A Bilingual, Open World Video Text Dataset and End-to-end Video Text Spotter with Transformer

Weijia Wu
Zhejiang University

Yuanqiang Cai
Chinese Academy of Sciences

Debing Zhang
Kuaishou Technology

Sibo Wang
Kuaishou Technology

Zhuang Li
Kuaishou Technology

Jiahong Li
Kuaishou Technology

Yejun Tang
Kuaishou Technology

Hong Zhou
Zhejiang University

Abstract

Most existing video text spotting benchmarks focus on evaluating a single language and scenario with limited data. In this work, we introduce a large-scale, **B**ilingual, **O**pen World **V**ideo text benchmark dataset(BOVText). There are four features for BOVText. Firstly, we provide **1,850+** videos with more than **1,650,000+** frames, **25** times larger than the existing largest dataset with incidental text in videos. Secondly, our dataset covers **30+** open categories with a wide selection of various scenarios, *e.g., Life Vlog, Driving, Movie, etc*. Thirdly, abundant text types annotation (*i.e., title, caption or scene text*) are provided for the different representational meanings in video. Fourthly, the BOVText provides bilingual text annotation to promote multiple cultures' live and communication.

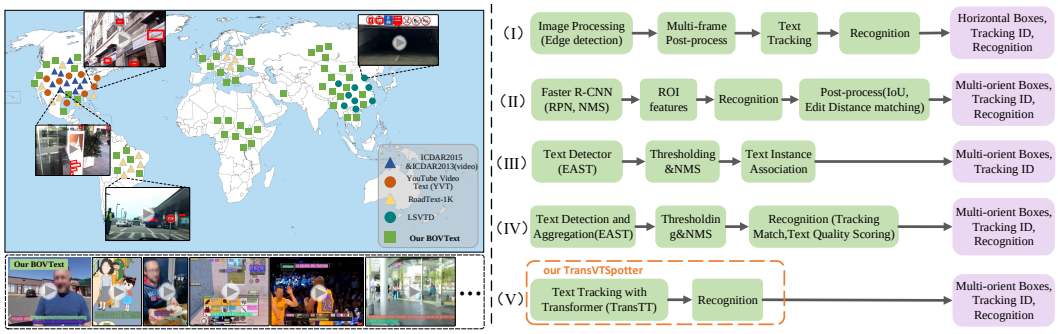
Besides, we propose an end-to-end video text spotting framework with Transformer, termed TransVTSpotter, which solves the multi-orient text spotting in video with a simple, but efficient attention-based query-key mechanism. It applies object features from the previous frame as a tracking query for the current frame and introduces a rotation angle prediction to fit the multi-orient text instance. On ICDAR2015(video), TransVTSpotter achieves the state-of-the-art performance with **44.2%** MOTA, **13** fps. The dataset and code of TransVTSpotter can be found at github.com/weijiawu/BOVText and github.com/weijiawu/TransVTSpotter, respectively.

1 Introduction

Text spotting [23, 15] has received increasing attention due to its numerous applications in computer vision, *e.g.,* document analysis, image-based translation, image retrieval [36, 28], etc. With the advent of deep learning and abundance in digital data, reading text from images has made extraordinary progress in recent years with a lot of great public datasets [11, 16, 6] and algorithms [47, 61, 26, 22]. By contrast, video text spotting almost remains at a standstill for the lack of large-scale multidimensional practical datasets, which limited numerous applications of video text, *e.g.,* video understanding [39], video retrieval [8], video text translation, and license plate recognition [1], etc.

Video text spotting(VTS) is the task that requires simultaneously classifying, detecting, tracking and recognizing text instances in a video sequence. There have been a few previous works [52, 50] and datasets [30, 17] in the community for attempting to develop video text spotting. ICDAR2015 (Text

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.



(a) Comparison of different data distributions (b) Comparison of different pipelines for video text spotting

Figure 1: **Comparison of dataset distribution and pipeline**. (a) Data distribution. BOVText provides various open world scenarios with unique *NBA, Game, etc.* (b) Pipelines: (I) Multi-stage pipeline in [51], [13], [53] *etc*; (II) Multi-orient video text spotting pipeline with Faster R-CNN [32] proposed by Wang *et al.* [48]; (III) Online text tracking pipeline in Yu *et al.* [56]; (IV) Fast video text spotting pipeline in Cheng *et al.* [5]; (V) An end-to-end pipeline with transformer in this work.

in Videos) [16] was introduced during the ICDAR Robust Reading Competition in 2015 and mainly includes a training set of 25 videos (13k frames) and a test set of 24 videos (14k frames). The videos were categorized into seven scenarios: walking outdoors, searching for a shop in a shopping street, etc. YouTube Video Text (YVT) [30] dataset harvested from YouTube, contains 30 videos with 13k frames, 15 for training, and 15 for testing. The text content in the dataset mainly includes overlay text and scene text (*e.g.,* street signs, business signs, words on shirt). RoadText-1K [31] are sampled from BDD100K [55], includes 700 videos (210k frames) for training and 300 videos for testing. The texts are all obtained from driving videos and match for driver assistance and self-driving systems. LSVTD [5] includes 100 text videos, 13 indoor (*e.g.,* bookstore, shopping mall) and 9 outdoor (*e.g.,* highway, city road) scenarios. However, as shown in Figure. 1 (a), most existing video text datasets are limited by the amount of training data (less than 300k frames), single video scenarios, and a single language. There are only a few outdoor scene text videos with 13k frames in ICDAR2015 (video). Similar situation for YVT, RoadText-1k, and LSVTD, the training set is limited and the dataset scenarios are single. This makes it difficult to evaluate the effectiveness of more advanced deep learning models for more open scenarios, such as *game, sport and news report*. Besides, most existing video text datasets are proposed before 2019 years, and some of them are no longer being maintained without an open-source evaluation script.

In this work, we contribute a large-scale, bilingual open-world benchmark dataset (BOVText) to the community for developing and testing video text spotting that can fare in a realistic setting. Our dataset has several advantages. **Firstly**, the large training set (*i.e.,* 1,850+video and 1,650,000+ video frames) from *KuaiShou* and *YouTube* enables the development of deep design specific for video text spotting. **Secondly**, unlike the existing datasets, BOVText support **30+** open scenarios, including many new scenarios such as *Sportscast(NBA, FIFA World Cup...), Life Vlog, Game, etc*, as shown in Figure. 1 (a). These data is collected from the worldwide user of *YouTube*[1] and *KuaiShou*[2], cover various daily scenarios without region limitation and virtual scenes. But the previous video text datasets usually are collected toward a special city or language from the hand-held camcorder. **Thirdly**, BOVText is the first benchmark for supporting abundant text types annotation. Caption, title, and scene text are separately tagged for the different representational meanings in the video. This made our BOVText has the potential to promote other video-and-language tasks, such as video understanding and video retrieval. **Fourthly,** bilingual text annotation(*i.e.,* Chinese, English) is provided in BOVText to promote multiple cultures' live and communication.

Except for the promising benchmark, we also proposed a simple, but effective video text spotter with transformer (TransVTSpotter). As shown in Figure. 1 (b), unlike previous methods that involve multiple steps, such as proposal generation, text aggregation, and post-processing(NMS), TransVTSpotter only requires two steps. 1) Text tracking: for each consecutive frame image, we obtain the multi-orient boxes tracking trajectories of text by boxes IoU matching between the predicted detection boxes [4] and the predicted tracking boxes [40], where the detection boxes are obtained by taking an object query as input, just like DETR [4]. And features from previously detected objects to form another

[1]https://www.youtube.com/

[2]https://www.kuaishou.com/en

"track query" to discover associated objects (*i.e.,* the predicted tracking boxes) on the current frames. Besides, an additional angle loss of multi-orient box and *Hungarian angle cost* are design to obtain the angle of multi-orient. 2) Text recognition: recognizing the tracked texts with attention-based text recognizer [25]. Without bells and whistles, TransVTSpotter achieves state-of-the-art performance on ICDAR2015 with **44.2%** MOTA, **13** fps. The main contributions of this work are three folds:

(1) We propose a large-scale, bilingual and open world video text spotting benchmark named BOVText. The proposed dataset provides **1,850+** videos, **1,650,000** frames, open videos scenarios (*e.g., Indoor, Outdoor, Game, Sport*), abundant text types (*i.e., title, caption or scene text*), multi-stage tasks and is **25** times the existing largest dataset with incidental text.

(2) We first propose a new video text spotting framework with Transformer, termed **TransVTSpotter**, which solves the video multi-orient text spotting with a simple, but effective pipeline based on the tracking query-key mechanism and rotated boxes angle prediction.

(3) We evaluate and compare TransVTSpotter and other techniques for scene text detection, recognition, text tracking, and end-to-end video text spotting on BOVText and other existing datasets. Besides, a thorough analysis of performance on the proposed dataset is provided.

2 Related Work

2.1 End-to-End Text Spotting

For image-level text spotting, various methods [19, 12, 26] based on deep learning have been proposed and have improved the performance considerably. Li *et al.* [19] proposed the first end-to-end trainable scene text spotting method. The method successfully uses a RoI Pooling [32] to joint detection and recognition features. Liao *et al.* [26] propose a Mask TextSpotter which subtly refines Mask R-CNN and uses character-level supervision to detect and recognize characters simultaneously. Compared to text spotting in a static image, video text spotting methods are rare. Yin *et al.* [54] provides a detailed survey, summarizes text detection, tracking and recognition methods in video. Wang *et al.* [48] introduced an end-to-end video text recognition method through associations of texts in the current frame and several previous frames to obtain final results. Cheng *et al.* [5] propose a video text spotting framework by only recognizing the localized text one-time. Nguyen *et al.* [30] improves detection and recognition performance by temporal redundancy and linearly interpolate to recover missing detection results. Rong *et al.* [35] tracked video text using tracking-by-detection. An MSER detector was used to locate scene text character, which was used as a constraint to optimize the trajectory search. To promote video text spotting, we attempt to establish a standardized benchmark (BOVText), covering various open scenarios and bilingual text annotation.

2.2 Text Spotting Datasets for Images and Videos

The various and practical benchmark datasets [16, 42, 17, 6, 17] contribute to the huge success of scene text detection and recognition at the image level. ICDAR2015 [16] was provided from the ICDAR2015 Robust Reading Competition. Google glasses capture these images without taking care of position, so text in the scene can be in arbitrary orientations. ICDAR2017MLT [29] is a large-scale bilingual text dataset, which is composed of complete scene images which come from 9 languages. The COCO-Text dataset [42] is currently the largest dataset for scene text detection and recognition. It contains 50,000+ images for training and testing.

The development of video text spotting is limited in recent years due to the lack of efficient data sets. ICDAR 2015 Video [17] consists of 28 videos lasting from 10 seconds to 1 minute in indoors or outdoors scenarios. Limited videos (*i.e.,* 13 videos) used for training and 15 for testing. Minetto Dataset [27] consists of 5 videos in outdoor scenes. The frame size is 640 x 480 and all videos are used for testing. YVT [30] contains 30 videos, 15 for training and 15 for testing. Different from the above two datasets, it contains web videos except for scene videos. USTB-VidTEXT [52] with only five videos mostly contain born-digital text sourced from Youtube. RoadText-1K [31] provides a driving videos dataset with 1000 videos. The 10-second long video clips in the dataset are sampled from BDD100K [55]. As shown in Table. 1, the existing datasets contain a limited training set and single video scenarios. To promote the development of video text spotting, we create a large-scale, bilingual open-world benchmark dataset.

(a) Open Category Distribution of BOVText

(b) Caption Text and Scene Text

Figure 2: **Distributions of BOVText**. (a) The benchmark dataset covers a wide and open range of life scenes (32 open-domain categories). (b) Caption text (blue box) and scene text (red box) are distinguished in BOVText, which is favorable for downstream tasks.

2.3 Transformers in Vision

Transformer is first proposed in [41] as a new paradigm for machine translation. But recently, there is a popularity of using transformer architecture in vision tasks, such as detection [4, 63], segmentation [60], 3D data processing [59], video object tracking [45] and even backbone construction [9]. Lately, some works [49, 58, 45] show using a transformer in processing sequential visual data also make remarkable shots. MOTR [58] introduce the concept of track query and the contiguous query passing mechanism for multiple-object Tracking. VisTR [49] solves instance segmentation by learning the pixel-level similarity and instance tracking is to learn the similarity between instances. But for video text spotting or multi-orient text tracking, to the best of our knowledge, there are still no transformer-based solutions while it is intuitive for its good capacity in temporal processing. Here, we propose the TransVTSpotter method and provide an affirmative answer to that, which shows convincingly high performance on the popular benchmark.

3 BOVText Benchmark

3.1 Data Collection and Annotation

Data Collection. To obtain abundant videos with various text types, we first start by acquiring a large list of video scenario category (*e.g., game scenario*, *travel scenario*) using *YouTube*[3] and *KuaiShou*[4] - an online resource that contains billions of videos with various scene text from cartoon movies to human relation. Then, we choose 31 open-domain categories with 1 unknown category, *i.e.,* , *Game*, *Home*, *Fashion*, and *Technology*, as shown in Figure. 2 (a). With each raw video category, we first choose the video clips with text, then make two rounds of manual screening to remove the ordinary videos without scene text and caption text. As a result, we obtain 1852 videos with $1,620,305$ video frames, as shown in Table 1. Finally, to fair evaluation, we divide the dataset into two parts: the training set with $1,124,213$ frames from 1297 videos, and the testing set with $496,091$ frames from 555 videos. As shown in Figure 2, different from the existing data sets, our dataset not only cares about scene text spotting in the real world, but also focuses on caption texts in the video. For the most part, caption text represents more global information than scene text, which is quite favorable for some downstream tasks, *e.g., video understanding, video caption translation, etc.*

Data Annotation. We invite a professional annotation team to label each video text with four kinds of description information: the rotated bounding box describing the location information, judging the tracking identification(ID) of the same text, identifying the content of the text information, and distinguishing the category of the caption, title or scene text. To save the annotation cost, we first sample the videos, annotate each sampled video frame, and then transform the annotation information from the sampled video frame to the unlabeled video frame by interpolation. Finally, we invite an audit team to carry out another round of annotation checks, and re-label part video frames with unqualified annotation. *For video sampling*, we use uniform sampling with a sampling

[3]https://www.youtube.com/

[4]https://www.kuaishou.com/en

Table 1: **Statistical Comparison.** 'D','R', 'T', 'S' and 'BI' denotes the Detection, Recognition, Tracking, Spotting and bi-lingual text, respectively. 'Incidental' denotes indoor and outdoor scenarios in daily life (*e.g., walking outdoors, driving)*. 'Open' refers to any scenarios, e.g., *Game, Sport(NBA).* In green refers to these scenarios only supported by BOVText.

Dataset	Category	BI	Task	Videos	Frames	Texts	Supported Scenario
AcTiV-D[57]	Caption	-	D	8	1,843	5,133	News video
UCAS-STLData[3]	Caption	-	D	3	57,070	41,195	Teleplay
USTB-VidTEXT[52]	Caption	-	D&S	5	27,670	41,932	Web video
YVT[30]	Scene, Caption	-	D&R&T&S	30	13,500	16,620	Incidental: Cartoon, Outdoor(supermarket, shopping street, driving...)
ICDAR2015 VT[62]	Scene	-	D&R&T&S	51	27,824	143,588	Incidental: Outdoor(walking, driving, supermarket, shopping street...)
LSVTD[5]	Scene	✓	D&R&T&S	100	66,700	569,300	Incidental: Indoor(shopping mall, supermarket, hotel...), Outdoor(driving...)
RoadText-1K[31]	Scene	-	D&R&T&S	1,000	300,000	1,280,613	Driving
BOVText(ours)	Scene, Caption, Title, Others	✓	D&R&T&S	**1,852**	**1,670,305**	**7,292,261**	Open: Cartoon, Vlog(supermarket, shopping street, driving), Travel (indoor and outdoor), Game(PUBG mobile...), Sport(NBA, world cup...), News , TV program, Education(campus, classroom, book...),Technology(introductory video, scientific propaganda...)...

frequency of 3 to sample all the videos in the dataset, and obtain the sampled video frame set. *For sampling video frame annotation*, each text instance is labeled in the same quadrilateral way as in the ICDAR2015 [62]. In addition, the text instance also will be marked with two description information: the category of the caption, title or scene, the recognition content, and the tracking ID. *For interpolation on unlabeled video frames*, each text instance is marked with tracking ID and recognition content, so we can judge whether the texts in adjacent sampling frames are the same text with the same ID. For the same text instance, we first determine whether the text annotation of the sampled video frame is the starting and end frame of the text instance. If not, we look forward and backward for the starting and end position of the text instance and label it. Then we use the linear interpolation way to calculate the position of the text object in the middle of the unmarked video frame, and give tracking ID, recognition content, and category. *For check and re-label bad cases,* the linear interpolation shows a dissatisfied performance in some cases, *e.g., the new text appears on starting frame, text suddenly disappears on ending frame,* which are difficult to capture. Therefore, we invite an audit team to carry out another round of annotation checks. Around $150,000$ video frames with unqualified annotation from $1,670,305$ video frames are selected to refine, taking 20 men in three weeks. As a labor-intensive job, the whole labeling process takes **30** men in three months, *i.e.,* **21,600** man-hours, to complete about **600,000** sampled video frame annotations.

3.2 Dataset Analysis

The statistic comparison between BOVText and other datasets are visualized in Figure 1 (a), YouTube, and summarized in Table 1. Besides, we provide more detailed information in supplementary material, such as 'data distribution for 32 open scenarios', 'text language and category distribution'.

To provide the community with unified text-level quantitative descriptions, we will compare our dataset with the previous datasets from four aspects, *i.e.,* text description, video scene, dataset size, and supported tasks. *For text description attribute* (*i.e., Category, MLingual*), our BOVText supports four types of text annotations(caption, title, scene, and other text) of video text with multi-language, which obviously has more extensive description ability than the existing dataset. *For video scene attribute* (*i.e., Scenario*), unlike the existing datasets, BOVText provides various open-world scenarios, including many new scenarios such as *Sportscast(NBA, FIFA World Cup...), Life Vlog, Game, etc.* We present the 31 open scenarios and an "Unknown" scenarios distribution on BOVText in three levels, *i.e.,* video, video frames, and text instances, as shown in Table 2. BOVText spans various video domains with these scenarios in the existing datasets (*e.g.,* Driving for RoadText-1k [31], Vlog(supermarket, shopping street, indoor), Travel(hotel, railway station) for LSVTD [5]) and more open domains that are not yet supported (*e.g.,* Game(PUBG mobile, Honor of Kings...), Sport(NBA, world cup...), News). *For the size of the dataset*(*i.e., Videos, Frames, Texts*), BOVText is **25** times larger than the existing largest dataset (*i.e.,* LSVTD [5]) with various scenarios($1,620,305$ v.s $66,700$ video frames). RoadText1k [31] contains 300k videos frames, but the supported scenario is too single for only supports driving video scenarios. *For the supported tasks*, the proposed BOVText supports all video text tasks: detection, recognition, video text tracking, end to end video text spotting. For comprehensive research, we not only focus the scale, location, recognition content, and tracking ID, but also additionally collect and annotate the category of caption, title, scene or other texts for

Table 2: **The Data Distribution for 32 Open Scenarios**. In green refers to these scenarios only supported by BOVText. "%" denotes the percentage of each scenario data for whole data.

Scenarios	Video	Video Frames	Text Instances	Scenarios	Video	Video Frames	Text Instances
Cartoon	61(3.2%)	60,395(3.6%)	123,191(2.1%)	Sport	91(4.8%)	72,469(4.3%)	266,996(4.6%)
Vlog	81(4.2%)	76,056(4.5%)	214,910(3.7%)	News Report	79(4.1%)	48,868(3.0%)	178,000(3.1%)
Driving	72(3.8%)	61,656(3.7%)	151,994(2.6%)	Celebrity	46(2.4%)	39,440(2.3%)	121,235(2.1%)
Advertising	33(1.7%)	29,329(1.0%)	91,090(1.0%)	Technology	59(3.1%)	52,072(3.1%)	140,172(2.4%)
Activity	31(1.6%)	23,585(1.4%)	67,879(1.2%)	Program	43(2.3%)	40,784(2.4%)	214,561(3.7%)
Comedy	86(4.5%)	79,404(4.7%)	317,865(5.5%)	Game	8(1.0%)	12,565(1.0%)	84,106(1.5%)
Interview	24(1.3%)	18,229(1.1%)	63,616(1.1%)	Livestreaming	59(3.1%)	60,494(3.6%)	211,569(3.6%)
Government	47(2.5%)	32,283(1.9%)	93,874(1.6%)	Speech	59(3.1%)	52,465(3.1%)	175,119(3.0%)
Travel	83(4.3%)	75,266(4.5%)	280,446(4.8%)	Movie	107(5.6%)	105,949(6.3%)	299,760(5.2%)
Campus	44(2.3%)	37,556(2.2%)	139,760(2.4%)	Photograph	53(2.8%)	52,771(3.1%)	173,832(3.0%)
International	60(3.1%)	60,486(3.6%)	132,117(2.3%)	Education	73(3.8%)	60,315(3.6%)	360,774(6.2%)
Short Video	84(4.4%)	79,148(4.7%)	326,930(5.6%)	Dance	37(1.9%)	22,264(1.3%)	71,740(1.2%)
Makeup	60(3.1%)	52,449(3.1%)	111,814(1.9%)	Fishery	86(4.5%)	79,750(4.7%)	230,085(4.0%)
Talent	79(4.1%)	66,038(3.9%)	339,382(5.9%)	Fashion	57(3.0%)	42,868(2.6%)	98,942(1.7%)
Beauty Industry	37(1.9%)	35,851(2.1%)	132,025(2.3%)	Introduction	73(3.8%)	42,086(4.2%)	236,721(4.1%)
Eating	52(2.7%)	65,609(3.9%)	191,035(3.3%)	Unknown	48(2.5%)	29,212(1.7%)	150,721(2.6%)

each text instance. As shown in Figure. 2 (b), in a video, different types of text instances may exist simultaneously, and they are helpful to understand videos synergistically. Caption text can directly show the dialogue between people in video scenes and represent the time or topic of the video scenes, scene text can unambiguously define the object and can identify important localization and road paths in video scenes. Therefore, the text category annotation is favoring downstream tasks (*e.g.,* video text translation, video understanding, and video retrieval), more details in the supplementary material.

3.3 Supported Tasks and Metrics

The proposed BOVText supports four tasks: (1) Text Detection. (2) Text Recognition. (3) Video Text Tracking. (4) End to End Text Spotting in Videos. Following ICDAR2015 [62], the evaluation protocols [44] are used for text detection and recognition task. For video text tracking and spotting task, the existing video text datasets such as ICDAR2015 (video) [17] and RoadText-1k [31] all adopted the MOT metrics (*i.e.,* Multiple Object Tracking Accuracy ($MOTA$) and Multiple Object Tracking Precision ($MOTP$)). However, there are two sets of measures for Multiple Object Tracking: the MOT metrics ($MOTA$,$MOTP$) [2] and ID metrics (ID_{F1}) [20, 34]. The CVPR19 MOTChallenge evaluation framework [7] presents that different measures serve different purposes. $Event\text{-}based$ measures like MOT help pinpoint the source of some errors and are thereby informative for the designer of certain system components. $Identity\text{-}based$ measure(ID_{F1}) is more favorable for evaluating how well computed identities conform to true identities. Except for using MOTA, MOTP, $Identity\text{-}based$ measures(ID_{F1}), as a new metric is adopted firstly for video text spotting task. More detailed information for metric can be obtained in the supplementary material.

3.4 Our Method: TransVTSpotter

Two ingredients are essential for direct text spotting for TransVTSpotter: (1) A set prediction loss that forces unique matching between predicted and ground truth multi-orient boxes. (2) An architecture that predicts a set of objects and associates the same objects during different frames.

Multi-orient Box Matching. Compare to DETR [4], the difference is that we propose an angle prediction and corresponding loss while only horizontal boxes prediction for DETR. Let us denote the ground truth set of objects by y, and $\hat{y} = \{\hat{y}_i\}_{i=1}^{N}$ the set of N predictions. y is as a set of size N padded with $\varnothing$ (no object). To find a bipartite matching between these two sets we search for a permutation of N elements $\sigma \in \mathfrak{S}_N$ with the lowest cost:

$$\hat{\sigma} = \underset{\sigma \in \mathfrak{S}_N}{\arg\min} \sum_{i}^{N} \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)}), \tag{1}$$

where $\mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$ is a pair-wise *matching cost* between ground truth y_i and a prediction with index $\sigma(i)$. The matching cost takes into account the class prediction, boxes prediction and the boxes rotated angle prediction. Each element i of the ground truth set can be seen as a $y_i = (c_i, b_i, a_i)$ where c_i is the target class label, $b_i \in [0, 1]^4$ is a vector that defines ground truth box center coordinates and its height and width relative to the image size, and a_i is rotation angle between the longest edge of ground truth multi-orient box and horizontal line (x-axis). For the prediction with index $\sigma(i)$ we define

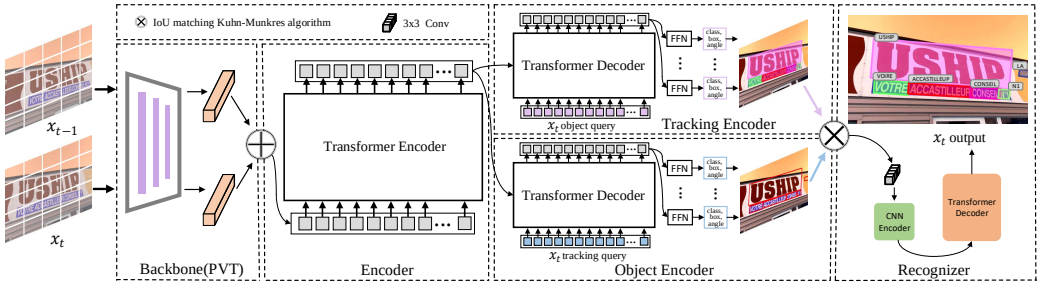


Figure 3: **Pipeline of TransVTSpotter**. It contains four main components: 1) A transformer backbone(PVT [46]) extracts feature representation of multiple images; 2) A transformer encoder models the relations of pixel-level features; 3) Two transformer decoders (shared weight) decode the instance-level features; 4) Attention-based recognizer [25] recognizes each text instance.

probability of class c_i as $\hat{p}_{\sigma(i)}(c_i)$, the predicted box as $\hat{b}_{\sigma(i)}$, and the predicted angle as $\hat{a}_{\sigma(i)}$. Thus we define $\mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$ as $-\mathbb{1}_{\{c_i \neq \varnothing\}}\hat{p}_{\sigma(i)}(c_i) + \mathbb{1}_{\{c_i \neq \varnothing\}}\mathcal{L}_{\text{box}}(b_i, \hat{b}_{\sigma(i)}) + \mathbb{1}_{\{c_i \neq \varnothing\}}\mathcal{L}_{\text{angle}}(a_i, \hat{a}_{\sigma(i)})$. Finally, we could compute the loss function with all pairs matched:

$$\mathcal{L}_{\text{Hungarian}}(y, \hat{y}) = \sum_{i=1}^{N} \left[-\log \hat{p}_{\hat{\sigma}(i)}(c_i) + \mathbb{1}_{\{c_i \neq \varnothing\}}\mathcal{L}_{\text{box}}(b_i, \hat{b}_{\hat{\sigma}}(i) + \mathbb{1}_{\{c_i \neq \varnothing\}}\mathcal{L}_{\text{angle}}(a_i, \hat{a}_{\sigma(i)})) \right], \tag{2}$$

where $\mathcal{L}_{\text{box}}(\cdot)$ a linear combination of the ℓ_1 loss and the generalized IoU loss [33, 4]. And $\mathcal{L}_{\text{angle}}(\cdot)$ refers to a cosine embedding loss with $1 - cos(\hat{a}_{\sigma(i)} - a_i)$.

TransVTSpotter architecture. The overall TransVTSpotter architecture is surprisingly simple and depicted in Figure. 3. A transformer-based backbone [46] is used to extract feature representation, transformer-based encoder-decoder framework learned current object query and previous frame tracking query as input and predicts *detection boxes* and *tracking boxes* [40]. With the detection boxes and tracking boxes, box IoU matching is used to obtain the final tracking result. Finally, attention-based recognizer [25] is utilized to obtain the final recognition results. **Text Tracking with Transformer.** Text Tracking with Transformer includes three components: backbone, transformer encoder and transformer decoder, as shown in Figure. 3. *Backbone*. Starting from the two consecutive frames $x_{\text{t}} \in \mathbb{R}^{3\times H_0 \times W_0}$ and $x_{\text{t-1}} \in \mathbb{R}^{3\times H_0 \times W_0}$, a transformer backbone [46] generates a lower-resolution activation map for the two frames($f_{\text{t}} \in \mathbb{R}^{C\times H\times W}$ and $f_{\text{t-1}} \in \mathbb{R}^{C\times H\times W}$), then a new feature sequence f_{t}^{*} can be obtained by simple concatenating f_{t} and $f_{\text{t-1}}$. The extracted features f_{t} of the current frame are temporarily saved and then re-used for the next frame. *Transformer Encoder.* We adopted deformable transformer encoder [63] to model the similarities among all the pixel level features for the extracted features f_{t}. *Transformer Decoder.* Two parallel decoders [40] are employed. The object decoder takes learned object query [4] as input and predicts detection rotated boxes. The tracking decoder takes the object feature from previous frames as input and predicts the corresponding tracking rotated boxes. Finally, with detection rotated boxes and tracking rotated boxes, TransTT obtains the final tracking result by box IoU matching and the Kuhn-Munkres(KM) algorithm [18]. **Recognizer.** Following MASTER [25] is utilized to predict output sequence with 2D-attention.

4 Experimental

In this section, we mainly conduct experiments on our BOVText. More experiments, such as the performance of TransVTSpotter in other datasets, would be provided in the supplementary material.

4.1 Implementation Details

BOVText. Except for the TransVTSpotter, we also adopt CRNN [37], RARE [38] as the recognition baseline and PSENet [47], EAST [61], DB [21] as the detection baseline to evaluate our BOVText. *Detection*: we train detectors directly with training set (*i.e.,* 641,049 frame images) of BOVText. *Recognition*: the network is pre-trained on the *chinese ocr*[5] and MJSynth [14], then fine-tuned on our BOVText. All of our experiments are conducted on 8 V100 GPUs. In the PSENet, EAST, DB, CRNN

[5]https://github.com/YCG09/chinese_ocr

Table 3: **Attribute Experiments for Scenarios.** In green are the gaps compare to training with RoadText-1k [31] and LSVTD [5].

Method	Training Set	Detection (F-score/%) on BOVText					Tracking (ID_{F1}/%) on BOVText				
		Cartoon	Travel	Game	Driving	Avg.	Cartoon	Travel	Game	Driving	Avg.
TransVTSpotter	LSVTD	54.3	42.7	40.5	70.2	59.8	42.5	25.3	35.5	66.9	41.6
	RoadText	13.6	23.8	10.3	13.0	17.7	3.2	1.6	4.2	3.5	2.4
	LSVTD&RoadText	56.2	45.1	40.1	72.1	61.6	40.2	23.5	35.2	68.8	43.2
	BOVText	**93.9**	**65.4**	**67.3**	**82.1**	**78.5**	**90.2**(+47.7)	**24.1**(+0.6)	**61.2**(+26.0)	**73.2**(+4.4)	**58.2**(+15.0)

and RARE experiments, all settings follow the original reports. **TransVTSpotter**. AdamW [24] as the optimizer and the batch size is set to be 16. The initial learning rate is 2e-4 for the transformer and 2e-5 for the backbone. The weight decay is 1e-4 All transformer weights are initialized with Xavier-init [10]. The data augmentation includes random horizontal, scale augmentation, resizing the input images whose shorter side is by 480-800 pixels while the longer side is by at most 1333 pixels. The model is first pre-trained on COCOText [42] and then fine-tuned on other video text training sets. For each iteration, two adjacent frames are randomly selected from one video from training set to train our model.

4.2 Attribute Experiments Analysis for BOVText

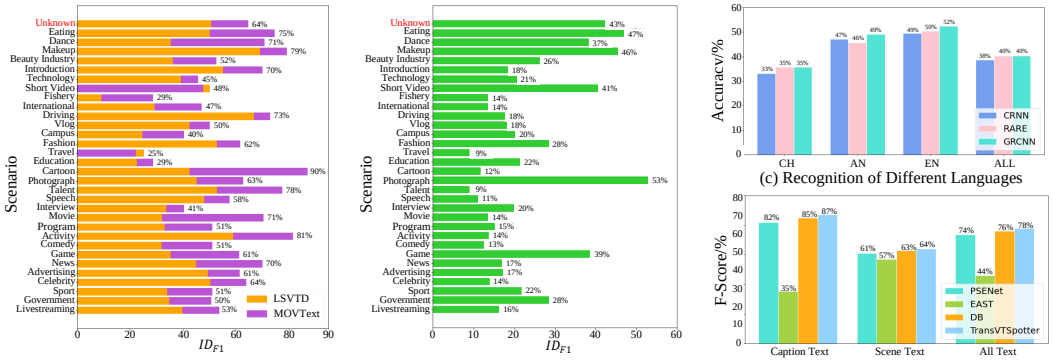
New Scenarios, New Challenge for Video Text Tasks. Figure. 4 (a) and Table. 3 gives the tracking performance ID_{F1} of TransVTSpotter in different scenarios of BOVText. Two new insights can be present from the figure and table: 1) The existing benchmark datasets cannot effectively test the effectiveness of advancing algorithm on some novel scenarios (*e.g.,* Game, Cartoon) for first proposed in BOVText. LSVTD [5] and RoadText-1k [31], as the two largest data sets on the existing video text datasets are used to compare with our BOVText, TransVTSpotter achieves a tracking performance $\mathrm{ID_{F1}}$ of 61.2% on $Game$ scenario with BOVText training set, 26 percent point improvement than training with LSVTD [5] and RoadText-1k [31]. We argue that the main cause is that there existing a mass of texts in $Game$ scenario, but LSVTD and RoadText almost no such dense text scenario, which is a new challenge for algorithms. A visualization example in YouTube is used to support the idea. Besides, training with only RoadText-1k obtains a low performance no matter which scenarios. There are two main causes for this. Firstly, the location annotation of RoadText-1k is an upright bounding box(two points), but the counterpart of BOVText is multi-orient boxes(four points). Secondly, the data domain is entirely different for the two datasets. Compare with various scenarios (*e.g.,* Game, Sports) and text types (*e.g.,* long caption text, big text), the scenario of RoadText-1k only contains small and low-resolution road signs, plate number on driving scenarios. 2) Huge performance gap existing during different scenarios. As shown in Figure. 4 (a), the model achieves the best performance with a ID_{F1} of 90.2% in $Cartoon$ videos, since the conspicuous text instances and simple background are designed in cartoon videos. By comparison, several scene categories obtain extremely dissatisfied performance due to complex background, various text appearance, and unsteady camera movements, such as $Campus$ of 40.2% and $Travel$ of 24.1%.

Bilingual Recognition, New Challenge. As shown in Figure. 4 (c), the text recognition results for different languages are provided. In summary, the alphanumeric recognition result (about 47%) is better than the Chinese recognition result (about 35%), regardless of the models. The final results (about 40%) for all characters are satisfactory, can not meet the requirement of the application. Unlike English, Chinese contains thousands of characters($3,856$ Chinese characters v.s. 26 English characters on BOVText), which are difficult to recognize.

Long Caption Text, New Challenge. As shown in Figure. 4 (d), for DB [21], PSENet [47] and our TransVTSpotter, the performance of caption text is better than the counterpart of scene text (around 80% vs. 60%) due to the more clear and bigger caption text. But for EAST [61], long caption text show a low performance with 35% F-score. The prime reason is that caption texts are all long text (average width-height ratio: 6.8 for caption $v.s.$ 2.3 for scene text on BOVText), a different case for EAST [61], as shown in YouTubeDemo. However, the existing video text datasets hardly contain long caption texts, our BOVText can fill out the gap for a more comprehensive evaluation of text types.

4.3 Text Detection, Recognition, Tracking and Spotting on BOVText

Video Text Detection and Recognition. As shown in Table. 4, image-based text detection on BOVText is not unsatisfactory, with lower results than these methods report on existing text datasets.



(a) Tracking of Different Scenario (b) Video Text Spotting of Different Scenario (c) Recognition of Different Languages (d) Detection of Different Text Category

Figure 4: **Attribute Experiments of BOVText**. (a) Tracking performance (*i.e.,* ID_{F1}) with TransVTSpotter in different scenarios, 'LSVTD' and 'BOVText' denotes training on LSVTD and BOVText, respectively. (b) End to end video text spotting performance (*i.e.,* ID_{F1}) with TransVTSpotter in different scenarios. (c) Recognition accuracy of different models in different languages. (d) Detection of different model in caption or scene text. 'CH', 'AN', 'EN' and 'ALL' refer to 'Chinese', 'Alphanumeric', 'English' and 'All Characters', respectively.

Table 4: **Detection and Recognition Performance on BOVText**.

Detection Performance/%				Recognition Performance/%								
Method	Precision	Recall	F-score	Method	Pretrained				Fine tuned			
					Chinese	Alphanumeric	English	All	Chinese	Alphanumeric	English	All
EAST [61]	52.2	38.1	44.1	CRNN [37]	26.0	32.1	36.1	23.2	33.2	47.1	49.5	38.6
PSENet [47]	74.3	73.2	73.5	RARE [38]	25.2	34.2	37.4	23.5	35.6	45.7	50.4	40.2
DB [21]	77.2	74.5	75.3	GRCNN [43]	23.1	39.8	40.4	26.7	35.6	49.2	52.4	40.3
TransVTSpotter	**80.2**	**76.4**	**78.5**	-	26.2	40.3	42.1	29.1	36.2	48.9	52.1	40.4

For example, EAST obtains an f-score of 44.1% compared to the F-score of 80.7% on icdar2015 [62], but our TransVTSpotter obtain an f-score of 78.5% on BOVText, at least 3% improvement compare to the image-based detectors (*i.e.,* DB, PSENet and EAST). For text recognition, CRNN [37] based on CTC loss, RARE [38] with attention mechanism and GRCNN [43] as the base text recognizers to test our BOVText. The text annotation in our BOVText covers two languages (*i.e.,* English and Chinese), thus we conduct several experiments for each language. The recognition model only yields about 40% accuracy on our dataset, but the same model reports at least 90% on most benchmark datasets [17] for text recognition. The main reasons have two points: (1) The proposed BOVText is bilingual, and the category number of Chinese characters in real-world is much larger than those of Latin languages. (2) The video texts are quite blurred, out-of-focus, and the distribution of characters is relatively smaller than the static image counterparts, which presents more challenges.

Video Text Tracking. As shown in Table. 5, $\mathrm{ID_{F1}}$ (58.2%) of our TransVTSpotter achieves the best performance, at least $20\%+$ improvement than other methods. Besides, without NMS and other post-processing, TransVTSpotter presents 13 fps no matter which dataset. More details and analysis concerning TransVTSpotter can be obtained in the supplementary material. And EAST shows the worst performance with a $\mathrm{ID_{F1}}$ of 23.2%. The $\mathrm{IDF_1}$ of EAST [61] is lower with 6.7% gap than that of PSENet [47]. The main reason is that MOVTtext contains a mass of long text instances, but regression-based EAST can not deal with the long text cases well. The performance of DB is similar to that of PSENet for both all are the segmentation-based methods.

End to End Text Spotting in Video. Detection and text tracking tasks are paving the way for the recognition task. Table. 5 shows the performance of text spotting on BOVText. Similar to the text tracking performance, our TransVTSpotter achieves the state-of-the-art performance with at least 6.2% $\mathrm{ID_{F1}}$ improvement compared to the other methods. Besides, the MOTP of TransVTSpotter achieves 0.781%, two percent points improvement than the counterpart of using DB and RARE. The great performance for 6.2% $\mathrm{ID_{F1}}$ and 0.781% MOTP present satisfactory tracking and recognition trajectory and detection results, respectively. The corresponding performance using EAST [61] as the detector in video text spotting is still not satisfied with around 5% $\mathrm{ID_{F1}}$ and -0.8 MOTA. Without TransVTSpotter, the combination of DB [21] and RARE [38] achieves the best performance with a 17.3% $\mathrm{ID_{F1}}$, but there is at least 6.2% gap compare to our method.

Table 5: **Text Tracking and End to End Video Text Spotting Performance on BOVText.** Text tracking trajectory id generation use a method proposed in [48]. In green is at least 2% improvement.

Method		Text Tracking on BOVText					End to End Text Spotting on BOVText				
Detection	Recognition	ID_P/%	ID_R/%	ID_{F1}/%	MOTA	MOTP	ID_P/%	ID_R/%	ID_{F1}/%	MOTA	MOTP
EAST [61]	CRNN [37]	23.5	22.9	23.2	-0.301	0.725	5.3	5.1	5.2	-0.835	0.743
	RARE [38]						3.0	3.6	3.2	-1.130	0.732
PSENet [47]	CRNN [37]	34.7	26.7	29.9	0.334	0.753	14.7	9.8	11.8	-0.300	0.752
	RARE [38]						15.2	10.4	12.4	-0.280	0.762
DB [21]	CRNN [37]	33.7	29.9	31.7	0.438	0.765	15.6	9.6	11.9	-0.284	0.760
	RARE [38]						20.1	15.2	17.3	-0.293	0.762
TransVTSpotter(ours)		**65.6**	**52.2**	**58.2**	**0.682**	**0.771**	**25.3**	**22.0**	**23.5**(+6.2)	**-0.207**(+0.09)	**0.781**(+0.02)

5 Conclusion and Future Work

In this paper, we establish a large-scale, bilingual open-world benchmark dataset for video text tracking and spotting, termed BOVText, with four description information, *i.e.,* , bounding box, tracking ID, recognition content, and text category label. Compare with the existing benchmarks, the proposed BOVText mainly contains four advantages: large-scale, open real scenarios, bilingual, and abundant text types annotation. Besides, we first propose an end-to-end video text spotting framework with Transformer, termed TransVTSpotter, which presents a simple, but efficient attention-based query-key pipeline. On ICDAR2015(video), TransVTSpotter achieves the state-of-the-art performance with **44.2%** MOTA, **13** fps. In general, we hope the proposed BOVText and TransVTSpotter would provide a standard benchmark to facilitate the advance of video-and-text research.

References

[1] Christos-Nikolaos E Anagnostopoulos, Ioannis E Anagnostopoulos, Ioannis D Psoroulas, Vassili Loumos, and Eleftherios Kayafas. License plate recognition from still images and video sequences: A survey. *IEEE Transactions on intelligent transportation systems*, 9(3):377–391, 2008.

[2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.

[3] Yuanqiang Cai, Weiqiang Wang, Shao Huang, Jin Ma, and Ke Lu. Spatiotemporal text localization for videos. *Multimedia Tools and Applications*, 77(22):29323–29345, 2018.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[5] Zhanzhan Cheng, Jing Lu, Yi Niu, Shiliang Pu, Fei Wu, and Shuigeng Zhou. You only recognize once: Towards fast video text spotting. In *ACM International Conference on Multimedia*, pages 855–863, 2019.

[6] Chee Kheng Ch'ng and Chee Seng Chan. Total-text: A comprehensive dataset for scene text detection and recognition. In *IEEE International Conference on Document Analysis and Recognition*, pages 935–942, 2017.

[7] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixe. Cvpr19 tracking and detection challenge: How crowded can it get? *arXiv preprint arXiv:1906.04567*, 2019.

[8] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[11] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2315–2324, 2016.

[12] Tong He, Zhi Tian, Weilin Huang, Chunhua Shen, Yu Qiao, and Changming Sun. An end-to-end textspotter with explicit alignment and attention. In *IEEE conference on computer vision and pattern recognition*, pages 5020–5029, 2018.

[13] Xian-Sheng Hua, Pei Yin, and Hong-Jiang Zhang. Efficient video text recognition using multiple frame integration. In *Proceedings. International Conference on Image Processing*, volume 2, pages II–II. IEEE, 2002.

[14] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, 2014.

[15] Keechul Jung, Kwang In Kim, and Anil K Jain. Text information extraction in images and video: a survey. *Pattern recognition*, 37(5):977–997, 2004.

[16] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *IEEE International Conference on Document Analysis and Recognition*, pages 1156–1160, 2015.

[17] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *IEEE International Conference on Document Analysis and Recognition*, pages 1484–1493, 2013.

[18] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[19] Hui Li, Peng Wang, and Chunhua Shen. Towards end-to-end text spotting with convolutional recurrent neural networks. In *IEEE International Conference on Computer Vision*, pages 5238–5246, 2017.

[20] Yuan Li, Chang Huang, and Ram Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *2009 IEEE conference on computer vision and pattern recognition*, pages 2953–2960. IEEE, 2009.

[21] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *AAAI Conference on Artificial Intelligence*, pages 11474–11481, 2020.

[22] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text spotting with a unified network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5676–5685, 2018.

[23] Shangbang Long, Xin He, and Cong Yao. Scene text detection and recognition: The deep learning era. *International Journal of Computer Vision*, 129(1):161–184, 2021.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[25] Ning Lu, Wenwen Yu, Xianbiao Qi, Yihao Chen, Ping Gong, Rong Xiao, and Xiang Bai. Master: Multi-aspect non-local network for scene text recognition. *Pattern Recognition*, 117:107980, 2021.

[26] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In *European Conference on Computer Vision*, pages 67–83, 2018.

[27] Rodrigo Minetto, Nicolas Thome, Matthieu Cord, Neucimar J Leite, and Jorge Stolfi. Snooper-track: Text detection and tracking for outdoor videos. In *IEEE International Conference on Image Processing*, pages 505–508, 2011.

[28] Anand Mishra, Karteek Alahari, and CV Jawahar. Image retrieval using textual cues. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3040–3047, 2013.

[29] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *IEEE International Conference on Document Analysis and Recognition*, volume 1, pages 1454–1459, 2017.

[30] Phuc Xuan Nguyen, Kai Wang, and Serge Belongie. Video text detection and recognition: Dataset and benchmark. In *IEEE winter conference on applications of computer vision*, pages 776–783, 2014.

[31] Sangeeth Reddy, Minesh Mathew, Lluis Gomez, Marçal Rusinol, Dimosthenis Karatzas, and CV Jawahar. Roadtext-1k: Text detection & recognition dataset for driving videos. In *IEEE International Conference on Robotics and Automation*, pages 11074–11080, 2020.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[33] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.

[34] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Workshops of European conference on computer vision*, pages 17–35, 2016.

[35] Xuejian Rong, Chucai Yi, Xiaodong Yang, and Yingli Tian. Scene text recognition in multiple frames based on text tracking. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2014.

[36] Georg Schroth, Sebastian Hilsenbeck, Robert Huitl, Florian Schweiger, and Eckehard Steinbach. Exploiting text-related features for content-based image retrieval. In *2011 IEEE international symposium on multimedia*, pages 77–84. IEEE, 2011.

[37] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2016.

[38] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In *IEEE conference on computer vision and pattern recognition*, pages 4168–4176, 2016.

[39] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International Conference on Machine Learning*, pages 843–852, 2015.

[40] Peize Sun, Yi Jiang, Rufeng Zhang, Enze Xie, Jinkun Cao, Xinting Hu, Tao Kong, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple-object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[42] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.

[43] Jianfeng Wang and Xiaolin Hu. Gated recurrent convolution neural network for ocr. In *Neural Information Processing Systems*, pages 334–343, 2017.

[44] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *2011 International Conference on Computer Vision*, pages 1457–1464. IEEE, 2011.

[45] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1571–1580, 2021.

[46] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021.

[47] Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. Shape robust text detection with progressive scale expansion network. In *IEEE conference on computer vision and pattern recognition*, pages 9336–9345, 2019.

[48] Xiaobing Wang, Yingying Jiang, Shuli Yang, Xiangyu Zhu, Wei Li, Pei Fu, Hua Wang, and Zhenbo Luo. End-to-end scene text recognition in videos based on multi frame tracking. In *IEEE International Conference on Document Analysis and Recognition*, pages 1255–1260, 2017.

[49] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8741–8750, 2021.

[50] Liang Wu, Palaiahnakote Shivakumara, Tong Lu, and Chew Lim Tan. A new technique for multi-oriented scene text line detection and tracking in video. *IEEE Transactions on multimedia*, 17(8):1137–1152, 2015.

[51] Jie Xi, Xian-Sheng Hua, Xiang-Rong Chen, Liu Wenyin, and Hong-Jiang Zhang. A video text detection and recognition system. In *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.*, pages 222–222. IEEE Computer Society, 2001.

[52] Chun Yang, Xu-Cheng Yin, Wei-Yi Pei, Shu Tian, Ze-Yu Zuo, Chao Zhu, and Junchi Yan. Tracking based multi-orientation scene text detection: A unified framework with dynamic programming. *IEEE Transactions on Image Processing*, 26(7):3235–3248, 2017.

[53] Jian Yi, Yuxin Peng, and Jianguo Xiao. Using multiple frame integration for the text recognition of video. In *2009 10th International Conference on Document Analysis and Recognition*, pages 71–75. IEEE, 2009.

[54] Xu-Cheng Yin, Ze-Yu Zuo, Shu Tian, and Cheng-Lin Liu. Text detection, tracking and recognition in video: a comprehensive survey. *IEEE Transactions on Image Processing*, 25(6):2752–2773, 2016.

[55] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018.

[56] Hongyuan Yu, Yan Huang, Lihong Pi, Chengquan Zhang, Xuan Li, and Liang Wang. End-to-end video text detection with online tracking. *Pattern Recognition*, 113:107791, 2021.

[57] Oussama Zayene, Mathias Seuret, Sameh Masmoudi Touj, Jean Hennebert, Rolf Ingold, and Najoua Essoukri Ben Amara. Text detection in arabic news video based on SWT operator and convolutional auto-encoders. In *Workshop on Document Analysis Systems*, pages 13–18, 2016.

[58] Fangao Zeng, Bin Dong, Tiancai Wang, Cheng Chen, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. *arXiv preprint arXiv:2105.03247*, 2021.

[59] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point transformer. *arXiv preprint arXiv:2012.09164*, 2020.

[60] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6881–6890, 2021.

[61] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *IEEE conference on computer vision and pattern recognition*, pages 5551–5560, 2017.

[62] Xinyu Zhou, Shuchang Zhou, Cong Yao, Zhimin Cao, and Qi Yin. Icdar 2015 text reading in the wild competition. *arXiv preprint arXiv:1506.03184*, 2015.

[63] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We describe the limitations in the supplementary material.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We present the potential negative societal impacts for the work in the supplementary material, and provide a solution.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We have provided the URL concerning the coding and the data to promote further research.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] We have blurred identifiable information or offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes]