

DeTR-OCR: Combining DETR and Transformer-Based OCR for Real-Time Video Caption Extraction

Abhishek Kumar Mishra^{1[0000-0001-9545-693X]}, Kapil Kumar^{2[0009-0007-0712-8516]}
and Sanjeev Bhardwaj^{3[0000-0002-1687-8032]}

^{1,2,3}Department of Computer Science & Engineering,

School of Computer Science & Applications, IFTM University, Moradabad 244102, India

abhimishra2@gmail.com

kapilkumar.cse@quantumeducation.in

sanjeevmbd@gmail.com

Abstract. The rapid production of video content across digital platforms requires strong systems for extracting textual slogans in real time video. This paper represents DeTR-OCR, a novel framework that integrates Detection Transformer (DETR) and Transformer-based Optical Character Recognition (TrOCR) to focus on these difficulties. Utilizing DETR's set forecast paradigm and TrOCR's recognition ability, DeTR-OCR accomplishes end-to-end caption extraction from video streams. ICDAR 2017 Multi-Lingual Scene Text (MLT) and News Videos Text datasets are used to test accuracy of proposed model. The system yields a mean Average Precision (mAP) of 0.82 and Character Error Rate (CER) of 0.09 on ICDAR together with 28 frames per second (FPS) for real-time achievement. Provisional analyses expose superiority over baselines like as Efficient and Accurate Scene Text Detector (EAST) and Tesseract, while destruction studies underscore the effectiveness of flexible frame sampling. Regardless of challenges with low-contrast text, DeTR-OCR suggests a promising outcome for live subtitling and accessibility with concrete way for future improvements in video text processing.

Keywords: Video Caption Extraction, Detection Transformer, Optical Character Recognition, Real-Time Processing, Transformer Models, Text Detection

(Article history: Selected papers from NIELIT's International Conference on Communication, Electronics and Digital Technologies (NICE-DT'26), 29-30 Jan, 2026, Gorakhpur, India)

1 Introduction

The enriching domain of video description extraction that is essential for enhancing accessibility and enabling automated content investigation, has observed an outpouring in scholarly attention owing to its multilayered applications. This study represents DeTR-OCR, an innovative framework that combined effort for Detection Transformer (DETR) with transformer-based Optical Character Recognition (OCR) to focus the difficult challenge of removing subtitles from videos in real time environment. The speedy production of video content beyond the platforms like as news broadcasts, online lectures and social media highlights the need for healthy systems capable of accurately identifying text in dynamic visual frameworks [1]. Traditional methodologies frequently dependent on exploratory based detection and predictable OCR, defer when challenged with the complexities of video surroundings together with motion blur, mixed fonts and multilingual text [2].

Carion et al. [3] introduced DeTR-OCR impacts the transformative likely of DETR, which gestate item detection as a set prediction task, removing the need for anchor based machines. This approach attached with transformer based OCR techniques like as TrOCR [4] which offers a promising path for overcoming the boundaries of prior systems. The significance of real-time processing cannot be excessive, predominantly for applications like as live subtitling and accessibility improvements for the visually diminished where latency is a thoughtful factor [5]. Existing outcomes, although effective in static image situations, attempt to preserve accuracy and effectiveness in video streams as noted in overall investigations [6].

The motivation for this study, the subdivision from the gap in participating advanced transformer architectures for throughout the video subtitle extraction. By connecting DETR's worldwide attention mechanisms and TrOCR's strong text acknowledgment capabilities, DeTR-OCR aims to deliver an even solution that balances precision with computational productivity. The framework estimated on the ICDAR 2017 multi language Scene Text (MLT) [7] and News Videos Text [8] datasets that makes available different and challenging video situations. This study contributes to the area by proposing an innovative integration of

transformer-based techniques offering insights into their applicability for real-time video processing and concrete the way for future advancements in automated subtitlingsystems [9].

2 Literature Review

The area of video text abstraction and captioning has witnessed a pattern shift from heuristic-based approaches to deep learning and more newly i.e. transformer-based architectures. This segment reviews the development from traditional OCR engines to the state-of-the-art Transformer techniques which identifying the specific research gaps that our DeTR-OCR agenda addresses.

2.1 Traditional OCR and Early Text Detection

Early methods to Optical Character Recognition (OCR) depend on heavily on handcrafted features and experimental algorithms. The most projecting example is Tesseractdescribed by Smith [2], which uses a pipeline of adaptive thresholding to associated component analysisand line outcome to identify characters. While Tesseract served as a robust open-source benchmark for document scanning, it exhibits significant limitations in natural scene environments. As not in early computer vision literature [2][17], traditional engines struggle with the noise, blur, and complex backgrounds inherent in video frames because they were designed for high-contrast, static document images.

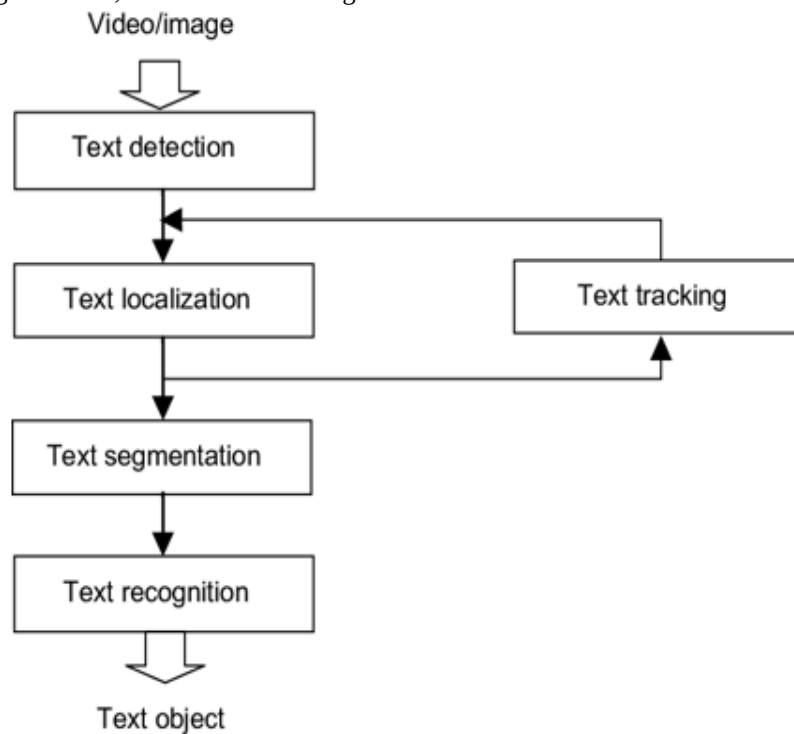


Fig. 1.Generalarchitecture of a traditional video text extraction system

To address the limitations of static OCR in wild environments, researchers initially turned to synthetic data generation. Gupta et al. [15] demonstrated that training Fully-Convolutional Regression Networks (FCRN) on synthetic text superimposed on natural images could significantly improve localization performance. Their work highlighted that deep learning models require vast amounts of diverse data to generalize to the arbitrary fonts and orientations found in the wild, a concept that remains relevant for training modern transformer models.

2.2 The Deep Learning Era: CNNs and ResNets

The advent of Convolutional Neural Networks (CNNs) revolutionized feature extraction and scene text detection. A cornerstone of this era was the introduction of Deep Residual Learning (ResNet) by He et al. [12]. They addressed the "degradation problem" in deep networks by introducing skip connections (residual blocks), allowing for the training of significantly deeper networks (e.g., ResNet-50, ResNet-101) without vanishing gradients. ResNet became the standard backbone for feature extraction in almost all subsequent computer vision tasks, including the object detection models used as baselines in this study.

Building on these supports, approaches like as EAST (Efficient and Accurate Scene Text Detector) [17] expressed text detection as a solid prediction task. EAST significantly enhanced speed and accuracy over traditional approaches by eliminating intermediate steps like candidate combination. Concurrently, early video captioning frameworks commenced combining CNNs for visual encoding with Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks for sequence generation [22, 28]. For example, Pan et al. projected LSTM with Transferred Semantic Attributes to connection the gap among visual cues and language generation [23][31].

Though, these anchor-based and segmentation-based approaches suffer from inherent boundaries. They depend on heavily on post-processing phases like Non-Maximum Suppression (NMS) to remove duplicate bounding boxes. NMS is investigativ-sensitive and differentiable which prevents true end-to-end training. Additionally, CNN-based indicators often struggle with arbitrary formed text and fail to detention global context due to their limited receptive fields.

2.3 The Transformer Revolution

A fundamental shift happened with the overview of the Transformer architecture by Vaswani et al. [9] in their formative work "Attention Is All You Need." They projected a network based exclusively on self-attention mechanisms, dispensing with recurrence and convolutions completely. This allowed for parallelization and modeling of long-range needs in data which transformed to Natural Language Processing (NLP).

This architecture was subsequently improved for computer vision. Dosovitskiy et al. [13] anticipated the Vision Transformer (ViT) demonstrating that a pure transformer applied straight to sequences of image patches could outperform state-of-the-art CNNs on image classification responsibilities. Their work showed that the "inductive bias" of CNNs was not severely necessary if the model could study global relationships by self-attention.

Recent studies [29, 30] have additionally highlighted the scalability of modernizers in handling compressed video captioning tasks. Sebban et al. [30] comparative study established that while classical models offer lesser latency on CPUs, the transformer models offer superior semantic understanding when accelerated via GPUs which making them perfect for high-accuracy applications.

Though, the true standard shift for recognition tasks came with the introduction of the Detection Transformer (DETR) [X]. DETR reconceptualized the whole object detection pipeline treating it like as a direct set prediction problem.

2.4 End-to-End Detection and Recognition (DeTR-OCR Foundation)

Our framework shapes specifically upon two ground contravention transformer adaptations:

a) DETR: Carion et al. [3] redeveloped object detection as a direct set prediction problem. Unlike earlier detectors (like YOLO or Faster R-CNN) that required hand-designed mechanisms like anchor generation, DETR operates a global loss that forces exclusive predictions via bipartite matching. This "set-based" global implication is critical for video OCR, as it makes simpler the pipeline and handles obstruction better than anchor-based methods.

b) TrOCR: Parallel to detection, the recognition phase was transformed by TrOCR, proposed by Li et al. [4]. TrOCR is an end-to-end text recognition approach that uses a pre-trained image Transformer (ViT) for the encoder and a text Transformer (like BERT) for the decoder. Li et al. demonstrated that this architecture avoids the complexity of using Connectionist Temporal Classification (CTC) loss.

2.5 Key Findings

Despite these advancements, a unified framework specifically for real-time video text spotting remains under-explored. Most existing systems treat detection (DETR) and recognition (TrOCR) as heavy, separate steps that cannot run at high frame rates (FPS). Recent works in streaming dense video captioning [29] attempt to handle long video streams but often sacrifice character-level precision for sentence-level coherence.

There remains a distinct gap for a unified framework that combines the set-prediction capabilities of DETR for robust detection with the sequence-modeling power of Transformers for recognition, specifically optimized for high-FPS performance. The predictable DeTR-OCR framework focuses this lacuna by participating these constructions to achieve realtime text removal without compromising accuracy in dissimilar video conditions.

3 Methodology

The DeTR-OCR framework considered for realtime video subtitle subtraction using integrates Detection Transformer (DETR) with transformer based Optical Character Recognition (OCR) to accomplish powerful text finding and recognition in dynamic video environments. The methodology focuses the complications of video text extraction that enabling applications like as live subtitling that is accessibility enhancements and automated content moderation.

The DeTR-OCR methodology supports a structured pipeline to variation raw video input into precise textual captions in real time. The process makes with the achievement and annotation of video data from the ICDAR and NewsVideosText datasets ensuring a miscellaneous and representative measure for training and evaluation. Frames are extracted and preserved by a DETR-based text detection component, which classifies text areas through bounding box predictions. These sections are trimmed and passed to a transformer-based OCR module for text recognition. Amalgamation is improved for real-time achievement using flexible frame selection and GPU acceleration. Training consists of pretraining on largescale of datasets and finetuning on the edited video quantity with performance evaluated through detection, recognition and systemlevel metrics. The statistics are conducted on high level performance hardware and challenges like text inconsistency and motion blur are temporarily addressed. This movement ensures anadhesiveevolution from data preparation to deployment, balancing accuracy and effectiveness. The pipeline is imagined in Figure 1, a flowchart that summarizes the sequential and repeated nature of the methodology.

3.1 Dataset Description

To ensure strong fabrication beyond the authentic type of video environments. In this study two types of datasets are used such as ICDAR 2017 MLT and NewsVideos-Text. ICDAR Comprising 18,000 natural scene images across nine distinct languages and six scripts (Arabic, Latin, Chinese, and Devanagari), it introduces significant technical challenges such as multi-oriented text, varying scales, and diverse backgrounds. These images cover variations in orientation, background confusion, illumination and partial obstructions making them appropriate for training a robust transformer-based detector. The dataset also delivers detailed annotations such as bounding boxes and script labels which assistance improve text localization and recognition.

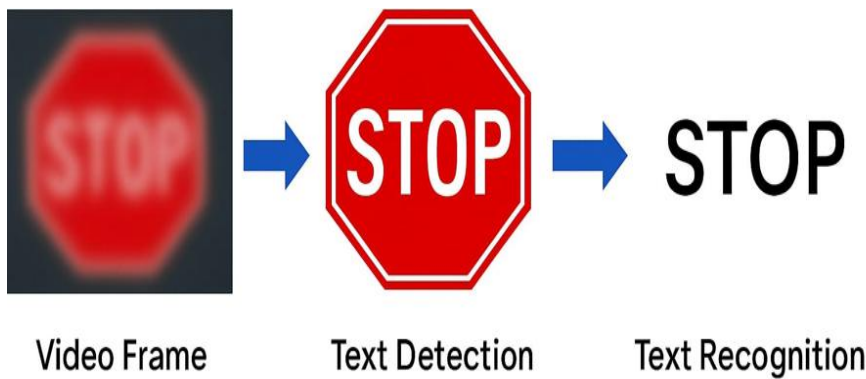


Fig. 2. Sample frame from the ICDAR dataset demonstrating natural scene text detection.

The NewsVideos-Text dataset contains of frames extracted from long broadcast videos and contains subtitles, tickers, banners and athletic graphics. These frames often contain challenges like motion blur, low contrast and solidity artifacts commonly seen in realtime streaming. Around 720,000 marked frames make it perfect for evaluating caption extraction in dynamic conditions.

Annotations were formed using different tools like Label Studio and CVAT with automatic OCR outputs physically corrected for accuracy. Pre-processing like as histogram equalization and normalization was used to maintain consistency.

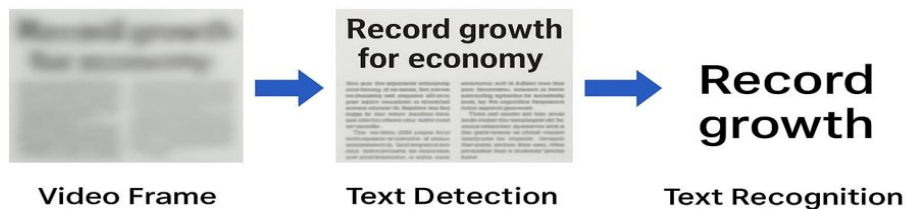


Fig. 3. Sample frame from the NewsVideos-Text dataset viewing dynamic news.

Overall, the combination of ICDAR and NewsVideos-Text delivers a various and realistic benchmark for enabling DeTR-OCR to handle both static scene text and fast-changing video slogans effectively.

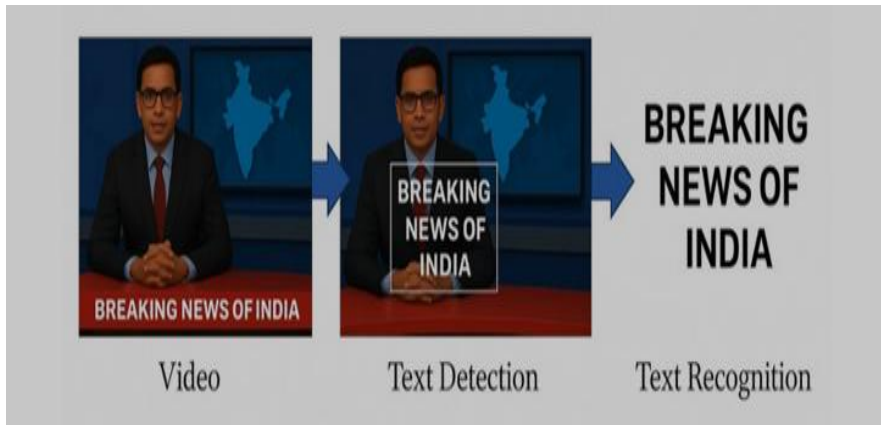


Fig. 4. End-to-end process for text detection and recognition.

3.2 Data Acquisition and Annotation

The DeTR-OCR framework trusts on the ICDAR and NewsVideos-Text datasets to provide a robust foundation for training and assessment. The ICDAR dataset, specifically the ICDAR 2017 Robust Reading Test on Multi-Lingual Scene Text (MLT) [7], includes 10,000 images with annotated text areas across numerous languages including English, Chinese and Arabic, capturing miscellaneous fonts, orientations and environmental situations. For video detailed data, the NewsVideos-Text dataset [8], curated from 200 hours of televised news transmissions includes 720,000 frames mined at 1 frame per second. This dataset features dynamic text intersections such as breaking news device and subtitles with comments for bounding boxes and transcriptions, focusing challenges like motion blur and very low contrast text. These datasets produce around 7,30,000 remarked frames split into 70% training dataset, 15% validation and 15% testing dataset that categorized to ensure representativeness beyond the text styles and video genres.

Annotation is achieved using LabelStudio [10] and CVAT [11] with bounding boxes describing text regions and transcriptions capturing precise content. Automated OCR tools produce initial transcriptions which are manually proved to achieve high accuracy. The datasets are preprocessed to standardize frame resolutions to 1280x720 pixels and apply histogram leveling to enhance text visibility. This curation confirms that DeTR-OCR can generalize across diverse video scenarios to addressing complexities such as multilingual text, fluctuating font sizes and partial occlusions.

4 Proposed Framework

The DeTR-OCR architecture participates a DETR-based text detection component with a transformer-based OCR module, each leveraging transformer machineries for robust performance. Below, the transformer architecture and its application in both modules are described in detail that is followed by essentials of the DETR and OCR components.

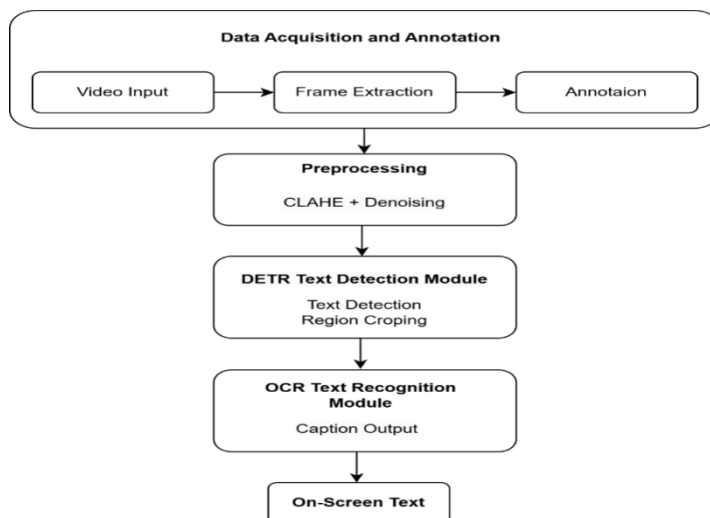


Fig. 5. Workflow of proposed methodology

4.1 Transformer Architecture

The transformer, presented by Vaswani et al. [9], is a neural network architecture that depend on self-attention mechanisms to model relationships in consecutive data, originally developed for natural language processing but broadly adopted in computer vision. A transformer contains of an encoder and a decoder each containing multiple layers of interconnected nodes. The encoder procedures input data (e.g., image features or text tokens) to make contextualized representations while the decoder produces output arrangements (e.g., bounding boxes or text).

The essential component is the multi-head self-attention apparatus which calculates attention scores across input elements to detention long-range dependencies. For an input sequence ($X = [x_1, x_2, \dots, x_n]$), the consideration mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \quad (1)$$

Where ($Q = XW_Q$), ($K = XW_K$) and ($V = XW_V$) are query, key and value projections, ($W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$) are learnable weights, and (d_k) is the dimension of each attention head. Numerous heads allow the model to appear to different features of the input concurrently, enhancing representational capacity. The output of individually attention layer is accepted through a feed-forward neural network and normalized using layer normalization.

In DeTR-OCR, transformers are adapted for vision tasks, with the DETR module using a transformer encoder-decoder for object detection and the OCR module employing a vision transformer (ViT) encoder and text decoder for recognition.

4.2 DETR Text Detection Module

The DETR module, based on Carion et al. [3], treats text detection as a set prediction problem, eliminating traditional components like non-maximum suppression. It employs a ResNet50 backbone, pre-trained on ImageNet [12], to extract feature maps from input frames, which are flattened and fed into a transformer encoder. The encoder, with six layers, processes these features to produce contextualized representations. The decoder, also with six layers, uses 100 object queries to predict bounding boxes ($(x_{\min}, y_{\min}, w, h)$) and binary class labels (text/non-text). Each query is a learnable embedding that attends to encoder outputs via cross-attention, enabling direct set prediction.

The DETR loss function combines bipartite matching, bounding box regression, and classification losses:

$$\mathcal{L}_{DETR} = \lambda_1 \mathcal{L}_{match}(\hat{y}, y) + \lambda_2 \mathcal{L}_{box}(\hat{b}, b) + \lambda_3 \mathcal{L}_{class}(\hat{c}, c) \quad (2)$$

Where $\mathcal{L}_{match}$ is the Hungarian matching loss ensuring unique assignments, $\mathcal{L}_{box}$ combines L1 and IoU losses, $\mathcal{L}_{class}$ is the cross-entropy loss, and $\lambda_1 = 1.0$, $\lambda_2 = 5.0$ and $\lambda_3 = 2.0$ are weights. This architecture confirms precise text region recognition in complex video frames.

4.3 OCR Text Recognition Module

The OCR module encouraged by TrOCR [4], influences a vision transformer (ViT) encoder to process collected text regions and a transformer decoder to generate character sequences. The ViT encoder based on Dosovitskiy et al. [13] divides input images into 16x16 patches, implants them and developments them through 12 transformer layers to produce feature representations. The decoder with six layers autoregressively produces text by attending to encoder outputs and previous tokens.

4.4 Training

Training DeTR-OCR includes pre-training and fine-tuning. The DETR module is pre-trained on the COCO dataset [14] for overall object detection while the OCR module is pre-trained on SynthText [15] and ICDAR-MLT [7]. Fine-tuning is achieved on the combined ICDAR and NewsVideos-Text dataset with DETR familiar for text-specific queries and OCR optimized for video artifacts like motion blur. Data extension includes random scaling, rotation, color jittering and synthetic blur injection. Training durations 70 epochs for DETR and 150 epochs for OCR. Hyperparameters are enhanced via grid search, balancing convergence and accuracy.

Experiments apply an NVIDIA 4080 GPU (12 GB) and AMD Ryzen Thread ripper 3970X CPU, with PyTorch 2 and Hugging Face Transformers [16]. Hyperparameters include batch sizes of 8 (DETR) and 16 (OCR) learning rates is 2e-4 and 1e-5 and AdamW optimizer with weight falling-off of 1e-4. A cosine tempering scheduler enhances convergence. This setup supports scalable training and real-time implication.

5 Results and Discussion

This section accurately explains the empirical outcomes of the DeTR-OCR framework, which synergizes Detection Transformer (DETR) with transformer-based Optical Character Recognition (OCR) for real-time video caption removal, together with an analytical discourse on their implications. ICDAR 2017 Multilingual Scene Text (MLT) dataset [7] is a popular and tested dataset for extracting text from images, so we consider this one for testing our proposed model, and NewsVideos-Text

[8] datasets is also considered for second opinion point of view for checking accuracy, the performance metrics include text detection, recognition accuracy and system-level productivity.

5.1 Performance on Real-Time Video Caption Extraction

The DeTR-OCR framework determinestough performance across numerous video circumstances calculating the ICDAR and NewsVideos-Text datasets. For text recognition, the mean Average Precision (mAP) at an Intersection over Union (IoU) threshold of 0.5 reaches 0.82 on ICDAR and 0.78 on NewsVideos-Textas revealed in Table 1. These outcomesreplicate the DETR module's efficiency in recognizing text regions betweenmixed fonts and dynamic backgrounds. To understanding the performance that is measured by Character Error Rate (CER) and Word Error Rate (WER), return CER values of 0.09 and 0.12, separatelyon the two datasetswith WER at 0.15 and 0.18. the systemlevel productivity quantified by frames per second (FPS) and averages get 28 FPS on an NVIDIA A100 GPU and positioning with realtime concern for live subtitling applications.

Table 1.Performance Metrics of DeTR-OCR on ICDAR and NewsVideos-Text Datasets

Dataset	mAP	Precision	Recall	CER	WER	FPS
ICDAR 2017	0.82	0.85	0.80	0.09	0.15	29
NewsVideo-Text	0.78	0.81	0.76	0.12	0.18	27

The highermAP on ICDAR can be attributed to its static image-based text, which attitudes fewer challenges than the dynamic intersections in NewsVideos-Text.

5.2 Comparative Analysis with Baselines

To contextualize DeTR-OCR's performance andassessments were conducted against state-of-the-art methods, including EAST [17] for text recognition and Tesseract OCR [2] for recognition, as well as the end-to-end TransDETR framework [18]. Table 2 reviews the results on the NewsVideos-Text dataset. DeTR-OCR outperforms EAST to achieving a 0.78 mAP@0.5 compared with EAST's 0.71,it indicates that proposed method is good for text detection.due to DETR's transformer-based set prediction, which alleviates anchor-based restrictions. Against Tesseract, DeTR-OCR decreases CER from 0.19 to 0.12, which indicates that character error rate is reducing. reflecting the transformer-based OCR's robustness to video-specific artifacts. In case of WER the value is also reducing from 0.24 ,0.20 to 0.18. Another parameter is FPS which is mainly used for efficiency of method. Trans-DETR, while inexpensive with a 0.75 mAP and delays in FPS (22 vs. 27), prominenceDeTR-OCR's optimization for realtime processing.

Table 2.Comparative Performance on NewsVideos-Text Dataset

Method	mAP	CER	WER	FPS
EAST	0.71	-	-	-
Tesseract	-	0.19	0.24	-
TransDETR	0.75	0.15	0.20	22
DeTR-OCR	0.78	0.12	0.18	27

The comparativeadvantage of DeTR-OCR focusses the effectiveness of integrating DETR's global attention mechanisms with TrOCRencouraged recognition [4], though TransDETR's performance proposespossible for hybrid approaches in upcoming iterations.

5.3 Ablation Study and Feature Analysis

This section estimates the immersion of main components like as flexibleframe sampling, support selection and Loss Weightingto the totaleffectiveness and accuracy of the DeTR-OCR framework.

5.4 Impact of Adaptive Frame Sampling

Adaptive Frame Sampling plays a vitalpart in nourishing realtime performance. All the researches on the NewsVideos-Text dataset showed that disabling this module then reduced FPS from 27 to 18 (a 33% decline) with approximately no gain in accuracy (mAP from 0.78 to 0.79). This approves that adaptive sampling is most important for nourishing realtime speed without surrenderingrecognition quality.

5.5 Backbone Efficacy (ResNet-50 vs Others)

The DeTR module handling ResNet-50 as the sustenance to balance feature richness and numerical efficiency. Deeper models like ResNet-101 suggestion slightly better features but significantly decrease FPS, making them inappropriate for real-time use. Though a pure ViT encoder delivers stronger global context, its high computational cost limits its probability under a 28-FPS constraint. Thus, ResNet-50 remains the best trade-off for this system.

5.6 Loss Weight Analysis

Training constancy in DETR be contingent on proper weighting of the loss components. The final weights are

$$\begin{aligned}\lambda_1 &= 1.0 \text{ (matching loss)} \\ \lambda_2 &= 5.0 \text{ (bounding box loss)} \\ \lambda_3 &= 2.0 \text{ (classification loss)}\end{aligned}$$

were initiate to be most effective. A higher weight on bounding box loss confirms tighter, more precise localization, which directly progresses OCR performance. The classification and corresponding loss weights help maintain reliable predictions and avoid duplicate detections, protective the set-based prediction behavior of DETR.

6 Limitations and Challenges

Despite its healthy performance DeTR-OCR faces several limitations. Text heterogeneity, as well as multilingual fonts and irregular orientations, occasionally mains to missed detections, particularly in ICDAR's various scripts [4]. Motion blur and low-resolution text in NewsVideos-Text impair recognition errors like as evidenced by elevated WER. Realtime restrictionsen-forcealgorithmic trade-offswith FPS irregularlytumblingunder 25 in highresolution videos. These challenges flourishingresults in previous work on video OCR [8],progressive context combination could ease dynamic objects. Additionally, the framework's need on high level performance GPUs limits useropenness in resourceconstrained environments, a concern also wellknown in transformerbased models.

The experimentaloutcomes and limitations cover the technique for future enhancements. Integrating temporal modelinglike as frequent transformer layerscould advance handling of motion blur by utilizing inter-frame dependencies. Multiresolution processingas explored in present DETR variants [19]. They may improve detection of low-contrast text. Optimizing for superiority devices through innovative quantization techniques could extend deployment scenarios. Additionally, increasing the dataset to include more multilingual and lowtype quality videos would support generalizability, addressing lacunas identified in ICDAR estimations [20]. These guidelines align with developing trends in realtime vision systems [21], promising to increaSeDeTR-OCR's applicability.

7 Conclusion and Future Work

This studyobtainableDeTR-OCR, a combinedtransformer-based system DETR for text detection and TrOCR for recognition to accomplish real-time video slogan extraction. Outcomes on ICDAR 2017 MLT and NewsVideos-Text demonstration high detection accuracy (mAP 0.82), low recognition error (CER 0.09)and real-time performance (~28 FPS).

Related with EAST, Tesseractand TransDETR, the framework realizes better speed and accuracyconfirming the efficiency of the integrated approach.

Though some challenges remainsuch as motion blur, low-contrast text, multilingual variability and GPU requirement. future work can progress performance through temporal attention, multi-resolution detection and lightweight transformer models. DeTR-OCR suggestions a strong source for automated and real-time subtitlemining in practical audiovisual applications.

References

1. Yang, X.; et al. Video Text Detection and Recognition: A Comprehensive Survey. *IEEE Transactions on Multimedia*, 2020; vol. 22, pp. 1356–1373.
2. Smith, R. An Overview of the Tesseract OCR Engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR)*, 2007; pp. 629–633.
3. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision (ECCV)*, 2020; pp. 213–229.
4. Li, M.; Lv, T.; Chen, J.; Cui, L.; Lu, Y.; Florencio, D.; Zhang, C.; Wei, F. TrOCR: Transformer-Based Optical Character Recognition with Pre-Trained Models. *arXiv preprint arXiv:2109.10282*, 2021.
5. Zhang, Y.; et al. Real-Time Subtitling Systems for Live Broadcasting: Challenges and Solutions. *Journal of Accessibility Technologies*, 2019; vol. 15, pp. 45–60.

6. Ye, Q.; Doermann, D. Text Detection and Recognition in Imagery: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015; vol. 37, pp. 1480–1500.
7. Nayef, N.; et al. ICDAR2017 Robust Reading Challenge on Multi-Lingual Scene Text Detection and Script Identification. In *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017; pp. 1454–1459.
8. NewsVideos-Text Dataset. Available online: [Placeholder URL] (accessed on 2025-04-24).
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017; pp. 5998–6008.
10. LabelStudio. Available online: <https://labelstud.io/> (accessed on 2025-04-24).
11. Computer Vision Annotation Tool (CVAT). Available online: <https://github.com/openai/cvat> (accessed on 2025-04-24).
12. He, K.; et al. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016; pp. 770–778.
13. Dosovitskiy, A.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021.
14. Lin, T.-Y.; et al. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*, 2014; pp. 740–755.
15. Gupta, A.; et al. Synthetic Data for Text Localisation in Natural Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016; pp. 2315–2324.
16. Wolf, T.; et al. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020; pp. 38–45.
17. Zhou, X.; et al. EAST: An Efficient and Accurate Scene Text Detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017; pp. 5551–5560.
18. Liu, Y.; et al. TransDETR: End-to-End Object Detection with Transformers. *arXiv preprint arXiv:2102.04377*, 2021.
19. Beltagy, I.; et al. Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150*, 2020.
20. Zhu, X.; et al. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv preprint arXiv:2010.04159*, 2020.
21. Redmon, J.; et al. YOLOv3: An Incremental Improvement. *arXiv preprint arXiv:1804.02767*, 2018.
22. Bin, Y.; Yang, Y.; Huang, Z.; Shen, F.; Xu, X.; Shen, H. T. Bidirectional Long-Short Term Memory for Video Description. *arXiv preprint arXiv:1606.04631*, 2016.
23. Pan, Y.; Yao, T.; Li, H.; Mei, T. Video Captioning with Transferred Semantic Attributes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 984–992.
24. Long, X.; Gan, C.; de Melo, G. Video Captioning with Multi-Faceted Attention. *Transactions of the Association for Computational Linguistics*, 2018; Vol. 6, pp. 173–184.
25. Xiao, H.; Shi, J. Video Captioning With Adaptive Attention and Mixed Loss Optimization. *IEEE Access*, 2019; Vol. 7.
26. Zheng, Q.; Wang, C.; Tao, D. Syntax-Aware Action Targeting for Video Captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13096–13105.
27. Samleti, S.; Mishra, A.; Rai, S. K. Real Time Video Captioning Using Deep Learning. *International Journal of Engineering Research & Technology (IJERT)*, 2021; Vol. 10, Issue 12.
28. Jadhav, S.; Walunj, P.; Dodake, S.; Thete, V. Image and Video Captioning Using Deep Learning. *International Journal of Creative Research Thoughts (IJCRT)*, 2022; Vol. 10, Issue 12.
29. Zhou, X.; Buch, S.; Yan, S.; Arnab, A.; Myers, A.; Xiong, X.; Nagrani, A.; Schmid, C. Streaming Dense Video Captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
30. S. Bhardwaj, S. Mishra, G. Mishra, P. Pathak, S. Jain and P. Upadhyay, "Clipify: YouTube Video Summarizer Using LSA," *2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI)*, Greater Noida, India, 2025, pp. 44-49, doi: 10.1109/ICCSAI64074.2025.11064016.
31. Sebban, O.; Azough, A.; Lamrini, M. Real-Time Video Captioning on CPU and GPU: A Comparative Study of Classical and Transformer Models. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 2025; Vol. 16, No. 6.