TWIST: Text-only Weakly Supervised Scene Text Spotting Labels

Lilong Wen 1,2  Xiu Tang 1,2,3 Dongxiang Zhang 1,2
1 The State Key Laboratory of Blockchain and Data Security, Zhejiang University 2 Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security 3 School of Software Technology, Zhejiang University

Abstract
Scene text spotting is crucial for image understanding but requires substantial annotated data. TWIST focuses on minimizing annotation efforts by using text-only annotations. Existing methods rely on attention maps, which are often inadequate. TWIST introduces a pseudo-label generation method enabling end-to-end training for both text recognition and location estimation, achieving state-of-the-art results in several benchmarks under text-only supervision.

Introduction
Et rutrum ex euismod vel. Pellentesque suscipit, velit in fermentum vestibulum, lectus nisi pretium nibh, sit amet aliquam lectus augue vel velit.

Importance: Essential for multimedia understanding, impacts self-driving vehicles, product search, image search, virtual assistants, and accessibility services.

Challenges: High accuracy requires comprehensive annotations, leading to labor-intensive data labeling. Synthetic datasets help but have domain gaps.

Solution: Weakly supervised methodologies reduce annotation complexity. TWIST introduces a novel approach to integrate pseudo-labels for end-to-end training.

Methodology
The system comprises two primary components: a pseudo-label generation module and a text spotting module. The pseudo-label generation process is critical as it uses text-only annotations to generate spatial labels (pseudo-labels) that help in training the spotting model.

Character-Level Pseudo-label Generation: Each text label is masked and processed through a transformer-based text encoder, generating a feature representation for each masked character. These features are used to predict the corresponding characters and derive spatial attention maps, which are then aggregated to form a comprehensive attention map for each character indexed by $i$.

Image $I$: $I_{n} = \{b_{i}, t_{i} \}$, where $b_{i}$ and $t_{i}$ are the bounding box and text content for each text instance.

Cross-Entropy Loss for Character Classification
$$L_{CE} = \frac{1}{n} \sum_{i=1}^{n} CE(y_{i}, \hat{y}_{i})$$

$CE$ is the classification loss using cross-entropy.

Aggregation of Attention Maps
$$A_{i}^{n} = \max_{j} (T (A_{j}^{n}))$$

$A_{i}^{n}$ is the aggregated attention map, $T$ is the threshold function applied to the attention map $A$ for each character indexed by $h_{i}$, $i$, and $j$.

Pseudo-label Generation
$$G(I, T) = \{ (h_{1}, t_{1}), \ldots , (h_{n}, t_{n}) \}$$

$G$ denotes the pseudo-label generation process for an image $I$ with text instances $T$, where $h_{i}$ and $t_{i}$ are the bounding box and text content for each text instance.

Contributions
A novel methodology that allows for text spotting without explicit spatial annotations, using pseudo-labels generated from text-only annotations.

Enhanced accuracy and reduced annotation requirements compared to previous methods.

This approach significantly simplifies the annotation process by relying solely on textual content, addressing the challenges of traditional text spotting methods that require detailed positional annotations. The method promises to reduce annotation efforts and computational costs while maintaining competitive spotting accuracy.

Visualization
Figure 3: Visualization of polygon results on Total-text dataset

Architecture Overview
Figure 1: The architecture of TWIST comprises two principal modules: Character-level Pseudo Label Generation generates character-level positions for each text instance through a masked character prediction task and aggregates these positions into pseudo-labels. Which are then served as training inputs for the Pseudo Label Driven Text Spotting module.

2. Cross-Entropy Loss for Character Classification
$$L_{CE} = \frac{1}{n} \sum_{i=1}^{n} CE(y_{i}, \hat{y}_{i})$$

$L_{CE}$ is the classification loss using cross-entropy.

3. Aggregation of Attention Maps
$$A_{i}^{n} = \max_{j} (T (A_{j}^{n}))$$

$A_{i}^{n}$ is the aggregated attention map, $T$ is the threshold function applied to the attention map $A$ for each character indexed by $h_{i}$, $i$, and $j$.

4. Pseudo-label Generation
$$G(I, T) = \{ (h_{1}, t_{1}), \ldots , (h_{n}, t_{n}) \}$$

$G$ denotes the pseudo-label generation process for an image $I$ with text instances $T$, where $h_{i}$ and $t_{i}$ are the bounding box and text content for each text instance.

This approach significantly simplifies the annotation process by relying solely on textual content, addressing the challenges of traditional text spotting methods that require detailed positional annotations. The method promises to reduce annotation efforts and computational costs while maintaining competitive spotting accuracy.

Figure 2: Comparing our method to fully supervised training. (a) A fully supervised method requires both location and pseudo labels. Which are then served as training inputs for the Pseudo Label Driven Text Spotting module.

[a] Supervised method

Character Level Masking

Text: "Bus", "STOP", "Since", "1900"

Supervised Text spotting

Location: $(x_{1}, y_{1})$

Supervised text spotting

Encoder

Pseudo Label Driven Text Spotting

Pseudo GT

Loss $L_{Dec}$

Augment

Polygons

Visual Backbone

Character-Level Pseudo-label Generation

Visual Backbone

Text

Query

G

Character Prediction

Mathematically, Character Prediction:
$$y_{i} = W \cdot E_{i} + b_{i}$$

Here, $y_{i}$ is the prediction output, $W$ is the weight matrix, $E_{i}$ is the encoded feature matrix from the text, and $b_{i}$ is the bias.