aligned} \tag{4}$$

where $\mathbf{W}_T$ and $\mathbf{W}_I$ are learnable transformation matrices. $\mathbf{F}_{I,(i,j)}$ denotes the feature of pixel at (i,j) in the image I.

The values of learned $\mathbf{M}$ are continuous values between $[0,1]$ instead of binary values while higher values indicate higher response to the transcription. Learning the activation map in such a soft modeling way eases the gradient propagation for optimization and can preserve richer similarity information than the hard representation by binary values. The most activated pixel with the peak value in the map can be identified as the anchor point for the transcription.

Following the formulation in Equation 2, the learned activation map $\mathbf{M}$ is further used to aggregate the correlated features in the image to the text transcription for subsequent contrastive learning:

$$\mathbf{F}_{I,T}^c = \sum_{i=0}^{w-1} \sum_{j=0}^{h-1} \mathbf{M}_{(i,j)} (\mathbf{W}_V^\top \mathbf{F}_{I,(i,j)}), \tag{5}$$

where $\mathbf{F}_{I,T}^c \in \mathbb{R}^C$ is the aggregated correlated visual features in the image I to the transcription T and $\mathbf{W}_V$ is a learnable matrix for feature transformation. Combining the soft modeling in Equation 4 and the aggregation of correlated features in Equation 5 essentially boils down to cross-modality cross-attention operation, where the encoded transcription feature $\mathbf{F}_T$ serves as the query while the all pixels of encoded image feature $\mathbf{F}_I$ serve as the keys and values.

Cross-Modality Contrastive Learning by Negative-Sample Mining. We perform cross-modality contrastive learning between the learned correlated visual feature $\mathbf{F}_{I,T}^c$ and the encoded transcription feature $\mathbf{F}_T$ to optimize all modules of *WeCromCL* jointly. Similar to CLIP, for a positive training pair between an image I and a transcription T, we construct negative pairs in two ways: either pair the image I to multiple unpaired transcriptions (termed as image-to-text construction) or pair the transcription T to multiple unpaired images (termed as text-to-image construction).

We maximize the Cosine similarity of positive pairs while minimizing the similarity of negative pairs. Formally, given a training batch of images $\{I_0, I_1, \ldots, I_{N-1}\}$ and their associated text transcriptions $\{T_0, T_1, \ldots, T_{N-1}\}$, the loss function for the positive pair (I_i, T_i) and negative pairs using the text-to-image construction is defined as:

$$\mathcal{L}_i^{T2I} = -\log \frac{\exp(\text{Cosine}(\mathbf{F}_{I_i,T_i}^c, \mathbf{F}_{T_i})/\tau)}{\sum_{j=0}^{N-1} \exp(\text{Cosine}(\mathbf{F}_{I_j,T_i}^c, \mathbf{F}_{T_i})/\tau)}, \tag{6}$$

Similarly, we can define the loss function for the positive pair (I_i, T_i) and negative pairs using the image-to-text construction. In particular, we devise a negative-sample mining scheme to introduce more challenging negative pairs and

thereby enhance the modeling robustness of *WeCromCL*. A straightforward way is to apply the hard-sample mining scheme that selects more similar but unpaired transcriptions with I_i to construct more hard negative pairs. Surprisingly, we observe that randomly selecting unpaired transcriptions from the training set can also yield similar performance gain compared to hard-sample mining scheme as long as sufficient unpaired samples are provided. Thus, the loss based on such negative-sample mining scheme is defined as:

$$\mathcal{L}_i^{I2T} = -\log \frac{\exp(\text{Cosine}(\mathbf{F}^c_{I_i,T_i}, \mathbf{F}_{T_i})/\tau)}{\sum_{j=0}^{N+N_{\text{aug}}-1} \exp(\text{Cosine}(\mathbf{F}^c_{I_i,T_j}, \mathbf{F}_{T_j})/\tau)}, \tag{7}$$

where N_{aug} is the number of augmented negative pairs. Note that we only augment the unpaired transcriptions during image-to-text construction of negative pairs instead of augmenting the unpaired images during text-to-image construction, such negative-sample mining scheme can be performed quite efficiently with negligible overhead. Integrating the losses in two ways of negative pair construction, the loss of contrastive learning for a batch of N images is:

$$\mathcal{L}_{\text{cm}} = \frac{1}{2N} \sum_{i=0}^{N-1} (\mathcal{L}_i^{T2I} + \mathcal{L}_i^{I2T}). \tag{8}$$

Rationale behind *WeCromCL*. Our *WeCromCL* learns an effective cross-modality character-wise consistency metric between a transcription and the visual appearance in a scene image based on atomistic contrastive learning. It is able to detect the correlated region in the image to the transcription in a weakly supervised mode. The rationale behind this is that a transcription acts as a cluster center that associates all paired images with it, and the model is optimized to learn the similar appearance pattern regarding this transcription among all the paired images to determine the activation map. Meanwhile, the optimization by minimizing the similarity between negative pairs can guide the model to learn discriminative appearance patterns for each transcription, thereby preventing the model from collapsing to a uniform pattern for different transcriptions.

3.3 Anchor-Guided Text Spotting

The most activated position in an activation map $\mathbf{M}$ learned from *WeCromCL* is identified as the anchor point for the corresponding transcription, which is further used as pseudo location label for learning the text spotter in the second stage. Theoretically, any existing single-point supervised text spotter can be readily applied to our framework. To validate the effectiveness of our *WeCromCL*, we conduct two instantiations of the text spotter in the second stage. We first instantiate it with SPTS [26], a prominent single-point supervised text spotter. Then we tailor a single-point text spotter specifically by adapting SRSTS v2 [35, 36], which is a state-of-the-art supervised text spotter, to construct an powerful transcription-only text spotting system.

Instantiation of Text Spotter with SPTS. SPTS performs text spotting using only one single point for each transcription as location supervision. It formulates text spotting as a sequence prediction task. We use the predicted anchor points by our *WeCromCL* as pseudo location labels to train SPTS and its performance on text spotting can reflect the performance of *WeCromCL*.
Instantiation by Adapting SRSTS. SRSTS v2 is initially designed using the text boundaries as location annotation for supervision, we adapt it to enable it to rely on only one single point during training and refer the adapted version as 'SRSTS-A'. We provide adaptation details including image encoding, model training and loss function design in the supplementary material. Integrating the proposed *WeCromCL* and SRSTS-A, we construct a powerful system for transcription-only supervised text spotting.

4 Experiments

4.1 Experimental Setup

Benchmarks. 1) ICDAR 2013 [12] contains 229 training images and 223 testing images, in which most text instances are horizontal or slightly rotated. It provides 'Strong', 'Weak', and 'Generic' lexicons, which are represented as 'S', 'W' and 'G' in Table 5. 'S' denotes a lexicon containing 100 words, including the groundtruth transcription, which is provided for each test image. 'W' means a lexicon that consists of all the words appearing in the test set. 'G' is a generic lexicon provided by Liao *et al.* [16]. 2) ICDAR 2015 [11] contains 1000 training images and 500 testing images. It involves oriented text instances annotated with quadrangles. 3) Total-Text [2] comprises 1255 and 300 images for training and test, respectively. Most samples in this dataset are curved and are annotated with polygons and word-level transcriptions. 'Full' lexicon is provided which includes all words in the testing set. 4) CTW1500 [21] consists of 1000 training images and 500 test images. The text instances are annotated at line-level and arbitrary-shaped. 'Full' lexicon is provided for evaluation.
Evaluation Protocol. Since our method only outputs transcriptions and corresponding anchor points, the evaluation protocol for fully supervised methods which relies on precise bounding box matching is not suitable for our method. We adopt the single-point and edit distance metrics, following SPTS [26]. For the single-point metric, we match each predicted anchor point with the nearest center point of groundtruth bounding boxes, and then check if their text content are consistent. As for the edit distance metric, matching is conducted solely based on the edit distance between predicted and groundtruth transcriptions.
Implementation details. Following the previous methods [20, 22, 26], we train our method on a joint training set which consists of training images from Curved Synthetic Dataset 150k [20], ICDAR 2017 MLT [25], ICDAR 2013, ICDAR 2015 and Total-Text. In the first stage, we employ *WeCromCL* to generate the pseudo location labels for all training images. The obtained pseudo location labels are further used as supervision in the text spotting stage. Detailed settings are illustrated in the supplementary material.

Table 1: Ablation about different text encoders. 'Token-wise' denotes the text encoder of CLIP focusing on learning the semantics of the entire text. 'Character-wise' denotes the text encoder of *WeCromCL*. Single-point metric is used for evaluation. 'P', 'R' and 'F' represent 'Precision', 'Recall' and 'F-measure' respectively.

Set	Text Encoder	ICDAR 2013			ICDAR 2015			Total-Text			CTW1500		
		P	R	F	P	R	F	P	R	F	P	R	F
Training	Token-wise	82.5	82.5	82.5	71.4	64.7	67.9	68.1	69.3	68.5	53.1	51.9	52.5
	Character-wise	**93.2**	**93.2**	**93.2**	**91.4**	**86.0**	**88.6**	**83.5**	**85.1**	**84.3**	**67.0**	**65.7**	**66.3**
Test	Token-wise	79.7	76.9	78.6	65.7	63.3	64.4	70.2	60.4	64.9	66.7	64.3	65.5
	Character-wise	**90.4**	**90.5**	**90.5**	**86.9**	**80.1**	**83.4**	**85.8**	**75.4**	**80.3**	**78.9**	**76.5**	**77.7**

Table 2: Ablation on the negative-sampling mining scheme for training *WeCromCL*.

Set	Negative-sampling mining scheme	ICDAR 2013			ICDAR 2015			Total-Text			CTW1500		
		P	R	F	P	R	F	P	R	F	P	R	F
Training	×	91.5	91.5	91.5	90.2	84.3	87.2	82.5	83.6	83.0	59.8	58.4	59.1
	✓	**93.2**	**93.2**	**93.2**	**91.4**	**86.0**	**88.6**	**83.5**	**85.1**	**84.3**	**67.0**	**65.7**	**66.3**
Test	×	87.1	86.6	86.8	86.2	78.5	82.1	85.3	74.5	79.5	68.4	65.9	67.1
	✓	**90.4**	**90.5**	**90.5**	**86.9**	**80.1**	**83.4**	**85.8**	**75.4**	**80.3**	**78.9**	**76.5**	**77.7**

4.2 Ablation Studies of WeCromCL

In this section, we conduct ablation studies to investigate the effectiveness of proposed method. Note that more ablation studies and qualitative results are provided in the supplementary materials.

Comparison between different text encoders. We further compare the character-wise text encoder of our *WeCromCL* with the token-wise text encoder of CLIP which focuses on learning semantics of the entire text. To be specific, we replace the text encoder of *WeCromCL* with the token-wise text encoder of CLIP and test its performance of transcription detection. We consistently use the prompt template "There is a word '*transcription*'" for text encoding by CLIP, where '*transcription*' corresponds to the input text.

Table 1 shows that our model performs substantially better on all three metrics when equipped with the designed character-wise text encoder than using the token-wise encoder of CLIP. The results demonstrate that, compared to a token-wise text encoder that prioritizes semantic matching, encoding text at the character level can facilitate the learning of character-level correlations between a transcription and its visual representation in the image.

Ablation on the negative-sample mining scheme. To investigate the effectiveness of the proposed negative-sample mining scheme in *WeCromCL*, we compare the performance of two variants of *WeCromCL*: training with the negative-sample mining scheme and without the negative-sample mining scheme. The results in Table 2 show that the negative-sample mining scheme yields large performance gain. Particularly, the F-measure on the training and test set of CTW1500 is improved by 7.2% and 10.2%, respectively.

Quantitative evaluation of transcription detection. As shown in Table 1 and Table 2, the obtained pseudo labels for training set are accurate and can

Table 3: Comparison between oCLIP using holistic contrastive learning and our proposed *WeCromCL* employing atomistic contrastive learning.

Set	Methods	ICDAR 2013			ICDAR 2015			Total-Text			CTW1500		
		P	R	F	P	R	F	P	R	F	P	R	F
Training	oCLIP	87.0	79.9	83.3	51.1	35.4	41.9	70.4	43.5	53.8	45.2	37.4	40.9
	WeCromCL	**93.2**	**93.2**	**93.2**	**91.4**	**86.0**	**88.6**	**83.5**	**85.1**	**84.3**	**67.0**	**65.7**	**66.3**
Test	oCLIP	76.6	68.9	72.5	51.6	35.0	41.7	53.9	35.5	42.8	49.2	43.1	45.9
	WeCromCL	**90.4**	**90.5**	**90.5**	**86.9**	**80.1**	**83.4**	**85.8**	**75.4**	**80.3**	**78.9**	**76.5**	**77.7**

Table 4: Performance comparison between NPTS and *WeCromCL* + SPTS. The performance is evaluated by edit distance metric.

Methods	ICDAR 2015			Total-Text	
	S	W	G	None	Full
NPTS	70.3	62.7	57.0	61.6	70.6
WeCromCL + SPTS	**71.8**	**64.7**	**59.7**	**63.2**	**70.7**

serve as supervision for the following text spotting stage. Impressively, although contrastive learning is only conducted on the training set, *WeCromCL* still achieves generally satisfactory results on the test set. In particular, our *WeCromCL* achieves 90.5%, 83.4%, 80.3% and 77.7% on the test set of four benchmarks in terms of F-measure respectively.

Atomistic contrastive learning VS. holistic contrastive learning: comparison with oCLIP. oCLIP performs holistic cross-modality contrastive learning between an entire scene image and all text appearing in the image to learn image encoder for OCR tasks. It employs an auxiliary task for optimization, which masks the characters for a transcription one by one and conducts prediction. We adapt it to weakly supervised text detection and compare it with our model. Specifically, we aggregate the predicted attention map for each masked character and the most activated pixel with the peak value in the aggregated map is identified as the location prediction for this transcription. Table 3 shows that our *WeCromCL* outperforms oCLIP significantly on all benchmarks, which reveals the superiority of atomistic contrastive learning of *WeCromCL* over the holistic contrastive learning of oCLIP.

Comparison between *WeCromCL* + SPTS and NPTS. As an indirect evaluation of our *WeCromCL*, we use the obtained pseudo location labels by *WeCromCL* to train SPTS and compare the performance with NPTS. SPTS is a prominent single-point text spotter while NPTS is its adapted transcription-only supervised variant. For fairness, both NPTS and *WeCromCL* + SPTS are implemented based on their official code and neither utilize Random Cropping operation because it requires bounding box information. As shown in Table 4, supervised by the pseudo labels from *WeCromCL*, *WeCromCL* + SPTS surpasses NPTS in all evaluation dimensions, particularly excelling in scenarios without the use of lexicons or use 'Generic' lexicon, where recognition accuracy is evident. Besides, the visualization results in Figure 3 also show the consistent

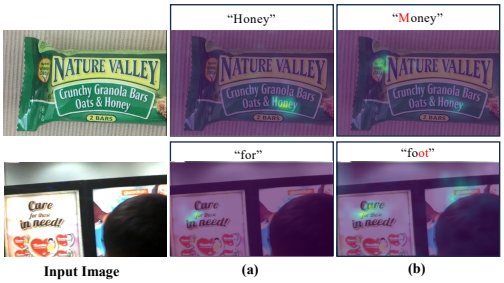

Fig. 3: Visual comparison of corresponding attention maps in the decoder of (a) *WeCromCL* + SPTS and (b) NPTS.

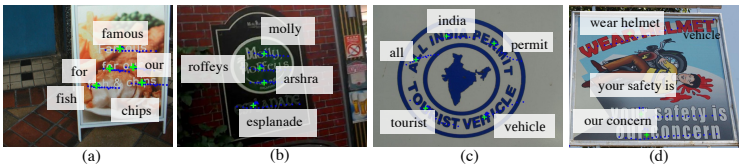

Fig. 4: Qualitative results of our transcription-only supervised text spotter on (a) ICDAR 2013, (b) ICDAR 2015, (c) Total-Text and (d) CTW1500. The green '+' represents the estimated anchor point while blue dots denote the sampled points.

Table 5: Quantitative results on ICDAR 2013, ICDAR 2015, Total-Text and CTW1500. '*' denotes the performance evaluated by single-point metric. '†' means the performance evaluated by edit-distance metric.

Methods	ICDAR 2013			ICDAR 2015			Total-Text		CTW1500	
	S	W	G	S	W	G	None	Full	None	Full
Fully Supervised Methods										
MTS v3 [17]	-	-	-	83.3	78.1	74.2	71.2	78.4	-	-
MANGO [27]	93.4	92.3	88.7	85.4	80.1	73.9	72.9	83.6	58.9	78.7
ABCNet v2 [22]	-	-	-	82.7	78.5	73.0	70.4	78.1	57.5	77.2
TESTR [42]	-	-	-	85.2	79.4	73.6	73.3	83.9	56.0	81.5
TTS [13]	-	-	-	85.2	81.7	77.4	78.2	86.3	-	-
ABINet++ [4]	-	-	-	84.1	80.4	75.4	77.6	84.5	60.2	80.3
Deepsolo [40]	-	-	-	86.8	81.9	76.9	79.7	87.0	64.2	81.4
ESTextSpotter [7]	-	-	-	87.5	83.0	78.1	80.8	87.1	64.9	83.9
SPTS* [26]	93.3	91.7	88.5	77.5	70.2	65.8	74.2	82.4	63.6	83.8
SPTS v2* [23]	93.9	91.8	88.6	82.3	77.7	72.6	75.5	84.0	63.6	84.3
Semi-supervised Methods										
TTS_{weak} [13]	-	-	-	78.7	75.2	70.1	75.1	83.5	-	-
Transcription-only Supervised Methods										
TOSS* [32]	86.4	85.1	82.2	65.9	59.6	52.4	65.1	74.8	**54.2**	65.3
WeCromCL+ SRSTS-A*	**89.9**	**88.1**	**83.7**	**82.1**	**76.1**	**68.8**	**70.1**	**81.4**	51.2	**75.7**
NPTS† [26]	89.6	86.4	83.2	70.3	62.7	57.0	61.6	70.6	50.9	70.5
WeCromCL+ SRSTS-A †	**91.2**	**89.8**	**84.6**	**79.5**	**72.8**	**66.2**	**68.1**	**79.1**	**52.7**	**79.9**

results. These results reveal 1) the effectiveness of our *WeCromCL* for detecting transcriptions without annotations and 2) the superiority of the two-stage modeling strategy of our method over the single-stage method like NPTS.

4.3 Transcription-only Supervised Text Spotting

In this section, we evaluate our optimized system for transcription-only supervised text spotting, namely the integration of *WeCromCL* + SRSTS-A.

Quantitative evaluation. We evaluate our transcription-only spotting system on four benchmarks. For a fair comparison, we remove Random Cropping operation and re-train NPTS. As shown in Table 5, we achieve superior performance when compared with other transcription-only supervised methods. In particular, our method surpasses NPTS and TOSS by 16.4% and 9.2% on ICDAR 2015 when evaluated with generic lexicon. Our method also performs well on the challenging Total-Text with curved text. Supervised by *WeCromCL*, SRSTS-A impressively outperforms TOSS by 5% and NPTS by 6.5% in the metric of 'None'.

TTS_{weak} [13] is a semi-supervised text spotter which follows DETR-based spotting framework. It is trained on fully-annotated synthetic data (including annotations of text boundaries) and transcription-annotated real-world data. As shown in Table 5, our method still achieves comparable performance compared to TTS_{weak} although our spotting system only uses pseudo location labels generated by our *WeCromCL* for all data.

Qualitative evaluation. We visualize the spotting results in Figure 4. As shown, our optimized system can handle various challenging scenarios, like tiny, fuzzy, curved and long text. The visualization results indirectly indicate the effectiveness and robustness of *WeCromCL*.

5 Conclusion

In this work, we decompose the transcription-only supervised text spotting into two stages: 1) detecting the anchor point for each transcription and 2) conducting text spotting guided by the obtained anchor points, among which the first stage is quite challenging and crucial. We formulate the detection of anchor points for text transcriptions as a weakly supervised atomistic contrastive learning problem across modalities, and devise a simple yet effective method dubbed *WeCromCL* for it. The detected anchor points are further used to guide the learning of text spotting. Extensive experiments on challenging benchmarks demonstrate the effectiveness and advantages of our proposed method.

Limitations. To measure the character-wise appearance consistency accurately between a transcription and the correlated region in the scene image, our *WeCromCL* requires high resolution of input images for detecting small texts.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (U2013210, 62372133), in part by Shenzhen Fundamental Research Program (Grant NO. JCYJ20220818102415032), in part by Guangdong Basic and Applied Basic Research Foundation (2024A1515011706), in part by the Shenzhen Key Technical Project (NO. JSGG20220831092805009, JSGG20201103153802006, KJZD20230923115117033).

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. Springer (2020)
2. Ch'ng, C.K., Chan, C.S.: Total-text: A comprehensive dataset for scene text detection and recognition. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR). vol. 1. IEEE (2017)
3. Duan, J., Chen, L., Tran, S., Yang, J., Xu, Y., Zeng, B., Chilimbi, T.: Multi-modal alignment using representation codebook. In: CVPR (2022)
4. Fang, S., Mao, Z., Xie, H., Wang, Y., Yan, C., Zhang, Y.: Abinet++: Autonomous, bidirectional and iterative language modeling for scene text spotting. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
5. Feng, W., He, W., Yin, F., Zhang, X.Y., Liu, C.L.: Textdragon: An end-to-end framework for arbitrary shaped text spotting. In: ICCV (2019)
6. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
7. Huang, M., Zhang, J., Peng, D., Lu, H., Huang, C., Liu, Y., Bai, X., Jin, L.: Estextspotter: Towards better scene text spotting with explicit synergy in transformer. In: ICCV (2023)
8. Huo, Y., Zhang, M., Liu, G., Lu, H., Gao, Y., Yang, G., Wen, J., Zhang, H., Xu, B., Zheng, W., et al.: Wenlan: Bridging vision and language by large-scale multi-modal pre-training. arXiv preprint arXiv:2103.06561 (2021)
9. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Reading text in the wild with convolutional neural networks. International journal of computer vision **116**(1) (2016)
10. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML. PMLR (2021)
11. 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). IEEE (2015)
12. Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L.G., Mestre, S.R., Mas, J., Mota, D.F., Almazan, J.A., De Las Heras, L.P.: Icdar 2013 robust reading competition. In: 2013 12th international conference on document analysis and recognition. IEEE (2013)
13. Kittenplon, Y., Lavi, I., Fogel, S., Bar, Y., Manmatha, R., Perona, P.: Towards weakly-supervised text spotting using a multi-task transformer. In: CVPR (2022)
14. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. NeurIPS **34** (2021)
15. Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., Wang, H.: Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. arXiv preprint arXiv:2012.15409 (2020)
16. 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) (2019)
17. Liao, M., Pang, G., Huang, J., Hassner, T., Bai, X.: Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In: ECCV. Springer (2020)

18. Liao, M., Shi, B., Bai, X.: Textboxes++: A single-shot oriented scene text detector. IEEE Transactions on Image Processing **27**(8) (2018)
19. Liao, M., Shi, B., Bai, X., Wang, X., Liu, W.: Textboxes: A fast text detector with a single deep neural network. In: AAAI (2017)
20. Liu, Y., Chen, H., Shen, C., He, T., Jin, L., Wang, L.: Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In: CVPR (2020)
21. Liu, Y., Jin, L., Zhang, S., Luo, C., Zhang, S.: Curved scene text detection via transverse and longitudinal sequence connection. Pattern Recognition **90** (2019)
22. 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 **44**(11) (2021)
23. Liu, Y., Zhang, J., Peng, D., Huang, M., Wang, X., Tang, J., Huang, C., Lin, D., Shen, C., Bai, X., et al.: Spts v2: single-point scene text spotting. arXiv preprint arXiv:2301.01635 (2023)
24. 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: ECCV (2018)
25. Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., Luo, Z., Pal, U., Rigaud, C., Chazalon, J., et al.: Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1. IEEE (2017)
26. Peng, D., Wang, X., Liu, Y., Zhang, J., Huang, M., Lai, S., Li, J., Zhu, S., Lin, D., Shen, C., et al.: Spts: single-point text spotting. In: ACM MM (2022)
27. Qiao, L., Chen, Y., Cheng, Z., Xu, Y., Niu, Y., Pu, S., Wu, F.: Mango: A mask attention guided one-stage scene text spotter. In: AAAI. vol. 35 (2021)
28. Qiao, L., Tang, S., Cheng, Z., Xu, Y., Niu, Y., Pu, S., Wu, F.: Text perceptron: Towards end-to-end arbitrary-shaped text spotting. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34 (2020)
29. 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: ICML. PMLR (2021)
30. Song, S., Wan, J., Yang, Z., Tang, J., Cheng, W., Bai, X., Yao, C.: Vision-language pre-training for boosting scene text detectors. In: CVPR (2022)
31. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: CVPR (2020)
32. Tang, J., Qiao, S., Cui, B., Ma, Y., Zhang, S., Kanoulas, D.: You can even annotate text with voice: Transcription-only-supervised text spotting. In: ACM MM (2022)
33. 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: AAAI. vol. 34 (2020)
34. 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)
35. Wu, J., Lyu, P., Lu, G., Zhang, C., Pei, W.: Single shot self-reliant scene text spotter by decoupled yet collaborative detection and recognition (2023)
36. Wu, J., Lyu, P., Lu, G., Zhang, C., Yao, K., Pei, W.: Decoupling recognition from detection: Single shot self-reliant scene text spotter. In: ACM MM (2022)
37. Xue, C., Zhang, W., Hao, Y., Lu, S., Torr, P.H.S., Bai, S.: Language matters: A weakly supervised vision-language pre-training approach for scene text detection and spotting. In: ECCV (2022)

38. Xue, C., Zhang, W., Hao, Y., Lu, S., Torr, P.H., Bai, S.: Language matters: A weakly supervised vision-language pre-training approach for scene text detection and spotting. In: ECCV. Springer (2022)
39. Yang, J., Duan, J., Tran, S., Xu, Y., Chanda, S., Chen, L., Zeng, B., Chilimbi, T., Huang, J.: Vision-language pre-training with triple contrastive learning. In: CVPR (2022)
40. Ye, M., Zhang, J., Zhao, S., Liu, J., Liu, T., Du, B., Tao, D.: Deepsolo: Let transformer decoder with explicit points solo for text spotting. In: CVPR (2023)
41. Yu, W., Liu, Y., Hua, W., Jiang, D., Ren, B., Bai, X.: Turning a clip model into a scene text detector. In: CVPR (2023)
42. Zhang, X., Su, Y., Tripathi, S., Tu, Z.: Text spotting transformers. In: CVPR (2022)
43. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: ICLR