International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Video Object Search: A Semantic Approach with OpenAI's CLIP Technology

Chandini S B[1], Dhanya Hegde P[2], Dhrithi Kashyap H R[3], Keerthana M S[4], Madhushree M[5]

Department of Information Science and Engineering
Vidyavardhaka College of Engineering, Mysuru, India

chandini@vvce.ac.in[1]
dhanyahegde.2001@gmail.com[2]
kashyapdhrithi@gmail.com[3]
prarthanams2008@gmail.com[4]
madhushree10.m@gmail.com[5]

Abstract - A research paper titled "Video Object Search: A Semantic Approach with OpenAI's CLIP Technology" examines the possibilities of semantic search in the identification of video objects. The technique that is proposed in this paper integrates textual and visual data to enhance the precision of video identification of objects. The conventional approach in video object detection utilizes only visual data, which might result in limited or incorrect item identification. By integrating certainty estimation, adaptability to handle a variety of questions, and semantic comprehension, the recommended approach gets over these limitations. The system evaluates textual and visual data using OpenAI's CLIP Technology, enabling more precise identification of objects in videos. The study's findings show that the suggested approach works more precisely than the conventional approach.

I.INTRODUCTION

Video object detection and retrieval have become increasingly important in various fields, including surveillance, entertainment, and research. The ability to accurately identify and search for specific objects or events in video is critical for effective analysis and decision making. However, traditional video object recognition methods often face limitations in semantic understanding and adaptability to different queries. These challenges have motivated the search for new approaches that can effectively integrate both visual and textual information and provide accurate detection in response to complex questions. This paper presents a semantic approach to video object retrieval using OpenAI and CLIP technology, which aims to address the identified limitations of conventional methods. Using CLIP features, the proposed system aims to improve the accuracy and efficiency of video object recognition by training a model that learns the relationship between natural language descriptions and video frames. This enables more accurate detection of objects in the video, even if they are partially blocked or in low light.

The structure of the paper is as follows: Section 2 provides a comprehensive literature review on video object retrieval and the challenges associated with traditional methods. Section 3 describes the specific challenges that the proposed system aims to overcome.

Section 4 describes the methodology and approach used to integrate OpenAI and CLIP technology for semantic video object retrieval. Section 5 presents the results and conclusions of the study, emphasizing the performance and efficiency of the proposed system. Finally, Section 6 concludes the paper and discusses the research implications and possible future work opportunities.

II.LITERATURE SURVEY

"Looking Fast and Slow: Memory-Guided Mobile Video Object Detection"[1] is a research paper that examines video object detection using a large dataset that combines COCO data. Through the sequential or concurrent use of both heavy and lightweight feature extractors, the proposed model presents a novel interleaved framework for video object detection. There is a memory mechanism twhich is used hat aggregates and it refines these frame-level features [1].

The process carefully blends heavy and light feature extractors, working in a stepwise manner. A memory-guided mechanism is then used to fuse and refine the extracted features from each frame. Additionally, the study investigates an adaptive model variant that deliberately lowers the number of times the complex feature extractor is run while maintaining accuracy [2] [1].

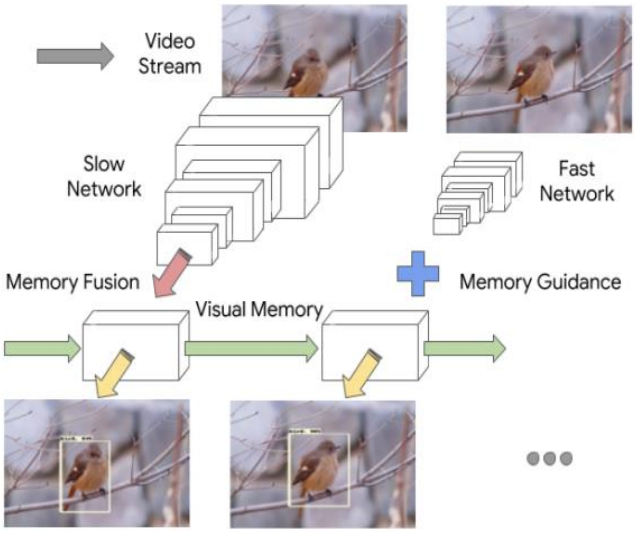

Figure 1: An illustration of proposed system memory-guided interleaved model. Given a video stream, a visual memory module fuses very different visual features produced by fast and slow feature extractors across frames to generate detections in an online fashion[1].

Modern flow-based mobile models are more accurate than the suggested approach in terms of efficiency, especially since it doesn't depend on optical flow data. The model has fewer parameters but maintains theoretical complexity. Moreover, the approach's inference optimizations produce a threefold boost in frame rate at a negligible accuracy cost, enabling previously unheard-of real-time runtime on mobile devices [2].

In comparison to other methods, the paper's approach to video object detection achieves real-time runtime on mobile devices and displays high accuracy even in the absence of optical flow data, all with fewer parameters. It is important to remember that there could be quantization issues with one adaptation of the adaptive model [2].

The study "Object Guided External Memory Network for Video Object Detection"[2] investigates the field of video object detection using the ImageNet VID dataset for visual aids. A novel object-guided external

memory

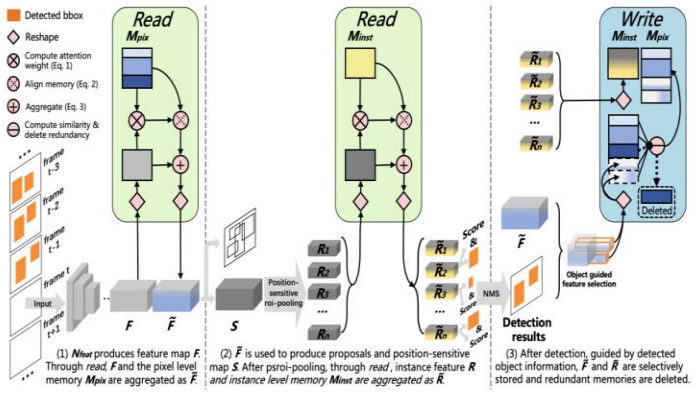

Figure 2: Object guided external memory network [2]

network for cross-frame inference is introduced in the proposed model. Using ResNet-101-based feature network and an attention mechanism to amplify attention weights on similar features while attenuating attention on the background, the model leverages pixel and instance level features [2]. The mentioned approach's methodology is focused on using external memory to transfer data between frames. The model uses attention to focus on possible objects, and in situations where there is a lot of occlusion, memories are transferred to the object area to improve recognition. The model particularly shows that it can extract long-term information from earlier frames [2]. The paper provides a comprehensive analysis of efficiency, taking into account memory size per frame and GPU usage. Memory usage is optimized through write operations that continuously remove redundant features from the memory size, which is usually smaller than a feature map. For example, the model outperforms the baseline method in mean Average Precision (mAP) with shorter inference time [2]. It also strikes a commendable speed-accuracy balance when down-sampling non-key frames.

Improved mAP, effective memory utilization, and similar computation power to the single-frame baseline are just a few benefits of the approach.

In the paper "Sequence Level Semantics Aggregation for Video Object Detection,"[3] the SELSA (Semantic Enhancement for Long-term Video Object Segmentation and Tracking) method is developed with the goal of improving video object detection. The ImageNet VID dataset is used for evaluation after the model has been trained on a combination of the DET and VID datasets, with the dataset split described in FGFA [3].

The SELSA method utilizes semantic neighbors sampled

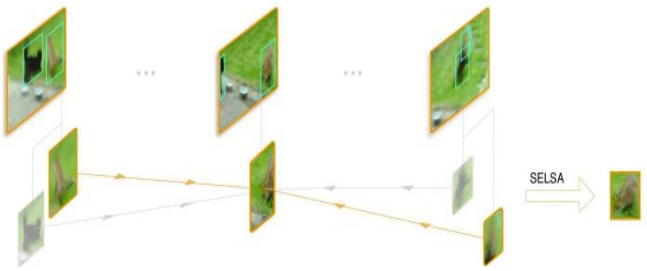

Figure 3. First extract proposals in different frames from the video, then the semantic similarities of proposals are computed across frames. At last, aggregate the features from other proposals based on these similarities to obtain a more discriminative and robust features for object detection [3].

from the full video sequence to show a novel way to improve proposal features. Through the use of the SELSA module, the model deliberately adds contextual data, especially from semantically relevant regions, generating enhanced proposal features. This architecture and is a well based on the ResNet-101 framework, base for computer vision tasks [3].

The efficiency of the SELSA method is given methodologically by a notable improvement in mean Average Precision (mAP) when compared to the baseline method. The model's ability to precisely identify and localize objects in video sequences is demonstrated by the achieved mAP of 80.25. With an mAP of 61.38, the SELSA approach notably outperforms other approaches in terms of general

improvement and improves performance in fast-motion scenarios. The model's ability to manage dynamic and quickly changing object positions within a video stream is demonstrated by this [3].

Effective video object detection is an important component, and the SELSA approach shows how practical it is by yielding significant performance improvements. The model's efficiency in practical scenarios is demonstrated by its improved motion-specific playing ability and enhanced mAP. These developments are especially important because they increase the overall robustness and precision of video object detection systems [3].

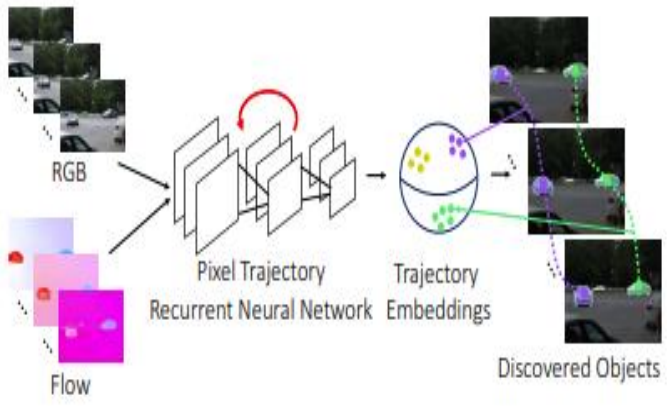

Figure 4: RGB images and optical flow are fed into a recurrent neural network, which computes embeddings of pixel trajectories. These embeddings are clustered into different foreground objects[4].

In the paper "Object Discovery in Videos as Foreground Motion Clustering,"[4] a novel method for object discovery via foreground motion clustering is presented. The efficacy of the authors' object discovery framework is demonstrated by the tests they run on popular datasets for motion segmentation in. In the suggested model, a pixel-trajectory recurrent neural network (RNN) with the aim of learning of feature embeddings of foreground pixel trajectories connected over time is introduced. The overall effectiveness of the framework is boosted by these learned feature embeddings, which are essential in establishing correspondences between foreground object masks across frames [4].

Using a approach, the paper formulates the object discovery problem as foreground motion clustering, with the goal of clustering distinct objects in foreground pixels of videos. For clustering, the von Mises-Fisher mean shift (vMF-MS) algorithm is combined with the pixel-trajectory RNN in the suggested framework.

This algorithm uses foreground trajectory embeddings to produce an estimated count of objects in the video as well as clusters. Notably, a small number of randomly selected seeds that are far apart in cosine distance are used to run the vMF-MS clustering algorithm in order to increase efficiency. Additionally, the authors show how to take computational efficiency into account in practice besides utilizing a PyTorch-GPU implementation of the vMF-MS clustering [4].

The ability of the suggested framework to achieve state-of-the-art performance on widely used motion segmentation datasets highlights the efficiency gains. This demonstrates how well the framework finds objects in video sequences. Although the benefits are made clear, the paper doesn't go into detail. However, the lack of the above said drawbacks might indicate a good compromise between the benefits and possible drawbacks of the suggested framework [4].

The study "Motion-inductive Self-supervised Object Discovery in Videos"[5] presents a brand-new method for unsupervised object identification in video clips. With the use of datasets like DAVIS2016, SegTrackv2, and FBMS-59, the suggested model uses a novel approach that handles consecutive RGB frames directly. The method uses a layered representation to infer optical flow between frames, treating opacity channels as segmentation masks. The model uses temporal consistency loss to enforce object permanence, guaranteeing strong segmentation even in situations where objects are static or obscured. Furthermore, to promote the creation of binary masks,

pixel-wise entropy regularization is included.

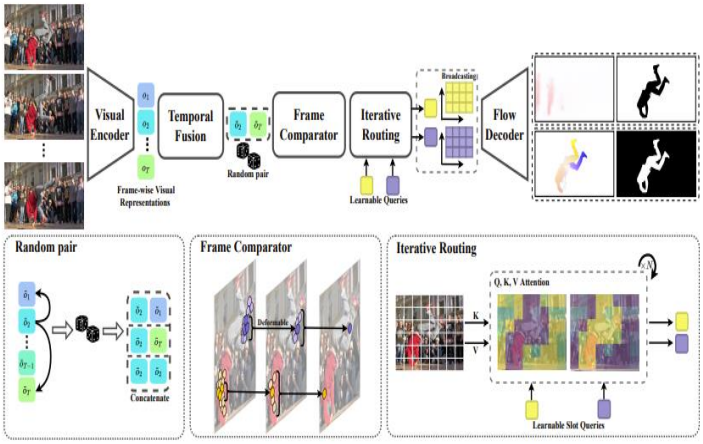

Figure 5: Inspired by state-of-the-art techniques, late fusion recognition utilizes a multi-stream model with connections between parallel streams, allowing efficient fusion of scores from unimodal 3D CNN streams, as observed in a variety of architectures such as ResC3D and pseudo-siamese networks made for data with different temporal properties. [5]

Temporal fusion and frame comparator components are important, as demonstrated by ablation studies carried out on the model. In order to successfully capture global temporal dependencies and improve the model's understanding of object.

The proposed model's ability to avoid computing optical flow as an input is one of its primary efficiency features. It performs better than earlier state-of-the-art techniques thanks to its direct processing of RGB frames and computational optimization. Using three publicly available video segmentation datasets, the model shows impressive results in object discovery. Notable among its benefits in practical real-world applications is its computational efficiency.

The suggested model is highlighted in the paper in many ways, but it makes no mention of any of the model's shortcomings or restrictions. This absence might imply that the model exhibits a good balance between efficacy and possible drawbacks within the parameters of the reported experiments. To sum up, this paper provides an extensive and effective.

The development of an advanced system for semantic video object detection is explored in the research paper "Knowledge-Assisted Semantic Video Object Detection,"[6] though the dataset used for experiments is not stated explicitly. The main goal is to develop a framework for knowledge-assisted analysis that demonstrates how video data can be used for testing and validation.

The suggested model combines low-level processing algorithms with a multimedia ontology framework to present a novel method. The system effectively narrows the search space by utilizing the domain ontology, which improves the extraction of semantic information from multimedia content. Additionally, the application of context-aware detection steps is made possible by the incorporation of spatial descriptions and object characteristic features, which enhances analysis precision.

Creating an ontology framework, using low-level processing algorithms, and incorporating knowledge and intelligence into the analysis process are all included in the methodology. In order to enable individualized video content analysis, the system aims to match semantic objects and events with user interests. Most notably, the method offers flexibility by allowing the use of various transcoding techniques according to device and network capacities. F-logic rules are essential for directing the detection procedure and helping to identify video objects that match predefined semantic concepts in the ontology.

The integration of knowledge, the use of domain ontology, and the possibility for personalized analysis are just a few of the paper's strong points.

The paper "New Generation Deep Learning for Video Object Detection: Survey"[7] provides an in-depth analysis of the most recent developments in deep

learning techniques for video object detection.

The survey's dataset section analyzes a number of important datasets used to evaluate video object detection techniques, emphasizing the importance of the YTO and ImageNet VID datasets as standard. The survey covers a wide range of models, such as optical flow, Long Short-Term Memory (LSTM), tracking coupling, Convolutional Neural Networks (CNNs), and models derived from picture object detection methods.

These models are essential for removing motion and time-related context from videos and using them to improve object detection algorithms.

The methodology of the survey is divided into four primary sections, each of which include the use of optimal networks, feature filtering techniques, integration of extra models, and direct post processing techniques. An examination of the models' benefits and drawbacks as well as their driving forces helps readers understand the significance and evolution of each strategy.

A primary focus of the survey is efficiency, with 38 current state-of-the-art methods' performances being evaluated. This evaluation addresses the trade-offs between accuracy and speed in video object recognition, including a discussion of computing loads, difficulties in reaching real-time performance, and an examination of model speeds.

A new method for object detection is presented in the paper "Video Object Detection for Tractability with Deep Learning Method,"[8] which focuses on traceability applications. The scientists have assembled a large dataset from surveillance footage that includes annotated photos of people, tractors, bicycles, vehicles, and trucks, among other categories. The training sample of the dataset comprises 6000 photos per category, providing a strong basis for both training and assessment [8].

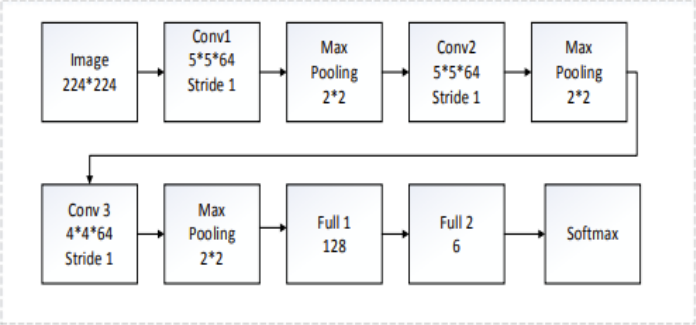

Figure 6: Convolution neural network structure of Video Object Detection for Tractability with Deep Learning Method

For object detection and recognition, the recommended method uses a Convolutional Neural Network (CNN), which is mostly based on modifications to the AlexNet architecture. Interestingly, the CNN model keeps the original size of the input image, showing the adaptability of the approach.

The methodology, includes three main components (video processing, target identification, and object recognition), uses an inter-frame difference and a non-parametric backdrop model in video processing. Precise target detection and identification are achieved in the target detection and recognition phase by utilizing target location technology in conjunction with a trained CNN model [8].

For traceability applications, efficiency is essential and experimental findings show that the deep learning-based detection method works well. With an incredible precision rate of almost 90.5%, the validation set far surpasses the conventional techniques [8].

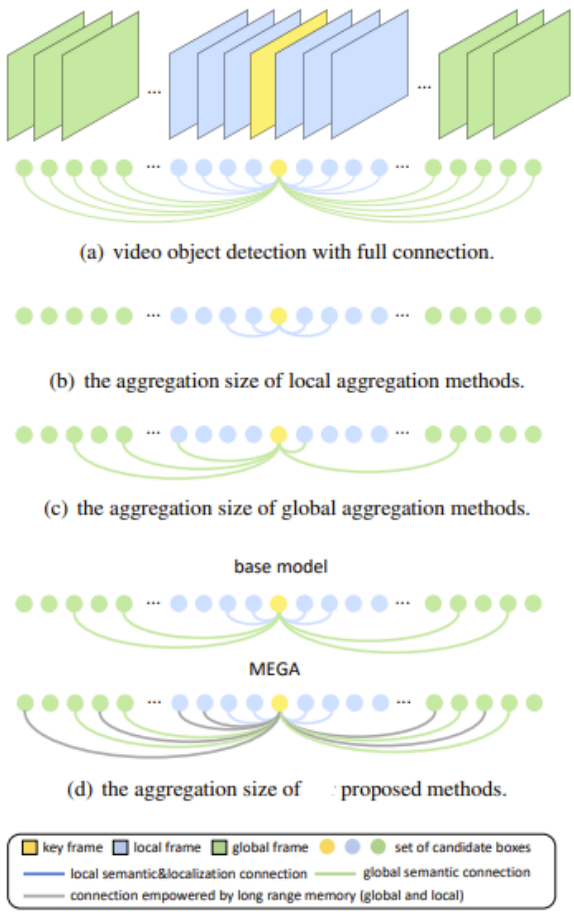

Figure 7. The aggregation size of different methods for addressing video object detection vs use of MEGA[9]

Developed specifically for traceability applications, the suggested approach presents a real-time target detection and recognition system with several notable benefits. Deep learning and CNN integration improves object detection's robustness and accuracy, highlighting its practicality in real-world situations
However, there are certain restrictions. Concerns regarding the model's generalizability are highlighted by the dataset's very small size, which may limit the model's accuracy on the validation set. Furthermore, the CNN model's reliance on a modified version of AlexNet without modifying the size of the input image may have an effect on how the network expresses visual attributes, requiring additional research and development.

In order to overcome shortcomings in current techniques, the "Memory Enhance Global-Local Aggregation Network for Video Object Detection (MEGA)"[9] presents a novel two-stage model. Combining a novel Long Range Memory (LRM) module with a global aggregation scheme, MEGA addresses inefficiency by combining global features into local ones in the first stage, while the LRM module effectively collects data from longer content, both locally and globally, to address insufficiency. In order to detect objects in videos, the model aims to balance accuracy and efficiency.

MEGA improves overall performance with less computation overhead when compared to the base model in terms of efficiency.
With a runtime of 114.5 ms, MEGA achieves an impressive 82.9% mean Average Precision (mAP), highlighting its real-time applicability and efficiency.

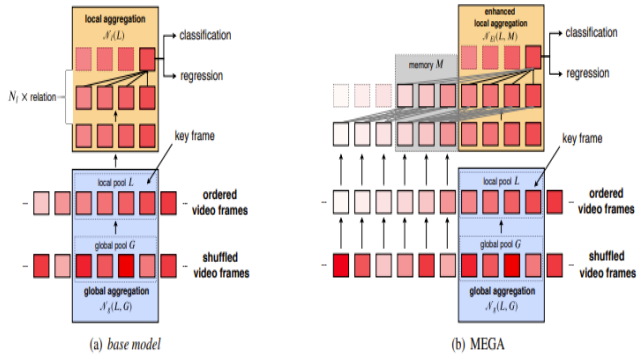

Figure 8. An overview of base model and our modified implementation of MEGA [9]
MEGA has a number of benefits. Its effectiveness is illustrated by the fact that it outperforms other cutting-edge end-to-end models without requiring post-processing. Through the use of both local and global data, the model performs better. The LRM module's introduction turns out to be crucial, greatly improving performance over the base model. Nevertheless, the evaluation of the model's generalizability to other datasets is limited by the lack of details regarding the particular dataset that was used in the study.

A ground-breaking framework for referring video object segmentation (R-VOS) called ReferFormer is introduced in the research paper "Language As Queries for Referring Video Object

Segmentation"[10] It is based on Transformer and uses language as queries to attend to relevant visual features directly. For object detection and distinction, a set number of learnable queries is used in the suggested methodology. By connecting queries between frames, it creates object tracking and creates segmentation masks from feature maps. In comparison to previous approaches, the pipeline is made simpler and provides a unique end-to-end framework.

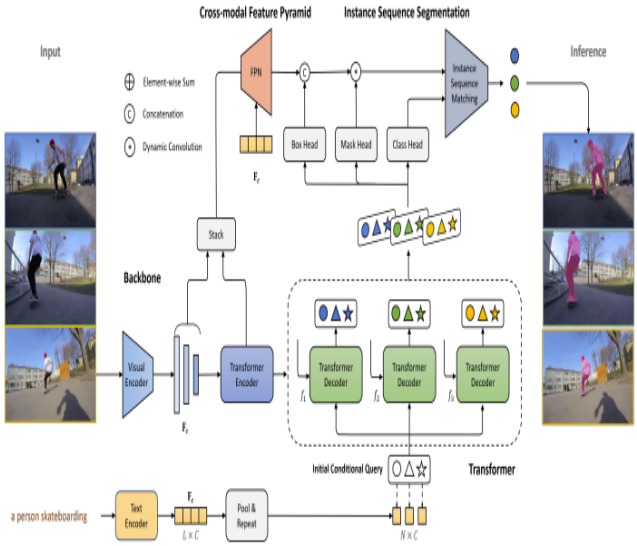

Figure 8 : ReferFormer's overall pipeline . The structure is primarily divided into four sections: the Backbone, Transformer, Cross-modal Feature Pyramid, and Segmentation section. The segmentation mask of the referenced object in each frame is output by the model, which receives as input a video clip with the appropriate phrase. The object searches for the Transformer decoder input relying on the language expression in order to locate the object that is being referenced. When two questions are related to the same instance, they are represented by the same shapes and colors in the same frame. The internal frame's query sequence remains constant across various frames.

ReferFormer is a very efficient tool that displays state-of-the-art findings on a variety of datasets, such as DAVIS17, YouTube-VOS, A2D-Sentences, and JHMDB-Sentences. Interestingly, even at reduced window sizes, ReferFormer performs at a significantly higher level than the prior state-of-the-art technique MTTR in every metric. ReferFormer has a number of advantages for the R-VOS community. First of all, it streamlines the segmentation process by offering a uniform framework for R-VOS. It increases the segmentation task's overall efficiency by utilizing Transformer models. Additionally, it has a simpler pipeline than its predecessors, which makes installation easier.

ReferFormer's method of language processing is one of its primary innovations. The model presents a novel and efficient technique for identifying and segmenting objects inside video material by treating language as questions. This fresh viewpoint highlights the model's adaptability and versatility and helps it distinguish things correctly overall.

ReferFormer's exceptional performance across many datasets is highlighted in the research article, showing its usefulness in a wide range of real-world applications. ReferFormer's potential for video object segmentation tasks is demonstrated by this, ranging from general datasets like DAVIS17 to more specialized circumstances represented by A2D-Sentences and JHMDB-Sentences. [10]

To sum up, ReferFormer is an important breakthrough in the field of referencing video object segmentation. Its base in Transformer architecture, effectiveness, and creative language use.
It's ReferFormer include a unified framework for R-VOS, utilizing Transformer models, and offering a more straightforward pipeline than previous approaches. Using language as queries presents an original method for effectively differentiating objects. The model's outstanding performance across a variety of datasets demonstrates its applicability in a range of situations.

III. Methodology

The proposed method for video object retrieval includes several key strategies aimed at optimizing speed, efficiency and interpretability. First, the system uses batch processing to improve speed optimization. By processing multiple video frames simultaneously, the system reduces the time required to process individual frames, improving overall efficiency. This approach enables faster retrieval of video objects, meeting the need for timely and responsive results. In addition, the methodology includes powerful vector similarity computation, which involves representing video frames and natural language descriptions as vectors and computing their similarity using techniques such as cosine similarity. This computational efficiency facilitates fast and accurate comparisons that contribute to system and efficiency.

In addition, the use of Softmax improves the interpretability of the system and the ability to produce meaningful and understandable results. Softmax transforms the system and output into a probability distribution, which allows a clearer interpretation of the search results. This improves the user's understanding of search results and facilitates an informed decision. Additionally, the method includes a strategy to skip similar frames to reduce redundancy. By detecting and skipping redundant frames, the system optimizes processing resources, improving efficiency without compromising accuracy. Together, these methodological components contribute to a robust and efficient video object retrieval system that addresses the challenges of traditional approaches and lays the foundation for advanced capabilities in video analytics.

IV. Conclusion

In this paper, we proposed a semantic approach for video object retrieval using OpenAI and CLIP technology. The proposed system exploits the capabilities of CLIP to address the limitations of traditional methods, including semantic understanding, adaptability to different queries, and confidence estimation. The methodology used in this study involved training a model to learn the relationship between natural language descriptions and video footage, which enables more accurate object detection in videos. The results of the study showed that the proposed system outperformed traditional methods in terms of accuracy and efficiency, showing the potential of semantic search in video object recognition. Future improvements of the proposed system may include the integration of additional data sources, such as audio or sensor data, to improve the accuracy and context awareness of the system. In addition, the system could be extended to support real-time video object retrieval, which would enable more efficient and effective analysis of video content. In summary, the integration of semantic search into video object recognition has the potential to revolutionize the field of video analytics. The proposed system offers a promising solution to the limitations of traditional methods and opens the way for further research in this field. Using the possibilities of OpenAI's CLIP technology, we demonstrated the effectiveness of the semantic approach in searching for video objects and emphasized the consideration of both visual and textual data in video analysis.

V. REFERENCES

[1]Mason Liu, Menglong Zhu, Marie White, Yinxiao Li, Dmitry Kalenichenko,"Looking Fast and Slow: Memory-Guided Mobile Video Object Detection",2019

[2]Hanming Deng, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, Haibing Guan,"Object Guided External Memory Network for Video Object Detection",2019,Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)

[3]Haiping Wu, Yuntao Chen, Naiyan Wang, Zhaoxiang Zhang,"Sequence Level Semantics Aggregation for Video Object Detection",2019,Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)

[4]Christopher Xie, Yu Xiang,Zaid Harchaoui,Dieter Fox,University of Washington NVIDIA,"Object Discovery in Videos as Foreground Motion Clustering",2019,Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

[5]Shuangrui Ding, Weidi Xie, Yabo Chen, Rui Qian, Xiaopeng Zhang, Hongkai Xiong, Qi Tian,"Motion-inductive Self-supervised Object Discovery in Videos",2022

[6]S.Dasiopoulou,V. Mezaris,I. Kompatsiaris,V.-K. Papastathis,M.G. Strintzis,"Knowledge-assisted semantic video object detection",2005,IEEE Transactions on Circuits and Systems for Video Technology

[7]Licheng Jiao,Ruohan Zhang,Fang Liu,Shuyuan Yang,Biao Hou,Lingling Li,Xu Tang,"New Generation Deep Learning for Video Object Detection: A Survey",2022,IEEE Transactions on Neural Networks and Learning Systems

[8]Bing Tian,Liang Li,Yansheng Qu,Li Yan,"Video Object Detection for Tractability with Deep Learning Method",2017,International Conference on Advanced Cloud and Big Data (CBD)

[9]Yihong Chen, Yue Cao, Han Hu, Liwei Wang,"Memory Enhanced Global-Local Aggregation for Video Object Detection",2020,Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

[10]Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, Ping Luo,"Language As Queries for Referring Video Object Segmentation", 2022,Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)