Role of Scene Text Understanding in Enhancing Driver Assistance

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science in ***Computer Science and Engineering*** *by Research*

by

George Tom
2018111004
george.tom@research.iiit.ac.in

International Institute of Information Technology, Hyderabad
(Deemed to be University)
Hyderabad - 500 032, INDIA
November 2024

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Role of Scene Text Understanding in Enhancing Driver Assistance" by George Tom, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Adviser: Prof. C V Jawahar
Date	Adviser: Prof. Dimosthenis Karatzas

To my family

Acknowledgments

This thesis would not have been possible without the invaluable support and guidance of countless individuals who crossed my path. I am deeply grateful for their contributions.

My most profound gratitude goes to my supervisors, Professor C.V. Jawahar and Professor Dimosthenis Karatzas. Prof. Jawahar played a pivotal role in my growth as a researcher, teaching me how to approach problems and persevere through challenges. His mentorship has been invaluable in shaping my research methodology and pushing me to where I am today. I'm deeply grateful for his unwavering support and guidance. I express my heartfelt appreciation to Prof. Dimosthenis Karatzas, my co-advisor, who provided me with insightful ideas and engaged in discussions that significantly contributed to this thesis. His patience and encouragement helped me with the confidence to navigate the initial confusion and uncertainty. Additionally, I would like to express my sincere appreciation for caring for me during my CVC visit.

I extend my sincere gratitude to Minesh Mathew for his exceptional mentorship. His extensive knowledge and expertise played an important part in guiding me through the uncertainties of the research process. He generously offered his time for insightful discussions that helped me navigate the challenges I faced. His mentorship played a key role in shaping the direction of my work, and I am truly grateful for the patience, wisdom, and encouragement he consistently provided.

My time at CVIT was fun because of the friends I made over there. I thank Soumya, Shivanshu, Zeeshan, Shubham, Siddhant, Madhav, Pranav, Varun, Darshan, Rudrabha, Ravi, Seshadri and Rupak for the laughter, chaos, and a healthy dose of peer pressure you brought to the table. I am also grateful to Jerin, who significantly influenced my decision to join CVIT. Thank you for the encouragement and invaluable guidance. Soumya deserves special mention for being an irreplaceable part of my project discussions and a wonderful companion during our travels. I extend my gratitude to Sergi, with whom I collaborated on multiple papers, and Ruben for their invaluable contributions and support. Meeting you both was a pleasure! Thinking back on my time at CVC, I realize that the friendships I made truly stand out - I thank Sanket, Andres, Ali, Dipam, Aarya, Esha, Andrea, Andrea(bro), Moha, Kevin, Lele, Ayan, Khan, and Laura.

Life at CVIT was made a lot easier by the supportive administrative and tech team. I thank Rohitha, Aradhana, Tessy, Sony, Mahathi and Cecilia for their support and assistance. I thank Varun and Ram for their IT expertise and for helping with the annotation tool and project pages. Additionally, my thanks go

to Mahender for his assistance with annotations. Your collective efforts have made a meaningful impact on my time at CVIT.

Next, I would like to thank all the misfits who made my monday mornings bearable. Ashwin, Dolton, Sambhav, Naren, Srinath, Sahil, Mohee, Koushik, Joseph, Giri, Bhuvanesh, Adarsh, Shinde, Siddarth, Amogh, Srivathsan, Sreevignesh, Thummala, Rhuthik, Pranoy, Vivek, Aswin, Ivin, Shiva and Rahul - here's to the crazy, chaotic, beautiful journey we shared, and to the lifelong connection that started amidst the late-night talks and shared caffeine.

To those I've remembered to thank, cheers! To those I haven't, my apologies! Consider this a virtual high-five. Thank you all for making this journey as epic as it has been.

Abstract

Scene text conveys important information to drivers and pedestrians, serving as an indispensable means of communication in various environments. Scene text contains information regarding the speed limit, route information, rest stops, and exits, among other important information. It is important for drivers and passengers to understand this information for a safe and efficient journey. However, outdoor scenes are cluttered with text, distracting drivers and making it hard to focus on what matters, potentially compromising their ability to focus on essential details and navigate safely. Recognising scene text in motion aggravates this challenge, as textual cues transiently appear and necessitate early detection at a distance. Driving scenarios introduce additional complexities, including occlusions, motion blur, perspective distortions, and varying text sizes, further complicating scene text understanding.

In this thesis, we look at improving scene text understanding in diving scenarios through video question answering and analysing the present state of scene text detection, recognition and tracking: (i) We introduce a new video questions answering task and dataset that requires an understanding of text and road signs in driving videos to answer the questions. (ii) We look at the current state of scene text detection, tracking and recognition in the driving domain through the RoadText-1K competition. (iii) We explore detection and recognition in special cases of occlusions, a common yet under-explored complication in real-world driving scenarios. By focusing on these areas, the thesis contributes to advancing scene text analysis methodologies, offering insights and solutions that are imperative for developing more intelligent and responsive driver assistance systems.

Contents

Chapter Page

1 Introduction . . . 1
1.1 Contributions . . . 4
1.2 Organization of Thesis . . . 4

2 Background and Related Works . . . 5
2.1 Scene Text Understanding . . . 5
2.1.1 Scene Text Detection and Recognition . . . 5
2.1.2 Scene Text Tracking . . . 6
2.1.3 Detection and Recognition under Occlusion . . . 6
2.2 Video Question Answering and Scene Text . . . 7
2.2.1 VideoQA . . . 7
2.2.2 VideoQA involving video text . . . 8
2.2.3 Scene Text VQA . . . 9

3 Reading Between the Lanes: Text VideoQA on the Road . . . 10
3.1 Introduction . . . 10
3.2 RoadTextVQA dataset . . . 11
3.2.1 Data Collection . . . 12
3.2.2 Annotation Tool . . . 14
3.2.3 Data Statistics and Analysis . . . 15
3.3 Baselines . . . 17
3.3.1 Heuristic Baselines and Upper Bounds . . . 18
3.3.2 M4C . . . 18
3.3.3 SINGULARITY . . . 18
3.3.4 GenerativeImage2Text . . . 19
3.4 Experiments and Results . . . 19
3.4.1 Experimental Setup . . . 19
3.4.2 Results . . . 21
3.5 Contextualizing Our Work in the LLM Era . . . 23
3.6 Summary and Future work . . . 24

4 Evaluating the current state of RoadText Detection, Tracking and Recognition . . . 25
4.1 Introduction . . . 25
4.2 The RoadText-1K Dataset . . . 27
4.3 RoadText-1K Challenge . . . 27

4.3.1 Evaluation Metrics . . . 27
4.3.2 Submitted Methods . . . 28
4.3.3 Analysis . . . 31
4.4 Summary . . . 33

5 Occluded RoadText . . . 34
5.1 Introduction . . . 35
5.2 RoadText Occluded Dataset . . . 36
5.2.1 Data Collection . . . 36
5.2.2 Data Statistics and Analysis . . . 37
5.3 Baseline: TESTR . . . 38
5.4 Experiments and Results . . . 39
5.4.1 Experimental Setup . . . 39
5.4.2 Results . . . 40
5.5 Summary . . . 41

6 Conclusions . . . 42

Bibliography . . . 44

List of Figures

Figure Page

1.1 The figure illustrates the progression of significant datasets from 2003 to 2024, showing the evolution from early scene text detection datasets like ICDAR 03 to advanced scene text related tasks . . . 3

3.1 Examples from our RoadTextVQA dataset. The question in the first clip is based on the speed limit road sign, so it is classified as a "road sign based" question. Meanwhile, the question for the clip in the second row draws information from the text on the van, making it a "text based" question. The ground truth answers and the baseline predictions are also presented. . . . 10

3.2 Distribution of the number of words in the question(left) and answer(right) of RoadTextVQA . . . 12

3.3 Top 10 questions in the dataset. . . . 13

3.4 Geographical distribution of videos in the RoadTextVQA dataset. . . . 13

3.5 An analysis of the distribution of questions based on their starting 4 grams has shown that a significant proportion of questions are aimed at obtaining the name and contact information of businesses located along roads, as well as obtaining the speed limit for the road. . . . 14

3.6 The number of occurrences of the answers in RoadTextVQA. The most recurring answer is "right", which makes up about 8% of the answers. . . . 15

3.7 The annotation tool features a dedicated workflow for adding and verifying annotations. The top screenshot shows the Annotation Page, where users directly add annotations to the content. The bottom screenshot features the Verification Page, allowing users to review annotations, add question flags and types, and validate submissions to ensure correctness and consistency. . . . 16

3.8 A visual representation of word frequency in the form of a word cloud, depicting the distribution of words in answers (left) and the distribution of OCR tokens from the videos (right). . . . 17

3.9 Distribution of number of videos vs number of tracks. . . . 17

3.10 Qualitative results showing predictions of M4C, SINGULARITY and GIT. The correct predictions are highlighted in green, whereas the incorrect ones are highlighted in red. . 23

4.1 Sample frames from RoadText-1K illustrating the various challenges and artifacts like glare, raindrops, out-of-focus, low contrast, and motion blur often encountered in driving videos. . . . 25

4.2 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 illegible text. 27
4.3 The chart illustrates the results for text detection and tracking, with MOTA, MOTP, and IDF1 represented by blue, red, and yellow bars, respectively. 30
4.4 The chart illustrates the results for text detection, tracking and recognition, with MOTA, MOTP, and IDF1 represented by blue, red, and yellow bars, respectively. 31
4.5 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. 32

5.1 Instances of occluded text from the Occluded RoadText dataset (left) are compared with instances of occluded text from another dataset (right). The occlusions in our dataset represent real-life scenes, whereas the occlusions in the other datasets are manually created and are less representative of real-life scenarios. 34
5.2 Cropped out test image in the middle and the additional images on the left and right help to recognize occluded text. 35
5.3 The label studio annotation tool is used to segment the video timeline to extract the occluded sections from the larger clip. 36
5.4 Screenshot of CVAT used for text instance annotation and tracking. This includes transcript labelling, bounding box creation, and instance-level tracking. 37
5.5 An illustration depicting various types of occlusions. On the left, both ”BOMBAY” and ”READYMADES” are occluded but visible in another text instance, falling into the ”Occluded Visible” category. In the middle, ”SAMSUNG” is occluded but can be inferred from the visual color and the font, categorizing it as ”Occluded Inferable.” On the right, the occluded text is unrecoverable(Occluded indeterminate) without the assistance of additional information, such as supplementary images. 38
5.6 Distribution of categories of text instances. 39
5.7 A word cloud showing the frequency of words in text transcriptions. 39
5.8 Qualitative results showing predictions of the TESTR model 40

List of Tables

Table Page

1.1 The table summarizes different datasets for scene text detection and recognition in images and videos. It includes information about the text attributes, language, and media type for each dataset. . . . 2

2.1 Comparison of RoadTextVQA with existing video question answering datasets. "Text-based" indicates whether the questions require an understanding of the text present in the videos to answer. "Road-based" questions are datasets which are based on the driving domain. "Synthetic questions" are questions that are not manually annotated and depend on automated methods for question-answer generation. Abbreviations used - OE: Open-ended questions, MC: Multiple choice questions. . . . 8

3.1 Comparison of average and maximum question and answer lengths with other text based video question answering datasets. . . . 13
3.2 Performance of various heuristic baselines and upper bounds that are commonly evaluated on text-based VQA datasets. . . . 21
3.3 Performance of RoadTextVQA on M4C. Abbreviations- TB: text-based questions, RSB: road sign-based questions, AP: questions where the answer is present in the video, ANP: questions where the answer is not present in the video. . . . 21
3.4 Performance of RoadTextVQA on SINGULARITY and GIT. Abbreviations- TB: text-based questions, RSB: road sign-based questions, AP: questions where the answer is present in the video, ANP: questions where the answer is not present in the video. . . . 22

4.1 Comparison of RoadText-1K with existing text video datasets. . . . 26
4.2 Results of RoadText video text detection, tracking . . . 30
4.3 Results of RoadText video text detection, tracking and recognition . . . 31

5.1 Performance of TESTR on text localization task . . . 40
5.2 Performance of TESTR in end-to-end recognition task . . . 40

Chapter 1

Introduction

Understanding the visual world around us is crucial for navigating our daily lives. This is particularly true for drivers, who rely on a constant stream of visual information to make informed decisions and ensure their safety. Text plays a vital role in this process, appearing on road signs, billboards, and other elements of the environment. By reading and comprehending this textual information, drivers can navigate efficiently, adhere to traffic regulations, and avoid potential hazards[22]. Whether it is a speed limit sign dictating our pace or a stop sign demanding our immediate attention, textual information provides clear and concise instructions that allow us to adapt our behaviour accordingly. It also serves as a constant reminder of the rules and regulations that govern our interactions with the physical environment, promoting order and safety on our roads. Furthermore, we can avoid potential hazards by reading and comprehending textual information. Road signs warning of upcoming turns, pedestrian crossings, or construction zones enable us to anticipate and adjust our driving, reducing the risk of accidents. In this way, textual information helps us to make informed choices, ensuring a smoother and safer journey for ourselves and others. Shop signage acts as a visual map, informing us of the businesses available in a particular area. Whether it's a restaurant logo or a brightly coloured store name, this information saves us valuable time and energy as we search for specific goods and services. It also adds visual interest to the landscape, transforming streets into vibrant thoroughfares buzzing with activity. Billboards and advertisements, while often overlooked, also play a role in our navigation. They serve as landmarks, helping us identify landmarks and orient ourselves within unfamiliar territory. While their primary purpose may be to promote products or services, they also contribute to the overall visual tapestry of the surroundings. However, an overabundance of textual information on the road can also be distracting and lead to accidents[65]. When drivers are bombarded with too many signs, billboards, and advertisements, it can become challenging to prioritize and process the most critical information quickly. This cognitive overload can lead to slower reaction times or even failure to notice essential warnings or directives.

These texts that are present within natural environments, such as on signs, billboards, posters, and shop signage, are referred to as "scene text". There is significant interest in understanding scene text owing to its relevance and the challenges it presents. The Table 1.1 list the different datasets of scene text detection and recognition datasets and the Figure 1.1 illustrates the progression of significant scene

Table 1.1 The table summarizes different datasets for scene text detection and recognition in images and videos. It includes information about the text attributes, language, and media type for each dataset.

Dataset	Text attributes	Language	Media type
ICDAR 2003[55]	Horizontal	English	Image
ICDAR 2013[34]	Horizontal	English	Image
CUTE80[71]	Curved	English	Image
ICDAR 2015[33]	Arbitrary oriented, blur	English	Image
Text in Videos[33]	Arbitrary oriented	English	Video
ICDAR 2017 MLT[61]	Arbitrary oriented	Multilingual	Image
COCO-Text[82]	Arbitrary oriented	English	Image
Total-Text[17]	Arbitrary oriented, Curved	English	Image
CTW[97]	Arbitrary oriented	Chinese	Image
CTW1500[98]	Curved	English & Chinese	Image
ICDAR 2019 MLT[60]	Arbitrary oriented	Multilingual	Image
ICDAR 2019 ArT[18]	Arbitrary oriented, Curved	English & Chinese	Image
RoadText-1K[68]	Arbitrary oriented	English	Video
IIIT-ILST[57]	Arbitrary oriented	Multilingual	Image
ISTD-OC[25]	Arbitrary oriented, Synthetic Occluded	English	Image
BOVText[90]	Arbitrary oriented	English & Chinese	Video
OCTT[67]	Arbitrary oriented, Synthetic Occluded	English	Image
RoadText-3K[24]	Arbitrary oriented	Multilingual	Video
Occluded RoadText	Arbitrary oriented, Occluded	Multilingual	Image

text related datasets from 2003 to 2024. Starting with early scene text detection and recognition datasets like ICDAR 2003[55], ICDAR 2013[34], and ICDAR 2015[33], it shows the evolution towards more complex datasets that incorporate curved texts, such as CUTE80[71] CTW-1500[98]. In 2017, ICDAR MLT[61] expanded the scope by providing a multilingual dataset, which was followed up by ICDAR 2019 MLT[60] and IIIT-ILST[57]. The RoadText-1K[68] scene text detection, recognition, and tracking dataset is 20 times larger than the Text in Videos[34] dataset. It was later expanded to RoadText-3K[24] to encompass multilingual scene text. Recent advancements in scene text understanding include the ST-VQA[8] dataset, which marked the introduction of the first Visual Question Answering (VQA) dataset

Figure 1.1 The figure illustrates the progression of significant datasets from 2003 to 2024, showing the evolution from early scene text detection datasets like ICDAR 03 to advanced scene text related tasks

based on scene text in images. Building upon this, ViteVQA[102] extended the concept to videos, laying the groundwork for VQA tasks involving the interpretation of text in video content.

Despite the importance of text in driving scenarios, current ADAS systems have largely ignored the need to read and understand text. In parallel with ViteVQA, our work RoadTextVQA was released, which models the textual understanding in driving scenarios as a Video Question-Answering (VideoQA) task. Early VideoQA primarily focused on the visual aspects of the scene, neglecting the rich information conveyed through text. To address this gap, we introduce a new dataset specifically designed for VQA on driving videos. In contrast to previous datasets that focus on diverse video content, our dataset consists exclusively of driving videos, ensuring that the questions and answers are relevant and practical for real-world driving scenarios. The dataset contains 10.5K questions and 3.2K videos based on road text and road signs and evaluates it on two state-of-the-art VideoQA models. However, the performance of VideoQA models on this dataset revealed a significant challenge: their struggle with reading scene text.

To further investigate the capabilities of existing models in reading text in driving videos, we organized a competition. This competition aimed to assess the performance of state-of-the-art systems in detecting, recognizing, and tracking text in driving scenarios. We separately assess the methods for the detection and tracking of text, and then we also evaluate them end-to-end, considering the recognition. The competition findings provided valuable insights into the strengths and weaknesses of different approaches and highlighted the need for further research and development in this area. The competition had participation from various industry leaders and universities.

Recognizing the limitations of existing models when it comes to reading text in real-world driving environments, we delved deeper into a specific scenario where models struggle: occluded scene text. Occlusions, which can arise from various sources, such as natural elements, moving objects, and cam-

era angles, significantly impact the visibility and readability of text. The current issue with existing datasets[88, 67, 26] is that their occlusions are mostly manually generated, which doesn't accurately reflect the complexities and unpredictability found in real-life situations. This artificial nature of occlusions can create a gap between a model's performance in controlled experimental settings and its effectiveness in real-world applications. We have developed a new dataset called Occluded RoadText specifically tailored for this challenging scenario, comprising over $1,000$ images and $3,659$ instances of occluded text of varying types. Additionally, the benchmark is publicly hosted to allow researchers to compare their methods.

1.1 Contributions

The contributions of the thesis are as follows:

- We introduce the task of scene text and road sign based VideoQA and its associated dataset, where models are required to comprehend and interpret scene text and road signs to answer questions. We provide a detailed aanalysis of the dataset, and the performance of various methods on the dataset
- We organise a competition to evaluate the current state of scene text detection, tracking and recognition in the driving domain.
- We introduce a new dataset for the special case of occlusion in scene text with 3,000+ instances of occluded text. We also introduce a new task of scene text detection involving multiple images for end-to-end scene text recognition.

1.2 Organization of Thesis

The thesis is organised as follows:

1. Chapter 2 lays out the background and literature review on scene text understanding, VideoQA, scene text recognition, and related works relevant to driving scenarios.
2. In Chapter 3, we look at the new RoadTextVQA and analyse the performance of VQA models on the RoadTextVQA dataset.
3. Chapter 4 discusses the organization and results of a competition to assess the state-of-the-art scene text recognition and tracking algorithms in driving videos.
4. Chapter 5 we explore the particular difficulty posed by occluded scene text, presenting a specialized dataset for this purpose.

Chapter 2

Background and Related Works

2.1 Scene Text Understanding

Scene text is the text that is present in the images or videos taken in outdoor environments. Scene text understanding involves detecting, recognizing and interpreting the text within these images. Researchers have come up with and explored various techniques for scene text detection, recognition, and understanding in driving videos. Apart from this, several pretraining techniques were introduced to improve the understanding of the scene text, which is not explicit to driving scenes. TAP[94] was the first to incorporate scene text in the pretraining task and proposed a "relative (spatial) position prediction" task, which involved predicting spatial relationships of scene texts. PreSTU[35] is a recent work where they introduce novel OCR-aware pretraining methods to connect scene text to the visual content. They introduce a SPLITOCR pretraining technique for pretraining, which involves providing a part of the scene text and model predicting the rest. The following subsections will go into detail on the prior works related to scene text understanding, specifically in the case of detection, recognition, and tracking.

2.1.1 Scene Text Detection and Recognition

Deep learning models, particularly Convolutional Neural Networks (CNNs), have enabled the development of highly accurate object detection models. Several CNN-based approaches, which consider text as objects, have demonstrated strong performance in localising text. EAST[104] was the first neural network-based approach that detected words but failed on curved and vertical texts. Horizontal bounding boxes are limited in representing scene text orientation. TextBoxes++[45] uses quadrilateral bounding boxes to represent text layout. Curved text detection was significantly improved by ABCNet models[50, 51] by utilizing Bezier-curve networks.

Text recognition algorithms can be broadly divided into character-based and word-based. Each character is individually recognised and grouped together in character-based methods, and word-based methods use CRNNs[75] to detect text words or lines. Liao et al.[45] introduced TextBoxes++, which com-

bines both CNN and RNN for word spotter and end-to-end recogniser. TESTR[100] is an end-to-end transformer-based end-to-end recogniser that works well with curved text instances.

The ICDAR-13[34], 15[33] datasets consist of 462 and 1500 images, respectively and contain text in various orientations and sizes. While both datasets offer word-level annotations, ICDAR-15 has a more robust multi-oriented ground truth. The Street Text View (STV)[87] dataset, derived from Google Street View, comprises 350 images and 674 cropped words. In contrast, the COCO-Text[83] dataset is notably larger, featuring 63,000 natural images with over 170,000 labelled text regions. Total-text[17] encompasses 1555 images, presenting diverse text orientations, including multi-oriented and curved. Meanwhile, the CTW1500[98] dataset, comprising 1500 images, has line-level annotations, mainly focusing on curved texts.

2.1.2 Scene Text Tracking

Text tracking methodologies fall into two primary categories. The first, Tracking by Detection (TbD), involves making detections in each frame and linking them to create a track. Wang et al. implement this approach by recognising text in each frame and forming a track by associating the detected texts. FREE[16], an end-to-end method, utilises spatial-temporal characteristics for simultaneous detection and tracking, performing text recognition once and employing a text recommender to select the highest quality text across frames. In Detection by Tracking, detection occurs in the initial frame and is then extended to subsequent frames. Gomez et al.[27] propose a method leveraging Maximally Stable Extremal Regions (MSER) for text detection and propagation. In the context of [78], recognition is conducted in each frame. The tracking trajectory is learned globally using dynamic programming using all detections and recognitions.

The ICDAR Text in Videos[34] dataset consists of 51 videos that were shot in a variety of contexts, such as roads, supermarkets, and indoor spaces. YouTube Video Text(YVT)[62] contains 30 videos taken from YouTube, each with scene text and digitally added text such as commentary or subtitles and scene text. RoadText-1K[68] is made up of 1,000 driving videos taken from the BDD[73] dataset that have been annotated for text detection, recognition, and tracking. BoVText[90], a recent addition, has 2,021 videos with both born-digital overlay text and scene text instances acquired from various video-sharing networks worldwide. The RoadText-3K[24] extends RoadText-1K with 2,000 videos taken from Europe and India and features texts in multiple languages.

2.1.3 Detection and Recognition under Occlusion

Text detection and recognition in occluded settings has been a difficult challenge in computer vision due to the diverse variety of occlusions that might occur in real-world scenarios. The Occlusion Scene Text(OCT)[67] dataset contains word images of occluded text. It contains 4,832 images with occlusions that are manually created, and the images are sourced from multiple other datasets, including ICDAR13[34], ICDAR15[33], CUTE80[71]. Occluded Character-level Total-Text (OCTT)[67] dataset

contains 300 images for occluded word text spotting and is taken from Total-Text[17] where at least one character is manually occluded. Raisi et al.[67] introduced an end-to-end architecture for scene text spotting which attains state-of-the-art performance in identifying occluded text instances by leveraging a pre-trained masked backbone. ISTD-OC[25] dataset 1500 images for detection and 4468 images for recognition and is specifically curated to provide occluded images. The images are taken from the ICDAR 2015 dataset. The dataset is designed to offer a variety of partially occluded images with a degree of randomness. Occlusion is applied to text regions by overlaying random parts of the original image generated using a Gaussian distribution.

2.2 Video Question Answering and Scene Text

2.2.1 VideoQA

In video question answering(VideoQA), the goal is to answer the question in the context of the video. Earlier approaches to VideoQA use LSTM to encode the question and videos[103, 44, 36, 85]. Several datasets have been created in recent years to assist research in the field of video question answering (VideoQA). Large datasets such as MSRVTT-QA[91] contain synthetically generated questions and answers where the questions require only an understanding of the visual scenes. MOVIE-QA[77] and TVQA[41] are based on scenes in movies and TV shows. Castro et al.[13] introduced a dataset with videos from the outside world for video understanding through VideoQA and Video Evidence Selection for interpretability. MOVIE-QA[77], TVQA[41], HowtoVQA69M[93] provide explicit text in the form of subtitles. Multiple-choice datasets[77, 41, 92] consist of a pre-defined set of options for answers. When compared to open-ended datasets, they can be considered limiting in the context of real-world applications. Synthetically generated datasets[91, 96, 13] contain questions that are generated through processing video descriptions, narration and template questions. MSRVTT-QA[91] exploits the video descriptions for QA creation. HowToVQA69M[93] uses cross-modal supervision and language models to generate question-answer pairs from narrated videos, whereas ActivityNetQA[96] uses template questions to generate the QA pairs. Xu et al. introduced the SUTD-TrafficQA[92] dataset and the Eclipse model for testing systems' ability to reason over complex traffic scenarios. The SUTD-TrafficQA[92] dataset contains multiple-choice questions that are based on different traffic events. RoadTextVQA is an open-ended dataset that deals with questions related to the text information found in road videos or the signs posted along roads. Recent studies[39, 40, 43, 42] on pretraining transformers on other vision and language tasks have shown excellent results for the VideoQA task. Lei et al. [39], in their study, uncovered the bias present in many video question-answering datasets, which only require information from a single frame to answer, and introduced new tasks aimed at training models to answer questions that necessitate the use of temporal information.

Table 2.1 Comparison of RoadTextVQA with existing video question answering datasets. "Text-based" indicates whether the questions require an understanding of the text present in the videos to answer. "Road-based" questions are datasets which are based on the driving domain. "Synthetic questions" are questions that are not manually annotated and depend on automated methods for question-answer generation. Abbreviations used - OE: Open-ended questions, MC: Multiple choice questions.

Dataset	Text-based	Road-based	Synthetic Questions	#Videos	#Questions	QA type
MovieQA[77]	✗	✗	✗	6.7K	6.4K	MC
MSRVTT-QA[91]	✗	✗	✓	10K	243.6K	OE
Activitynet-QA[96]	✗	✗	✓	5.8K	58K	OE
TVQA[41]	✗	✗	✗	21.7K	152.5K	MC
WildQA[13]	✗	✗	✗	0.4K	0.9K	OE
HowtoVQA69M[93]	✗	✗	✓	69M	69MK	OE
SUTD-TrafficQA[92]	✗	✓	✗	10K	62.5K	MC
NewsVideoQA[31]	✓	✗	✗	3K	8.6K	OE
M4-ViteVQA[102]	✓	✗	✗	7.6K	25.1K	OE
RoadTextVQA	✓	✓	✗	3.2K	10.5K	OE

2.2.2 VideoQA involving video text

NewsVideoQA[30] and M4-ViteVQA[102] are two recently introduced datasets that include videos with embedded born-digital text and scene text, respectively. Both datasets require an understanding of the text in videos to answer the questions. Embedded text, sometimes called video text in news videos, is often displayed with good contrast and in an easy-to-read style. Scene text in the RoadTextVQA dataset can be challenging to read due to factors such as occlusion, blur, and perspective distortion. M4-ViteVQA contains videos from different domains, a few of them being shopping, driving, sports, movies and vlogs. The size of RoadTextVQA is more than three times the size driving subset of M4-ViteVQA. Additionally, a subset of questions in RoadTextVQA also requires domain knowledge to answer questions related to road signs. Few recent works[86, 14] on vision and language transformers have shown to work well with text-based VQA tasks. Kil et al.[35] introduced PreSTU, a pretraining method that improves text recognition and connects the recognized text with the rest of the image. GIT(GenerativeImage2Text)[86] is a transformer-based model for vision and language tasks with a simple architecture that does not depend on external OCR or object detectors.

2.2.3 Scene Text VQA

Our work in Chapter 3, which focuses on VQA requiring text comprehension within videos, shares similarities with other studies dealing with text in natural images, commonly known as Scene Text VQA. The ST-VQA[9] and TextVQA[76] datasets were the first to incorporate questions requiring understanding textual information from natural images. LoRRa[76] and M4C[29] utilized pointer networks[84] that generate answers from a fixed vocabulary and OCR tokens. In addition, M4C used a multimodal transformer[81] to integrate different modalities. TAP[94] employed a similar architecture to M4C and incorporated a pretraining task based on scene text, improving the model's alignment among the three modalities. Another study, LaTr[7], focused on pretraining on text and layout information from document images and found that incorporating layout information from scanned documents improves the model's understanding of scene text.

Chapter 3

Reading Between the Lanes: Text VideoQA on the Road

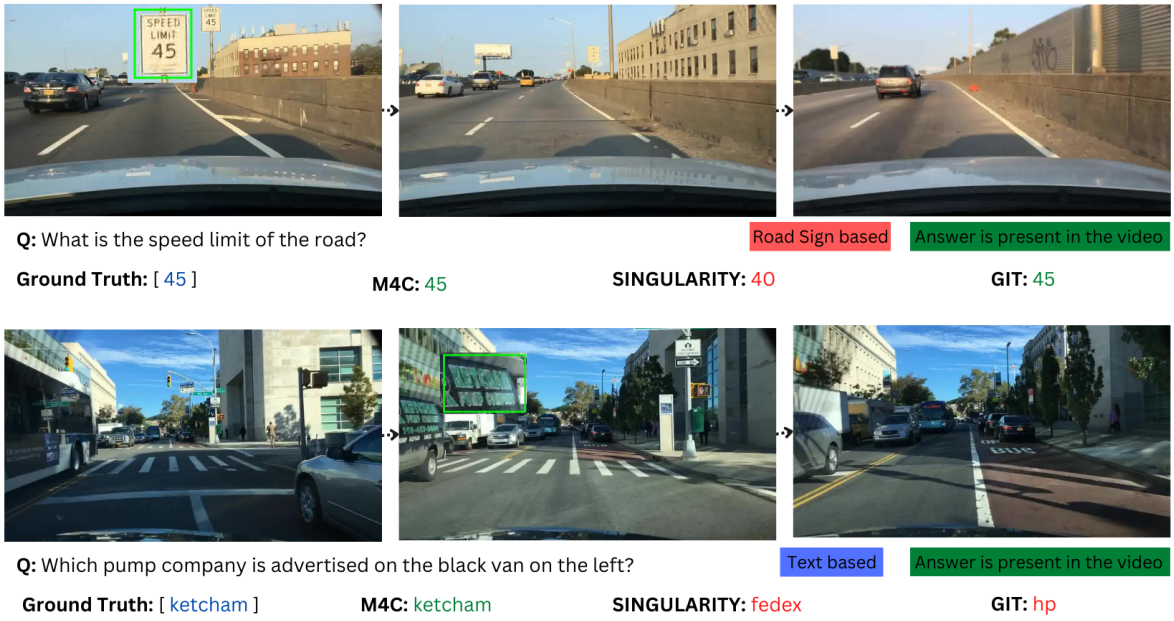


Figure 3.1 Examples from our RoadTextVQA dataset. The question in the first clip is based on the speed limit road sign, so it is classified as a "road sign based" question. Meanwhile, the question for the clip in the second row draws information from the text on the van, making it a "text based" question. The ground truth answers and the baseline predictions are also presented.

3.1 Introduction

We propose a new dataset for Visual Question Answering (VQA) on driving videos, with a focus on questions that require reading text seen on the roads and understanding road signs. Text and road signs

provide important information to the driver or a driver assistance system and help to make informed decisions about their route, including how to reach their destination safely and efficiently. Text on roads can also provide directions, such as turn-by-turn directions or the distance to a destination. Road signs can indicate the location of exits, rest stops, and potential hazards, such as road construction or detours. Reading text and understanding road signs is also important for following traffic laws and regulations. Speed limit signs, yield signs, and stop signs provide important information that drivers must follow to ensure their safety and the safety of others on the road.

VQA is often dubbed as the Turing test for image/video understanding. The early datasets for VQA on images/vqa and videos [2, 77, 91] largely ignored the need for reading and comprehending text on images/vqa and videos, and questions were mostly focused on the visual aspects of the given image or video. For example, questions focused on the type, attributes and names of objects, things or people. However, the text is ubiquitous in outdoor scenes, and this is evident from the fact that nearly 50% of the images/vqa in the MS-COCO dataset have text in them [82]. Realizing the importance of reading text in understanding visual scenes, two datasets—Scene text VQA [9] and Text VQA [76] were introduced that focus exclusively on VQA involving scene text in natural images/vqa. Two recent works called NewsVideoQA[31], and M4-ViteVQA[39] extend text-based VQA works to videos by proposing VQA tasks that exclusively focus on question-answers that require systems to read the text in the videos.

Similar to these works that focus on text VQA on videos, our work proposes a new dataset where all the questions need to be answered by watching driving videos and reading the text in them. However, in contrast to NewsVideoQA which contains news videos where question-answer pairs are based on video text (born-digital embedded text) appearing on news tickers and headlines, the text in videos in our dataset is scene text. The text in the road or driving videos are subjected to blur, poor contrast, lighting conditions and distortions. Text while driving goes by fast and tends to be heavily occluded. Often, multiple frames needs to be combined to reconstruct the full text, or a good frame with readable text needs to be retrieved. These difficulties made researchers focus on road-text recognition exclusively, and there have been works that focus exclusively on the detection, recognition and tracking of road text videos [68, 24]. On the other hand M4-ViteVQA contains varied type of videos such as sports videos, outdoor videos and movie clips. A subset of these videos are driving videos. In contrast, our dataset is exclusively for VQA on driving videos and contains at least three times more questions than in the driving subset of M4-ViteQA. Additionally, questions in our dataset require both reading road text and understanding road signs, while M4-ViteVQA's focus is purely on text-based VQA.

3.2 RoadTextVQA dataset

This section looks at the data collection and annotation procedure, data analysis, and statistics.

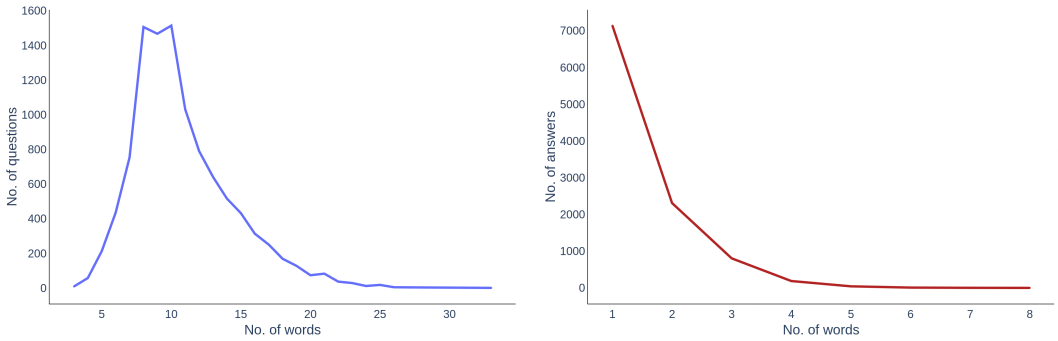


Figure 3.2 Distribution of the number of words in the question(left) and answer(right) of RoadTextVQA

3.2.1 Data Collection

The videos used in the dataset are taken from the RoadText-3K[24] dataset and YouTube. The RoadText-3K dataset includes 3,000 ten-second road videos that are well-suited for annotation because they have a considerable quantity of text. The RoadText-3K dataset includes videos recorded in the USA, Europe, and India and features text in various languages such as English, Spanish, Catalan, Telugu and Hindi. Each video contains an average of 31 tracks. However, the European subset is excluded from the annotation process for RoadTextVQA as it is dominated by texts in Spanish/Catalan, and the RoadTextVQA is designed specifically for English road-text. In addition to the videos from RoadText-3K, additional dashcam videos were sourced from the YouTube channel J Utah[1]. 252 videos from the USA and the UK were selected, and clips with a substantial amount of text were further selected by running a text detector over the video frames. Being a free and open-source text detector popular for scene text detection, we went with EasyOCR[32] as our choice of text detector. The RoadText-3K videos have a resolution of 1280x720 with a frame rate of 30 frames per second. The YouTube clips were downsampled to the same resolution and frame rate of 1280X720 at 30fps to keep the data consistent.

Individuals who are proficient in the English language were hired to create the question-answer pairs. To ensure the quality of the applicants, an initial training session was conducted, followed by a filtering mechanism in the form of a comprehensive quiz. The quiz was designed to ensure that the question-answer pairs were created by individuals who had a solid grasp of the English language and a good understanding of the task, thereby enabling us to maintain a high standard of quality in the annotations. The annotation process involved two stages, and a specifically designed web-based annotation tool was used. In the initial stage, annotators add the question, answers and timestamp triads for videos shown to them. All the questions have to be based on either some text present in the video or on any road sign. In cases where a question could have multiple answers in a non-ambiguous way, the annotators were given the option to enter several answers. The timestamp is an additional data point which is collected, and it

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

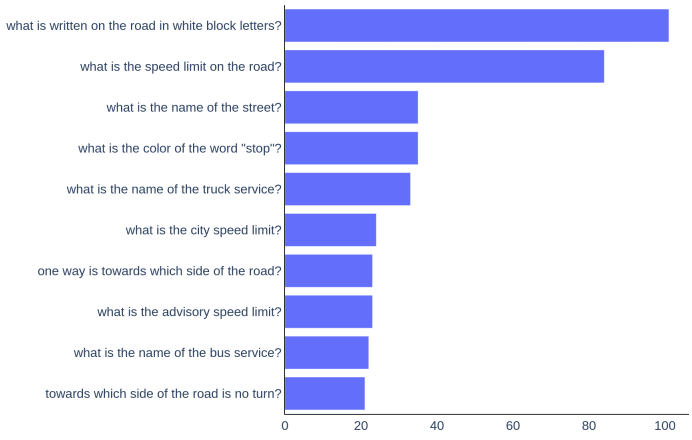


Figure 3.3 Top 10 questions in the dataset.

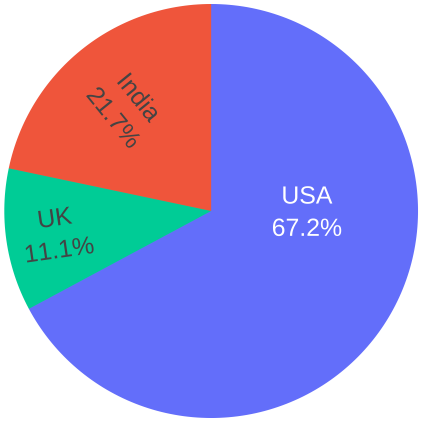


Figure 3.4 Geographical distribution of videos in the RoadTextVQA dataset.

is the aptest point in the video at which the question is answerable. The annotators were instructed to limit the number of questions to not more than ten per video and to avoid asking any questions related to the vehicle license plate numbers. If there were no possible questions that could be asked from the video, then the annotators were given the option to reject it. In the verification stage, the video and the questions are shown, and the annotators had to add the answers and the timestamps. We made sure that verification is done by an annotator different from the one who has annotated it in the first stage.

Table 3.1 Comparison of average and maximum question and answer lengths with other text based video question answering datasets.

Dataset	Average Length		Max Length	
	Question	Answer	Question	Answer
M4-ViteVQA[102]	6.75	1.94	24	26
NewsVideoQA[31]	7.04	2.02	20	19
RoadTextVQA	10.78	1.45	33	8

The question is flagged and rejected if it is incorrect or does not follow the annotation guidelines. If there are common answers in the annotation stage and verification stage for a question, then that question is considered valid. All the common answers are considered valid answers to the question. In the verification stage, additional data regarding the question-answers are also collected. The questions are categorically tagged into two distinct classes. Firstly, based on the type of question— text-based or

traffic sign-based. The second classification captures whether the answer for a question, i.e., the text that makes up the answer, is present in the video or not.

3.2.2 Annotation Tool

We built upon the DocVQA and InfographicVQA annotation tools, extending its video data collection functionality. The tool, built using Django and SQLite, was enhanced with the Video.js library for seamless video playback, improving the annotation experience for both users and annotators. Dedicated queues managed the data to be annotated and validated. Items assigned to users disappeared from their queue and were automatically re-added if not annotated within three days. Further enhancements included video timestamp collection and additional verification checks. These checks ensured data accuracy by mandating checks like the selection of either "Answer is a text present in the video" or "Answer is a text not present in the video."

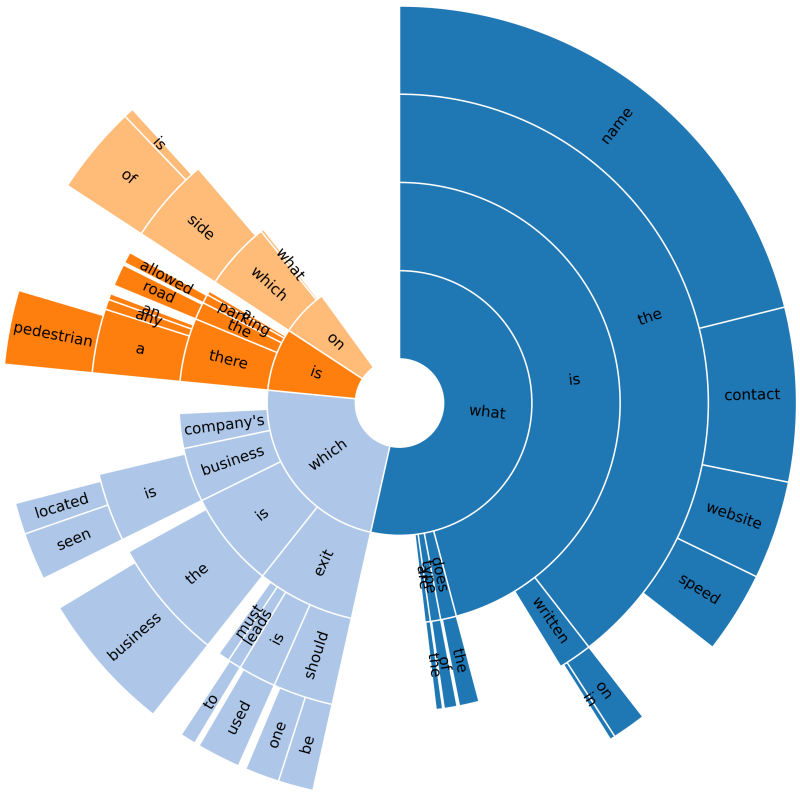


Figure 3.5 An analysis of the distribution of questions based on their starting 4 grams has shown that a significant proportion of questions are aimed at obtaining the name and contact information of businesses located along roads, as well as obtaining the speed limit for the road.

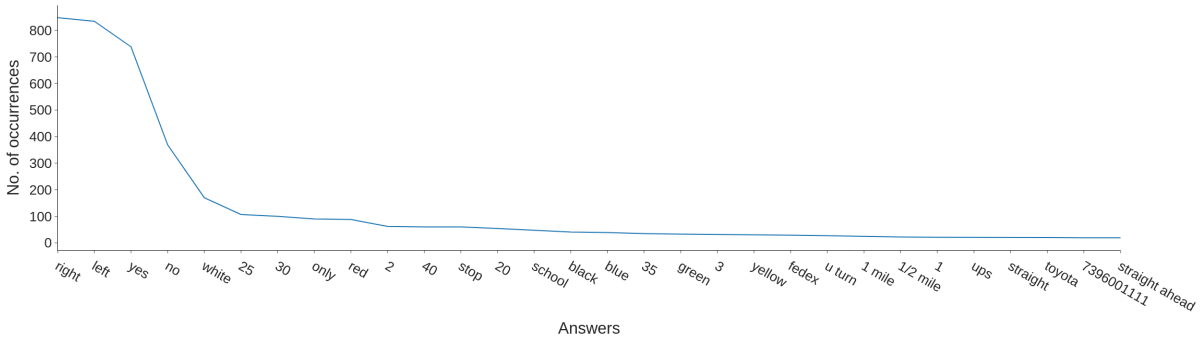


Figure 3.6 The number of occurrences of the answers in RoadTextVQA. The most recurring answer is "right", which makes up about 8% of the answers.

3.2.3 Data Statistics and Analysis

The RoadTextVQA dataset contains 3,222 videos and 10,500 question-answer pairs. Among the 3,222 videos, 1,532 videos are taken from the RoadText-3K dataset and the rest are from YouTube. The data is randomly split into 2,557 videos and 8,393 questions in the train set, 329 videos and 1,052 questions in the test, and 336 videos and 1,055 questions in the validation set.

The videos for the test and validation sets were randomly chosen from the RoadText-3K split, as it has ground truth annotations for text tracking. Methods that use OCR data can take advantage of the accurate annotations provided by RoadText-3K.

We present statistics related to the questions in RoadTextVQA through Figure 3.2, and Figure 3.3. Figure 3.3 shows the most frequent questions and their frequencies. "What is written on the road with white block letters?" is the most recurrent, followed by questions regarding the speed limits on the roads.

Figure 3.5 provides a comprehensive overview of the question distribution in RoadTextVQA, with the majority of the questions being centred around details of shops located along the road. Figure 3.2 depicts the word count in the questions and answers, respectively. The average number of words in the questions in RoadTextVQA is 10.8, while the average number in the answers is 1.45. The average number of words in questions is much higher when compared to other text-based VideoQA datasets, as seen in Table 3.1. The percentage of unique questions stands at 86.6%, while the percentage of unique answers is 40.7%. Figure 3.6 shows the top 30 answers and the number of occurrences. Figure 3.8, in the form of a word cloud, illustrates the most frequently occurring answers and OCR tokens. The most popular answers are "right", "left", "yes", and "no". The most prevalent OCR tokens in the videos are "stop", "only", and "one way". The distribution of the videos in the dataset based on the geographic location where it was captured is shown in Figure 3.4. More than two-thirds of the videos in the dataset are captured from roads in the USA.

The majority of questions are grounded on text seen in the video (61.8%), and the rest are based on road signs. Road signs can also contain text, such as speed limit signs or interchange exit signs. 68% of

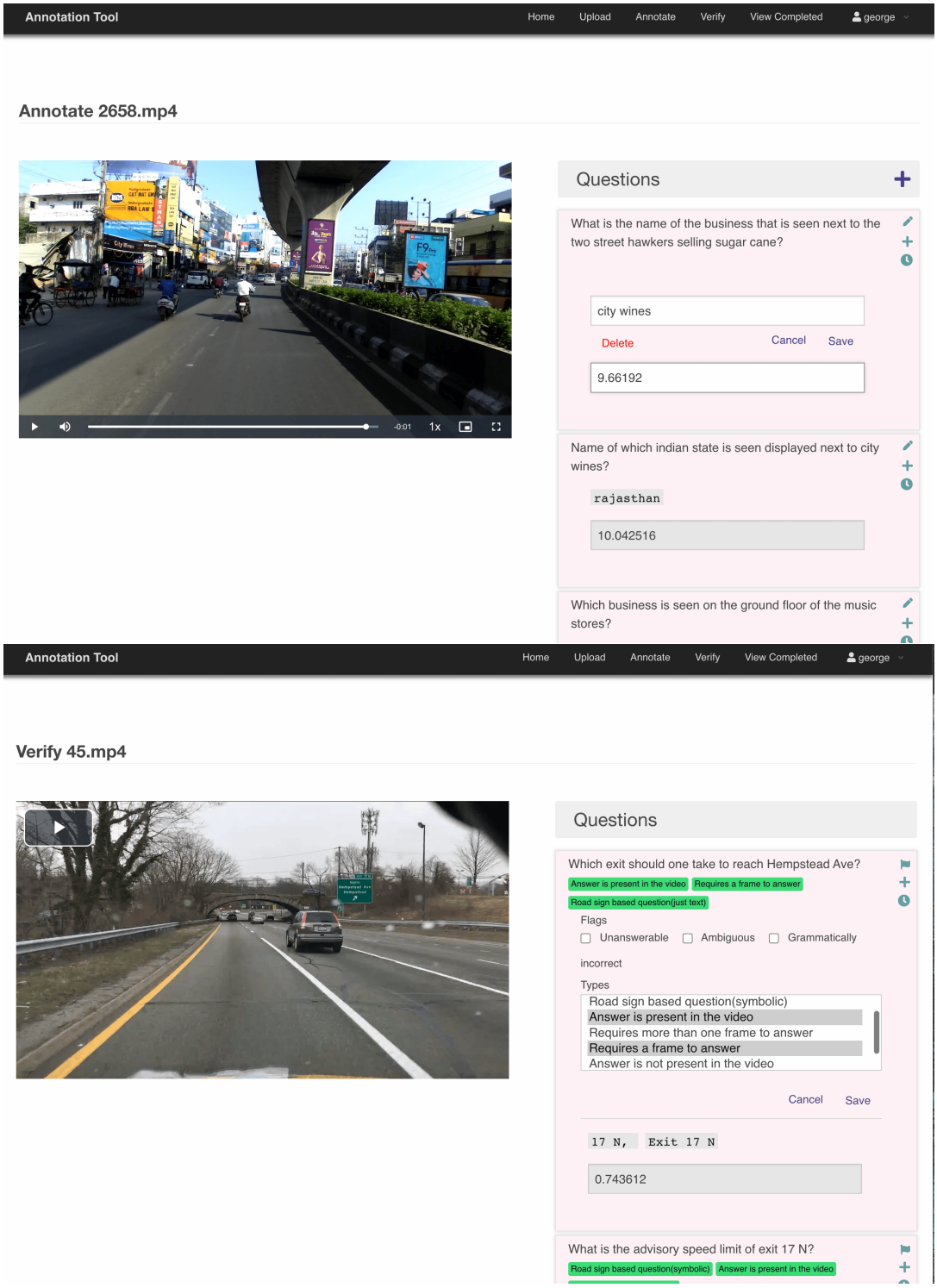


Figure 3.7 The annotation tool features a dedicated workflow for adding and verifying annotations. The top screenshot shows the Annotation Page, where users directly add annotations to the content. The bottom screenshot features the Verification Page, allowing users to review annotations, add question flags and types, and validate submissions to ensure correctness and consistency.

Figure 3.8 A visual representation of word frequency in the form of a word cloud, depicting the distribution of words in answers (left) and the distribution of OCR tokens from the videos (right).

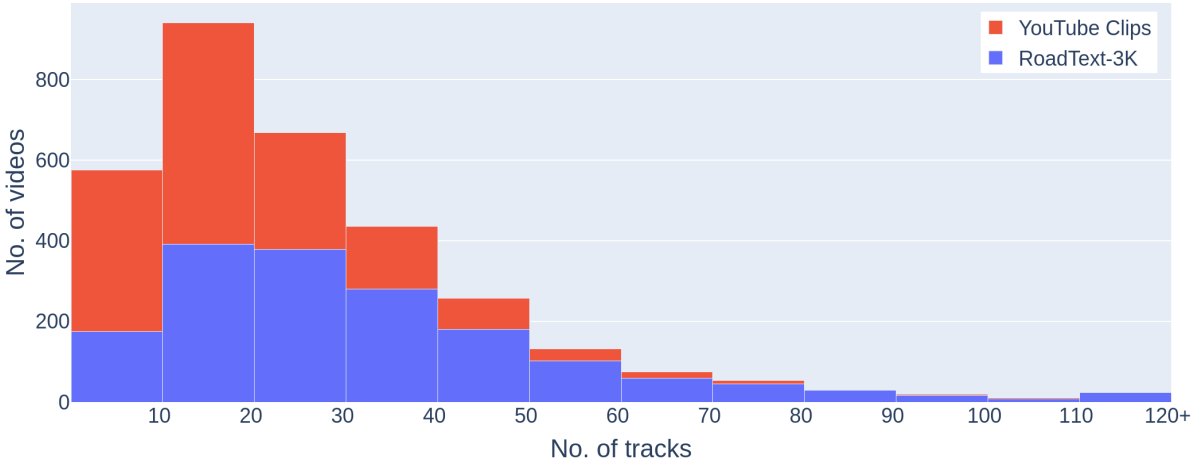


Figure 3.9 Distribution of number of videos vs number of tracks.

questions have answers that can be found within the text present in the video, while the remaining 32% of questions require an answer that is not a text present in the video.

3.3 Baselines

This section presents details of the baselines we evaluate on the proposed RoadTextVQA dataset.

3.3.1 Heuristic Baselines and Upper Bounds

We evaluate several heuristic baselines and upper bounds on the dataset. These heuristics and upper bounds are similar to those used in other VQA benchmarks, such as TextVQA[76] and DocVQA[58]. The following heuristic baselines are evaluated: (i) **Random Answer:** performance when answers to questions are randomly selected from the train split. (ii) **Random OCR token:** performance when a random OCR token from the video is picked as the answer. (iii) **Majority Answer:** performance when the most common answer in the train split is considered as the answer for all the questions. The following upper bounds are evaluated (i) **Vocab UB:** the upper bound on predicting the correct answer if it is present in the vocabulary of all the answers from the train split. (ii) **OCR UB:** the upper bound on performance if the answer corresponds to an OCR token present in the video. (iii) **Vocab UB + OCR UB**: this metric reflects the proportion of questions for which answers can be found in the vocabulary or the OCR transcriptions of the video.

3.3.2 M4C

The M4C[29] model uses a transformer-based architecture to integrate representations of the image, question and OCR tokens. The question is embedded using a pretrained BERT[19] model. Faster R-CNN[70] visual features are extracted for the objects detected and the OCR tokens in the image. The representation of an OCR token is formed from the FastText[10] vector, PHOC[1] vector, bounding box location feature, and Faster R-CNN feature of the token. A multi-head self-attention mechanism in transformers is employed, enabling all entities to interact with each other and model inter- and intra-modal relationships uniformly using the same set of transformer parameters. During answer prediction, the M4C model employs an iterative, auto-regressive decoder that predicts one word at a time. The decoder can use either a fixed vocabulary or the OCR tokens detected in the image to generate the answer.

3.3.3 SINGULARITY

The architecture of SINGULARITY[39] is made up of three major components: a vision encoder using ViT[21], a language encoder utilizing BERT[19], and a multi-modal encoder using a transformer encoder[81]. The multi-modal encoder uses cross-attention to collect information from visual representations using text as the key. Each video or image is paired with its corresponding caption during the pretraining phase, and the model is trained to align the vision and text representations using three losses (i) Vision-Text Contrastive: a contrastive loss that aligns the representations of vision and language encoders, (ii) Masked Language Modeling[19]: masked tokens are predicted (iii) Vision-Text Matching: using the multi-modal encoder, predict the matching score of a vision-text pair. We use the SINGULARITY-temporal model, which is pretrained on 17M vision caption pairs[82, 38, 64, 74, 15, 4]. The SINGULARITY-temporal model contains a two-layer temporal encoder that feeds its outputs into

the multi-modal encoder. SINGULARITY-temporal makes use of two new datasets named SSv2-Template Retrieval, and SSv2-Label Retrieval created from the action recognition dataset Something-Something v2 (SSv2)[28]. The pretraining is a video retrieval task using text queries. An additional multi-modal decoder is added for open-ended QA tasks and is initialised from the pretrained multi-modal encoder, which takes the multi-modal encoder's output as input and generates answer text with [CLS] as the start token.

3.3.4 GenerativeImage2Text

GIT(GenerativeImage2Text)[86] is a transformer-based architecture aimed at unifying all vision-language tasks using a simple architecture pretrained on 0.8 billion image text pairs. GIT consists of an image encoder and a text decoder and is pretrained on a large dataset of image text pairs. The image encoder is a Swin-like[52] transformer based on the contrastive pretrained model, which eliminates the need for other object detectors or OCR. As for the text decoder, the GIT uses a transformer with a self-attention and feed-forward layer to generate text output. The visual features and the text embeddings are concatenated and used as inputs to the decoder. GIT is able to gradually learn how to read the scene text with large-scale pretraining and hence achieves SoTA performance on scene-text-related VQA tasks such as ST-VQA. For video question answering, GIT employs a method of selecting multiple frames from the video and separately embeds each frame with a learnable temporal embedding which is initialized as zeros, and the image features are concatenated and used similarly to the image representation. The question and the correct answer are combined and used in as a special caption, and the language model loss is computed solely on the answer and the [EOS] token.

3.4 Experiments and Results

This section covers the evaluation metrics, the experimental setup, and the experiment results.

3.4.1 Experimental Setup

Evaluation metrics. We use two evaluation metrics to evaluate the model's performance: Accuracy and Average Normalized Levenshtein Similarity (ANLS)[9]. The Accuracy metric calculates the percentage of questions where the predicted answer exactly matches any of the target answers. If the ground truth contains multiple answers, a prediction is considered correct if the prediction exactly matches any of the ground truth answers. Unlike accuracy metric, which would give zero if a model makes a few character errors during recognition, ANLS gives a score between 0.5 and 1, gently penalizing OCR mistakes. The score was originally proposed to act softly on cases where the predicted answer differs slightly from the actual. ANLS assigns a score of 0 if the normalized Levenshtein distance exceeds 0.5, indicating that more than half of the answer is incorrect. Conversely, the ANLS score for the prediction is determined by taking one minus the normalized Levenshtein distance. In cases with multiple

predictions, the highest ANLS score from the prediction-ground truth pair is selected as the final ANLS score.

OCR transcriptions. The ground truth annotations were utilized for the videos in the RoadText-3K set, while for the remaining videos, the OCR transcriptions were sourced using the Google Cloud Video Intelligence API. Both RoadText-3K ground truth annotations, and the Google API provide text transcriptions at the line level. We use the line-level text transcriptions as the OCR tokens for the calculation of OCR upper bounds and OCR-based heuristics as given in the Table 3.2. When a text track gets cut off from the frame or partially occluded by other objects in a video, the Google Cloud Video Intelligence API treats it as a new track, whereas RoadText-3K annotations ignore the partially occluded tracks. This is why in the Figure 3.9, the number of videos vs the number of tracks is a bit inflated for the YouTube clips when compared to RoadText-3K clips.

Experimental setup for M4C. The M4C[29] model is trained using the official implementation, and the training parameters and implementation details remain consistent with those used in the original paper. We used a fixed vocabulary of size 3926 generated from the train set. The training data consists of image question-answer pairs where the image selected for training is the one on which the questions are based, specifically the timestamp frame. After training, the model is evaluated using two approaches. Firstly, it is tested on the timestamp QA pairs of the test set, and secondly, it is evaluated on the video level by sampling ten frames from the respective video for each QA pair and obtaining the model prediction for every frame individually. The final answer is determined by taking the most common answer from the ten individual frame predictions.

Experimental setup for SINGULARITY. We fine-tuned the pretrained SINGULARITY-temporal 17M model on four NVIDIA Geforce RTX 2080 Ti. The fine-tuning process was run for 20 epochs with a batch size of 16, starting with an initial learning rate of 1e-5 and increasing linearly in the first half epoch, followed by cosine decay[54] to 1e-6. The other parameters used for training are the same as the official implementation. The video frames were resized to 224x224, and a single frame with random resize, crop and flip augmentations were utilised during training, whereas 12 frames were used during testing. Additionally, we fine-tuned the SINGULARITY model, which has been pretrained on the MSRVTT-QA[91] dataset.

Experimental setup for GIT. The training process for GIT was carried out using a single Tesla T4 GPU for 20 epochs with a batch size of 2. We use an Adam[37] optimizer with an initial learning rate starting at 1e-5 and gradually decreasing to 1e-6 through the use of cosine decay. The GIT model was trained using the official VideoQA configuration used for MSRVTT-QA training. We fine-tuned the pretrained GIT-large model on our dataset, using six frames that were evenly spaced as inputs during both training and testing. In addition, we further fine-tuned the GIT model that was pretrained on the MSRVTT-QA[91] dataset.

Table 3.2 Performance of various heuristic baselines and upper bounds that are commonly evaluated on text-based VQA datasets.

Baseline	Test	
	ANLS	Acc.(%)
Random Answer	0.09	0
Random OCR token	3.20	1.98
Majority Answer	-	3.49
Vocab UB	-	59.26
OCR UB	-	36.67
Vocab + OCR UB	-	76.18

3.4.2 Results

Heuristic baselines and upper bound results are presented in the Table 3.2. The heuristic baselines yield very low accuracy, which indicates the absence of any bias due to the repetition of answers.

Random OCR heuristic gives close to 2% accuracy, meaning that there is enough text present in the video that selecting a random OCR from the video will not yield high accuracy. The OCR upper bound is 36.6% which is low when compared to the percentage of questions which have the answers present in the video. The low OCR UB can be attributed to how the text detection and how ground truth annotation is done. The response to a question may be split into multiple lines within the video, leading to the representation of the answer as separate tokens in the OCR output. This happens because the annotations in the OCR process were carried out on a line level. From the upper bound result of Vocab + OCR UB, we can see that more than three-quarters of the answers are present in either the vocabulary or in the OCR tokens of the video.

Table 3.3 Performance of RoadTextVQA on M4C. Abbreviations- TB: text-based questions, RSB: road sign-based questions, AP: questions where the answer is present in the video, ANP: questions where the answer is not present in the video.

Test Frames	TB		RSB		AP		ANP		All	
	ANLS	Acc. (%)	ANLS	Acc. (%)	ANLS	Acc. (%)	ANLS	Acc. (%)	ANLS	Acc. (%)
1	35.28	29.27	55.49	49.46	37.55	29.70	63.85	63.19	44.22	38.20
10	23.92	21.48	42.83	38.32	20.38	15.96	67.12	66.91	32.27	28.92

Table 3.4 Performance of RoadTextVQA on SINGULARITY and GIT. Abbreviations- TB: text-based questions, RSB: road sign-based questions, AP: questions where the answer is present in the video, ANP: questions where the answer is not present in the video.

Method	Pretrain Data	TB		RSB		AP		ANP		All	
		ANLS	Acc. (%)	ANLS	Acc. (%)	ANLS	Acc. (%)	ANLS	Acc. (%)	ANLS	Acc. (%)
SINGULARITY	-	15.38	14.04	45.29	33.04	17.36	9.22	61.71	61.33	28.62	22.45
SINGULARITY	MSVRTT-QA	17.25	15.22	47.84	36.46	19.50	11.50	63.98	63.19	30.79	24.62
GIT	-	18.09	14.38	50.36	39.82	20.98	12.16	65.65	65.05	32.34	25.61
GIT	MSRVTT-QA	**22.61**	**19.62**	**51.20**	**42.18**	**23.40**	**15.96**	**69.93**	**69.51**	**35.23**	**29.58**

The results on M4C are shown on Table 3.3. The frame level results, where we evaluate on the timestamp frame, show an accuracy of 38.20% and the video level results, where we evaluate on ten frames, give an accuracy of 28.92%. The results show that answering the question is still a challenging task, even when we reduce the complexity of the problem by providing the aptest frame for answering the question and ground truth OCR tokens.
We show the results after fine-tuning on SINGULARITY and GIT in Table 3.4. The accuracy of the questions requiring answers to be extracted from the video (AP) is comparatively lower, while the accuracy of the questions where the answer is not present in the video is comparatively higher. Compared to AP, ANP is less complex to answer because it involves a fixed set of answers. In contrast, AP requires dynamic extraction from OCR tokens, resulting in the ANP set having better accuracy than AP. Additionally, fine-tuning the model that has been pretrained on the MSRVTT-QA dataset shows improvement in accuracy across all categories(TB, RSB, AP, and ANP).

Fine-tuning GIT results in better performance compared to SINGULARITY. GIT also shows a similar trend when fine-tuned on pretrained MSRVTT-QA dataset. The "answer is present in the videos(AP)" subset has an improvement of 3.9% in accuracy when compared with SINGULARITY, whereas the "answer is not present(ANP)" in the videos subset has a gain of 6.3%. M4C tested on a single frame shows better results compared to VideoQA models. This can be attributed to the fact that we explicitly provide the OCR tokens and the correct frame on which the question is framed to the model. M4C tested on ten frames gives comparable results to GIT.

We show some of the qualitative results in Figure 3.10. As the complexity of the scene and the obscurity of the scene text increase, it becomes more and more difficult for the model to predict the correct answer. VideoQA baselines achieve better results on questions that do not require the extraction of answers from the video.

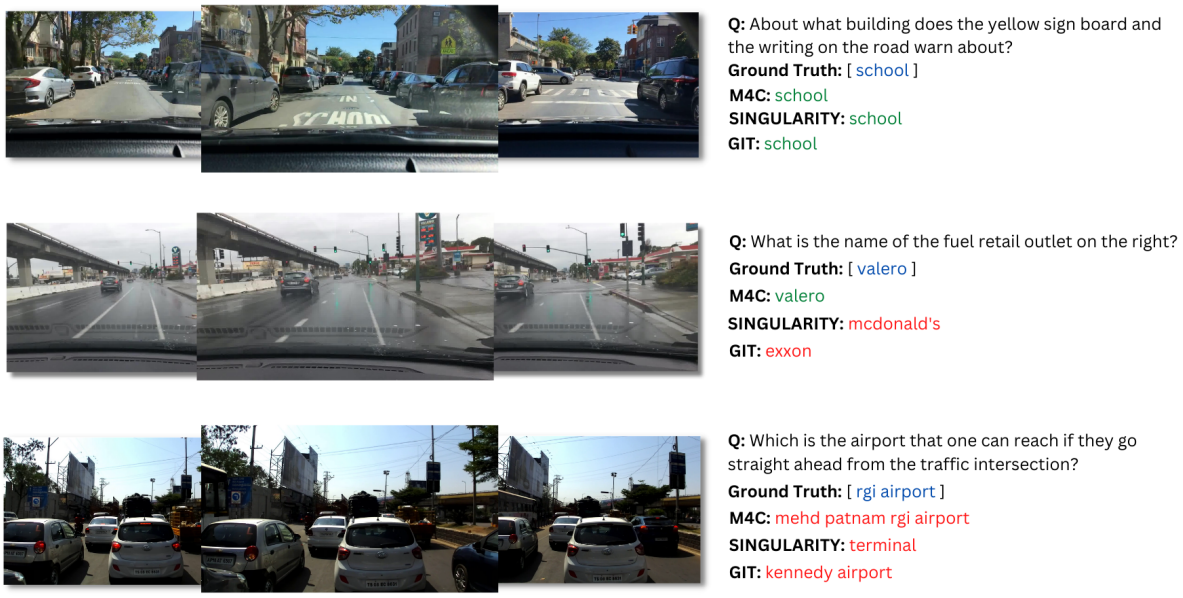


Figure 3.10 Qualitative results showing predictions of M4C, SINGULARITY and GIT. The correct predictions are highlighted in green, whereas the incorrect ones are highlighted in red.

3.5 Contextualizing Our Work in the LLM Era

Large language models and Vision language models have significantly transformed the landscape of natural language processing and computer vision. The M4C[29] model is inspired by a transformer, and BERT[19] architectures are only meant to work on VQA tasks. Different from these models, GPT3[11], LLaMA[80], and LLaVA[49] are pre-trained on a huge amount of data and show impressive zero-shot performance in a variety of tasks using prompts. Vision LLMs like GPT4[2], and LLaVA[49] have demonstrated remarkable proficiency in understanding and generating language contextualized by visual input.

Recent advancements in video understanding have led to the development of several sophisticated models that integrate large language models (LLMs) with video analysis capabilities. Video-LLaMA[99], Video-ChatGPT[56], and Video-LLaVA[48] are some of the recent works. These models leverage the power of LLMs to provide more nuanced and context-aware interpretations of video content. Video-LLaVA utilizes a unique combination of visual feature extraction and language model integration to tackle video question-answering (VideoQA) tasks. It uses models like CLIP[66] to encode frames and aligns these with textual queries to enhance performance in temporal and causal reasoning. Video-ChatGPT[56] and Video-LLaVA[48] similarly enhance video comprehension by integrating advanced visual encoders and language models. Video-LLaVA uses models like CLIP to encode frames and aligns these with textual queries to enhance performance in temporal and causal reasoning.

[2]https://openai.com/index/gpt-4/

Our dataset builds upon this foundation by specifically addressing the need for reading and understanding text in driving videos. Unlike earlier datasets that primarily focused on the visual attributes of scenes or objects, our dataset emphasizes the critical task of text recognition and comprehension in dynamic environments requiring models that can dynamically interpret and integrate information across multiple frames. This focus aligns with the advancements brought forth by models like Video-LLaMA, Video-ChatGPT, and Video-LLaVA, which blend temporal video analysis with sophisticated language understanding to tackle complex VQA tasks. Although the work conducted in this chapter predates the introduction of Vision LLMs, our findings remain relevant as it establish a benchmark for Vision Language Models, particularly in their capacity to comprehend and interpret text and visual cues in driving videos.

3.6 Summary and Future work

In this chapter, we look at RoadTextVQA, a new Video Question Answering dataset and its associated task, where the questions are grounded on the text and road signs present in the road videos. We offer a thorough comparison of our dataset with other datasets. We evaluate different models, including M4C, Singularity, GIT and many heuristic baselines. Our findings from the baseline models' performance indicate a need for improvement in existing VideoQA approaches for text-aware multimodal question answering. Currently, there are no Visual Question Answering models that explicitly incorporate road signs. Models can integrate road signs as an additional input or pretrain on road sign-description pairs to enhance their ability to respond to questions that require domain knowledge. We believe this work would encourage researchers to develop better models that incorporate scene text and road signs and are resilient to the challenges posed by driving videos. Additionally, drive further research in the area of scene text VideoQA and the development of advanced in-vehicle support systems.

Chapter 4

Evaluating the current state of RoadText Detection, Tracking and Recognition

Figure 4.1 Sample frames from RoadText-1K illustrating the various challenges and artifacts like glare, raindrops, out-of-focus, low contrast, and motion blur often encountered in driving videos.

4.1 Introduction

Text in driving videos exhibits a multifaceted complexity. Unlike their static counterparts, road signs, billboards, and other textual elements are subject to camera motion, varying lighting, occlusions, diverse scales, orientations, and fonts. The text's sheer volume and heterogeneity also necessitate adaptive and robust algorithms that interpret standardized road signage and non-standard text formats. Accurately detecting, tracking, and recognizing text amidst these challenges unlocks many opportunities for autonomous driving, traffic monitoring, and accessibility applications. The last text-tracking competition occurred nearly a decade ago and introduced the Text in Videos[34] dataset, which comprises 51 egocentric videos encompassing indoor and outdoor scenarios. The USTB-VidTEXT[79] and YouTube Video Text(YVT)[63] datasets contain videos sourced from YouTube. The USTB-VidTEXT dataset primarily consists of text in the form of overlaid captions, whereas the YouTube Video Text dataset 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 previous Robust Reading Challenge on text detection in videos, the ICDAR 2013-15 Text-in-Videos Challenge, the RoadText-1K[69] dataset used in our competition is significantly larger and more diverse. Specifically, the RoadText-1K dataset contains 1000 10-second driving videos, which is 20 times larger than the ICDAR 2013 dataset. The text objects in driving videos typically have short lifetimes, which require models tolerant to occlusions, able to 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 caused by vehicle movement. As a result, approaches developed on existing text-centric datasets tend to be challenging to adapt to real-world applications, such as driver assistance and self-driving systems.

The competition aimed to provide a robust benchmark and platform for researchers and developers to evaluate and refine their text-processing algorithms in a realistic driving video setting. The competition attracted multiple participants, showcasing the immense potential of computer vision and deep learning in tackling this complex task. This report delves into the performance of the submissions, analyzing their strengths and weaknesses. The Robust Reading Competition (RRC)[1] portal serves as the host platform for the Challenge. Submissions are assessed through automated methods, and the outcomes depicted in this chapter represent the state of submissions at the conclusion of the RoadText 2023 Challenge. However, the challenge will remain open to accept new submissions that are not considered part of the official competition.

Table 4.1 Comparison of RoadText-1K with existing text video datasets.

Dataset	Text in Videos [34]	USTB-VidTEXT [79]	YouTube Video Text [63]	RoadText-1K [69]
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

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

4.2 The RoadText-1K Dataset

The RoadText-1K[69] dataset comprises 10-second video clips extracted from the BDD100K[95] dataset. The videos are 720p and 30 fps, and capture diverse locations, weather conditions (such as sunny, overcast, and rainy), as well as 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. In the English category, it further distinguished between English text and license plate text. 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. License plates are separately tagged in the ground truth.

4.3 RoadText-1K Challenge

4.3.1 Evaluation Metrics

The evaluation is based on an adaptation of the CLEAR-MOT [6, 59] and DukeM-TMC[72] framework, designed to track multiple objects. Each approach is evaluated using three different metrics, namely Multiple Object Tracking Precision (MOTP), Multiple Object Tracking Accuracy (MOTA), and IDF1 score. Additionally, the number of objects tracked for at least 80 percent of their lifespan is noted as "Mostly Matched," while those tracked between 20 and 80 percent of their lifespan are "Partially Matched," and those tracked for less than 20 percent of their lifespan are categorised as "Mostly Lost.". MOTA metric will be used to rank the submissions. 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 word recognition is correct. The assessment of word recognition is case-insensitive and is only done for English category tracks. Leading and trailing spaces are disregarded, and instances of two or

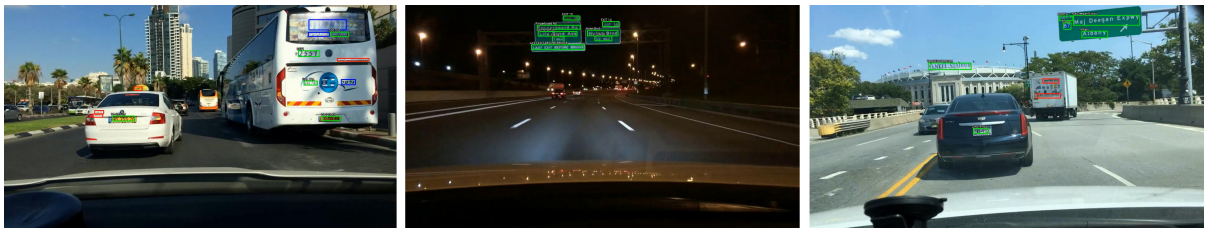


Figure 4.2 These are sample frames from clips in RoadText-1K, and they have annotations indicating the location and transcription of the text overlaid on them. The boxes that are colored green indicate English text, the ones in blue represent non-English text, and the red boxes represent illegible text.

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 will 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 will 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.

$$\text{MOTA} = 1 - \frac{\sum_t (m_t + fp_t + mm_t)}{\sum_t g_t}$$

MOTA measures overall tracking performance, considering both false positives, missed detections, and mismatches. Where t represents a frame in the video sequence, m_t denotes the number of missed objects in frame t, fp_t denotes the number of false positives in frame t, mm_t denotes the number of mismatches (incorrect ID assignments) in frame t, g_t denotes the total number of ground truth objects in frame t.

$$\text{MOTP} = \frac{\sum_{i,t} d_{i,t}}{\sum_t c_t}$$

MOTP measures localization accuracy of tracked objects. it is calculated as the average overlap score $(d_{i,t})$ between estimated and ground truth bounding boxes across all tracked objects i and frames t, divided by the total number of correctly identified and localized objects (c_t) in each frame t. Higher MOTP indicates better localization precision.

$$IDF1 = \frac{2 * IDTP}{2 * IDTP + IDFP + IDFN}$$

IDF1 measures the ratio of correctly identified detections to the average number of ground-truth and computed detections. Unlike MOTA, IDF1 focuses on the duration for which the tracker correctly identifies an object rather than simply counting errors. The global assignment is determined using the Hungarian algorithm, which selects the best combination of prediction and ground truth to maximize the IDF1 score for the entire video. IDTP is the number of correctly identified detections. IDFP is the tracker predictions that don't match any ground truth. IDFN is the Ground truth trajectories that aren't tracked.

4.3.2 Submitted Methods

The challenge received a total of 16 submissions, out of which 6 were unique. 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 - The TH-DL method provides an integrated approach for text detection, recognition, and tracking in driving videos. For text detection and recognition, the algorithm adopts TESTR[100] based on Transformer and finetunes the pre-trained TESTR model on the training set of the Roadtext Challenge. For multi-object tracking, ByteTrack[101] 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 - TencentOCR integrates the detection results of DBNet[46] and Cascade MaskRCNN[12], built with multiple Backbone architectures, with the Parseq[5] English recognition model for recognition and further improves the end-to-end tracking with OCSort. The result is end-to-end tracking and trajectory recognition.

TransDetr - The method used in this submission is TransDETR[89]. The approach involves pretraining the network weights on the ICDAR2015 video[33] and fine-tuning the network on the RoadText-3K[23] and BOVText[90] datasets for 20 epochs each. Finally, the network is fine-tuned on the RoadText-1K dataset for 20 epochs.

RoadText DRTE - EasyOCR[32] is used to perform the subtasks of detection and recognition on the RoadText-1K[69] dataset. The algorithm uses the CRAFT[3] algorithm for detection and the CRNN[75] 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.

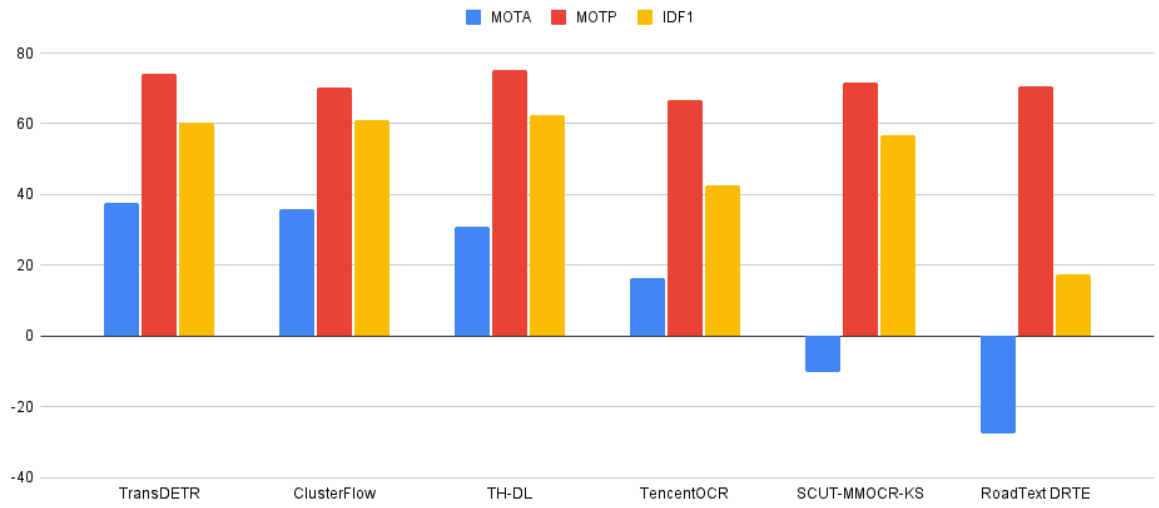


Figure 4.3 The chart illustrates the results for text detection and tracking, with MOTA, MOTP, and IDF1 represented by blue, red, and yellow bars, respectively.

Table 4.2 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

SCUT-MMOCR-KS - This submission utilizes DBNet++[47] for text detection, which is first pre-trained on a collection of TextOCR, HierText[53], DSText, YVT[63], ICDAR2015-Video[33], and Minetto before being fine-tuned on DSText. For text recognition, a ViT-based[20] recognizer is used, which is pre-trained on 10M unlabeled real STR LaTex/images/tracking and fine-tuned on 4M labelled real STR LaTex/images/tracking. CoText tracking module is used for text tracking.

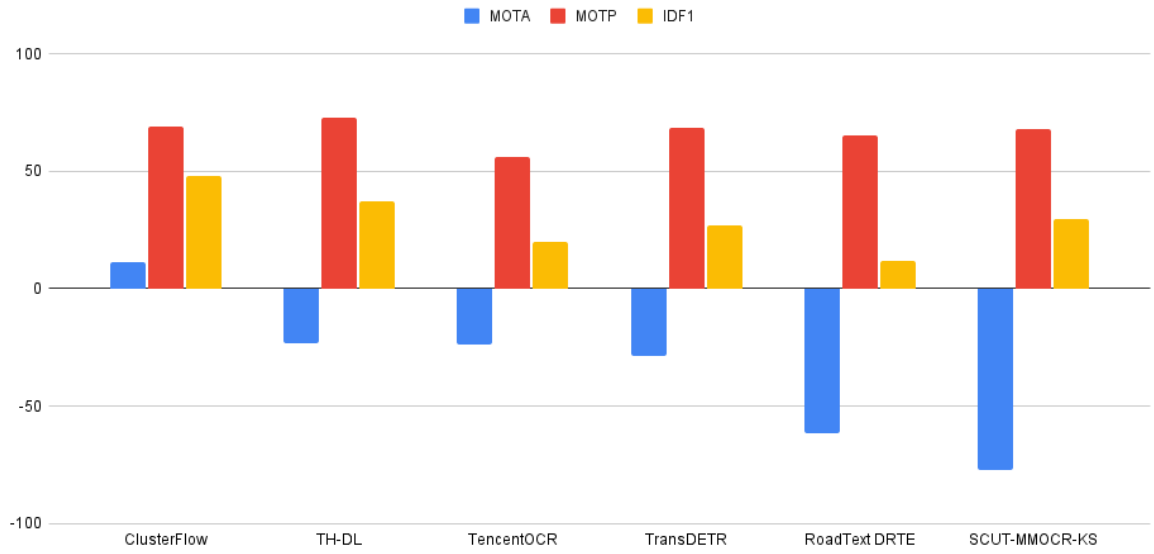


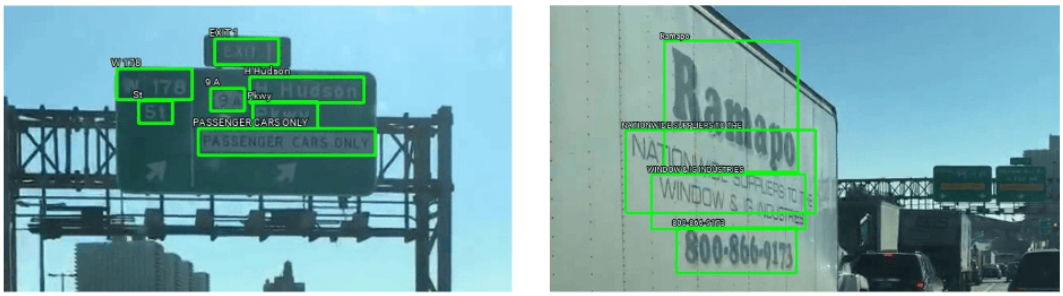
Figure 4.4 The chart illustrates the results for text detection, tracking and recognition, with MOTA, MOTP, and IDF1 represented by blue, red, and yellow bars, respectively.

Table 4.3 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

4.3.3 Analysis

The results of the evaluation are presented in Table 4.2 and Table 4.3, 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. Compared to TESTR, ViT-based, and Parseq methods used for recognition, the commer-



(i) Ground Truth

(ii) ClusterFlow

(iii) TH-DL

(iii) TencentOCR

Figure 4.5 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.

cial Google OCR performs better. 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, other 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 methods 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.

4.4 Summary

We present the analysis and results of the RoadText challenge in this chapter. Unlike previous text video tracking challenges, our challenge involves more complex scenarios, including driving videos not originally designed for text recognition. This adds a new level of complexity to the task, as these videos are much more realistic and present new challenges for text recognition systems. The report summarizes the unique features of the RoadText-1K dataset, detailing why it is particularly challenging and different from previous datasets. Additionally, it provides an analysis of the submissions. The RoadText challenge will remain open for new submissions in the future, providing a platform for researchers to benchmark and showcase their methods.

Chapter 5

Occluded RoadText

Figure 5.1 Instances of occluded text from the Occluded RoadText dataset (left) are compared with instances of occluded text from another dataset (right). The occlusions in our dataset represent real-life scenes, whereas the occlusions in the other datasets are manually created and are less representative of real-life scenarios.

5.1 Introduction

The rapid advancement in computer vision and deep learning has impacted automated driving, augmented reality, and information retrieval. However, robust scene text detection and recognition, particularly in the presence of occlusions, pose a significant challenge. This difficulty is of practical importance in interpreting real-world environments. The task is not only challenging but also holds a substantial impact in various real applications.

Scene text often appears in unstructured environments with varying degrees of occlusion, making its detection and recognition a complex task. This complexity is often amplified in urban landscapes, where texts on signs, billboards, and storefronts are frequently partially obscured by objects, lighting conditions, or angles. In Occluded RoadText we focus on scene text seen during driving. The existing datasets[88, 67, 26] for occluded scene text often suffer from a significant limitation: their occlusions are predominantly manually created, diverging considerably from the complexities and unpredictability encountered in real-life scenarios. This artificiality in the creation of occlusions can lead to a gap between the model's performance in controlled, experimental settings and its efficacy in real-world applications. In reality, occlusions can arise from a multitude of sources, including natural elements, urban infrastructure, dynamic crowd movement, and more. Each source presents unique challenges in terms of shape, size, texture, and positioning relative to the text.

Figure 5.2 Cropped out test image in the middle and the additional images on the left and right help to recognize occluded text.

To address these challenges, we introduce the Occluded RoadText dataset, focusing on scene text seen while driving. It comprises of 1,019 images with 3,659 instances of occluded text. These images were captured using side-mounted car cameras in India, providing a diverse and realistic set of occlusion scenarios. The typical scene text detection evaluation consists of two tasks: (i) Text localization, which involves locating all text instances within an image, and (ii) End-to-end recognition, which involves both localizing and recognizing the text. Additionally, we introduce a new task called "Multi-Image

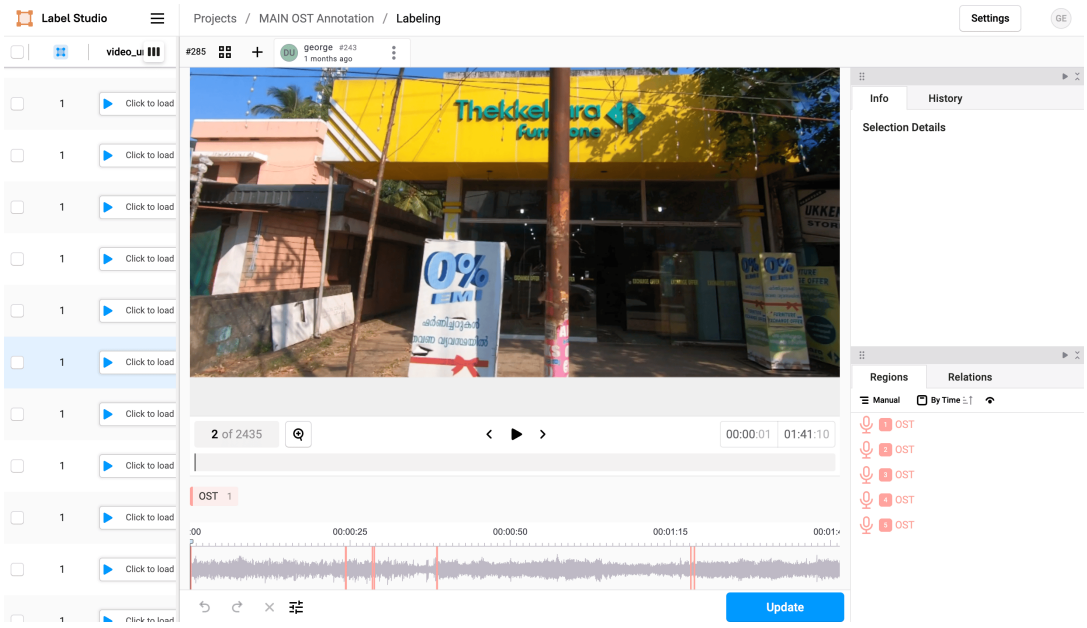


Figure 5.3 The label studio annotation tool is used to segment the video timeline to extract the occluded sections from the larger clip.

End-to-End Recognition," where two supplementary images of the same scene but from different angles are provided to aid in recognition. We evaluate our dataset using the TESTR[100] baseline and report the results. Additionally, the benchmark for evaluation is hosted at https://rrc.cvc.uab.es/?ch=29. Custom metrics are calculated to evaluate the methods and rank them based on how well they localize and recognize occluded texts.

5.2 RoadText Occluded Dataset

5.2.1 Data Collection

The images are extracted from videos captured in India using side-mounted cameras on a car, with cameras positioned on both sides. The recorded footage was subsequently edited into shorter 2-minute clips. Annotators were then tasked with identifying segments of text where at least one instance of text was obscured. These segments include at least one track where text is consistently hidden throughout the clip, while all characters in the track are visible at some point in the video. It is important to note that the occluded track must be in English. The annotation process employed the label studio tool for this initial phase.

After the initial annotation, the next stage involved annotating polygons and transcribing the text using the CVAT tool. A single image from the set was chosen, and all the text instances were annotated using polygons. The text instances were also categorized during annotation into four groups: (i) Oc-

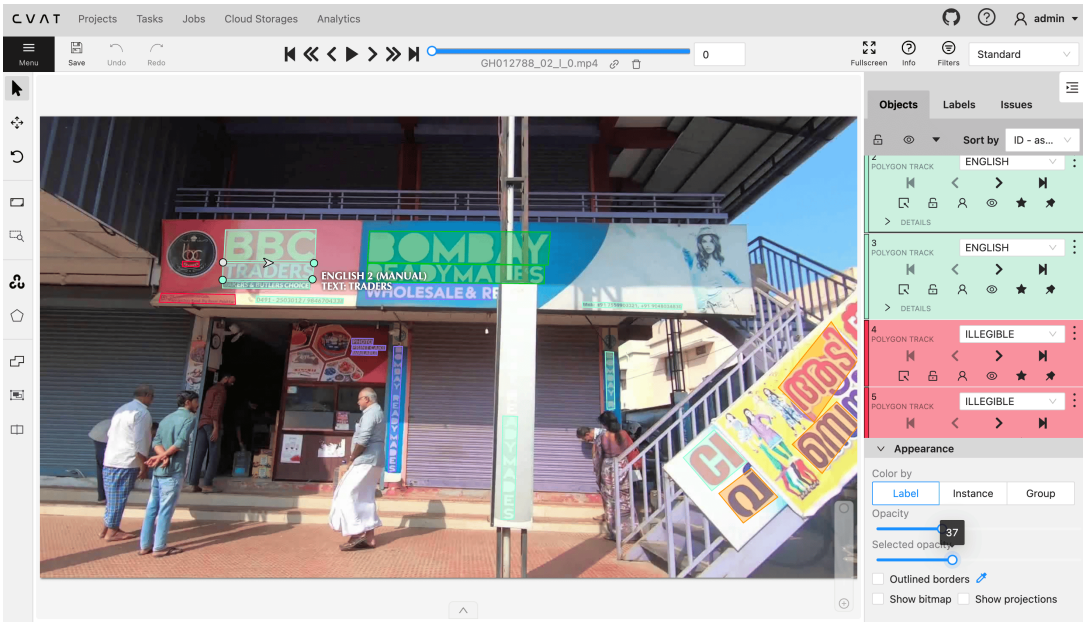


Figure 5.4 Screenshot of CVAT used for text instance annotation and tracking. This includes transcript labelling, bounding box creation, and instance-level tracking.

cluded English: English text instances that are occluded, (ii) English: English text instances that are not occluded and are easily recognizable, (iii) Non-English: Texts that are not in English, and (iv) Illegible: Text that is unreadable even to human annotators. Text transcriptions were collected only for English and Occluded English. Additional information was collected on how the occluded text can be recognized, with categories including Occluded Visible where the text is occluded but visible in another text instance, Occluded Inferable where the occluded text can be inferred from visual cues, and Occluded Indeterminate where the occluded text can only be inferred using supplementary images. Two additional images were selected from the video as supplementary images, and they are different from the annotated image. One of the supplementary image comes before the annotated image, and the other one comes after the annotated image.

5.2.2 Data Statistics and Analysis

The Occluded Road dataset contains $1,019$ images, with a total of $35,261$ text instances. Each image contains at least one instance of occluded text. Out of the $1,019$ images, 202 images belong to the validation set, and the remaining 817 images belong to the test set. The split is done randomly.

Figure 5.6 shows the distribution of the different categories of text. Out of the $35,261$ instances, $16,641$ are illegible, $10,077$ are in English, $4,884$ are Non-English, and $3,659$ are Occluded English. Transcription is only available for both English and Occluded English categories. In the test set, the

Figure 5.5 An illustration depicting various types of occlusions. On the left, both ”BOMBAY” and ”READYMADES” are occluded but visible in another text instance, falling into the ”Occluded Visible” category. In the middle, ”SAMSUNG” is occluded but can be inferred from the visual color and the font, categorizing it as ”Occluded Inferable.” On the right, the occluded text is unrecoverable(Occluded indeterminate) without the assistance of additional information, such as supplementary images.

occluded text is further categorized. 387 instances belong to occluded visible, 264 instances belong to occluded inferable, and 2286 instances belong to occluded indeterminate.

Each image contains an average of 34 text instances and 3 instances of occluded text. The images are dense in text since they are captured from roadsides, and they contain a lot of signboards, shop signage, and advertisements. The word cloud of the transcriptions is shown in Figure Figure 5.7.

5.3 Baseline: TESTR

The TExt Spotting TRansformers(TESTR)[100] model, is an end-to-end framework designed for text spotting. It is a generic text spotting framework using Transformers for text detection and recognition in various environments. TESTR utilizes a single encoder and dual decoders for joint text-box control point regression and character recognition. Notably, it is free from Region-of-Interest operations and heuristics-driven post-processing procedures. TESTR is particularly effective for dealing with curved text boxes, where it addresses the need for adapting traditional bounding-box representations. Key components of TESTR include Multi-Scale Deformable Attention, which leverages multi-scale feature maps to handle small text instances, Dual Decoders for text detection and character recognition, and Box-to-Polygon Detection, which represents text instances using Bezier curves or polygons, making it suitable for arbitrarily-shaped text.

5.4 Experiments and Results

This section covers the evaluation metrics, the experimental setup, and the experiment results.

5.4.1 Experimental Setup

Evaluation Metrics: For text localization and end-to-end recognition, we calculate precision, recall, and F1 score. Additionally, recall is calculated for occluded text instances and subcategories of Occluded English. We calculate a customized F1 score with the recall of Occluded English categories and overall precision. Our approach prioritizes localizing and recognizing occluded text while minimizing false positives. In text localization, the condition for a ground truth (gt) and a prediction (pred) to match is if the Intersection over Union (IoU) of gt and pred is greater than 0.7. In end-to-end recognition, both IoU greater than 0.7 and transcriptions should match if transcriptions exist for the ground truth. We calculate the possible set of matches and then optimize to maximize the true positives. The illegible category is treated as "don't care" and is filtered out from detections. This means that a method will not be penalized if it fails to detect any illegible words, while a method that does detect them will not receive a higher score.

Experimental setup for TESTR: We evaluate the images using the TESTR[100] model, which has been fine-tuned on the CTW1500[98] dataset with polygonal annotations. The CTW1500 dataset includes annotations at the line level and encompasses a variety of horizontal and multi-oriented text. We filtered out predictions with scores lower than 0.3 before evaluation.

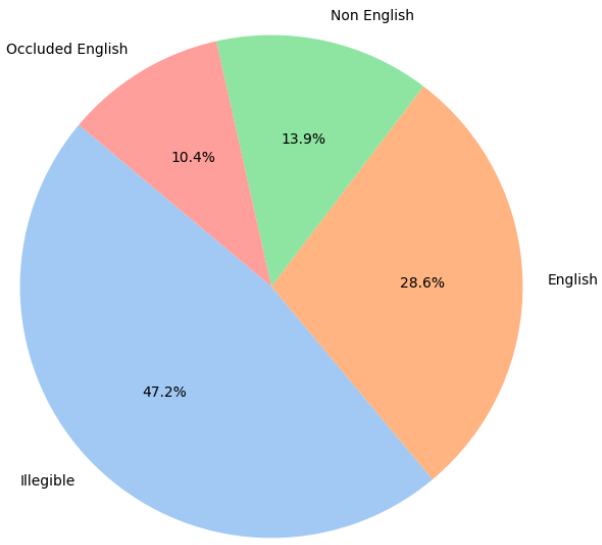


Figure 5.6 Distribution of categories of text instances.

Figure 5.7 A word cloud showing the frequency of words in text transcriptions.

5.4.2 Results

Occluded		Overall		
F-score	Recall	Precision	Recall	F-score
32.49%	23.39%	53.16%	28.27%	36.91%

Table 5.1 Performance of TESTR on text localization task

Occluded		Overall			Occluded Subcategory Recall		
F-score	Recall	Precision	Recall	F-score	Occluded Visible	Occluded Inferable	Occluded Indeterminate
4.42%	2.45%	22.66%	12.05%	15.73%	3.10%	4.92%	2.06%

Table 5.2 Performance of TESTR in end-to-end recognition task

The results on TESTR are shown in Table 5.1 and Table 5.2. Table 5.1 summarizes the performance of our text localization model under occluded and general conditions. The F-score for the occluded scenario is 32.49%, indicating the model's balanced performance considering both precision and recall. For the overall performance, the model's F-score is marginally higher at 36.91%, with a recall of 28.27% and a precision of 53.16%. The recall of occluded categories is lower when the match is calculated with an IoU of 0.7 compared to the overall recall. When matching is done with an IoU greater than 0.5, the

Figure 5.8 Qualitative results showing predictions of the TESTR model

overall recall (40.02%) and occluded recall (40.01%) were similar. This suggests that in most occluded cases, only one part of the occluded text is detected, and it's mostly the larger part.

Table 5.2 details the performance of TESTR on the end-to-end recognition task. For the occluded scenario, the model achieves an F-score of 4.42% and a recall of 2.45%. These results indicate a significant challenge in recognizing text when occlusion is present, with low recall suggesting the model struggles to detect and correctly recognize most occluded text instances. Additionally, we calculate recall metrics for three subcategories within the occluded category: Occluded Visible, Occluded Inferable, and Occluded Indeterminate. The recall values are 3.10%, 4.92%, and 2.06%, respectively, highlighting varying degrees of difficulty in recognizing text depending on the level and nature of occlusion.

Some test results are visualized in Figure 5.8. Text localization is easier for small degrees of occlusion where the occluding object is typically small. The model is also able to recognize correctly when a character is only partially occluded, as seen in Figure 5.8 on the bottom right, but fails when the character is completely occluded, as seen in Figure 5.8 on the bottom left.

5.5 Summary

The chapter discusses the challenges and advancements in scene text detection and recognition, particularly in conditions where text is partially obscured. This is important for applications in automated driving, augmented reality, and information retrieval. Despite significant progress in computer vision and deep learning, robust scene text detection and recognition remain challenging, especially when various elements in real-world environments partially obscure text. The chapter introduces the Occluded RoadText dataset, which consists of 1,019 images capturing over 3,659 instances of occluded text from side-mounted car cameras in India. This dataset presents a diverse and realistic set of occlusion scenarios encountered while driving. Additionally, a new "Multi-Image End-to-End Recognition" task is introduced, providing supplementary images of the same scene from different angles to assist in text recognition. The Occluded RoadText dataset is evaluated using the TESTR baseline, and the results are benchmarked on an evaluation platform hosted on the Robust Reading Competition website.

Chapter 6

Conclusions

In this thesis, we examine the crucial role of textual information in ensuring safe and efficient navigation in driving scenarios. Text on road signs, billboards, and advertisements provides drivers with essential information that guides their behaviour and helps them avoid potential hazards. Despite its importance, current Advanced Driver Assistance Systems (ADAS) have mainly overlooked the need to read and understand this textual information, focusing primarily on visual scene analysis.

In Chapter 2, we explore existing datasets and methods for Video Question Answering, scene text spotting, recognition, and tracking.

In Chapter 3, We proposed modelling textual understanding in driving scenarios as a Video Question Answering (VideoQA) task. Our new dataset for VQA on driving videos emphasizes the practical relevance of questions and answers in real-world driving contexts. The dataset's performance evaluation on state-of-the-art VideoQA models revealed significant challenges in reading scene text, highlighting the need for improved methods.

Further investigation through a competition involving industry leaders and academic institutions provided valuable insights into the capabilities of existing models for text detection, recognition, and tracking in driving videos(Chapter 4). This competition underscored the strengths and weaknesses of current approaches and emphasized the necessity for further research and development in this field.

Recognizing the limitations of existing models, particularly in reading occluded text, we developed the Occluded RoadText dataset. We present this work in chapter 5. This dataset addresses the challenge of occlusions in real-world scenarios, providing a more accurate reflection of the complexities and unpredictability drivers face. By offering this benchmark publicly, we aim to foster collaboration and innovation among researchers, enabling the development of more robust models capable of handling occluded text in driving environments.

In conclusion, the work discussed in this thesis represents momentous steps toward integrating textual understanding into ADAS. The development of specialized datasets and the insights gained from our investigations pave the way for future enhancements in text detection and recognition technologies. These improvements will ultimately contribute to safer and more efficient driving experiences, better equipping drivers to navigate the visual world around them.

Related Publications

- **George Tom**, Minesh Mathew, Sergi Garcia-Bordils, Dimosthenis Karatzas, and C. V. Jawahar. "Reading Between the Lanes: Text VideoQA on the Road." In International Conference on Document Analysis and Recognition, 2023.
- **George Tom**, Minesh Mathew, Sergi Garcia-Bordils, Dimosthenis Karatzas, and C. V. Jawahar. "ICDAR 2023 Competition on RoadText Video Text Detection, Tracking and Recognition." In International Conference on Document Analysis and Recognition, 2023.
- Garcia-Bordils, Sergi, **George Tom**, Sangeeth Reddy, Minesh Mathew, Marçal Rusiñol, C. V. Jawahar, and Dimosthenis Karatzas. "Read While You Drive-Multilingual Text Tracking on the Road." In International Workshop on Document Analysis Systems, 2022. (not part of this thesis)

Bibliography

[1] J. Almazán, A. Gordo, A. Fornés, and E. Valveny. Word spotting and recognition with embedded attributes. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2552–2566, 2014. 18

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 11

[3] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee. Character region awareness for text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9365–9374, 2019. 29

[4] M. Bain, A. Nagrani, G. Varol, and A. Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 18

[5] D. Bautista and R. Atienza. Scene text recognition with permuted autoregressive sequence models. In *European Conference on Computer Vision*, pages 178–196, Cham, 10 2022. Springer Nature Switzerland. 29

[6] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 27

[7] A. F. Biten, R. Litman, Y. Xie, S. Appalaraju, and R. Manmatha. Latr: Layout-aware transformer for scene-text vqa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16548–16558, 2022. 9

[8] A. F. Biten, R. Tito, A. Mafla, L. Gomez, M. Rusinol, E. Valveny, C. Jawahar, and D. Karatzas. Scene text visual question answering. In *ICCV*, pages 4291–4301, 2019. 2

[9] A. F. Biten, R. Tito, A. Mafla, L. Gomez, M. Rusinol, E. Valveny, C. Jawahar, and D. Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 9, 11, 19

[10] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017. 18

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 23

[12] Z. Cai and N. Vasconcelos. Cascade r-cnn: high quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1483–1498, 2019. 29

[13] S. Castro, N. Deng, P. Huang, M. Burzo, and R. Mihalcea. In-the-wild video question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5613–5635, 2022. 7, 8

[14] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 8

[15] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 18

[16] Z. Cheng, J. Lu, B. Zou, L. Qiao, Y. Xu, S. Pu, Y. Niu, F. Wu, and S. Zhou. Free: A fast and robust end-to-end video text spotter. *IEEE Transactions on Image Processing*, 30:822–837, 2020. 6

[17] C. K. Ch'ng and C. S. Chan. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 935–942. IEEE, 2017. 2, 6, 7

[18] C. K. Chng, Y. Liu, Y. Sun, C. C. Ng, C. Luo, Z. Ni, C. Fang, S. Zhang, J. Han, E. Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1571–1576. IEEE, 2019. 2

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 18, 23

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 30

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 18

[22] J. D. Ewald. An investigation of the communicative purposes of road signs: How signs do things. *Advances in Language and Literary Studies*, 11(6):72–82, 2020. 1

[23] S. Garcia-Bordils, G. Tom, S. Reddy, M. Mathew, M. Rusiñol, C. Jawahar, and D. Karatzas. 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*, pages 756–770. Springer, 2022. 29

[24] S. Garcia-Bordils, G. Tom, S. Reddy, M. Mathew, M. Rusiñol, C. V. Jawahar, and D. Karatzas. Read while you drive - multilingual text tracking on the road. In S. Uchida, E. Barney, and V. Eglin, editors, *Document Analysis Systems*, pages 756–770, Cham, 2022. Springer International Publishing. 2, 6, 11, 12

[25] A. Geovanna Soares, B. Leite Dantas Bezerra, and E. Baptista Lima. How far deep learning systems for text detection and recognition in natural scenes are affected by occlusion? In *International Conference on Document Analysis and Recognition*, pages 198–212. Springer, 2021. 2, 7

[26] A. Geovanna Soares, B. Leite Dantas Bezerra, and E. Baptista Lima. How far deep learning systems for text detection and recognition in natural scenes are affected by occlusion? In *International Conference on Document Analysis and Recognition*, pages 198–212. Springer, 2021. 4, 35

[27] L. Gomez and D. Karatzas. Mser-based real-time text detection and tracking. In *2014 22nd International Conference on Pattern Recognition*, pages 3110–3115. IEEE, 2014. 6

[28] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 19

[29] R. Hu, A. Singh, T. Darrell, and M. Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9992–10002, 2020. 9, 18, 20, 23

[30] S. Jahagirdar, M. Mathew, D. Karatzas, and C. Jawahar. Watching the news: Towards videoqa models that can read. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4441–4450, 2023. 8

[31] S. Jahagirdar, M. Mathew, D. Karatzas, and C. V. Jawahar. Watching the news: Towards videoqa models that can read, 2022. 8, 11, 13

[32] JaidedAI. Easyocr. https://github.com/JaidedAI/EasyOCR. 12, 29

[33] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1156–1160. IEEE, 2015. 2, 6, 29, 30

[34] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. De Las Heras. Icdar 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*, pages 1484–1493. IEEE, 2013. 2, 6, 25, 26

[35] J. Kil, S. Changpinyo, X. Chen, H. Hu, S. Goodman, W.-L. Chao, and R. Soricut. Prestu: Pre-training for scene-text understanding. *arXiv preprint arXiv:2209.05534*, 2022. 5, 8

[36] J. Kim, M. Ma, T. Pham, K. Kim, and C. D. Yoo. Modality shifting attention network for multi-modal video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10106–10115, 2020. 7

[37] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 20

[38] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 18

[39] J. Lei, T. L. Berg, and M. Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022. 7, 11, 18

[40] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 7

[41] J. Lei, L. Yu, M. Bansal, and T. L. Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 7, 8

[42] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 7

[43] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, 2020. 7

[44] X. Li, J. Song, L. Gao, X. Liu, W. Huang, X. He, and C. Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 7

[45] M. Liao, B. Shi, and X. Bai. Textboxes++: A single-shot oriented scene text detector. *IEEE transactions on image processing*, 27(8):3676–3690, 2018. 5

[46] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11474–11481, 2020. 29

[47] M. Liao, Z. Zou, Z. Wan, C. Yao, and X. Bai. 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. 30

[48] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 23

[49] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *NeurIPS*, 2023. 23

[50] Y. Liu, H. Chen, C. Shen, T. He, L. Jin, and L. Wang. Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9809–9818, 2020. 5

[51] Y. Liu, C. Shen, L. Jin, T. He, P. Chen, C. Liu, and H. Chen. Abcnet v2: Adaptive bezier-curve network for real-time end-to-end text spotting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8048–8064, 2021. 5

[52] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 19

[53] S. Long, S. Qin, D. Panteleev, A. Bissacco, Y. Fujii, and M. Raptis. 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. 30

[54] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 20

[55] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto, et al. Icdar 2003 robust reading competitions: entries, results, and future directions. *International Journal of Document Analysis and Recognition (IJDAR)*, 7:105–122, 2005. 2

[56] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 23

[57] M. Mathew, M. Jain, and C. Jawahar. Benchmarking scene text recognition in devanagari, telugu and malayalam. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 7, pages 42–46. IEEE, 2017. 2

[58] M. Mathew, D. Karatzas, and C. Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, pages 2200–2209, 2021. 18

[59] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 27

[60] N. Nayef, Y. Patel, M. Busta, P. N. Chowdhury, D. Karatzas, W. Khlif, J. Matas, U. Pal, J.-C. Burie, C.-l. Liu, et al. Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1582–1587. IEEE, 2019. 2

[61] N. Nayef, F. Yin, I. Bizid, H. Choi, Y. Feng, D. Karatzas, Z. Luo, U. Pal, C. Rigaud, J. Chazalon, 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)*, volume 1, pages 1454–1459. IEEE, 2017. 2

[62] P. X. Nguyen, K. Wang, and S. Belongie. Video text detection and recognition: Dataset and benchmark. In *IEEE winter conference on applications of computer vision*, pages 776–783. IEEE, 2014. 6

[63] P. X. Nguyen, K. Wang, and S. J. Belongie. Video text detection and recognition: Dataset and benchmark. In *WACV*, 2014. 25, 26, 30

[64] V. Ordonez, G. Kulkarni, and T. Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 18

[65] O. Oviedo-Trespalacios, V. Truelove, B. Watson, and J. A. Hinton. The impact of road advertising signs on driver behaviour and implications for road safety: A critical systematic review. *Transportation research part A: policy and practice*, 122:85–98, 2019. 1

[66] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 23

[67] Z. Raisi and J. Zelek. Occluded text detection and recognition in the wild. In *2022 19th Conference on Robots and Vision (CRV)*, pages 140–150. IEEE, 2022. 2, 4, 6, 7, 35

[68] S. Reddy, M. Mathew, L. Gomez, M. Rusinol, D. Karatzas, and C. Jawahar. Roadtext-1k: Text detection & recognition dataset for driving videos. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11074–11080. IEEE, 2020. 2, 6, 11

[69] S. Reddy, M. Mathew, L. Gomez, M. Rusinol, D. Karatzas, and C. Jawahar. Roadtext-1k: Text detection & recognition dataset for driving videos. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11074–11080. IEEE, 2020. 26, 27, 29

[70] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 18

[71] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, 41(18):8027–8048, 2014. 2, 6

[72] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. 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*, pages 17–35. Springer, 2016. 27

[73] D. Seita. Bdd100k: A large-scale diverse driving video database. *The Berkeley Artificial Intelligence Research Blog. Version*, 511:41, 2018. 6

[74] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 18

[75] B. Shi, X. Bai, and C. 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. 5, 29

[76] A. Singh, V. Natarjan, M. Shah, Y. Jiang, X. Chen, D. Parikh, and M. Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 9, 11, 18

[77] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 7, 8, 11

[78] S. Tian, W.-Y. Pei, Z.-Y. Zuo, and X.-C. Yin. Scene text detection in video by learning locally and globally. In *IJCAI*, pages 2647–2653, 2016. 6

[79] S. Tian, X. Yin, Y. Su, and H. W. Hao. A unified framework for tracking based text detection and recognition from web videos. *TPAMI*, 2018. 25, 26

[80] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 23

[81] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 9, 18

[82] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. In *arXiv preprint arXiv:1601.07140*, 2016. 2, 11, 18

[83] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. In *arXiv preprint arXiv:1601.07140*, 2016. 6

[84] O. Vinyals, M. Fortunato, and N. Jaitly. Pointer networks. *Advances in neural information processing systems*, 28, 2015. 9

[85] B. Wang, Y. Xu, Y. Han, and R. Hong. Movie question answering: Remembering the textual cues for layered visual contents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 7

[86] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 8, 19

[87] K. Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition. In *2011 International conference on computer vision*, pages 1457–1464. IEEE, 2011. 6

[88] Y. Wang, H. Xie, S. Fang, J. Wang, S. Zhu, and Y. Zhang. From two to one: A new scene text recognizer with visual language modeling network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14194–14203, 2021. 4, 35

[89] W. Wu, Y. Cai, C. Shen, D. Zhang, Y. Fu, H. Zhou, and P. Luo. End-to-end video text spotting with transformer. *arXiv preprint arXiv:2203.10539*, 2022. 29

[90] W. Wu, D. Zhang, Y. Cai, S. Wang, J. Li, Z. Li, Y. Tang, and H. Zhou. Bovtext: A large-scale, multidimensional multilingual dataset for video text spotting. *Organization*, 2021. 2, 6, 29

[91] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 7, 8, 11, 20

[92] L. Xu, H. Huang, and J. Liu. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9878–9888, 2021. 7, 8

[93] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021. 7, 8

[94] Z. Yang, Y. Lu, J. Wang, X. Yin, D. Florencio, L. Wang, C. Zhang, L. Zhang, and J. Luo. Tap: Text-aware pre-training for text-vqa and text-caption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8751–8761, 2021. 5, 9

[95] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 27

[96] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134, 2019. 7, 8

[97] T.-L. Yuan, Z. Zhu, K. Xu, C.-J. Li, and S.-M. Hu. Chinese text in the wild. *arXiv preprint arXiv:1803.00085*, 2018. 2

[98] L. Yuliang, J. Lianwen, Z. Shuaitao, and Z. Sheng. Detecting curve text in the wild: New dataset and new solution. *arXiv preprint arXiv:1712.02170*, 2017. 2, 6, 39

[99] H. Zhang, X. Li, and L. Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 23

[100] X. Zhang, Y. Su, S. Tripathi, and Z. Tu. Text spotting transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9519–9528, 2022. 6, 29, 36, 38, 39

[101] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang. 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*, pages 1–21. Springer, 2022. 29

[102] M. Zhao, B. Li, J. Wang, W. Li, W. Zhou, L. Zhang, S. Xuyang, Z. Yu, X. Yu, G. Li, et al. Towards video text visual question answering: Benchmark and baseline. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 3, 8, 13

[103] Z. Zhao, X. Jiang, D. Cai, J. Xiao, X. He, and S. Pu. Multi-turn video question answering via multi-stream hierarchical attention context network. In *IJCAI*, volume 2018, page 27th, 2018. 7

[104] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560, 2017. 5