ICDAR 2023 Competition on RoadText Video Text Detection, Tracking and Recognition

George Tom$^{1[0009-0002-7343-1680]}$, Minesh Mathew$^{1[0000-0002-0809-2590]}$, Sergi Garcia-Bordils$^{2,3[0000-0002-4222-8367]}$, Dimosthenis Karatzas$^{2[0000-0001-8762-4454]}$, and C.V. Jawahar$^{1[0000-0001-6767-7057]}$

[1] Center for Visual Information Technology (CVIT), IIIT Hyderabad, India
{george.tom,minesh.mathew}@research.iiit.ac.in, jawahar@iiit.ac.in
[2] Computer Vision Center (CVC), UAB, Spain {sergi.garcia,dimos}@cvc.uab.cat
[3] AllRead Machine Learning Technologies

Abstract. In this report, we present the final results of the ICDAR 2023 Competition on RoadText Video Text Detection, Tracking and Recognition. The RoadText challenge is based on the RoadText-1K dataset and aims to assess and enhance current methods for scene text detection, recognition, and tracking in videos. The RoadText-1K dataset contains 1000 dash cam videos with annotations for text bounding boxes and transcriptions in every frame. The competition features an end-to-end task, requiring systems to accurately detect, track, and recognize text in dash cam videos. The paper presents a comprehensive review of the submitted methods along with a detailed analysis of the results obtained by the methods. The analysis provides valuable insights into the current capabilities and limitations of video text detection, tracking, and recognition systems for dashcam videos.

Keywords: Scene text · Tracking · Recognition.

1 Introduction

Text detection and recognition in videos have traditionally been explored by the document analysis community. The last text-tracking competition was held nearly a decade ago and introduced the Text in Videos[9] dataset, which comprises 51 egocentric videos encompassing indoor and outdoor scenarios. Othe rpopular datasets that deal with text in videos are USTB-VidTEXT[18] and YouTube Video Text(YVT)[14]. They contain videos sourced from YouTube. The USTB-VidTEXT dataset primarily consists of text in the form of overlaid captions, whereas the YVT includes both born-digital text and scene text. These datasets contain videos with text that are incidental and widely dispersed across the scene.

Compared to the ICDAR 2013-15 Text-in-Videos Challenge that used a dataset containing 50 videos, our challenge uses the RoadText1K[15] dataset having much larger and diverse set of videos. The text objects in driving videos typically have short lifetimes, which require models tolerant to occlusions, able to

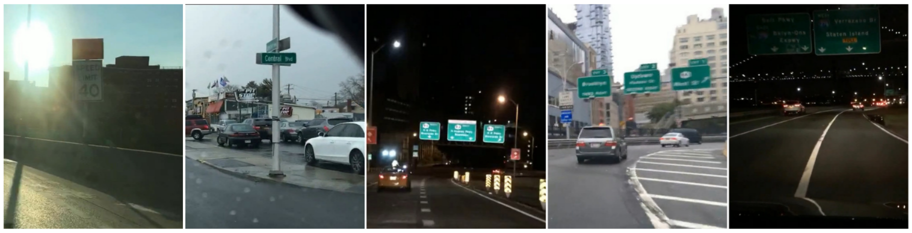


Fig. 1. Sample frames from RoadText-1K illustrating the various challenges and artefacts like glare, raindrops, out-of-focus, low contrast, and motion blur often encountered in driving videos.

handle tiny text instances, and robust to motion blur and significant perspective distortions. Additionally, text instances may not be fully readable in any single frame, necessitating the combination of detections across various frames to transcribe them successfully. Furthermore, camera movement during driving introduces distortions, such as motion blur. As a result, the approaches developed for existing video text datasets tend to be challenging to adapt to real-world applications, such as driver assistance and self-driving systems.

2 Competition Protocol

The competition took place between December 2022 and March 2023. The training and validation data were made available at the end of December 2022, while the test data was released in mid-February 2023. Submissions were accepted between March 1st and March 27th. The expectation was that participating authors would adhere to the established rules of the challenge, to which they had agreed when registering at the Robust Reading Competition (RRC) portal[4], as a means of ensuring scientific integrity throughout the competition.

The RRC portal serves as the host platform for the challenge. Submissions are assessed through automated methods, and the outcomes reported in this report represent the state of submissions at the conclusion of the challenge. However, the challenge will remain open to accept new submissions. But the submissions made after the official challenge period are not considered as an official challenge entry.

3 The RoadText-1K Dataset

The RoadText-1K[15] dataset comprises 10-second video clips extracted from the BDD100K[21] dataset. The videos are 720p and 30 fps, and capture diverse locations, weather conditions (such as sunny, overcast, and rainy), as well as

[4] https://rrc.cvc.uab.es/?ch=25

Dataset	Text in Videos [9]	USTB-VidTEXT [18]	YouTube Video Text [14]	RoadText-1K [15]
Source	Egocentric	Youtube	Youtube	car-mounted
Size (Videos)	51	5	30	**1000**
Length (Seconds)	varying	varying	15	10
Resolution	720 × 480	480 × 320	1280 × 720	1280 × 720
Annotated Frames	27,824	27,670	13,500	**300,000**
Total Text Instances	143,588	41,932	16,620	**1,280,613**
Text type	Scene Text	Digital (captions)	Scene Text and Digital	Scene Text
Unique Words	3,563	306	224	8,263
Avg. text frequency per frame	5.1	1.5	1.23	4.2
Avg. Text Track length	46	161	72	48

Table 1. Comparison of RoadText-1K with existing text video datasets.

different times of day. To identify videos with a significant number of text instances, an off-the-shelf text detector was utilised to scan through the frames of the videos in BDD100K. The dataset was randomly partitioned into train, validation and test sets of 500, 200 and 300 videos, respectively.

The bounding boxes and their transcriptions are provided at line level for all the frames in the dataset. The tracks are classified into English, Non-English, and Illegible. Ground truth ext transcriptions are provided only for text instances of the English category. In contrast to most scene text datasets, text lines rather than individual "words" (separated by spaces) were annotated to expedite annotation and avoid ambiguity in cases involving numbers or abbreviations.

4 RoadText-1K Challenge

4.1 Evaluation Metrics

The evaluation is based on an adaptation of the CLEAR-MOT [3,13] and ID[16] frameworks, designed for tracking multiple objects. Each submission is evaluated using three different metrics, namely Multiple Object Tracking Precision (MOTP), Multiple Object Tracking Accuracy (MOTA), and IDF1 score. The number of objects tracked for at least 80 per cent of their lifespan are considered as "Mostly Matched". Those objects that are tracked between 20 and 80 per

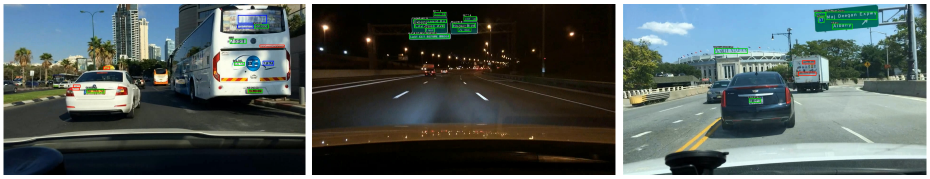


Fig. 2. These are sample frames from clips in RoadText-1K, and they have annotations indicating the location and transcription of the text overlaid on them. The boxes that are colored green indicate English text, the ones in blue represent non-English text, and the red boxes represent text that is illegible.

Table 2. Affiliations and the methods of the competition participants.

Method	Affiliation
ClusterFlow	Google
TH-DL	Tsinghua University
TencentOCR	TencentOCR
TransDETR	ByteDance Inc
RoadText DRTE	KLE Technological University
SCUT-MMOCR-KS	South China University of Technology, Shanghai AI Laboratory and KingSoft Office CV R&D Department

cent of their lifespan fall under“Partially Matched”, and those tracked for less than 20 per cent of their lifespan are categorised as “Mostly Lost”. For ranking the submissions, MOTA is used. During the evaluation process, a predicted word is classified as a true positive if its intersection over union with a ground-truth word is greater than 0.5 and the predicted transcription matches the ground truth transcription. The assessment of transcription is case insensitive and it is only done for English category tracks. Leading and trailing spaces are disregarded, and instances of two or more spaces are treated as a single space. The recognition of punctuation marks at the start or end of a ground truth word is discretionary and does not influence the evaluation. The evaluation process does not consider areas that contain illegible or non-English legible text. As a result, if a method fails to detect such words, it will not be penalised. Similarly, a method that is successful in detecting such words does not receive a higher score. Even though we only have a single end-to-end task, we also provide results of detection and tracking without taking recognition into account.

4.2 Submitted Methods

The challenge received a total of 16 submissions, out of which 6 were unique and had fulfilled all the competition criteria. The contestants were permitted to submit multiple entries, but they were required to select a single submission as their official entry for the competition. This selection had to be made blindly, without access to the evaluation scores for the submissions. Table 2 presents the names of the submitted methods and affiliations. A brief description of the 6 submitted methods is provided below:

ClusterFlow - ClusterFlow benefits from merging multiple algorithms, including optical character recognition (OCR), optical flow, clustering, and decision trees. The approach involves using a cloud API to extract OCR results at the line level for every image frame of each video, followed by calculating a dense optical flow field using a modern RAFT implementation. The optical flow field is then used to temporally extend the OCR line results to generate tubes or tracklets of lines, which are then grouped into clusters across the entire video using an unsupervised clustering algorithm. To achieve this, the algorithm searches for the optimal distance metric between tracklets, clustering algorithm, and hyperparameters using the training dataset. Once the tracklets are clustered, the algorithm selects geometry and text from the tracklet to create tracked lines that appear at most once within any video frame. This is accomplished by generating a set of features from each line appearance, tracklet, and cluster, which are then inputted into a classification algorithm. The classification algorithm is trained to select the appearances of the cluster that match the ground truth in the training set. During inference, the classification probabilities are used to choose the most suitable line text appearance within a cluster at any video frame.

TH-DL - It uses an integrated approach for text detection, recognition, and tracking in driving videos. For text detection and recognition, the algorithm adopts TESTR[22] based on Transformer and finetunes the pre-trained TESTR model on the training set of the Roadtext Challenge. For multi-object tracking, ByteTrack[23] is employed, which uses similarities with tracklets to recover true objects from low score detection boxes. A post-processing module is included to filter duplicate instances of text detection and recognition.

TencentOCR - It integrates the detection results of DBNet[10] and Cascade MaskRCNN[4], built with multiple backbone architectures, with the Parseq[2] English recognition model for recognition and further improves the end-to-end tracking with OCSort[5]. The result is end-to-end tracking and trajectory recognition.

TransDetr - The method used in this submission is TransDETR[19]. The approach involves pre-training the network weights on the ICDAR2015 video[9] and fine-tuning the network on the RoadText-3K[7] and BOVText[20] datasets for 20 epochs each. Finally, the network is fine-tuned on the RoadText-1K dataset for 20 epochs.

RoadText DRTE - EasyOCR[8] is used to perform the subtasks of detection and recognition on the RoadText-1K[15] dataset. The algorithm uses the CRAFT[1] algorithm for detection and the CRNN[17] model for recognition. Once the video is processed frame by frame, the algorithm performs the tracking subtask by assigning a unique ID to each unique transcription in the video. Instances of the same unique transcription are assigned the same ID throughout the video.

SCUT-MMOCR-KS - This submission utilizes DBNet++[11] for text detection, which is first pre-trained on a collection of TextOCR, HierText[12], DSText, YVT[14], ICDAR2015-Video[9], and Minetto before being fine-tuned on DSText. For text recognition, a ViT-based[6] recognizer is used, which is pre-trained on 10M unlabeled real STR images and fine-tuned on 4M labelled real STR images. CoText tracking module is used for text tracking.

Table 3. Results of RoadText video text detection, tracking

Method	MOTA	MOTP	IDF1	Mostly Matched	Partially Matched	Mostly Lost
TransDETR	37.53	74.18%	60.27%	1665	1762	1563
ClusterFlow	36.01	70.29%	61.19%	1757	1194	2029
TH-DL	31.07	75.20%	62.35%	2180	1495	1317
TencentOCR	16.40	66.59%	42.58%	746	894	3231
SCUT-MMOCR-KS	-10.27	71.84%	56.91%	2354	1660	978
RoadText DRTE	-27.61	70.46%	17.42%	1083	1692	2214

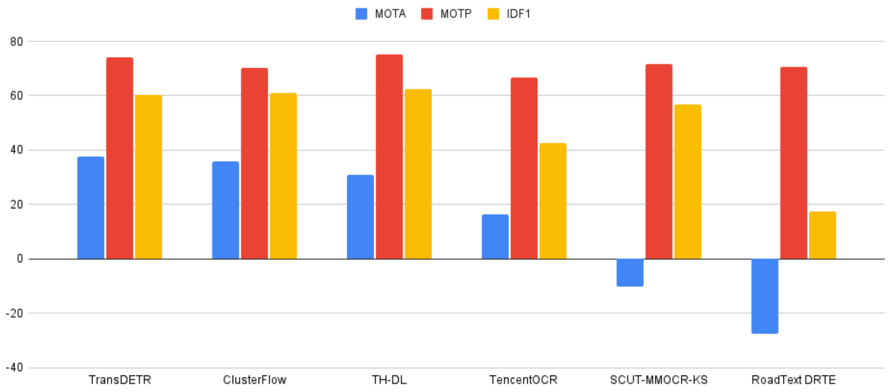


Fig. 3. The chart illustrates the results for text detection and tracking, with MOTA, MOTP, and IDF1 represented by blue, red, and yellow bars, respectively.

Table 4. Results of RoadText video text detection, tracking and recognition

Method	MOTA	MOTP	IDF1	Mostly Matched	Partially Matched	Mostly Lost
ClusterFlow	11.09	69.04%	48.07%	1392	920	2668
TH-DL	-23.10	72.83%	37.34%	1235	737	3020
TencentOCR	-23.87	56.19%	19.71%	315	454	4102
TransDETR	-28.50	68.74%	26.87%	660	741	3589
RoadText DRTE	-61.39	65.47%	12.08%	146	823	4020
SCUT-MMOCR-KS	-77.1	67.83%	29.6%	1196	918	2878

The participants could use any dataset for training their methods, except the RoadText-1K test set.

4.3 Analysis

The results of the evaluation are presented in Table 3 and Table 4, with the first one focusing on text detection and tracking and the second one displaying text tracking results with recognition. In the absence of recognition, the method with the highest MOTA score was TransDETR, while TH-DL achieved the highest MOTP score and IDF1 score for text tracking. However, in the presence of recognition, ClusterFlow is the winner of the competition and the only method with a positive MOTA value and also achieved the highest IDF1 score, while TH-DL maintained its position for the highest MOTP value. The commercial

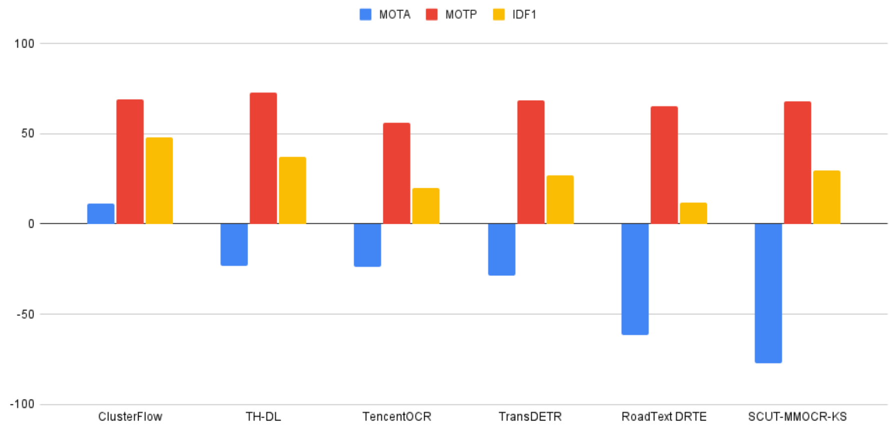


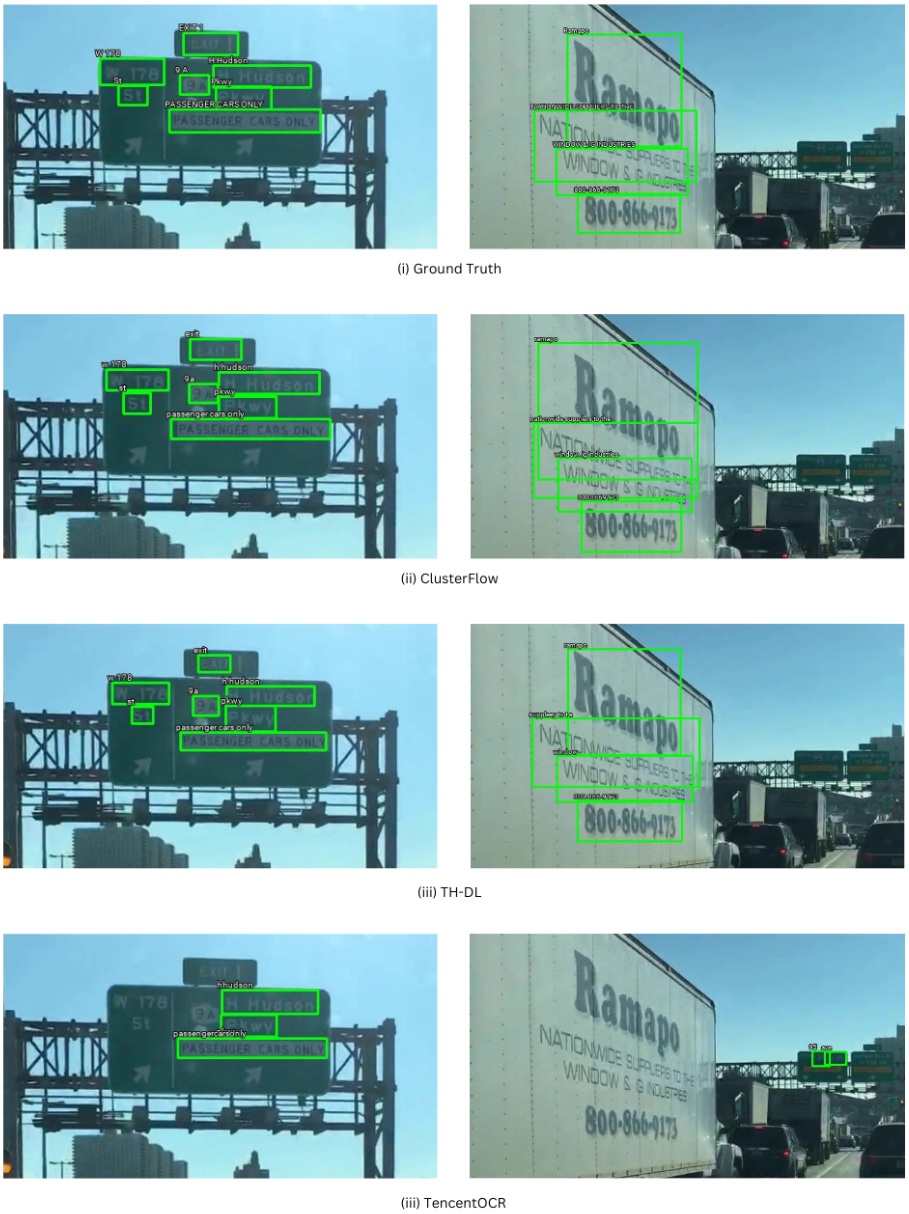
Fig. 4. The chart illustrates the results for text detection, tracking and recognition, with MOTA, MOTP, and IDF1 represented by blue, red, and yellow bars, respectively.

Google OCR performs better in comparison to TH-DL, which utilizes TESTR, SCUT-MMOCR-KS, employing ViT-based, and TencentOCR, which relies on Parseq methods for recognition. In the evaluation process, predicted words are only considered true positives when they match the ground truth. This means that if the recognition fails to identify a word, the corresponding track will be considered a false positive, leading to negative MOTA values. Text that appears in a frontal or head-on position is relatively easy to detect. However, text detection methods appear to struggle when presented with text instances such as fancy shop signage or text situated on distant portions of the road beyond the driver's lane.

The participants utilized various approaches and strategies to enhance the effectiveness of their methods. These include pre-training and fine-tuning models on diverse datasets, implementing post-processing steps like filtering out repeated text detection and recognition instances to improve outcomes, and merging multiple algorithms and methods. Despite these efforts, the detection, tracking, and recognition still have significant room for improvement, particularly recognition in challenging scenarios presented by the dataset.

5 Conclusions and Future Work

The text detection, tracking and recognition challenge introduces a robust benchmark based on driving videos. The challenge is based on the already existing RoadText-1K dataset and has received a total of 16 submissions from multiple teams. In this report, we have summarised the unique features of the RoadText-1K dataset, which make it particularly challenging and different from previous datasets. The report also details a concise overview and an analysis of the sub-



(i) Ground Truth

(ii) ClusterFlow

(iii) TH-DL

(iii) TencentOCR

Fig. 5. Sample visualisation of the detected text and the recognition are shown for the ground truth and the top three methods. Green bounding boxes are drawn over detected text, and the recognised text is displayed over the bounding box.

missions. The RoadText challenge will remain open for new submissions in the future, thereby providing a platform for researchers to benchmark and showcase their methods. Looking ahead, we plan to expand the RoadText challenge further by gaining deeper insights into the results and incorporating additional tasks that encompass multilingual settings.

6 Acknowledgement

This work has been supported by IHub-Data at IIIT-Hyderabad, and MeitY, and grants PDC2021-121512-I00, and PID2020-116298GB-I00 funded by MCIN/AEI/ 10.13039/501100011033 and the European Union NextGenerationEU/PRTR.

References

1. Baek, Y., Lee, B., Han, D., Yun, S., Lee, H.: Character region awareness for text detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9365–9374 (2019)
2. Bautista, D., Atienza, R.: Scene text recognition with permuted autoregressive sequence models. In: European Conference on Computer Vision. pp. 178–196. Springer Nature Switzerland, Cham (10 2022). https://doi.org/10.1007/978-3-031-19815-1$_1$1, `https://doi.org/10.1007/978-3-031-19815-1_11`
3. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing **2008**, 1–10 (2008)
4. Cai, Z., Vasconcelos, N.: Cascade r-cnn: high quality object detection and instance segmentation. IEEE transactions on pattern analysis and machine intelligence **43**(5), 1483–1498 (2019)
5. Cao, J., Weng, X., Khirodkar, R., Pang, J., Kitani, K.: Observation-centric sort: Rethinking sort for robust multi-object tracking. arXiv preprint arXiv:2203.14360 (2022)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Garcia-Bordils, S., Tom, G., Reddy, S., Mathew, M., Rusiñol, M., Jawahar, C., Karatzas, D.: Read while you drive-multilingual text tracking on the road. In: Document Analysis Systems: 15th IAPR International Workshop, DAS 2022, La Rochelle, France, May 22–25, 2022, Proceedings. pp. 756–770. Springer (2022)
8. JaidedAI: Easyocr. `https://github.com/JaidedAI/EasyOCR`
9. Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L.G., Mestre, S.R., Mas, J., Mota, D.F., Almazán, J., de las Heras, L.: ICDAR 2013 robust reading competition. In: ICDAR (2013)
10. Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X.: Real-time scene text detection with differentiable binarization. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 11474–11481 (2020)
11. Liao, M., Zou, Z., Wan, Z., Yao, C., Bai, X.: Real-time scene text detection with differentiable binarization and adaptive scale fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 919–931 (2022)

12. Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., Raptis, M.: Towards end-to-end unified scene text detection and layout analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
13. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
14. Nguyen, P.X., Wang, K., Belongie, S.J.: Video text detection and recognition: Dataset and benchmark. In: WACV (2014)
15. Reddy, S., Mathew, M., Gomez, L., Rusinol, M., Karatzas, D., Jawahar, C.: Roadtext-1k: Text detection & recognition dataset for driving videos. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 11074–11080. IEEE (2020)
16. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II. pp. 17–35. Springer (2016)
17. Shi, B., Bai, X., Yao, C.: 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)
18. Tian, S., Yin, X., Su, Y., Hao, H.W.: A unified framework for tracking based text detection and recognition from web videos. TPAMI (2018)
19. Weijia Wu, Chunhua Shen, Y.C.D.Z.Y.F.P.L.H.Z.: End-to-end video text spotting with transformer. arxiv (2022)
20. Wu, W., Zhang, D., Cai, Y., Wang, S., Li, J., Li, Z., Tang, Y., Zhou, H.: Bov-text: A large-scale, multidimensional multilingual dataset for video text spotting. Organization (2021)
21. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2636–2645 (2020)
22. Zhang, X., Su, Y., Tripathi, S., Tu, Z.: Text spotting transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9519–9528 (2022)
23. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII. pp. 1–21. Springer (2022)